NTNU

Innovation and Creativity

Multilevel Analysis Applied to Fetal Growth Data with Missing Values.

Eystein Widar Bråthen

Master of Science in Physics and Mathematics

Submission date:	June 2006
Supervisor:	Mette Langaas, MATH
Co-supervisor:	Stian Lydersen, Det medisinske fakultet Geir Jacobsen, Det medisinske fakultet

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem Description

The goals of this assignment are to study how growth of a fetus depends on different background variables, describe and use methods for handling with missing data, and compare these results with the results from the "complete-case" analysis.

Assignment given: 20. January 2006
Supervisor: Mette Langaas, MATH

Preface

This master thesis is the result of my five years study at the Norwegian University of Science and Technology where the courses have spanned from very general to this specialized project. It has been an interesting period with a lot of hard work, challenges and during these years I also got many new friends.

Acknowledgments

First and foremost I must thank Professor Stian Lydersen, Associate professor Mette Langaas and Professor Geir Jacobsen for being my teaching supervisors in this project. They have provided academic support and guidance throughout the entire process. I would also thank medical student Silje Forseth Eilertsen, who have worked with the same data set in her master thesis, for her contribution of medical information and advice during this semester.

Eystein Widar Bråthen
Trondheim, June, 2006.

Abstract

Intrauterine growth retardation means that the growth of a fetus is restricted as compared with its biological growth potential. This contributes to an increased risk for illnesses or death of the newborn. Therefore it is important to characterize, detect and to follow up clinically any suspected or confirmed growth restriction of the fetus. In this master thesis we aim to describe the course of growth during the pregnancy based on repeated ultrasound measurements and study how the growth depends on different background variables of the mother in analyzing the data from the SGA (small-for-getational age) - project. The SGA-project contains data from 5722 pregnancies that took place in Trondheim, Bergen and Uppsala from $1986 - 1988$, named *The Scandinavian SGA-studies*. In this thesis we have confined ourselves to a random sample of 561 pregnancies.

A problem with many studies of this kind is that the data set contain missing values. In the SGA data set under study there were missing values from one or more of the ultrasound measurements for approximately 40% of the women. Until recently, the most popular used missing-data method available has been complete case analysis, where only subjects with a complete set of data are being analysed. There exist a number of alternative ways of dealing with missing data. Bayesian multiple imputation (MI) has become a highly useful paradigm for handling missing values in many settings. In this paper we compare 2 general approaches that come highly recommended: Bayesian MI and maximum likelihood (ML), and point out some of its unique features. One aspect of MI is the separation of the imputation phase from the analysis phase. It can be advantageous in settings where the models underlying the two phases are different.

We have used a multilevel analysis for the course of fetal growth. Multilevel analysis has a hierarchic structure with two levels of variation: variation between points in time for the same fetus (level 1) and variation between fetuses (level 2). Level 1 is modeled by regression analysis with gestational age as the independent variable and level 2 is modeled by regarding the regression coefficients as stochastic with a set of (non directly observed) values for individual fetuses and some background variables of the mother.

The model we ended up with describes the devolopment in time of the abdominal diameter (MAD) of the fetus. It had several "significant" covariates ($p-\text{value} < 0.05$), they were gestational age (Time-variable), the body-mass index (BMI), age of the mother, an index varible wich tells if a mother has given birth to a low-weight child in an earlier pregnancy and the gender of the fetus. The last covariate was not significant in a strictly mathematical way, but since it is well known that the gender of the fetus has an important effect we included gender in the model as well. The growth model for MAD is

$$
\begin{aligned}
y_{ij} = {} & \beta_{0_j} + \beta_{1_j} x^*_{ij_{Time}} + \beta_{2_j} x^{*^2}_{ij_{Time}} + \beta_3 x_{ij_{Gender}} + \beta_4 x_{ij_{AGE}} + \beta_5 x_{ij_{BMI}} + \beta_6 x_{ij_{LBW}} \\
& + \beta_7 x_{ij_{AGE}} x^*_{ij_{Time}} + \beta_8 x_{ij_{BMI}} x^*_{ij_{Time}} + \beta_9 x_{ij_{LBW}} x^*_{ij_{Time}} + \epsilon_{ij}.
\end{aligned}
$$

When we used the MI-method on the random sample (561) with missing values, the estimated standard deviations of the parameters have been reduced compared to those obtained from the complete case analysis. There were not a significant change in the parameter estimates except for the coefficient for the age of the mother.

We also have found a procedure to verify if the MI-method gives us reasonable imputed values for the missing values by following the MCAR-procedure defined in Section 6.2. Another interesting observation from a simulation study is that estimates of the coefficients for variables used to generate the MAR and MNAR missing mechanism are "suffering" because they tend to be more biased compared to the values from the complete case analysis on the random sample (320) than the other variables. According to the *MAR* assumption such a procedure should give unbiased parameter estimates.

Key Words: Longitudinal data, multilevel analysis, missing data, multiple imputation (MI), Gibbs sampling, linear mixed-effects model and maximum likelihood (ML)-procedure.

Contents

1 Introduction

Intrauterine growth retardation means that a fetus does not live up to its growth potential. This represents a seriously increased risk for illnesses or death at birth and later in life. Therefore it is important to characterize, detect and follow up restricted growth of fetuses.

In my project (Bråthen, 2005) last semester I studied the observed measurements of the different variables of the subjects from the SGA-study, and I also described some of the relevant statistical models which are suitable for repeated measurements. The intention is to use the suggested methods in analyzing the data from the SGA-project.

This project will analyze a sub-sample of data from 5722 pregnancies that took place in Trondheim, Bergen and Uppsala named *the Scandinavian SGA* (small-for-gestational age) - *studies* (SGA-Scandinavia, 1986-1988). Background variables of the mother were registered, such as age, parity, weight, height and smoking status, outcome of previous gestations, such as length of the pregnancy, length and weight of fetuses and its condition at birth. 1384 of the women were defined in a high risk group, which had an increased risk to give birth to smaller-than-expected children. This was defined as similar outcome in earlier pregnancies, if they had experienced a stillbirth, smoked during pregnancy, had a pre-pregnancy weight below 50 kg or had been diagnosed with certain chronic diseases (chronic renal disease, essential hypertension or heart disease). The pregnant women of the risk group were followed by 4 ultrasound measurements during pregnancy; week 17, 25, 33 and 37. That included among others measurement of the femur length of the fetus, its abdomen and bi-partial diameter of the head. A similar procedure was followed for a *random sample* of 561 of the 5722 women. This random sample is the data set studied in this thesis. As defined in Section 2.2 and the remaining part of this master thesis we use the form 'SGA Data Set' to refer to the random sample of size 561.

One problem in many studies of this kind is that the data set is not complete because participants may be present for some portion of data collection and missing for others. For the SGA study data were missing for one or more ultrasound measurements from approximately 40% of the female participants. Traditionally in such situations one have used “complete-case” analysis, where only the data from women with a complete set of data is analyzed. However, parts of the data is discarded, which may cause the results to be *biased* (skewness in the expectation of the results).

The goals of this assignment is to study how that growth depends on different background variables, and to describe methods for dealing with missing data, especially “maximum likelihood” and “multiple imputation”. Finally, these methods will be used to generate a complete data set for the SGA Data Set and compare this results to the “complete-case” analysis. During this thesis I have collaborated with medical student, Silje Forseth Eilertsen. She has contributed with the biological and the human medical knowledge which has been essential to perform a correct analysis and interpretation of the results.

The outline of this master thesis is as follows

- In Chapter 2 we will describe the collecting procedure of the data material from the Scandinavian SGA-project.
- In Chapter 3 and 4 we present background theory on the relevant statistical models suitable in the repeated measurements situation and describe different methods available for imputations of missing values. We also look at the relation between the response variable and the covariates in the raw data for the three different missing mechanisms; MAR, MCAR and MNAR (Chapter 4).
- In Chapter 5 we study marginal and joint descriptions of the missing values in the data set. We investigate different imputation models to be used for the SGA data.
- In Chapter 6 we through a simulation study assess the accuracy of the estimates of the MI method.
- In Chapter 7 we compare two recommended methods for dealing with missing data; "maximum likelihood" and "multiple imputation". We also perform a simulation study to study the bias (skewness from the expectation values of the parameter) and the standard deviation of the parameters for the growth model constructed from data sets with different missing mechanisms.
- In Chapter 8 we describe the selection procedure for variables to be included in the growth model for MAD (mean abdominal circumference) of a fetus, and perform a complete data analysis for the growth model for the SGA Data Set (561 subjects) with imputed values for the missing observations, and compare the results to the complete case analysis of SGA Data Set (320 subjects with complete data).
- In Chapter 9 we discuss the main results from the master thesis and suggest further work within this topic.
- The Appendix consists of two parts, A and B. In Appendix A we have described the methods we have used for Multilevel analysis and dealing with missing data; LME, MI-method and the ECME-method. The program code is displayed in Appendix B.

2 The Scandinavian SGA Project

2.1 Definitions

The following definitions are taken from SGA-Scandinavia (1986-1988) study.

Gestational Age: was based on the first day of the last menstrual period (LMP) if it was accurately recalled to within ±3 days. A sonogram of the biparietal diameter (BPD) were used at the first visit (approximately 17 weeks) to date the pregnancy if the discrepancy between BPD and LMP gestational age was more than ±14 days, or if the LMP could not be recalled accurately.

Parity: According to the WHO (World Health Organization) parity is defined as the number of pregnancies that a woman delivers past 28 week or more duration.

Pre-Pregnancy Weight: was reported by the mother. A weight less than 50 kilograms was the cut-off used to identify women with a low pre-pregnancy weight.

Previous Low Birth Weight (LWB) Infant: was defined as the prior first birth of a baby girl below 2700 grams or a baby boy below 2800 grams, or a prior second baby girl below 2800 grams or a baby boy below 2900 grams, regardless of the gestational age at the delivery. These limits were chosen since they correspond approximately to the lowest 10th percentile of weight at term. Births below these limits represent a group of small for gestational age births and preterm births. For practical reasons, when women were included in the study those who were responsible for the study had to rely on the mothers own recall of the weight of their previous birth(s). As it turned out, the mothers tended to remember the birth weight of their previous infants extremely well. Birth weights reported by the mothers were validated subsequently against birth weights recorded in the Norwegian and Swedish medical birth registers. In 97% of the cases, mothers had recalled the birth weights of previous infants to within 50 gram. The exact birth weight was reported by 89% of the mothers.

Small for Gestational Age (SGA) Birth: the index pregnancy was defined as an infant with a low birth weight at each specific gestational week. The reference standards were sex specific for parous women and based on last menstrual period dating, as previously published using data from the Norwegian Medical Birth Registry. Because ultrasound dating of gestational age is much more common now than when these standards were derived, both the nominal 10th and 15th percentiles of birth weight for gestational age were used in defining a SGA birth. Since ultrasound standards employ gestational ages that, on average are shifted from three to seven days earlier than LMP based estimates, the nominal 15th percentile (based on the original LMP based standards) corresponds better to a current population based birth weight percentile. This so-called 15th percentile standard for SGA also compares favorably with a recently published update of the Swedish standards. The LMP based estimate of gestational age was chosen unless the ultrasound and LMP-based expected dates of delivery differed by more than two weeks, in which case the ultrasound based estimate of gestational age was substituted.

Smokers: Women who at the first visit reported that they smoked daily at the time of conception.

2.2 Collecting Procedure

The NICHD Study of Successive Small-for-Gestational-Age births is a multicentre study (SGA-Scandinavia, 1986-1988) organized in Scandinavia at the Universities of Trondheim and Bergen in Norway, and Uppsala in Sweden. The University hospital in each site was the basis for the collection of prenatal, delivery and follow up data. In this project, the prenatal data will be used supplemented by the delivery and newborn data. In each of the three geographical areas there was only one obstetrical and pediatric department and practically all women delivered at the University hospital. Recruitment of pregnant women in Norway was based on referrals from general practitioners and obstetrician in Trondheim and Bergen who had agreed to refer their eligible patients to the antenatal clinic at the University hospital for four special antenatal study visits. They were arranged in addition to the antenatal visits by their ordinary practitioner. In Sweden women were recruited from all antenatal care centers in Uppsala County, which were under direct supervision by the University hospital. Recruitment took place over a 27 months period. Women who were invited to participate in the study were given a consent form, which was discussed and signed before they entered into the study. In the present project SGA was defined as birth weight-for-gestation below the 15th percentile of the background population reference, adjusted for gender and gestational age.

There were 6354 women who were eligible for enrolment in the study at the first antenatal visit. They were women of Caucasian origin who spoke one of the Scandinavian languages and expected their second or third child. 432 of the women were excluded because they did not fulfill the study criteria. 34 had a multiple pregnancy and 229 aborted. In addition 169 women were found ineligible due to ethnic or language incompatibilities, were not expecting their second or third child, or their pregnancy had gone beyond 20 weeks of gestation. Another 200 women failed to come to their first scheduled appointment. A overview of this process can modeled as follows:

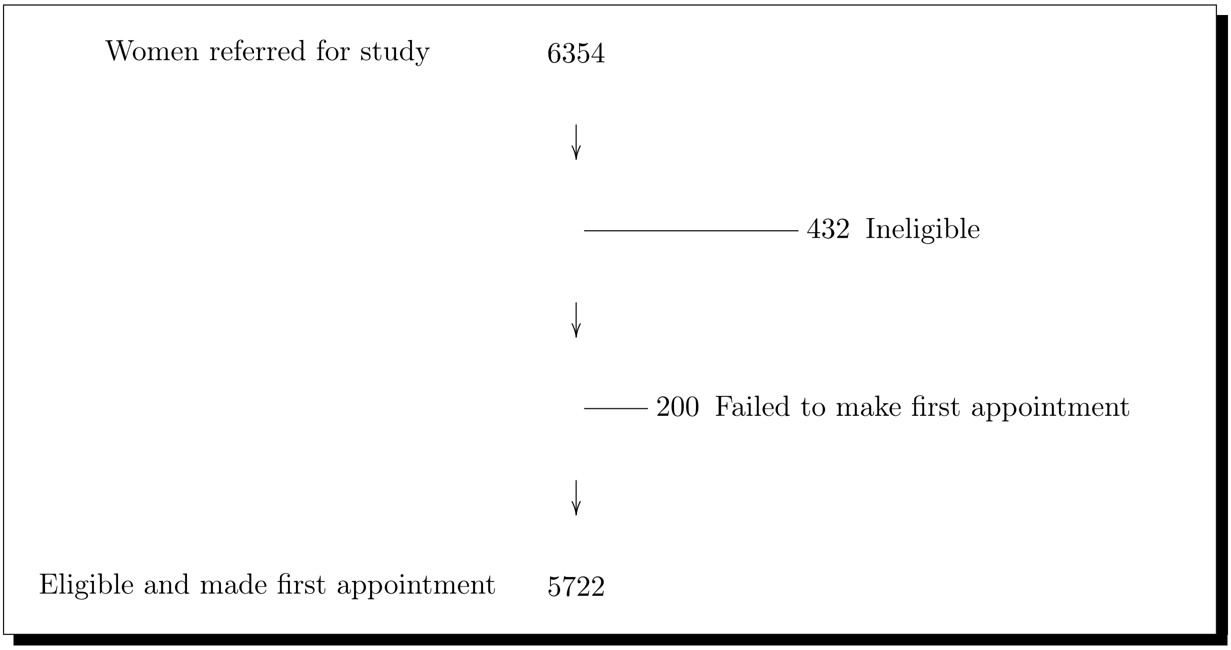
Women referred for study 6354
432 Ineligible
200 Failed to make first appointment
Eligible and made first appointment 5722

Figure 1: This figure shows the first step of the selection procedure of women eligible for the SGA-project.

The 5722 eligible women who remained, were divided into three different groups

1. A 10% random sample, ($n = 561$).
2. A high risk group, ($n = 1384$).
3. A rest population, ($n = 3777$).

First, the random sample was selected using the sealed envelope method[1], and this sample serve as a control group that is representative of the parous pregnant population at each study site. The analyses in this project are limited to study the random sample (561 pregnancies). The selected women were examined at four defined intervals throughout pregnancy, at 17, 25, 33 and 37 weeks of gestation. The information collected at each of the four antenatal study visits and at birth, which is used in the analysis is presented in Table 1.

[1] Randomly selected of the 5722 women who were eligible for the study

Variable	*Variables*	*Week 17*	*Week 25*	*Week 33*	*Week 37*	*Delivery*
Eligibility and study entry	V0001-V0068	X				
Prenatal record 1	V0103A-V0168	X				
Ultrasound 1	V02001-V0230	X				
Medical history	V0502A-V0533M	X				
Prenatal record 2	V1103A-V1164		X			
Ultrasound 2	V1201-V1218		X			
Prenatal record 3	V2103A-V2171			X		
Ultrasound 3	V2201-V2220			X		
Prenatal record	V3103A-V3166				X	
Ultrasound 4	V3201-V3220				X	
Prenatal record 5	V4102B-V4165C					X
Ultrasound 5	V4201-V4229					X
Newborn record (part A)	V51002-V51147D					X

Table 1: Collected data in the SGA-project.

2.2.1 High risk group

The high-risk group was selected after the selection of random sample from the remaining women if they fulfilled one or more of the defined risk criteria for SGA birth. These risk factors were

1. A prior Low-Birth-Weight birth (LBW) .
2. Maternal cigarette smoking at conception.
3. Low pre-pregnancy weight ($<$ 50 kg).
4. A previous prenatal or perinatal death.
5. The presence of chronic maternal disease (chronic renal disease, essential hypertension or heart disease).

This high risk group was selected in order to enrich the study sample with a statistically sufficient number of SGA births. Selection criteria of the high risk group were used to determine which women in addition to the 561 women in group 1, who should be offered the detailed follow up. Women who met one or more of the high risk group criteria were included with one notable exception. If they reported smoking around time of conception, then 50% were randomly selected, using the sealed envelope method, to be included in the detailed follow up group, regardless of the fact that they might have other risk factors in addition. Overall, the gestational age was estimated by ultrasound in 18.8% of the “high risk group”, of which an equal proportion (9.4% each) was due to an uncertain LMP versus discrepancies of more than ±14 days. The corresponding proportion of gestational ages estimated by ultrasound in the random sample was 16.6% of which 7.0% was due to uncertain dates and 9.6% was due to discrepancies of more than ±14 days. After the selection of the random sample had taken place, 598 of the remaining women had smoking as the only risk factor. A total of 1384 women, in example the randomly selected smokers and those who fulfilled one or more of the other risk criteria were studied prospectively as the high risk group. The remaining 3777 women were defined as a “rest population”.

Their newborn was entered into the follow up study of the children if it was diagnosed as a SGA birth. The selection process of the three different groups can be modeled as follows

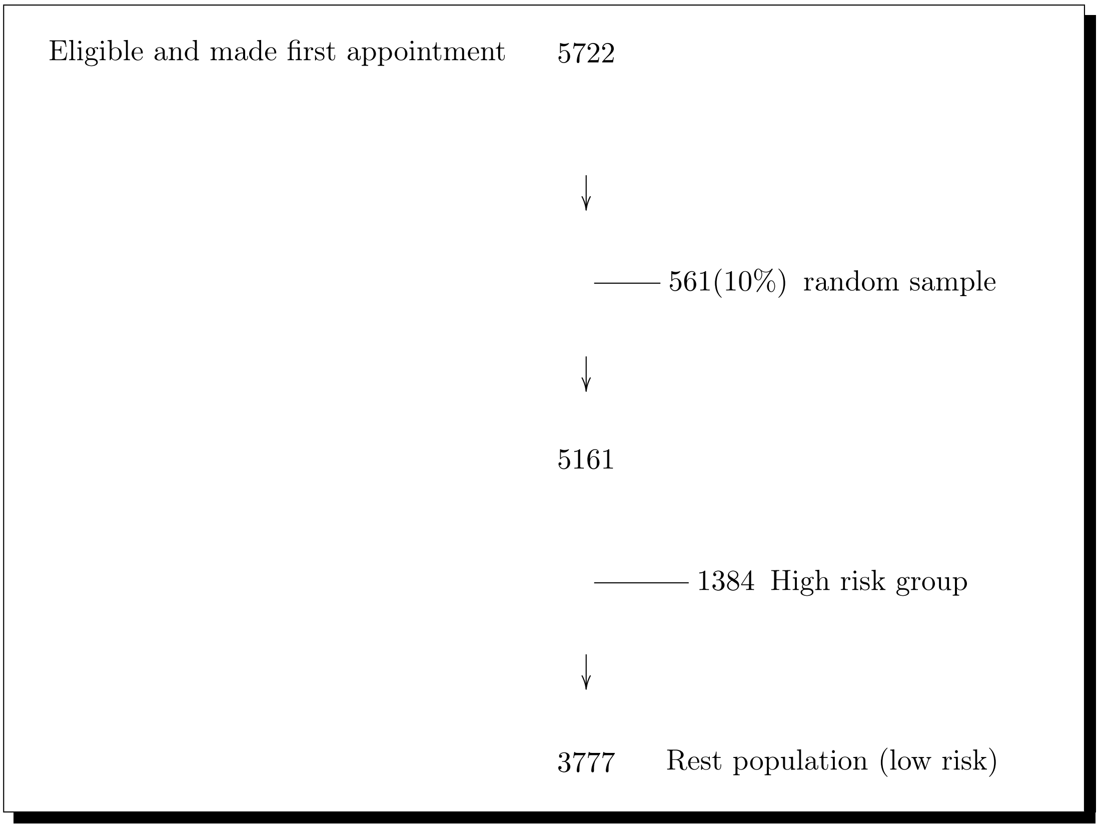
Eligible and made first appointment 5722
561(10%) random sample
5161
1384 High risk group
3777 Rest population (low risk)

Figure 2: This figure is illustrating the allocation of the study objects. That is the para 1 and para 2 pregnant women from the time of initial refferal by clinics, obstetrician or family practitioners until assignment to one of the three study groups

From the 1945 (High risk group + 10% random sample) women who were invited to the detailed study, 393 (20, 2%) failed to complete more than one of the four study examinations that took place during the pregnancy. Those women were defined as “drop outs”. Most of them gave “social inconvenience” as reason for not continuing with the study which included no available baby-sitter, no possibility to obtain a leave from work and long travel times to the university hospital. This was in particular the case in Uppsala where women were recruited from the whole country. Of the 393 drop outs, 358 (91, 1%) came from the Uppsala part. Of the remaining women who did not complete the study as planned, 15 (0, 8%) had moved, 60 (3, 1%) “refused” and 15 (0, 8%) gave no reason. The data used in this project is an extract of the complete data file and consists of the random sample of 561 pregnancies. We will in the remaining part of this master thesis use the form ’SGA-Data set’ to refer to the random sample of size 561.

3 Multilevel Analysis

This presentation of theory of multilevel models is largely based on Goldstein (2003).

3.1 Notation

All covariates used in the equations in this project is defined in a “unusual” form, but I find it very informative and a natural way of defining them. For example if we use time measured in weeks of gestations as a covariate, it is being defined as: $X_{ij_{Time}}$. This means that measurement i of fetus j from gestational age is used as a covariate.

3.2 The Basic Two-Level Model

We find it necessary to state the fact that multilevel models are the same as Mixed-effects models. Mixed-effects models provide a flexible and powerful tool for the analysis of grouped data. Examples of such data include longitudinal data, repeated measures, blocked design and multilevel data. The increasing popularity of mixed-effects models is explained by the flexibility that they offer in the modeling of the within-group correlation, the number of “levels” are the same as the number of nested levels of random effects. Mixed-effects models assumes that both the random effects and the errors follow Gaussian distributions.

To illustrate the basic two-level model we use the following data set which consists of two occasions of measurements of scores from mathematics tests. The data set (JSP = Junior School Project) consists of 728 pupils in 48 different elementary schools. The first occasion of measurement took place in their fourth year of schooling (when the pupils were 8 years old), and the second measurement took place three years later at age 11 years. Goldstein (2003) uses the results from mathematics tests together with information collected on the social background of the pupils and the gender to predict the 11-year scores. A simple model for one school which relates the 11 year score, y_i, to the 8 year score, x_i, is for pupil i

$$y_i = \alpha + \beta x_i + e_i, \tag{1}$$

where the standard interpretations are given to the intercept (α), slope (β) and to the residuals (e_i). Goldstein (2003) also use the typically assumption that the residuals follows a *Gaussian distribution* with a zero mean and a common variance, $e_i \sim N(0, \sigma_e^2)$. To describe simultaneously the relationships for several schools for all pupils

$$y_{ij} = \alpha_j + \beta_j x_{ij} + e_{ij}, \tag{2}$$

where j refers to the level two unit which in this case are the different schools and i refers to the different pupils from each school which is the level one unit. Equation (2) is as it stands essentially a single-level model which describes separate relationships for each of the j schools. To make Equation (2) into a genuine two level model we let α_j and β_j become random variables, and we replace α_j by β_{0_j} and β_j by β_{1_j} such that β_{0_j}

can be written as $\beta_0 + u_{0_j}$ and β_{1_j} can be written as $\beta_1 + u_{1_j}$, where u_{0_j},u_{1_j} are random variables. Again Goldstein (2003) uses the usual assumption that the random parameters are *normally distributed* with

$$E(u_{0_j}) = E(u_{1_j}) = 0 \tag{3}$$
$$\text{var}(u_{0_j}) = \sigma^2_{u0}, \;\; \text{var}(u_{1_j}) = \sigma^2_{u1}, \;\; \text{cov}(u_{0_j}, u_{1_j}) = \sigma_{u01}. \tag{4}$$

Equation (2) may now be written as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0_j} + u_{1_j} x_{ij} + e_{0_{ij}}) \tag{5}$$
$$\text{var}(e_{0_{ij}}) = \sigma^2_{e0}. \tag{6}$$

Where the response variable (y_{ij}) can be written as a sum of a fixed and a random part. The fixed part of Equation (5) can generally be written in the matrix-form as in Equation (7).

$$E(Y) = X\beta, \;\; \text{with}\, Y = \{y\}_{ij} \tag{7}$$
$$E(y_{ij}) = (X\beta)_{ij}. \tag{8}$$

The feature of Equation (5) which distinguishes it from the standard linear models of the regression or analysis of variance type is the presence of more than one residual term and this implies that special procedures are required to obtain satisfactory parameter estimates. *Note* that it is the structure of the *random part* of the model which is the key factor. The fixed part of the variables can be measured at any level.

3.3 The variance components model

Equation (5) requires estimation of two fixed coefficients β_0, β_1 and four random parameters $\sigma^2_{u0}, \sigma^2_{u1}, \sigma^2_{u01}$and σ^2_{e0}. First we consider the simplest 2-level model $y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{0_{ij}}$, which only includes the random parameters σ^2_{u0} and σ^2_{e0}. The variance of the response about the fixed component is

$$\text{var}(y_{ij} \mid \beta_0, \beta_1, x_{ij}) = \text{var}(u_o + e_{0_{ij}}) = \sigma^2_{u0} + \sigma^2_{e0},$$

which is equal to the sum of the level one and level two variances. The JSP-data in Goldstein's model implies that the total variance for each student is constant and that the covariance between two students in the same school are given by

$$\text{cov}(u_{o_j} + e_{0_{i_1 j}}, u_{o_j} + e_{0_{i_2 j}}) = \text{cov}(u_{0_j}, u_{0_j}) = \sigma^2_{u0}.$$

Since the level 1 residuals are assumed to be independent, the correlation between two students in the same school is given by

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}}.$$

Goldstein (2003) calls this the *intra-level-two-unit correlation*, in this case the intra-school correlation. For the variance components model this also measures the proportion of total variance between the different schools.

The block-diagonal covariance matrix for the response-vector Y for a two-level variance components model is derived from the expressions above between two different schools. Let A be the covariance matrix for the scores of three students in one school and B the covariance matrix for the scores of two students in a another school.

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

$$\mathrm{A} = \begin{bmatrix} \sigma^2_{u0} + \sigma^2_{e0} & \sigma^2_{u0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{e0} + \sigma^2_{u0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{e0} & \sigma^2_{u0} + \sigma^2_{e0} \end{bmatrix}, \quad \mathrm{B} = \begin{bmatrix} \sigma^2_{u0} + \sigma^2_{e0} & \sigma^2_{u0} \\ \sigma^2_{u0} & \sigma^2_{e0} + \sigma^2_{u0} \end{bmatrix}$$

This “block-diagonal” structure reflects the fact that the covariance between students in different schools are zero and extends to any number of level 2 units.

3.4 The General 2-Level Model With Random Coefficients

It is possible to extend Equation (5) in a standard way to include further fixed explanatory variables in addition to the exiting ones

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{h=2}^{n} \beta_h x_{h_{ij}} + (u_{0_j} + u_{1_j} x_{ij} + e_{0_{ij}})$$

or written in a more compact form

$$y_{ij} = \beta X_{ij} + \sum_{h=0}^{1} u_{h_j} z_{h_{ij}} + e_{0_{ij}} z_{0_{ij}}, \tag{9}$$

where Goldstein (2003) uses the new explanatory variables for the random part of the model to write these more generally.

$$\begin{aligned} Z &= \{Z_0, Z_1\} \\ Z_0 &= \{1\} \text{ , a vector of 1's} \\ Z_1 &= \{x_{1_{ij}}\}. \end{aligned}$$

Any of the explanatory variables may be measured at any levels, for example we may have pupil characteristics at level 1 or school characteristics at level 2. In this model is the coefficient of X_1 random at level 2, and that give rise to the following block structure for a level 2 block with two level 1 units. Ω_2 is the covariance matrix of the random intercept and slope at level 2. It is necessary to distinguish between the covariance matrix of the responses and the random coefficients, Ω_1 is the covariance matrix for the set of the level one random coefficients and in this case it is just a single variance term at level one.

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix}$$

where

$$\begin{aligned} A &= (\sigma^2_{u_0} + 2\sigma_{u_{01}} x_{1_j} + \sigma^2_{u_{01}} x^2_{1_j} + \sigma^2_{e_0}) \\ B &= (\sigma^2_{u_0} + \sigma_{u_{01}}(x_{1_j} + x_{2_j}) + \sigma^2_{u_1} x_{1_j} x_{2_j}) \\ C &= (\sigma^2_{u_0} + 2\sigma_{u_{01}} x_{2_j} + \sigma^2_{u_1} x^2_{2_j} + \sigma^2_{e_0}) \end{aligned}$$

giving

$$\begin{aligned} \begin{bmatrix} A & B \\ B & C \end{bmatrix} &= X_j \Omega_2 X_j^T + \begin{bmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{bmatrix} \\ X_j = \begin{bmatrix} 1 & x_{1_j} \\ 1 & x_{2_j} \end{bmatrix}, \ \Omega_2 &= \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_{01}} & \sigma^2_{u_1} \end{bmatrix} \quad, \ \Omega_1 = \sigma^2_{e_0} \end{aligned}$$

From the equations above we see the covariance matrix for a level 2 unit with two level 1 units for a 2-level model with a random intercept and regression coefficient at level 2. We also see the general pattern for constructing the response covariance matrix which generalizes to both higher order models and to complex variation structures at level 1. In the next Section 3.5, is the Maximum likelihood (ML) procedure for obtaining estimates presented.

3.5 Maximum likelihood estimation using iterative Generalized Least Squares (IGLS)

Suppose we knew the values of variances and could construct the block-diagonal matrix, then we could apply Generalized Least Squares (GLS) estimation procedure to obtain the estimator for the fixed coefficients.

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y,$$

where in this case

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n_m m} \end{bmatrix}, \quad \mathrm{Y} = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_m m} \end{bmatrix}$$

with m level 2 units and n_j level 1 units in the j-th level 2 unit. Since we have assumed that the residuals have $Normal$ distributions, also yields maximum likelihood estimates. This estimation procedure is iterative, we can usually start from reasonable estimates of the fixed parameters. These will be chosen from an initial ordinary least squares (OLS) fit. Assuming $\sigma^2_{u_0} = 0$ giving OLS estimates of the fixed coefficients ($\hat{\beta}^{OLS}$). From these we can form the "raw"residuals

$$\tilde{y}_{ij} = y_{ij} - \hat{\beta}_0^{OLS} - \hat{\beta}_1^{OLS} x_{ij},$$

where the vector and raw residuals is written $\tilde{Y} = \{\tilde{y}_{ij}\}$. If we form the cross-product matrix $\tilde{Y}\tilde{Y}^T$ we see that the expected value of this is simply the covariance matrix. This has been used again in $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ to give a better estimate of β's until the estimated values of β have converged.The maximum likelihood procedure produces biased estimates of the random parameters because it takes no account of sampling variation of the fixed parameters. This may be important in small samples, and we can produce unbiased estimates by using a modification known as Restricted Maximum Likelihood (REML), which means that certain conditions is being imposed before the estimates of the coefficients are estimated. Then we can form the raw residuals to get the covariance matrix and use this to estimate the coefficients again until the values have converged.

3.6 Models for Repeated Measures Data

When measurements are repeated on the same subjects, for example as in our *SGA-data set* where different measures of the pregnant women are repeated at four different antenatal visits during the study, then a two-level hierarchy is established with measurements/repetitions as level one units and subjects as level two units. Such data are often referred to as "longitudinal" as opposed to "cross-sectional" where each subject is measured only once. It is important to distinguish between two classes of models which use repeated measurements on the same subject, Goldstein (2003) uses this two classes:

1. In the first model are earlier measurements being treated as covariates rather than responses, this is appropriate when there are a small number of discrete occasions and where different measures are used at each one.
2. The second model is usually referred to as "repeated measure" models, where all the measurements are treated as responses.

We may also have repetitions at higher levels of the data hierarchy, for example annual examination data on successive cohorts of 16 year old students in a sample of schools. In this case the school is the level 3 unit, year is the level 2 unit and the students are the level 1 unit. In repeated measures models typically most of the variation is at level 2, so the proper specification of a multilevel model for the data is of particular importance. In the models considered so far Goldstein have assumed that the level 1 residuals are uncorrelated, but for some kind of repeated measures data this assumption will not be reasonable.

3.7 A 2-Level Repeated Measures Model

Consider a data set consisting of repeated measurements of heights of a random sample of children, then we can write this model as follows:

$$y_{ij} \;= \beta_{0_j} + \beta_{1_j} x_{ij} + e_{ij}. \tag{10}$$

In this model, Y_{height} is linearly related to age where both *intercept* and the *slope* are treated as random effects. That will cause each subject to have their own intercept and slope such that

$$\begin{aligned} &E(\beta_{0_j}) \;= \beta_0, \quad E(\beta_{1_j}) \;= \beta_1, \\ &\text{var}(\beta_{0_j}) = \sigma^2_{u0}, \ \text{var}(\beta_{1_j}) = \sigma^2_{u1}, \ \text{cov}(\beta_{0_j}, \beta_{1_j}) = \sigma_{u01}, \ \text{var}(e_{ij}) = \sigma^2_e. \end{aligned}$$

Note that there is no restriction on the number of ages, which means that we can fit a single model to subjects who may have only one or several measurements. We can extend Equation (10) to include further explanatory variables measured either at the occasion level (level 1) such as time of year and state of health, or at the subject level (level 2) such as birth weight and gender. It is also possible to extend the basic linear function (10) to include higher order terms, in an attempt to elaborate the model even further it is possible to model the level 1 residual as a function of age.

3.8 LME (Linear mixed-effects) - Model Formulation

R is a free software environment for statistical computing and graphics for performing statistical analysis. The method we used to perform our analysis is multilevel analysis, to perform such an analysis in *R* we used the *nlme* package, which fits and compare Gaussian linear and nonlinear mixed-effects models. The fitting function for linear mixed effects models is *lme*. The lme-procedure has two different types of maximum likelihood fits, 'ML' and 'REML'. Using 'REML' is the model fitted by maximizing the restricted log-likelihood and with 'ML' is the usual log-likelihood maximized. We used 'REML' as the maximum likelihood fits in the analysis in this thesis. Several optional arguments can be used with this function, but a typical call is

$$\text{lme}(\mathit{fixed},\ \mathit{data},\ \mathit{random})$$

The first two arguments to lme, fixed and data, give the model for the expected response (the fixed-effects part of the model) and the object containing the data to which the model should be fit. The third argument, random, is a one sided formula describing the random effects and the grouping structure of the model. For the SGA-data using MAD (mean abdominal diameter) as a response variable and with gestational time centered at the average gestational time for the 4 antenatal visits, these formulas are:

$$\mathit{fixed} = \mathit{MAD} \sim \text{Covariable}_1 + ... + \text{Covariable}_n, \mathit{random} = \ \text{Covariable}_1|\text{fetus}$$

Note that the response variable is specified only in the *fixed* formula. Here is Covariable$_1$ chosen as a level 1 random effect and fetus as a level 2 random effect. The final growth model for MAD is defined in Section 8, the R documentation for the lme-function used in this master thesis is reproduced in Appendix A.

4 Methods for Analyzing Missing Data

4.1 Introduction

I find it necessary to clarify the two different kind of data set used in the analysis. *Complete-case* analysis means that only subjects with a complete set of variables are used in the analysis, and *complete-data* method contain imputed values which is used in the analysis (observed and missing values). The complete case data set is thus a subset of the complete data set.

Missing data can seriously affect the result the analysis, if the missing data are ignored or if one assume that excluding missing data is sufficient there is a risk getting invalid and insignificant results. A missing value is a data value that should have been recorded, but for some reason was not. Missing data are a part of almost all data, and statisticians all have to decide how to deal with it from time to time. Missing values in a data set is a problem because most statistical methods assume that every case has information on all the variables to be included in the analysis. There are alternative ways of dealing with missing data, and this document is an attempt to outline those approaches. Our focus will be on the two general approaches that are highly recommended: Bayesian multiple imputation (MI) and maximum likelihood (ML).

4.2 Types and Patterns of Nonresponse

In longitudinal studies participants may be present for some portion of data collection and missing for others. The kind of missingness may be called *nonresponse* or *dropouts* (a subject leaves the study at some time after which no more measurements are taken), which is a special case of nonresponse and occurs when one leaves the study and does not return, dropouts or attrition may be the most common type of nonresponse. It is however not uncommon for participants to be absent from one time during a study and subsequently reappear

Many data sets can be arranged in a rectangular or matrix form, where the rows correspond to observational units or participants and the columns correspond to items or variables. With rectangular data there are several important classes of overall missing data patterns. Consider Figure 3 from Schafer and Graham (2002) which shows different patterns of nonresponse. Figure 3a in which missing values occur on an item Y but is completely observed on a set of p other items $X_1, ..., X_p$ in a data set, this type of missingness is called a *univariate pattern.* The univariate pattern is also meant to include situations in which Y represents a group of items that is either entirely observed or entirely missing for each unit.

In Figure 3b, subjects or subjects groups $Y_1, ..., Y_p$ may be ordered in such a way that if variable Y_j is missing for a unit, then the subsequent variables $Y_{j+1}, ..., Y_p$ are missing as well. This type of missingness is called a *monotone pattern.* Monotone pattern may arise in longitudinal studies with attrition when Y_j is representing variables collected at the jth occasion.

The third kind of missing pattern that we see in Figure 3c shows an *arbitrary pattern* in which any set of variables may be missing for any unit.

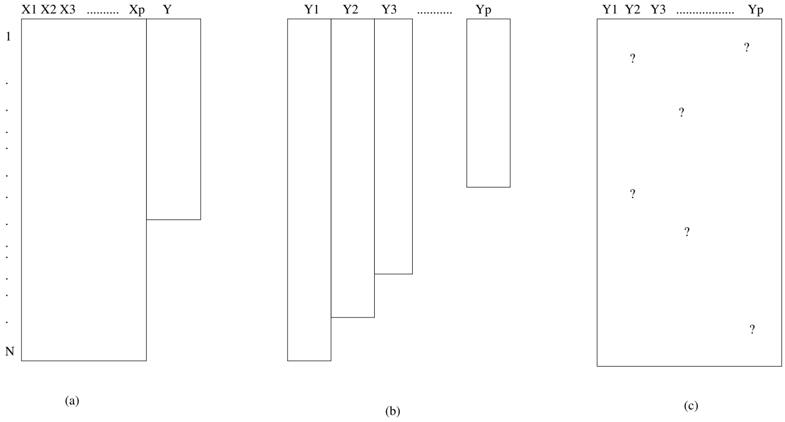

Figure 3: Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables

4.3 Missingness Mechanism

For any data set one can define different indicator variables (Z) which identity what is known and what is missing, therefore is Z being referred to as *the missingness.* The form of the missingness depends on the complexity of the pattern. In Figure 3c Z can be a matrix of binary indicators with the same dimension as the data matrix with elements of Z set to be either one or zero according to whether the corresponding data values are observed or missing. The following terminology were introduced by Rubin (1987) and Little and Rubin (2002), where the missing-data mechanism can be classified according to the probability of response, the missing data are said to be as follows

- **MCAR** - missing completely at random. MCAR means that the probability that Y is missing for a participant does not depend on his/her own values of X or Y, and by independence does not depend on the X or Y of other participants either. If, for instance the probability that income is recorded is the same for all individuals regardless of their age or income itself, then the data are said to be MCAR.

- **MAR** - missing at random. MAR means that the probability that Y is missing may depend on the other variables X_i but not on Y itself.

- **MNAR** - missing not at random. MNAR means that the probability that Y is missing depends on both the Y variable and the other variables X_i.

When Equation (11) is violated and the distribution depends on Y_{mis} then the missing data are said to be *missing not at random* (MNAR). MAR is also called *ignorable non-response* and MNAR is called *non-ignorable* (NI).

Because we often consider real-world reasons why data become missing, if one could code all the different reasons for missingness into a set of variables. This might include variables that explain why some participants were physically unable to show up fore example age and health status, variables that explain the tendency to say "I don't know" or "I'm not sure" and so on. These causes of missingness are not likely to be present in the data set but some of them are possibly related to X and Y and thus by omission induce relationships between X, Y and Z. Other causes may be entirely unrelated to X and Y and may be viewed as external noise. If we let K denote the component of the cause that is unrelated to X and Y, then MCAR, MAR and MNAR may be represented by the following graphical relationships from Schafer and Graham (2002)

The relationships between the different options for MCAR-mechanism shows that MCAR requires the causes of missingness to be entirely contained within the unrelated part K.

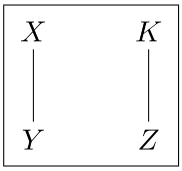

Figure 4: Graphical representation of *missing completely at random* (MCAR) in a univariate missing-data pattern. X represents the variables that are completely observed within the data set, Y represents a variable that is partly missing, K represents the component of the causes of missingness unrelated to X and Y and Z represents the missingness.

The graphical representations of missing at random shows that MAR allows some causes to be related to the observed values X_i.

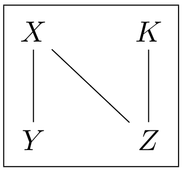

Figure 5: Graphical representation of *missing at random* (MAR) in a univariate missing-data pattern. X represents the variables that are completely observed within the data set, Y represents a variable that is partly missing, K represents the component of the causes of missingness unrelated to X and Y and Z represents the missingness.

The graphical representations of missing not at random shows that MNAR requires some causes to be residually related to Y after the relationships between X and Z are taken into account

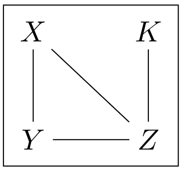

Figure 6: Graphical representation of *missing not at random* (MNAR) in a univariate missing-data pattern. X represents the variables that are completely observed within the data set, Y represents a variable that is partly missing, K represents the component of the causes of missingness unrelated to X and Y and Z represents the missingness.

Note that under MAR there could be a relationship between missingness and Y induced by their mutual relationship to X, but there must be no residual relationship between them once X is taken into account. Under MNAR some residual dependence between missingness and Y remains after accounting for X. Z is being treated as a set of random variables having a joint probably distribution, but it is not necessary to specify a particular distribution. To describe accurately all potential causes or reasons for missingness is not realistic. The distribution of Z is best regarded as a mathematical device to describe the rates and patterns of missing values and to capture roughly possible relationships between the missingness and the values of the missing items themselves.

Because missingness may be related to the data we classify distributions for Z according to the characteristics of the relationship. Let the complete set of data be denoted as Y_{com} and portion it as $Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} and Y_{mis} are the observed and missing parts of the data. Rubin (1976) defined the missing data to be *MAR* if the distribution of missingness does not depend on Y_{mis}, which means that *MAR* allows the probabilities of missingness to depend on observed data but not on missing data.

$$P(Z|Y_{\text{com}}) = P(Z|Y_{\text{obs}}). \tag{11}$$

An important special case of MAR, called *missing completely at random* (MCAR), occurs when the distribution does not depend on Y_{obs} either such that

$$P(Z|Y_{\text{com}}) = P(Z).$$

Processes that are neither *MCAR* or *MAR* are called *MNAR*, in which the probability of dropouts depends on the unobserved measurements.

4.4 Single Imputation

Both complete-case and available-case analysis make no use of cases where missing values (Y_j) occurs when the marginal distribution of Y_j or measures of covariation between Y_j and the other variables are estimated. When a unit provides partial information it is tempting to replace the missing items with plausible values and proceed with the desired analysis rather than discard the unit entirely. The *imputation-method* which are filling in missing items has several desirable features, it is potentially more efficient than case deletion because no units are sacrificed performing the analysis. Retaining the full sample-size helps us to prevent loss of power compared with a reduced sample size, which may give parameter estimates which are biased and may have bigger or smaller standard deviation than one gets performing the analysis with the full sample-size. If the observed data contain useful information for predicting the missing values, the imputation procedure would make use of this information estimating the missing values and maintain high precision.

The methods that impute the values of items that are missing will now be discussed. These methods can be applied to impute one value for each missing item (single imputation) or in some cases impute more than one value to allow appropriate assessment of imputation uncertainty (multiple imputation). Note that with single imputation there is no simple way to reflect the missing data uncertainty. Imputation is a general and flexible method for handling missing-data problems, but it has pitfalls. In the words of Dempster and Rubin (1983);

The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can legitimately handled in this way and situations where the standard estimators applied to the real and imputated data have substantial biases.

Imputations which are used are means or values drawn from a predictive distribution of the missing values and require an imputation-method of creating a predictive distribution for the imputation based on the observed data. There are *two* generic approaches to generating this distribution:

Little and Rubin (2002) defined *Explicit modeling* as the predictive distribution which is based on a formal statistical model (e.g. multivariate normal) and hence the assumptions are explicit. There exists several explicit imputation methods

- *Mean imputation* - where means from the responding units in the sample are substituted, the means may be formed within subjects or classes analogous to the weighting classes. Mean imputation then leads to estimates similar to those found by weighting provided the sampling weights are constant within weighting classes.

- *Regression imputation* - replaces missing values by predicted values from a regression method on missing items observed for the unit. This method is usually calculated from units with both observed and missing variables present. The model is first fit for the subjects for which Y is known, then the values of X from the non-correspondents are plugged into the regression-method to obtain predicted values for $\hat{Y}$. Note that mean imputation method can be regarded as a special case of the regression-method where the predictor variables are dummy indicator variables for the observations within which the means are imputed.

- *Stochastic regression imputation* - replaces missing values by a value predicted by regression plus a residual which is drawn to reflect the uncertainty in the predicted value. With normal linear regression models the residuals will be normally distributed with zero mean and variance equal to the residual variance in the regression.

Little and Rubin (2002) defined *Implicit modeling* as follows: the focus is on an algorithm which implies an underlying model and the assumptions are implicit but they still need to be carefully assessed to ensure that they are reasonable. There exists several implicit imputation methods:

- *Hot deck imputation* - involves substituting individual values which is drawn from "similar" responding units. Hot deck imputation is common in survey practice and may involve very elaborate schemes for selecting units that are similar for imputation, it has no parametric model.

- *Substitution* - is a method for dealing with nonresponse units at the "fieldwork" stage of a survey, and replaces the non-responding units with alternative units which originally were not selected into the sample. For example if a household cannot be contacted, then a previously nonselected household in the same housing block may be substituted. The tendency to treat the resulting sample as complete should be resisted since the substituted units are respondents and hence may differ systematically from nonrespondents who were originally selected. Hence at the analysis stage substituted values should be regarded as imputed values of a particular type.

- *Cold deck imputation* - replaces a missing values of an item by a constant value from an external source such as a value from a previous realization of the same survey. As with substitution, current practice usually treats the resulting data as a complete sample which ignores the consequences of imputation.

- *Composite methods* - can also be defined as a combination from different methods above. For example hot deck and regression imputation can be combined by calculating predicted means from a regression but then adding a residual randomly chosen from the empirical residuals to the predicted value when forming values for the imputation.

4.5 Multiple Imputation (MI)

Since the theoretical motivation for multiple imputation is Bayesian will a short introduction to the Bayesian way of argumentation be given here first.

4.5.1 Bayesian Methodology

The full process of an typical Bayesian analysis can be described as consisting of three main steps (Gelman et al., 1995a)

(a) setting up a full probability model (the joint distribution) which captures the relation ships among all the variables (e.g observed data, missing data and the unknown parameters) in the consideration.

(b) summarizing the findings for particular quantities of interest by appropriate posterior distributions which is a conditional distribution of the quantities of interest given the observed data.

(c) evaluating the appropareness of the model and suggesting improvements.

The Bayesian statistics is based on specifying a probability model for the observed data U with a joint density $f_{u|\Theta}(\theta|u)$ giving a vector of the unknown parameters $\Theta = \theta$ which is identical to the likelihood function $L(\theta; u)$ understood as a function of θ. Then we assume that Θ is a random parameter instead of treating θ as an unknown constant as in a frequentist approach, and it has a prior density or probability functions f_Θ. This prior is typically regarded as known to the researcher independently of the data under the analysis. Inference about Θ is then summarized in the function $f_{\Theta|U}$, which is called the posterior distribution of Θ given the data. The posterior distribution is derived from the joint distribution $f_{U,\Theta} = f_{U|\Theta} f_\Theta$ according to Bayes' formula

$$f_{u|\Theta}(\theta|u) = \frac{f_{u,\Theta}(\theta, u)}{f_U(u)} = \frac{f_{U|\Theta}(u|\theta) f_\Theta(\theta)}{\int_\Omega f_{\Theta,U}(\theta, u) d\theta} = \frac{L(\theta; u) f_\Theta(\theta)}{\int_\Omega L(\theta; u) f_\Theta(\theta) d\theta},$$

where Ω denotes the parameter space of Θ. Notice that from a Bayesian perspective the joint distribution $f_{U|\Theta}(u|\theta)$ equates the likelihood $L(\theta; u)$ when the data are observed and only Θ is still variable. The Bayesian approach has at least two advantages

- First, through the prior distribution where we can inject our prior knowledge and information on the value of Θ

- Second, treating all the variables in the system as random variables will greatly clarifies the methods of analysis.

If we are interested in only one of the components in Θ from the posterior distribution we only have to integrate out the other remaining components.

4.5.2 Multiple Imputation Paradigm

MI has emerged as a flexible alternative to likelihood methods for a wide variety of missing-data problems. MI retains much of the attractiveness of single imputation from a conditional distribution but solves the problem of understating uncertainty. When the MI-method is executed each missing value is replaced by a list of $m > 1$ simulated values as shown in Figure 7 from Schafer and Graham (2002) p. 165. By substituting the jth element of each list for the corresponding missing value, $j = 1, ..., M$ produces M plausible alternative versions of the complete data set, and each of the M data sets are then analyzed by the same complete-data method. The results which may vary, are then combined by simple arithmetic to obtain the overall estimates and standard errors that reflect missing-data uncertainty as well as finite sample variation.

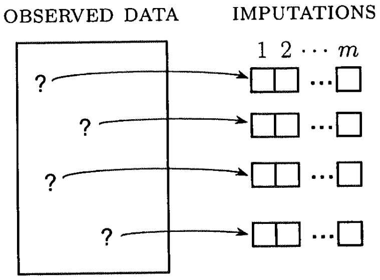

Figure 7: Schematic representation of multiple imputation from Figure 4 in Schafer and Graham (2002), where $m = M$ is the number of imputations which is estimated by each missing value in the original data set.

In Schafer (2003) he states that all currently available *MI* programs assumes that the missingnes is *MAR*. That means that the MI-method which is included in the *Pan* package used in this master thesis uses the *MAR* assumptions.

4.5.3 Rules for MI Inference

The analysis of a multiply-imputed data set is quite direct. Each of the M data set which are completed by imputation is analyzed using the same *complete-data* method that would be used in the *complete-case* situation. Let $\hat{\theta}_j$, W_j, $j = 1, ..., M$ be M *complete-data* estimates and their associated variances for an estimated parameter θ, which is computed from each of the M repeated imputations under one model.

Note that the MI-method assumes that the sample is large enough so that $\sqrt{W}(\hat{\theta} - \theta)$ has approximately a standard normal distribution, so that $\hat{\theta} \pm 1.96\sqrt{W}$ has about 95% coverage. We can not compute $\hat{\theta}$ and W, because we have M different versions of them, $[\hat{\theta}^{(j)}, W^{(j)}]$, $j = 1, ..., M$. By using Rubin's (1987) rules, we proceed as follows: the MI estimate or the overall estimate is the average of the M different estimates,

$$\overline{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \qquad (12)$$

Since the imputations involved in MI are conditional draws rather than conditional means will they under a good imputation model provide valid estimates for a wide range of estimates. Averaging over M imputations from data sets in Equation (12) increases the efficiency of an estimate obtained from single a data set with imputed conditional draws. The variability associated with this estimate has two components. The first component is the average within-imputation variance

$$\overline{W}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{W}_m, \tag{13}$$

and between-imputation component

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\hat{\theta}})^2. \tag{14}$$

The total variability associated with $\overline{\theta}_M$ is a modified sum of the two components.

$$T_M = \overline{W}_M + \frac{M+1}{M}B_M, \tag{15}$$

where $(1+\frac{1}{M})$ is an adjustment for finite M. Hence

$$\hat{\gamma}_M = \frac{1+\frac{1}{M}B_M}{T_M} \tag{16}$$

is an estimate of the fraction of information about θ missing due to nonresponse. For large sample sizes and a scalar θ will the reference distribution for interval estimates and significance tests be a t distribution

$$(\hat{\theta}_m - \overline{\theta}_M)T_M^{-\frac{1}{2}} \sim t_\nu, \tag{17}$$

where the degrades of freedom are given by

$$\nu = (M-1)(1+\frac{1}{M+1}\frac{\overline{W}_M}{B_M})^2. \tag{18}$$

Barnard and Rubin (1999) have later improved the expression for degrees of freedom for small data sets, they relax the assumption of a normal reference distribution of $\sqrt{W}(\hat{\theta}-\theta)$ for the complete-data interval estimates and tests to allow for a t distribution, and they derive the corresponding degrees of freedom for the MI-inference to replace the formula, Equation (18), given here. In their article they had made a simulation study which demonstrates the superior frequentist performance when they used $\tilde{\nu}_m$ rather than ν_m. MI has many attractive features, like single imputation it allows the analyst to proceed with familiar complete-data techniques and software. One good set of M imputations may effectively solve the missing data problems in many analyses, and one does not necessarily need to re-impute before computing every new analysis.

4.5.4 Why Only a Few Imputations Are Needed

Unlike other Monte Carlo methods, with MI we do not need a large number of repetitions to obtain precise estimates. There are two fundamental reasons for this. First, like Rao-Blackwellzation, multiple imputation relies on simulation to solve only the missing-data aspect of the problem. As with any simulation method one could effectively eliminate Monte Carlo error by choosing M to be very large, but with multiple imputation the resulting gain in efficiency would typically be unimportant because the Monte Carlo error is relatively small portion of the overall inferential uncertainty. Little and Rubin (2002) p.114 shows that the efficiency of an estimate based on M imputations, relative to one based on an infinite number is $(1+\frac{\lambda}{M})^{(-1)}$, where λ is the rate of missing information. λ is distinct from the rate of missing observations and measures the increase in the large-sample variance of a parameter estimate due to missing values. It may be greater or smaller than the rate of missing values in any given problem. For example with 50% missing information an estimate based on $M = 3$ imputations has a standard error which is about 8% higher than one based on $M = \infty$, because $\sqrt{1+\frac{0.5}{3}} = 1.0801$. Schafer (1999) states that unless the fraction of missing information is unusually high (i.e., far more than 50%), there is little benefit in using more than five to ten imputations.

The second reason why we can often obtain valid inferences with a very small M is that the rules for combining the M complete-data analyses explicitly account for Monte Carlo error. A multiple imputation interval estimate makes provisions for the fact that both the point and variance estimates contain a predictable amount of simulation error due to finiteness of M, and the width of the interval in accordingly adjusted to maintain the appropriate probability of coverage.

4.5.5 Proper MI-Procedure

An essential part of the MI-method is treating the parameters as random rather than fixed, and the validity of *MI* rests on how the imputation-procedures are created and how the procedure relates the model which is used to subsequently analyze the complete dataset. Creating complete data set with the MI-procedure often requires special algorithms for estimating the missing-values. In general the missing-values should be drawn from a distribution for the missing data which reflects uncertainty about the parameters of the model. With single imputation it is desirable to impute from the conditional distribution $P(Y_{\text{mis}}|Y_{\text{obs}};\hat{\theta})$, where $\hat{\theta}$ is an estimate derived from the observed data. MI extends this by first simulating M different and independent plausible values for the parameters $\theta^{(1)},...,\theta^{(M)}$, and then drawing the missing data $Y_{\text{mis}}^{(t)}$ from $P[Y_{\text{mis}}|Y_{\text{obs}};\ \theta^{(t)}]$ for $t = 1,....,M$.

4.5.6 Using Auxiliary Variables in the MI-Method

Let $Y_1,\ Y_2,...,\ Y_p$ denote the most significant variables observed in the *SGA-project* used in the MI-method, if missing values occur in any of these variables there will be other variables (partially or fully observed) $X_1,\ X_2, ...,\ X_p$ which may potentially contain useful information for predicting the missing values. If so, they may be included in the MI-procedure but they will be excluded from the subsequent analysis. In Collins et al. (2001) they classify the auxiliary variables into three different types

- Type A-variables, which are correlated with the outcomes $Y_1,\ Y_2, ...,\ Y_p$ and may help to explain why $Y_1,\ Y_2, ...,\ Y_p$ are missing. That means that they are related to the *missingness.*
- Type B-variables are correlated with the outcomes $Y_1,\ Y_2, ...,\ Y_p$, but they are unrelated to the *missingness.*
- Type C-variables are unrelated to any of $Y_1,\ Y_2, ...,\ Y_p$.

From Section 4.5 we know that Schafer (2003) has verified that all current available MI programs assumes that the missingness is *MAR*. Some of the major findings Collins et al. (2001) made in their article were; Because the Type A variables were correlated with the missingness it will cause that the *MAR* assumption is violated if they are excluded from the imputation procedure, and therefore by include them in the imputation procedure may help reducing the bias. Type B variables will not reduce bias under *MAR*, but they may increase the precision of the parameter estimates because they contain useful information for predicting the missing values for the $Y_1,\ Y_2, ...,\ Y_p$ variables. Type B variables can under MNAR conditions instead of the MAR situation may help reducing the bias. Type C variables which is the last of the three groups will neither reduce the bias or the precision of the parameter estimates. They will only make the imputation model unnecessary large and complicated, therefore it is no use taking they into consideration.

In Collins et al. (2001) they discuss when it is beneficial to use the auxiliary variables in this fashion and if there are any potential dangers in doing so in depth, using different simulation studies. In one set of the simulations Collins et al. (2001) showed that the biases incurred by failing to include Type A variables in the imputation procedure were not as serious as some have previously thought. By including or excluding a type A variable X_k made little difference unless the correlation between X_k and an the outcome Y_j was unusually strong (much greater than 0.4), and the rate of missing values in Y_j was very high (50% or more missing). With weaker correlations and lower rates of missing values were biases in parameters related to Y_j and its relationships to other variables barely noticeable when X_k was excluded from the imputation procedure.

In another set of simulation Collins et al. (2001) showed that by including a type B variable can substantially increase the precisions under *MAR* conditions if its correlation with the outcomes is very strong (approximately 0.9). This situation is very simular to our project, because in longitudinal studies there are some measured variables which are repeated on individuals over time, which will cause the variables to be highly correlated. Responses at one occasion may be very useful for imputing missing responses at another.

Finally, Collins et al. (2001) showed that the costs of unnecessarily including type C variables into the imputation procedure tend to be minimal. Overall there were potentially important gains and small risks associated with auxiliarly variables in the MI-method.

4.6 Imputation-Model Used in MI

Linear mixed-effects models are mixed-effects models in which both the fixed and the random effects occur linearly in the model function. An extension of linear models may be obtained by incorporating random effects, which can be regarded as additional error terms to account for the correlation among the observations within the same group. We have used the same procedure in estimating the missing values using a Gibbs sampler as Schafer and Yucel (2002).

In the imputation model y_i is an $n_i * r$ matrix of multivariate responses for sample unit i, $i = \{1, 2, ..., M\}$, where each row of y_i is a joint realization of the variables $Y_1, Y_2, ..., Y_r$. The model we have used for the complete data is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{19}$$

where X_i $(n_i * p)$ and Z_i $(n_i * q)$ are known covariate matrices, β $(p * r)$ is a matrix containing the regression coefficients which is common to all the units in the data-set, and b_i $(q * r)$ is a matrix of coefficients which are specific for each unit i. β and b_i may also be called fixed and random effects. We assume that the n_i rows of ϵ_i are independently distributed as $N(0, \Sigma)$, and the random effects are distributed as $\text{vec}(b_i) \sim N(0, \Psi)$ independently for $i = \{1, 2, ..., n\}$. (Note that the "vec" operator vectorizes a matrix by stacking its columns). The DAG of the hierarchic model made in Winbugs [2] can be modeled as follows

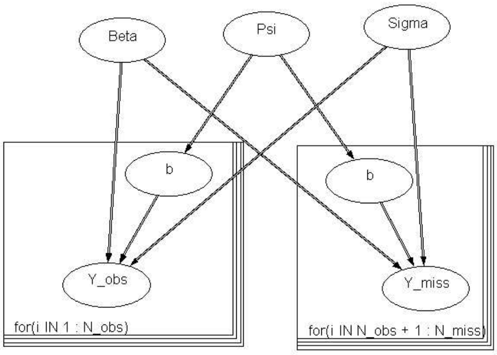

Figure 8: DAG of the model ($y_i = X_i\beta + Z_i b_i + \epsilon_i$) for the complete data.

[2] Winbugs is a freely available software for constructing Bayesian statistical models and evaluating them using MCMC methods

Where

$$b_i|pa(i) = f(b_i|\Psi) = \frac{f(b_i,\Psi)}{f_\Psi(\Psi)} \sim N(0,\Psi) \tag{20}$$

$$y_i|pa(i) = f(y_i|\beta,\Sigma,\Psi,b_i) = \frac{f(y_i,\beta,\Sigma,\Psi,b_i)}{f(b_i|\Psi)f(\Psi)f(\beta)f(\Sigma)} \tag{21}$$

Without conditioning on $b_1, ..., b_m$ the implied model used for $\text{vec}(y_i)$ is normally distributed with mean equal to $\text{vec}(X_i\beta)$ and a covariance matrix

$$W_i^{-1} = (I_r \otimes Z_i)\Psi_i(I_r \otimes Z_i)^T \; + \; (\Sigma \otimes I_{n_i}). \tag{22}$$

Note because of β, Ψ, Σ is being drawn from a multivariate distribution, this will result in that the mixing is getting better. In longitudinal applications such in our situation, gestational times of measurement may be incorporated into the X_i matrix used as a predictor and the Z_i matrix (or a vector) which allows the relevant aspects of the growth curves (e.g. intercepts and slopes) to vary by each subject in the data-set.

4.6.1 The Gibbs Sampler Used in the MI-Method

A Gibbs sampler is a *MCMC* procedure and an iterative simulation algorithm in which current values of the unknown parameters are drawn from the conditional distribution of the parameter given the last updated values of all the other parameters, this is being called the full-conditional distribution. The missing values $Y_{\text{mis}}^{(t)}$ are updated in the three following steps, given the starting values for $\beta^{(0)}, \Sigma^{(0)}, \Psi^{(0)}$ and the missing values $Y_{\text{mis}}^{(0)}$. First,

$$\begin{aligned} b_i^{(t+1)} &= \pi(b_i^{(t+1)}|Y_{\text{obs}}, Y_{\text{mis}}^{(t)}, \beta^{(t)}, \Sigma^{(t)}, \Psi^{(t)}) \\ &\propto \pi(Y_{\text{obs}}, Y_{\text{mis}}^{(t)}|b_i^{(t+1)}, \beta^{(t)}, \Sigma^{(t)})\pi(b_i^{(t+1)}|\Psi^{(t)}) \end{aligned} \tag{23}$$

independently for $i = \{1, 2, ..., m\}$;next

$$\begin{aligned} \beta_i^{(t+1)} &= \pi(\beta_i^{(t+1)}|Y_{\text{obs}}, Y_{\text{mis}}^{(t)}, B^{(t+1)}, \Sigma^{(t)}, \Psi^{(t)}) \\ &\propto \pi(Y_{\text{obs}}, Y_{\text{mis}}^{(t)}|B_i^{(t+1)}, \beta^{(t+1)}, \Sigma^{(t)})\pi(\beta_i^{(t+1)}) \end{aligned} \tag{24}$$

$$\begin{aligned} \Psi_i^{(t+1)} &= \pi(\Psi_i^{(t+1)}|Y_{\text{obs}}, Y_{\text{mis}}^{(t)}, B^{(t+1)}, \beta^{(t+1)}, \Sigma^{(t)}) \\ &\propto \pi(B^{(t+1)}|\Psi^{(t+1)})\pi(\Psi^{(t+1)}) \end{aligned} \tag{25}$$

$$\begin{aligned} \Sigma_i^{(t+1)} &= \pi(\Sigma_i^{(t+1)}|Y_{\text{obs}}, Y_{\text{mis}}^{(t)}, \beta^{(t+1)}, B^{(t+1)}, \Psi^{(t+1)}) \\ &\propto \pi(Y_{\text{obs}}, Y_{\text{mis}}^{(t)}|B_i^{(t+1)}, \beta^{(t+1)}, \Psi^{(t+1)})\pi(\Sigma^{(t+1)}) \end{aligned} \tag{26}$$

and finally,

$$\begin{aligned} Y_{\text{mis}}^{(t+1)} &= \pi(Y_{\text{mis}}^{(t+1)}|Y_{\text{obs}}, B^{(t+1)}, \beta^{(t+1)}, \Sigma^{(t+1)}, \Psi^{(t+1)}) \\ &\propto \pi(Y_{\text{mis}}^{(t+1)}, Y_{\text{obs}}|B^{(t+1)}, \beta^{(t+1)}, \Sigma^{(t+1)}, \Psi^{(t+1)}) \end{aligned} \tag{27}$$

for $i = \{1, 2, ..., m\}$. *These three steps form one cycle of the Gibbs sampler algorithm.* Executing the *cycle* repeatedly creates sequences of the parameters contained in $\{\theta^{(1)}, \theta^{(2)}, ..., \theta^{(m)}\}$,where $\theta^{(t)}$ consists of $\beta^{(t)}, \Sigma^{(t)}$and $\Psi^{(t)}$, and for the missing values $\{Y_{\text{mis}}^{(1)}, Y_{\text{mis}}^{(2)}, ..., Y_{\text{mis}}^{(m)}\}$, whose limiting distributions are $P(\theta|Y_{\text{obs}})$ and $P(Y_{\text{mis}}|Y_{\text{obs}})$.

Implementing the second step of the cycle requires a prior distribution for θ. It is known from their article and in general that mixed-effects models with improper prior distributions for the covariance components may lead to Gibbs samplers that do not converge to proper posterior distributions, even though if each step of the *cycle* is well defined. Therefore are proper prior distributions for the covariance matrices are important to achieve satisfactory results.

We also need to specify the prior distributions for the covariance matrices Sigma and Psi. Schafer and Yucel (2002) applied the independent inverted Wishart priors for Σ^{-1} $W(\nu_1, \Lambda_1)$ and Φ^{-1} are distributed as $W(\nu_2, \Lambda_2)$, where $W(\nu, \Lambda)$ denotes a Wishart distribution with $\nu > 0$ degrees of freedom and mean $\nu\Lambda > 0$ because this priors are appropriate for a model with unstructured Ψ. These priors exist provided that $\Lambda_1, \Lambda_2 > 0$, $\nu_1 \geq r$ and $\nu_2 \geq qr$, where q are the number of random effects and r are the number of response variables (in our case $q = 2$ and $r = 2$). Choosing values for the hyperparameters it is helpful to regard $\nu^{-1}\Lambda^{-1}$ and $\nu^{-2}\Lambda^{-2}$ as prior guesses for Σ and Ψ with confidence equivalent to ν_1 and ν_2 degrees of freedom. Note that small prior guesses for the values for ν_1 and ν_2 make the prior densities relatively diffuse and are reducing their impact on the final inferences. For the β-parameters Schafer and Yucel (2002) uses an improper uniform density over $\mathcal{R}^{\text{pr}}$.

Under these priors guesses each of the steps in the *cycle*, Equations (23) - (27), is derived by straightforward application of Bayes' theorem. In our model are the pairs (y_i, b_i) distributed as:

$$\begin{aligned} \text{vec}(y_i)|b_i, \theta &\sim N(\text{vec}(X_i\beta + Z_i b_i), (\Sigma \otimes I_{n_i})), \\ \text{vec}(b_i)|\theta &\sim N(0, \Psi) \end{aligned}$$

independently for $i = \{1, 2, ..., m\}$. It follows that

$$\text{vec}(b_i|y_i, \theta) \sim N(\text{vec}(\tilde{b}_i), U_i),$$

where

$$\begin{aligned} \text{vec}(\tilde{b}_i) &= U_i(\Sigma^{-1} \otimes Z_i^T)\text{vec}(y_i - X_i\beta), \\ U_i &= (\Psi^{-1} + (\Sigma^{-1} \otimes Z_i^T Z_i))^{-1}. \end{aligned}$$

Simulation of the parameters contained in θ (step two of the cycle) proceeds as follows:

- First draw Ψ^{-1} from a Wishart distribution with $\nu_2^{'} = \nu_2 + m$ degrees of freedom and using the scale $\Lambda_2^{'} = (\Lambda_2^{-1} + B^T B)^{-1}$.

- Next, the ordinary least-squares coefficients $\hat{\beta} = (\sum_{i=1}^{m} X_i^T X_i)^{-1}(\sum_{i=1}^{m} X_i^T(y_i - Z_i b_i))$ and the residuals $\hat{\epsilon}_i = y_i - X_i\hat{\beta} - Z_i b_i$ will be calculated. Then Σ^{-1} will be drawn from a Wishart distribution with $\nu_1^{'} = \nu_1 - p + \sum_{i=1}^{m} n_i$ degrees of freedom and scale $\Lambda_1^{'}$ equal to $(\Lambda_1^{-1} + \sum_{i=1}^{m} \hat{\epsilon}_i^T \hat{\epsilon}_i)^{-1}$.

- Finally will β be drawn from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\Sigma \otimes V$, where $V = (\sum_{i=1}^{m} X_i^T X_i)^{-1}$.

For simulating the values of the parameters (β) it is very helpful to note that if G and H are upper-triangular square roots of Σ and V. ($G^T G = \Sigma$ and $H^T H = V$), then $G \otimes H$ is an upper-triangular square root of $\Sigma \otimes V$.

Notice from executing the final step, Equation (27), of the Gibbs sampler that the rows of $\epsilon_i = y_i - X_i\beta - Z_i b_i$ are independent and normally distributed with mean zero and covariance matrix Σ. Therefore in any row of ϵ_i the missing elements have an intercept-free multivariate normal regression on the observed elements; the slopes and the residual covariances for this regression can be quickly calculated by inverting the square sub-matrix of Σ corresponding to the observed variables. Drawing the missing elements in ϵ_i from these regressions and adding them to the corresponding elements of $X_i\beta + Z_i b_i$ completes the simulation of $y_{i_{\text{mis}}}$.

4.6.2 Implementation Issues

If any of the steps, Equations (23) - (27), from the *cycle* could be carried out without conditioning on the simulated values of Y_{mis} or B, would cause the algorithm to converge in fewer iterations. With modern computers iterations contained in the *cycle* can perform quickly even with large datasets provided that sufficient physical memory is available to store the data, which consists of $Y_{\text{obs}}, Y_{\text{mis}}^{(t)}$ and the covariate matrices X_i and Z_i. The convergence behavior of this algorithm is governed by two factors:

- the amount of information about θ carried in the missing values relative to observed ones

- and the degree to which the random effects b_i can be estimated from the values contained in the y_i matrix.

If the missing portions of y_i exert a high leverage over the components contained in θ or if the the random effects b_i are poorly estimated for example if the within-unit precision matrices of $\Sigma^{-1} \otimes Z_i^T Z_i$ tend to be small relative to Ψ^{-1}) then convergence can be slow. Convergence may also be slow when the number of subjects m is large because for large m the posterior distribution for Ψ given $b_1, b_2, ..., b_m$ becomes very tight and causing the drawn value for Ψ to be correlated with its previous value. If the parameters and values produced from the multiple imputations method have slow convergence, it is not catastrophic because in most situations we only need a few independent draws of Y_{mis}. Assessing convergence of the MI-method can done by examining time-series plots of the parameters and autocorrelation-plots (ACF) for individual elements or functions of θ, one should in particular pay close attention to the elements of Ψ matrix (random effects) because these parameters tend to exhibit high correlations between the values estimated at each iteration.

Any rows of the y_i matrix that is completely missing may be omitted from consideration along with the corresponding rows of X_i and Z_i without changing the form of complete-data model, Equation (19), by ignoring these rows we will eliminate any unnecessary computation at each of the three steps in the *cycle* and simultaneously reducing the rate of missing information and speeding the overall rate of convergence. The deleted rows may be restored at the final imputation step, Equation (27), to produce a fully completed dataset to be used in the further analysis.

4.6.3 Prior Guesses and Alternative Covariance Structures

When specifying values for the hyper-parameters the usual practice is to set $\nu_1 = r$ and $\nu_2 = qr$ to make the priors as dispersed as possible and minimize their subjective influence. We will typically set the initial values of Λ_1^{-1} equal to $\nu_1\hat{\Sigma}$ and Λ_2^{-1} equal to $\nu_2\hat{\Psi}$ where $\hat{\Sigma}$ and $\hat{\Psi}$ are reasonable guesses for Σ and Ψ. As a prior guess estimate of Σ we will use the $(r * r)$ identity matrix and for Ψ we use the $(rq * rq)$ identity matrix.

When modeling a large number of response variables it may be advantageous to restrict Ψ to have a block-diagonal structure not only for the purpose of obtaining prior guesses but also when running the Gibbs sampler itself. If Ψ is block-diagonal then independent inverted Wishart prior distributions may be applied to the $q * q$ nonzero blocks, $\Psi_j^{-1} \sim W(\nu_j, \Lambda_j)$ for $j = 1, 2, ..., r$. We obtain weak priors by setting $\nu_j = q$ and $\Lambda_j^{-1} = \nu_j\hat{\Psi}_j$, where $\hat{\Psi}_j$ is an estimate or our prior guess for Ψ_j. The distribution for these blocks in step two of the cycle become $\Psi_j^{-1} \sim W(\nu_j, \Lambda_j)$, where $\nu_j^{'} = \nu_j + m$, $\Lambda_j^{'-1} = \Lambda_j^{-1} + \sum_{i=1}^{m} b_{ij}b_{ij}^{T}$ and b_{ij} is the jth column of b_i.
The choice of selecting between an unstructured or a block-diagonal Ψ will depend on both theoretical and practical considerations. A block diagonal structure indicates that there exists no a priori associations between the random effects for any two response variables in the y_j matrix and in $y_{j'}$.

4.7 Maximum Likelihood (ML) Estimation)

Most of the theory from this section is based upon Little and Rubin (2002) (Chapther 8) and by Schafer and Graham (2002).
The principle of drawing inferences from a likelihood function is widely accepted. Under the MAR assumption will the marginal distribution of the observed data provide the correct likelihood for the unknown parameter θ, provided that the model for the complete data is realistic ($P(Y_{\text{obs}}; \theta) = \int P(Y_{\text{com}}; \theta)dY_{\text{mis}}$). Schafer and Graham (2002) defines this

as the *observed-data likelihood*, and the logarithm of this function,

$$l(\theta; Y_{\text{obs}}) = \log L(\theta; Y_{\text{obs}}), \tag{28}$$

plays a vital role in the estimation-procedure. The *ML* estimate of $\hat{\theta}$ is the value of θ for which Equation (28) has it highest values, and it has attractive theoretical properties just as it does in the complete-data problems. Under rather general regularity conditions $\hat{\theta}$ tends to be approximately unbiased and highly efficient when the sample size is large, and as the sample size increase its variance approaches the theoretical lower bound of what is achievable by any unbiased estimator (e.g., Casella and Berger (2001) p. 337). Confidence intervals and regions are often computed by the assumptions that $\hat{\theta}$ is approximately normally distributed about the true parameter θ with approximate covariance matrix

$$V(\hat{\theta}) \approx [-\, l\,''(\hat{\theta})]^{-1}, \tag{29}$$

where $l\,''(\hat{\theta})$ is the matrix of second partial derivate of Equation (28) with respect to the elements of θ. In Schafer and Graham (2002) the matrix $-\, l\,''(\hat{\theta})$ is referred to as the *observed information*, which describes how quickly the *log-likelihood* function drops as we move away from the *ML* estimate. Sometimes this matrix is replaced by its *expected information* or *Fisher information* because the expected value is sometimes easier to compute. In complete-data problems the approximation, Equation (29), is still valid when the observed information is replaced by the expected information, earlier Kenward and Molenberghs (1998) have shown that this is not necessary true with missing data. Expected information implicitly uses Equation (28) as the sampling distribution for Y_{obs}, which is only valid if the missing data are *MCAR*. If we want to obtain standard errors and confidence intervals that are valid under the general *MAR* condition in missing-data problems, Schafer and Graham (2002) suggests that we should base them on the observed rather than the expected information matrix.

Log-likelihood also provides methods for testing hypotheses about elements or functions of θ. If we wish to test the null hypothesis, θ lies in a certain area or region of the parameter space versus the alternative that it does not. Under suitable regularity conditions this test may be performed by comparing a difference in the *log-likelihood* with a chi-square distribution. We would reject the null hypothesis at the designed α-level if $2[l(\hat{\theta}; Y_{\text{obs}}) \,-\, l(\tilde{\theta}; Y_{\text{obs}})]$ exceeds the $100(1-\,\alpha)$ percentile of the chi-square distribution. The degrees of freedom are given by the difference in number of free parameters under the null and alternative hypothesis, that is the number if restrictions that must be placed on the elements of θ to ensure that it lies in the null region. These likelihood-ratio tests are very attractive for missing-data problems because they only require that we are being able to compute the maximizers $\hat{\theta}$ and $\tilde{\theta}$, and no second derivatives are needed. In a few problems the maximizer of the log likelihood can be computed directly.

4.8 The Expectation Maximation (EM) - Method

The key idea behind *EM* is to solve a difficult incomplete-data estimation problem by repeatedly solving tractable complete-data problems. We “fill in the missing data” with a best guess under the current estimate of the unknown parameters from the observed and filled in data. EM algorithm is an alternative computing strategy for incomplete-data problems which does not require second derivatives to be calculated or approximated. The fact that Y_{mis} contains information which is relevant to estimate θ and θ in turn helps us to find likely values of the missing values (Y_{mis}) which helps us to suggest the following scheme for estimating θ in the presence of Y_{obs} alone:

1. Replace the missing values by estimated values.

2. Estimate the parameters.

3. Re-estimate the missing values assuming that the new parameter estimates are correct.

4. Re-estimate the parameters, and so forth iterating until convergence.

In any incomplete-data problem can the distribution of the complete-data Y_{com} be factored as follows

$$f(Y|\theta) = f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) = f(Y_{\text{obs}}|\theta)f(Y_{\text{mis}}|Y_{\text{obs}}, \theta),$$

where $f(Y_{\text{obs}}|\theta)$ is the density of the observed data Y_{obs} and $f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ is the density of the missing data given the observed data. The corresponding decompositions of the log-likelihood as a function of θ can be written as

$$l(\theta|Y) = l(\theta|Y_{\text{obs}}, Y_{\text{mis}}) = l(\theta|Y_{\text{obs}}) + \log f(Y_{\text{mis}}|Y_{\text{obs}}, \theta), \tag{30}$$

where $l(\theta|Y) = \log f(Y|\theta)$ denotes the complete-data log-likelihood , $l(\theta|Y_{\text{obs}}) = \log f(\theta|Y_{\text{obs}})$ denotes the observed-data log-likelihood and the term $f(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ may be called *the predictive distribution of the missing data given* θ which plays a central role in the *EM-method* because it captures the uavhengighet between the missing values and the parameters (θ). When it is viewed as a probability distribution it summaries the knowledge about Y_{mis} for any assumed value of θ, and when it is viewed as a function of θ it brings the evidence about θ contained in the missing values beyond that it is already provided by the observed values (Y_{obs}). The aim is to estimate the parameters by maximizing the incomplete-data log-likelihood $l(\theta|Y_{\text{obs}})$ with respect to θ for the observed values Y_{obs}, because Y_{mis} is unknown we cannot calculate the second term in the right-hand side of Equation (30). Rubin solves this problem by performing the following procedure, first he rewrites Equation (30) as

$$l(\theta|Y_{\text{obs}}) = l(\theta|Y) - \ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta),$$

then taking the expectation of both sides over the distribution of the missing data (Y_{mis}) given the observed data Y_{obs} and the current estimate of theta $\theta^{(t)}$ Rubin gets

$$l(\theta|Y_{\text{obs}}) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}),$$

where

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y_{\text{obs}}, Y_{\text{mis}}) f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})\, dY_{\text{mis}}$$

and

$$H(\theta|\theta^{(t)}) = \int \ln f(Y_{\text{mis}}|Y_{\text{obs}}, \theta) f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})\, dY_{\text{mis}}.$$

A central result of Dempster et al. (1977) is that if we let $\theta^{(t+1)}$ be the value of θ that maximizes $Q(\theta|\theta^{(t)})$, then $\theta^{(t+1)}$ is a better estimate than $\theta^{(t)}$ in the sense that its observed-data log-likelihood is at least as high as that of $\theta^{(t)}$,

$$l(\theta^{(t+1)}|Y_{\text{obs}}) \geq l(\theta^{(t)}|Y_{\text{obs}}).$$

This can be seen by writing the difference in the values of Y_{obs} at successive iterates as follows

$$l(\theta^{(t+1)}|Y_{\text{obs}}) - l(\theta^{(t)}|Y_{\text{obs}}) = Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}).$$

The quantity $Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$ is non-negative because θ has been chosen to satisfy

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) \text{ for all } \theta. \tag{31}$$

The remainder $H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})$ which can be written

$$\int \log \left[\frac{P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})}{P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t+1)})} \right] P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})\, dY_{\text{mis}},$$

is easily shown to be non-negative by Jensen's inequality for any two probability distributions ($\pi_1(x)$, $\pi_2(x)$) and convexity of the function $x\text{log}x$. Let

$$\int \log \left[\frac{\pi_2(x)}{\pi_1(x)} \right] \pi_1(x)dx \leq \log \int \left[\frac{\pi_2(x)}{\pi_1(x)} \right] \pi_1(x)dx = 0.$$

Hence

$$S(\theta^{(t)}|\theta^{(t)}) \geq S(\theta^{(t+1)}|\theta^{(t)}). \tag{32}$$

By putting (31) and (32) together, we have proven that $F(\theta^{(t+1)}) \geq F(\theta^{(t)})$. From the proof we can see that any $\theta^{(t+1)}$ that increases the Q-function will increase $F(\theta)$. A proper design of the EM-method is to think of one iteration of EM defined by (31) as a consisting of the two following steps

1. *The Expectation* or E-step, in which the function $Q(\theta|\theta^{(t)})$ is calculated by averaging the complete-data log-likelihood $l(\theta|Y)$ over $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$

2. *The maximization* or M-step, in which $\theta^{(t+1)}$ is found by maximizing $Q(\theta|\theta^{(t)})$.

Lately has the *EM-method* also been applied to many situations that are not necessarily thought of as missing-data problems but it can be formulated as one. For example in multilevel linear models for unbalanced repeated measures data, such as in our project, where not all participants are measured at all time points.

4.9 Extension of EM

The ECM algorithm defined by Meng and Rubin (1993) generalizes the EM algorithm by replacing the M-step with a sequence of simple constrained maximization steps which is shortened *CM-steps*, indexed by $i = 1, ..., k$ each of which fixes some function of the parameters θ to be maximized. A newer development of the ECM algorithm which itself is an extension of the EM algorithm can be obtained by replacing some *CM-steps* of ECM which maximize the correspondingly constrained actual likelihood function with steps that maximize the correspondingly actual likelihood function. This algorithm which Chuanhai Liu and Donald Rubin call ECME algorithm for Expectation/Conditional shares with both EM and ECM their stable monotone convergence and basic simplicity of implementation relative to competing faster converging methods. ECME can have a substantially faster convergence rate than either EM or ECM, measured using either the number of iterations or actual computer time. There are two reasons for this improvement. First, in some if ECME's maximization steps is the actual likelihood being conditionally maximized rather than a current approximation to it as with EM and ECM. Secondly ECME allows faster converging numerical methods to be used on only those constrained maximizations where they are most efficient.

4.9.1 The Expectation Conditional Maximization Either (ECME) method

The ECME algorithm defined by Liu and Rubin (1995) extends the ECM algorithm by allowing *CM-steps* to maximize either the constrained expected log-likelihood or the correspondingly constrained the actual log-likelihood function $L(\theta)$. The E-step of ECME is the same as the E-step of EM and ECM. The *CM-step* 1 of ECME is the same as the *CM-step* of ECM, but the *CM-step* 2 of ECME maximizes the actual likelihood, Equation (30), and the constraint functions which correspond to a different conjugate linear combinations of the parameters across iterations. Code for this maximization involves a one-dimensional search as with EM, where Fisher scoring are incorporated into the M-step. This procedure may be used when the response variable are partially missing. The ECME algorithm used in *Pan*[3] is based upon the multivariate model in Equation (19), the likelihood function from the marginal normal distribution for y_i may be expressed as

$$L(\theta) \propto \prod_{i=1}^{m} |W_i|^{\frac{1}{2}} \exp\{-\frac{1}{2}\delta_i^T W_i \delta_i\}, \tag{33}$$

where $\delta_i = \text{vec}(y_i - X_i\beta)$ and W_i is defined by Equation (22). Schafer and Yucel (2002) uses the relationship $|W_i| = |\Sigma \otimes I_{n_i}|^{-1}|\Psi|^{-1}|U_i|$ and ignore the constants of proportionality. Then the logarithm of L becomes

$$l(\theta) = -\frac{N}{2}\text{log}|\Sigma| - \frac{m}{2}\text{log}|\Psi| + \frac{1}{2}\sum_{i=1}^{m}\text{log}|U_i| - \frac{m}{2}\sum_{i=1}^{m}\delta_i^T W_i \delta_i \tag{34}$$

[3] defined in Section 4.10

Fischer scoring updates the current estimate of $\theta(t)$ by solving the linear system $C(\theta^{(t+1)}) = d$, where $C = -El^{''}(\theta^{(t)})$ and $d = C\theta^{(t)} + l^{'}(\theta^{(t)})$. Upon convergence will the final value of C^{-1} provide an estimated covariance matrix for $\hat{\theta}$. For further details see Schafer and Yucel (2002).

The E-step calculates the expectation of the complete data log-likelihood function, Equation (33), with respect to the conditional distribution of Y_{mis} given Y_{obs} under a current estimate of θ and the M-step updates the estimate of θ, maximizing this expected log-likelihood by scoring. For the E-step note that Equation (33) is a linear function of the sufficient statistics $\text{vec}(y_i)$ and $\text{vec}(y_i)\text{vec}(y_i)^T$. It follows from Equation (19) that $\text{vec}(y_i)$ and $\text{vec}(b_i)$ are jointly normal with covariance matrix

$$\begin{bmatrix} (I_r \otimes Z_i)\Psi(I_r \otimes Z_i)^T & (I_r \otimes Z_i)\Psi \\ \Psi(I_r \otimes Z_i)^T & \Psi \end{bmatrix} \tag{35}$$

Finding the expectations, Schafer and Yucel use Equation (35) as a basis whose dimension is $(rq + rn_i) \times (rq + rn_i)$, where $r = 1$ in the ecme-method in the *PAN* package, and then apply an orthogonalization method for $i = 1, 2, ..., m$. Q are equal to the number of random variables in the *pred* matrix and n_i are equal to the number of rows in the y vector and the *pred*-matrix. Since this procedure is best fitted to small samples Schafer and Yucel changed their strategy, and assumed that the rows of y_i are conditionally independent given b_i with a constant variance. They expressed the expectation of the first complete-data sufficient statistic as follows

$$E(y_i|Y_{\text{obs}}) = E(E(y_i|Y_{\text{obs}}, b_i)|Y_{\text{obs}}). \tag{36}$$

Equation (36) require access to the distribution of the missing values given the observed values and the random effects $(y_{(\text{obs})}, b_i)$ and the random effects given the observed values. The previous is quite simple since, given b_i, the rows of $\epsilon_i = y_i - X_i\beta - Z_ib_i$ are independent and identically distributed as $N(0, \Sigma)$. Therefore will the missing values in any row of ϵ have, given the observed values and the random effects, an intercept-free regression on the observed values. The parameters of this regression can be obtained by inverting the square sub-matrix of Σ which corresponds to the observed values. Schafer and Yucel (2002) divided the variable y_i^* into two parts, where $y^*_{ij_{(\text{mis})}}$ is the missing portion and $y^*_{ij_{(\text{obs})}}$ are the observed portions of the jth row of y_i^*. Then they expressed the expectation of the missing values given the observed ones and the random effects as

$$E(y^*_{ij_{(\text{mis})}}|Y_{(\text{obs})}, b_i) = \Sigma_{21}\Sigma_{11}^{-1}y_{ij_{(\text{obs})}},$$

where Σ_{11} is the square sub-matrix of Σ corresponding to the observed elements and Σ_{21} is the rectangular sub-matrix of covariances between the missing and observed elements. Since y_i^* is a linear function of b_i will the expectation of y_i without conditioning on b_i be obtained by a direct substitution of $E(b_i|y_{i_{(\text{obs})}})$ for b_i. The value of $\Sigma_{21}\Sigma_{11}$ varies by the different missingness pattern, but not by observational units $i = 1, 2, ..., m$.

For the second sufficient statistic $\text{vec}(y_i)\text{vec}(y_i)^T$ Schafer and Yucel (2002) applies a similar argument by first calculating the conditional expectation given the random effects and the observed values, and then averaging over the distribution of b_i given $y_{i_{(\text{obs})}}$. Let y_{ijk} denote the kth element of the jth row of y_i, and the formula for the expectation of $y_{ijk}y_{ij'k'}$ depends on whether y_{ijk} and $y_{ij'k'}$ are observed or missing and whether they are in the same $(j = j^{'})$ or different $(j \neq j^{'})$ rows. Schafer and Yucel (2002) have shown that the expectation of $y_{ijk}y_{ij'k'}$ given $y_{i_{(\text{obs})}}$ is given by: $y_{ijk}y_{ij'k'}$ if both are observed, $y_{ijk}E(y_{ij'k'}|y_{i_{(\text{obs})}}$ if y_{ijk} is observed and $y_{ij'k'}$ is missing, and

$$E(y_{ijk}|Y_{(\text{obs})})E(y_{ij'k'}|Y_{(\text{obs})}) + cov(y_{ijk}y_{ij'k'}|y_{i_{(\text{obs})}})$$

if both are missing. The covariance between y_{ijk} and $y_{ij'k'}$ given $y_{i_{(\text{obs})}}$ is equal to

$$\text{cov}(A_{ijk}, A_{ij'k'}|y_{i_{(\text{obs})}}) + [\Sigma_{22\cdot1}]_{kk'}$$

if they are in the same row, and

$$\text{cov}(A_{ijk}, A_{ij'k'}|y_{i_{(\text{obs})}})$$

if they are in different rows, where

$$A_{ijk} = E(y_{ijk}|b_i, y_{i_{(\text{obs})}})$$

comes from the regression predictions for the missing elements in the jth row of y_i given the observed values.

In the M-step will the expected log-likelihood which are computed in the E-step be maximized, and this can be performed by a slightly modification of the Fischer scoring procedure. This has almost the same form as Equation (34), but smaller changes must be made to the log-likelihood function and it derivatives. The expected second derivatives are the same. Schafer and Yucel (2002) write the first derivatives of $l_\text{e} = E(l|Y_{\text{mis}})$ with respect to the elements of θ as following:

$$\begin{aligned}
\frac{\partial l_e}{\partial \text{vec}(\beta)} &= -\left(\sum_{i=1}^{m}(I_r \otimes X_i)^T W_i (I_r \otimes X_i)\right)\text{vec}(\beta - \tilde{\beta}) \\
\frac{\partial l_e}{\partial \omega_j} &= \frac{1}{2}\sum_{i=1}^{m}\text{tr}(\Psi - U_i - (\Sigma^{-1} \otimes Z_i^T Z_i)U_i T_i U_i(\Sigma^{-1} \otimes Z_i^T Z_i))G_j, \\
\frac{\partial l_e}{\partial \sigma_l} &= \frac{1}{2}\sum_{i=1}^{m}\text{tr}(n_i \Sigma F_l - (F_l \otimes Z_i^T Z_i)U_i - W_i(\Sigma F_j \Sigma \otimes I_{n_i})W_i T_i, \\
\text{vec}(\beta) &= \Gamma \sum_{i=1}^{m}(I_r \otimes X_i)^T W_i E(\text{vec}(y_i)|\theta, y_{i_{(\text{obs})}}), \\
T_i &= E\{\text{vec}(y_i - X_i\beta)\text{vec}(y_i - X_i\beta)^T|y_{i_{(\text{obs})}}, \theta\}.
\end{aligned}$$

After these derivatives have been calculated will the parameters be updated in the same way as in Section 3.2 from Schafer and Yucel (2002) until the estimates have converged.

4.10 Pan - Imputation of Multivariate Panel or Clustered Data

Pan (for PANel data) is a software package for *R*, written by Joseph L. Schafer. The package can be downloaded freely from the Comprehensive R Archive Network at *http://www.cran.r-project.org/*. It is designed for use with clustered sampling and longitudinal data sets. Pan uses a multivariate extension of a two-level linear regression model commonly applied to multilevel data. This package takes full advantage of the known design characters to perform more efficient imputations and is particularly useful for the situation where time-varying predictors of change are partially unobserved. More details of these models are given by Schafer (1997). The two main functions used from the Pan package are

- pan – A Gibbs sampler for the multivariate linear mixed models with incomplete data. This function will be used to produce multiple imputations of missing values in multivariate data. The input is a data set with missing values.
- ecme – Performs maximum-likelihood estimations for generalized mixed models. The input is a data set with missing values.

The R documentation for the pan-function used in this master thesis is reproduced in Appendix A.

5 Missing Values in the SGA-Data set

In this chapter we

(a) describe the marginal and joint description of the missing values in the SGA Data Set.

(b) investigate different imputation models in the MI-procedure to be used for the SGA Data Set.

The selection procedure of variables included in the growth model for MAD is described in Section 8.1. The "SGA-data Set" is the 561 random sample defined in Section 2.2.

5.1 Marginal description of missing values

The variables included in the analysis and their rates of missingness are reported in Table 2. Notice that MAD and time of gestation are missing from approximately 25% of the observed measurements from the three last antenatal visits. The gender of the fetus, the LBW variable and the age of the mother are the only covariates with a complete set of measurements from all the subjects.

Name	*Description*	*Missing (%)*
SEX	Gender of the fetus (0 = female, 1 = male)	0.0
LBW	Previous low birth weight infant of the mother (0 = No, 1 = Yes)	0.0
Height	Height of the mother	0.5
Age	Age of the mother	0.0
Time1	Gestational age measurued in weeks, based on ultrasound	3.2
MAD17	Mean abdominal diameter at Time_1 weeks of gestation	3.0
Weight17	Weight of the mother at Time_1 weeks of gestation	0.9
Time2	Gestational age measurued in weeks, based on ultrasound	23.5
MAD25	Mean abdominal diameter at Time_2 weeks of gestation	23.2
Weight25	Weight of the mother at Time_2 weeks of gestation	8.0
Time3	Gestational age measurued in weeks, based on ultrasound	20.9
MAD33	Mean abdominal diameter at Time_3 weeks of gestation	20.9
Weight33	Weight of the mother at Time_3 weeks of gestation	8.4
Time4	Gestational age measurued in weeks, based on ultrasound	23.9
MAD37	Mean abdominal diameter at Time_4 weeks of gestation	24.1
Weight37	Weight of the mother at Time_4 weeks of gestation	10.7
Total	561 pregnant women	44.00

Table 2: Variables from the SGA-data set used in the analysis, where the marginal rates of missingness for each variable is $= \frac{\text{Total number of subjects with at least 1 missing value}}{561} * 100\%$

Histograms of these 16 variables are displayed in Figures 9 and 10.

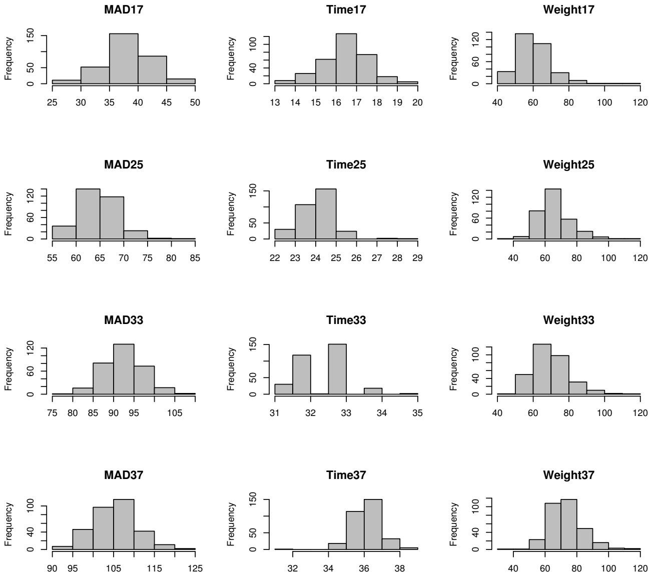

Figure 9: Histogram of gestational age (measured in weeks), weight of the pregnant woman and MAD at each of the four antenatal visits during the study.

In Figure 9 we see that MAD, Weight and gestational age have equal distribution of observed values at each of the four antenatal visits during the study. As a measure of weight gain during the pregnancy the weight variable is used together with the height variable to calculate the BMI (Body mass per index $(\frac{\text{kg}}{m^2})$). From the histogram of gender in Figure 10 we see that there is approximately as many female and males fetuses. Most of the women who participated in the *SGA*-study are between the age of 25 and 30, and have a height between $1.6 - 1.7$ meters. We also see that 30 of the 561 women have earlier given birth to a low-weight child (LBW-variable).

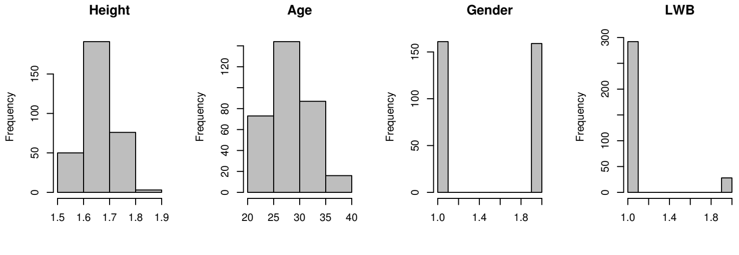

Figure 10: Histogram of age, height, LBW and gender of the fetus. For the gender of the fetus 1 denote girls and 2 denote boys.

If analysis involving all these variables were performed using standard statistical packages, the built-in case deletion procedures would discard up to 44.00% of the subjects, resulting in a substantial loss of power.

5.2 Joint Descriptions of Missing Data

In order to describe the joint pattern of missing values in the SGA-data set we divided the data set into groups, such that in each group every member is missing the same set of variable(s). In Table 3 we see that there are 314 subjects with no missing values. There are 59 subjects which are missing values of MAD and Time from the last three antenatal visits, and 21 subjects are missing values of MAD, Time and Weight from the last antenatal visit during the study. We also notice that there are 26 groups which consists of only one subject each, whose missing pattern is different from all the other groups. In Total there are 53 different groups, each with a different set of missing variables.

Group	$n_{(\text{missing variables})}$	Fetus	Gender	LBW	Height	Age	MAD17	Time1	Weight17	MAD25	Time2	Weight25	MAD33	Time3	Weight33	MAD37	Time4	Weight37	Freq
1	0																		314
2	1				X														2
3	1						X												4
4	1							X											4
5	2						X	X											1
6	1								X										2
7	1										X								1
8	2									X	X								16
9	1											X							16
10	3									X	X	X							3
11	2											X	X						1
12	2							X						X					1
13	2												X	X					8
14	4									X	X		X	X					9
15	1														X				15
16	2											X			X				3
17	4						X	X				X			X				1
18	3											X	X	X					6
19	1															X			1
20	3									X	X					X			1
21	3									X	X						X		2
22	2															X	X		3
23	4									X	X					X	X		10
24	4												X	X		X	X		3
25	5							X					X	X		X	X		1
26	6									X	X		X	X		X	X		59
27	7				X					X	X		X	X		X	X		1

Continued on the next page

Group	$n_{(\text{missing variables})}$	Fetus	Gender	LBW	Height	Age	MAD17	Time1	Weight17	MAD25	Time2	Weight25	MAD33	Time3	Weight33	MAD37	Time4	Weight37	Freq
28	7						X			X	X		X	X		X	X		3
29	7							X		X	X		X	X		X	X		1
30	8						X	X		X	X		X	X		X	X		1
31	7								X	X	X		X	X		X	X		3
32	7									X	X	X	X	X		X	X		2
33	3														X	X	X		1
34	6											X	X	X	X	X	X		1
35	8									X	X	X	X	X	X	X	X		1
36	1																	X	10
37	2							X										X	1
38	3									X	X							X	1
39	2											X						X	2
40	3												X	X				X	1
41	2														X			X	1
42	4									X	X				X			X	1
43	2															X		X	1
44	3															X	X	X	21
45	4							X								X	X	X	1
46	4										X					X	X	X	1
47	5									X	X					X	X	X	2
48	4														X	X	X	X	1
49	6												X	X	X	X	X	X	1
50	7											X	X	X	X	X	X	X	1
51	9									X	X	X	X	X	X	X	X	X	7
52	10						X			X	X	X	X	X	X	X	X	X	1
53	11						X	X		X	X	X	X	X	X	X	X	X	6
Sum:																			561

Table 3: Patterns of missing values in the SGA-data set with 516 subjects.

5.3 Imputing missing values of Gestational Age

Since *Time* is one of the best predictors in the model for both *MAD* and *BMI*, we had to make some decisions with respect to missing values in the time measurements. Since the ultrasound measurements were planned at given gestational ages, one solution to this problem is to fill inn the average gestational time at each of the 4 antenatal visits of the missing values. The number of subjects who are missing time measurements are presented in Table 4.

time of gestation*	*Min*	*Mean*	*Max*	*Number missing*
17	12	16.76	20	18 (3.21%)
25	22	24.62	29	132 (23.53%)
33	29	32.47	36	117 (20.86%)
37	31	36.63	39	134 (23.89%)
SGA-data set :	561			

Table 4: Summary measurements on gestational age, where time* is the expected time among the sample at each of the four antenatal visits during the study.

As we see from Table 4 3.21% of the time measurements are missing for the first antenatal visit, and approximately 20% of the time measurements are missing at each of the last three antenatal visits. For these missing values of time we have chosen to impute the mean values from each of the four antenatal visits. By doing this we believe that we are not introducing bias, as we see from Figure 9 clearly gestational age is for most of the women measured within one week from the planned time of gestation. In the next Section (Section 5.4.3) we will look at an alternative strategy for handling missing values in the time measurements. We see from Table 5 when the *Time-variable* is imputed there are now 320 subject who have a complete set of measurements, and there are $561-320 = 241$ subjects who are missing one or more variables. The number of groups with different combinations of missing-data are reduced from 53 to 38, where members from each group are missing measurements from only *MAD* and *BMI* variables.

Group	$n_{(\text{tot. var. missing})}$	Fetus	Gender	LBW	Height	Age	MAD17	Time1	Weight17	MAD25	Time2	Weight25	MAD33	Time3	Weight33	MAD37	Time4	Weight37	Freq
1	0																		320
2	1				X														2
3	1						X												5
4	1								X										2
5	1									X									18
6	1											X							16
7	2									X		X							3
8	1												X						8
9	2									X			X						9
10	2											X	X						1
11	1														X				15
12	2											X			X				3
13	3						X					X			X				1
14	2												X		X				6
15	1															X			4
16	2									X						X			11
17	2												X			X			4

Continued on the next page

Group	$n_{(\text{tot. var. missing})}$	Fetus	Gender	LBW	Height	Age	MAD17	Time1	Weight17	MAD25	Time2	Weight25	MAD33	Time3	Weight33	MAD37	Time4	Weight37	Freq
18	3									X			X			X			60
19	4				X					X			X			X			1
20	4						X			X			X			X			4
21	4								X	X			X			X			3
22	4									X		X	X			X			2
23	2														X	X			1
24	4											X	X		X	X			1
25	5									X		X	X		X	X			1
26	1																	X	11
27	2									X								X	1
28	2											X						X	2
29	2												X					X	1
30	2														X			X	1
31	3									X					X			X	1
32	2															X		X	24
33	3									X						X		X	2
34	3														X	X		X	1
35	3														X	X		X	1
36	5											X	X		X	X		X	1
37	6									X		X	X		X	X		X	7
38	7						X			X		X	X		X	X		X	7
Sum:																			561

Table 5: Pattern of missing values when the time have been imputed with the mean values from each of the four antenatal visits.

5.4 Using the Multivariate Linear Mixed Effects Model to Produce Multiple Imputation

5.4.1 The Imputation Model

We are using the linear mixed-effects model, Equation (19), defined in Section 4.6 as our choice of imputation model. In the imputation model, the response matrix y, is a $n_i * 2$ matrix of multivariate responses for sample unit i (where $i = \{1, 2, ..., 561\}$), and consists of MAD and the BMI variables. It is important to note that the imputation model and the model to be used for modeling MAD model differ. In the analysis model for MAD the BMI variable is treated as a covariate in the growth model for MAD, but in the imputation model both MAD and BMI are treated as responses with the same set of covariates.

The missing values in the response will be predicted from an *covariate-matrix* which consists of a constant, LBW, Time and Time^2 with gestational age and fetus as random effects. Note that we use the result from Section 5.3 with complete time measurements.

Our data set contains both continuous and binary variables and we construct dummy indicators for LBW to preserve program effects (LBW will be treated as a factor and not a continuous variable). The missing values will be imputated using the Gibbs sampler defined in Section 4.6.1, where Equations (23) - (27) denote one *cycle*. The cycle repeatedly creates sequences of the parameters, and estimates of the missing values $\{Y_{\text{mis}}^{(1)}, Y_{\text{mis}}^{(2)}, ..., Y_{\text{mis}}^{(M)}\}$ for both MAD and BMI.

5.4.2 Assessing Convergence

We can examine the behavior of the imputation model by making time-series plot and plot of the acfs (autocorrelation function) for the model parameters to assess the convergence of the MCMC-algorithm. All plots are plotted after the burn-in period is removed. In the time series, the value of the parameter (vertical axis) is plotted against the iteration number (horizontal axis). Using the definition from Brockwell and Davis (2002), let $\{X_t\}$ be a stationary time series. The *autocorrelation function* (ACF[4]) of $\{X_t\}$ at lag h is defined as

$$\rho_x(h) \equiv \frac{\gamma_x(h)}{\gamma_x(0)} = \text{Cor}(X_{t+h}, X_t), \tag{37}$$

where $\gamma_x(t) = \text{Cov}(X_i(t+h), X_j(h))$. The approximate 95% confidence limits (or acceptance area) are shown as dotted lines, these are for an independent serie for which $\rho_t = I(t = 0)$. As with a time series *a priori* one is expecting autocorrelation, these limits must be viewed with caution. In particular, if any ρ_x is non-zero then all the limits are invalid. In Figure 11 we see the time-series plot of the covariance matrix between MAD and BMI. We see that $\sigma_{\text{MAD}}, \sigma_{\text{BMI}}$ and $\sigma_{\text{MAD,BMI}}$ have converged very well and they are slightly positive correlated, which is expected. If the MAD increases it means that the fetus has grown which will cause an increase in the weight of the mother.

[4] The *acf-method* in *R* chooses the number of lags to plot unless this is specified by the argument lag.max. The default is $10 * log_{10}(\frac{N}{m})$ where N is the number of observations and m the number of series, in our case N is equal to 561 and m is equal to 10.

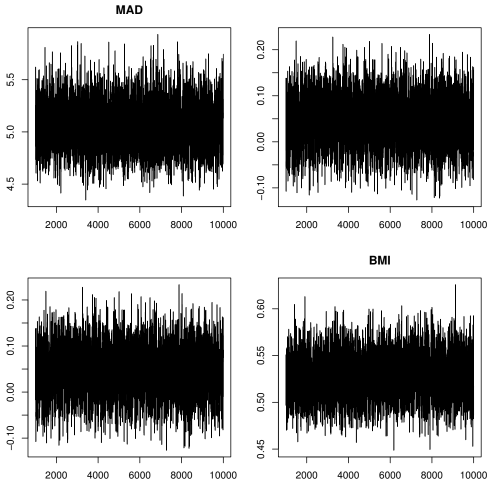

Figure 11: Time-series plot of Σ, which is the covariance-matrix of MAD and BMI.

From the ACF-plot of Σ in Figure 12, we see that the estimates at lag 100 are approximately independent.

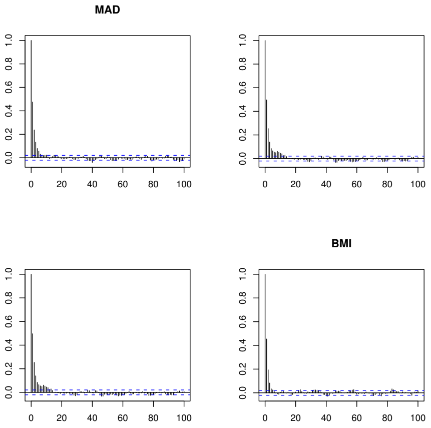

Figure 12: ACF plots of the covariance matrix, Σ, between MAD and BMI.

From the covariance matrix of the random effects (Figure 13) we see that the there is a slightly larger variation at the intercept for MAD than there is for the intercept for BMI. The intercept for MAD and BMI are positively correlated which means that an increase in MAD would result in a larger BMI value. The slope in the model for MAD increases more than the slope for BMI, which is natural since MAD is increasing much more for a fetus during a pregnancy than the increase in BMI by the mother during the same period of time.

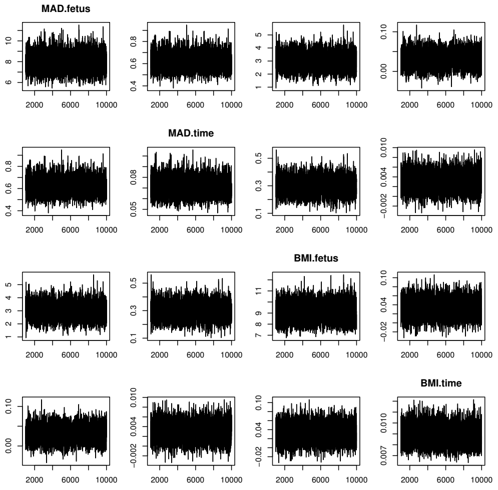

Figure 13: Time-series plot of Ψ, which is the covariance matrix of the random effects.

From the ACF-plot of Ψ (Figure 14), we see that the estimates are un correlated at lag 100.

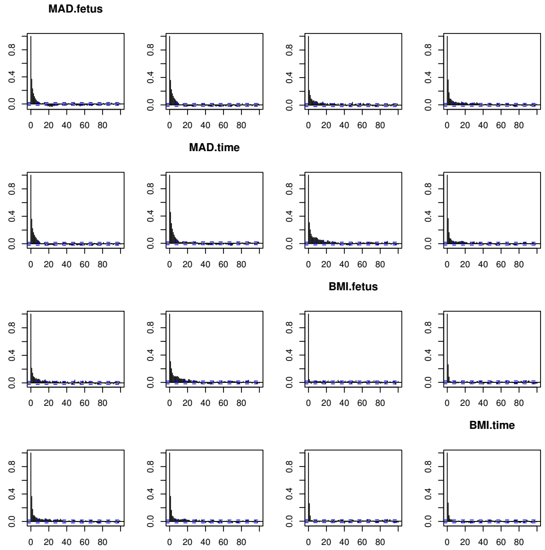

Figure 14: ACF plots of the covariance matrix Ψ, with Time and fetus as random effects.

From the time-series plot of the estimated fixed parameters of the imputation model in Figure 15 we see that the estimated values for both models (MAD and BMI) have converged very well. All four estimates for the β values are significant in each model, because they have converged very well and the estimated values are clearly not zero.

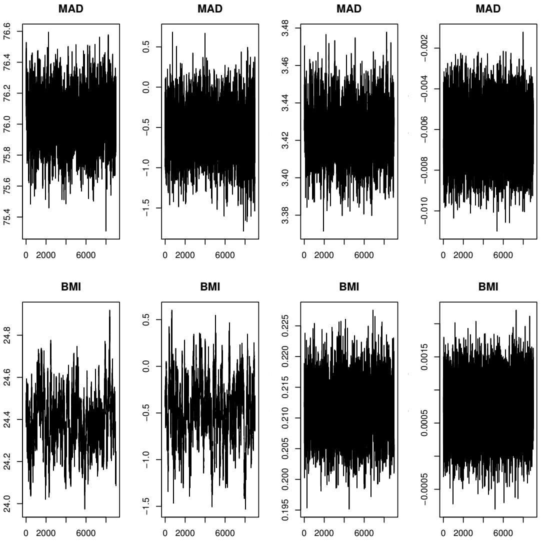

Figure 15: Time-series plot of the estimated parameters in the imputation model for MAD and BMI, in the first column β_0 = constant, the second column β_1 = LBW-variable, the third column β_2 = Time-variable and in the fourth column $\beta_3 =$ Time$^2-$variable.

From the ACF-plot of β (Figure 16), we see that the estimates are approximately uncorrelated at lag 100 in the model for MAD. In the model for BMI we see that the β_0 and β_1 values have correlated estimates, but the parameter estimates for Time and Time^2 seems to be uncorrelated at lag 100. That is because the mixing may not be very good since the imputation model we have used for both MAD and BMI is based upon the model we got from backwards elimination by the *lme* model for MAD. Because of that there is a possibility that there exist some covariates which are better for predicting missing-values for BMI, but after all these results are quite satisfying. Based upon the results from Section 5.4.2 we have decided to use values at lag 1000, by doing this we are sure that the estimates are uncorrelated.

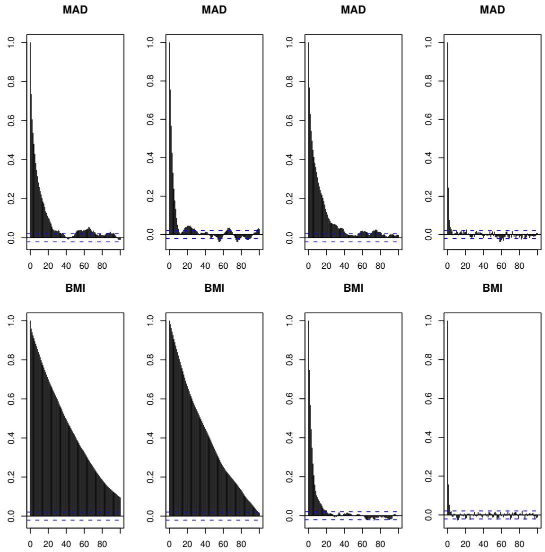

Figure 16: Acf plots of the estimated parameters in the imputation model for MAD and BMI, in the first column β_0 = constant, the second column β_1 = LBW-variable, the third column β_2 = Time-variable and in the fourth column $\beta_3 =$ Time$^2-$variable.

5.4.3 Imputing Missing Values of Gestational Age

In Section 5.3-5.4.2 we used the average values of gestational age from each of the 4 antenatal visits in our imputation model. An alternative approach is to estimate the time measurements which are missing. The time-variable will be included into the response-matrix together with the MAD and BMI variables in the imputation model, and then the missing values will be predicted by a covariate-matrix which consists of a constant and LWB and with fetus as a random effect.

From Figure 17 we see that the estimated values of the standard deviations of the response variables (MAD and BMI) in the response-matrix are very high compared to the previous alternative. This is not surprising since gestational age is not used as a covariate to model MAD and BMI. Therefore the covariate-matrix will not contain enough information to give us reasonable estimates.

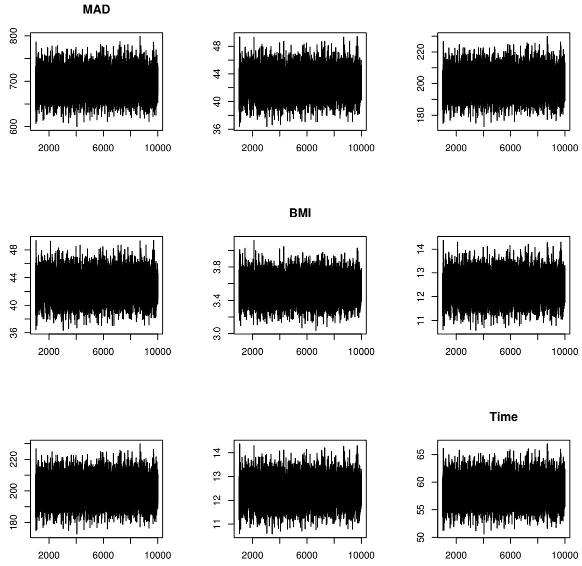

Figure 17: Time series plot of the covariance matrix (Σ) of the response variables (MAD, BMI and gestational time) in the response-matrix of the MI-procedure.

The only random effects which is included into the model is the person effect. From Figure 18 we see that the estimated values for the covariance matrix of the random effects (Ψ) have not converged and the mixing could have been better.

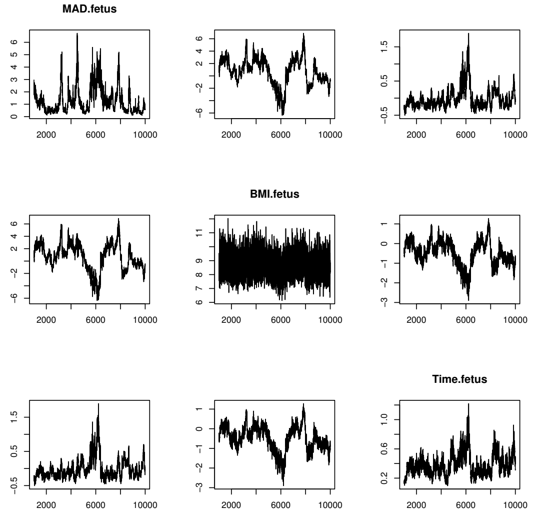

Figure 18: Time series plot for the covariance matrix of random effects (Ψ) when only fetus is being used as a random effect.

From the ACF[5]-plot (Figure 19) we see that the correlations between the estimated values from each iteration are very high, which means that all the values show a significant correlation.

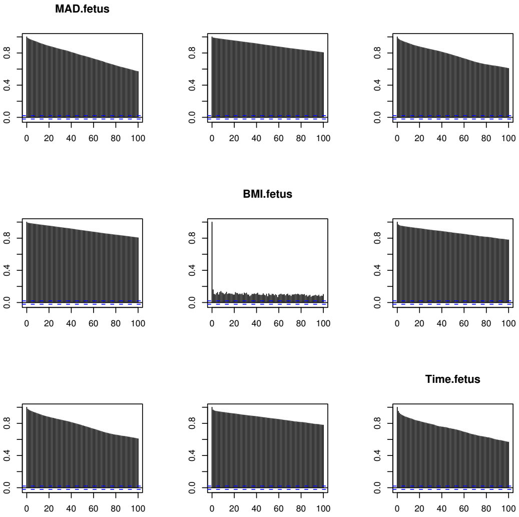

Figure 19: ACF-plot for the covariance matrix of the random effects (Ψ), when only fetus is being used as a random effect.

5.4.4 Conclusions

From the results shown in Figures 17, 18 and 19 compared with Figures 11, 12, 13 and 14 we see that the imputation procedure where missing values of gestational age are imputed instead of modeled in the imputation model performs the best. Therefore we use the MI-procedure where the gestational age variable is included as covariates in the imputation model. This model has much smaller variance and the ACF-plot of the parameters in the imputation-model clearly shows that the estimates of the parameters are much more uncorrelated and the mixing is much better.

[5] acf = autocorrelation function is defined in Equation (37)

6 Evaluation of Multiple Imputation method using a Simulation Study

6.1 Introduction

In this chapter we

through a simulation study assess the accuracy of the estimates of the MI method.

6.2 Assessing the Accuracy of the MI-Method

Procedures for assessing model fit have not yet been implemented under multiple imputation. Imputation variance serves to increase the "noise" in assessing the fit of a specific model, and use of the full sample size also serves to inflate the model chi-square statistic. Note that the MI-method consist of the MI (used to impute missing values) and the growth model for MAD (perform the multilevel analysis with a complete set of data) method. In an attempt to assess the accuracy of the model we used the following idea.

*We only use subjects with a complete set of data (*320*) as a basis, from this data set we delete values following a MCAR-procedure such that we get a simular pattern of missing values as in the full data set. Then we perform the MI-method (MI + lme method) for each data set with missing values, and compare the estimates of imputed observations with the true values. We also compare the summary statistics, the values it selves are not the most important.*

6.2.1 MCAR-procedure

From Table 5 we removed all groups with a frequency less than 10, the remaining eight groups were

Group	$n_{(\text{total missing})}$	Fetus	Gender	LBW	Height	Age	MAD17	Time1	Weight17	MAD25	Time2	Weight25	MAD33	Time3	Weight33	MAD37	Time4	Weight37	Freq
1	0																		320
2	1									X									18
3	1											X							16
4	1														X				15
5	2									X						X			11
6	3									X			X			X			60
7	1																	X	11
8	2															X		X	24
Sum:																			475

Table 6: Missing pattern of groups from Table 5 with a frequency $\geq$ 10 in each group.

From Table 6 we see that there are 320 subjects which have no missing values and 155 are missing *MAD* or *BMI* from one ore more antenatal visits. This will contain 85% (475 of 561) of all the subjects in the original data set.

The eight different groups from Table 6 formed a basis, with a probability $\frac{\text{Freq}_{(\text{group}_j)}}{\sum_{i=1}^{8}\text{Freq}_i}$ to draw one of the groups. We sampled each subject in the complete set of data (320) to decide which of the eight group they will be a member of, and deleted the observed values which were missing in that specific group. The distribution for the groups from the first simulation is displayed in Table 7.

	Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*	*Group 6*	*Group 7*	*Group 8*
Freq	195	17	13	9	9	43	9	25

Table 7: Results from the first simulation following the outlined MCAR-procedure.

From Table 7 we see that there are 195 subjects with a complete set of data, and $\frac{125}{320} = 39.06\%$ of the subjects have missing values in the simulated data set. In the original 561 SGA data set $\frac{561-320}{561} = 42.96\%$ were missing. This means that the percentage of records with missing values are approximately the same in the simulated data set as in the SGA data set.

The 195 subjects with no missing values were used to perform the complete case analysis resulting in the estimates for the fixed coefficients in Table 8

Coefficient	*Value*	*Std.dev*	*DF*	*t-value*	*p-value*
β_0	65.4585	1.9195	579	34.1024	0.0000
β_1	2.3534	0.1967	579	11.9668	0.0000
β_2	-0.0120	0.0023	579	-5.1778	0.0000
β_3	-0.2652	0.2724	191	-0.973	0.3316
β_4	0.2032	0.0503	191	4.0403	0.0001
β_5	0.1904	0.0561	579	3.3968	0.0007
β_6	-1.4182	0.7536	191	-1.8818	0.0614
β_7	0.0160	0.0049	579	3.2464	0.0012
β_8	0.0228	0.0057	579	4.0331	0.0001
β_9	-0.1651	0.0738	579	-2.2361	0.0257

Table 8: Estimates of the fixed effects from the complete case by the MCAR-procedure, here 195 subjects are randomly chosen from those who originally 320 had a complete set of values.

If we compare the estimated values from the original data set (Table 18) with the simulated one (Table 8), we see that most of the parameters are almost unchanged except β_3 which represents effect of gender on the fetus and β_6 which represents the effect of low weighted fetuses from earlier pregnancies. The covariate β_3 have changed from -0.00152 to -0.26519, which means that the gender of the fetus have become more significant, β_6 have changed from -2.34333 to -1.41817 which means that result of earlier pregnancies with respect by LWB have slightly reduced its significance. We see that none of the parameter estimates have changed significantly (lies outside the 95% confidence interval of the parameter estimates from Table 18) and all the estimated standard deviations have increased, this is as expected since the data set from the MCAR-procedure contains as subset of the original complete case ($195 < 320$).

Performing the MI-procedure defined in Section 4.5.2 on the data set we got from the MCAR-procedure, we get M imputations of each missing value which gives us M different imputed data sets. We here choose $M = 10$. Each of the imputed data sets are then analyzed by the same complete-data method (lme), and we use Rubin's rules from Section 4.5.3 to combine the parameter estimates. Before we study the parameter estimates we compare the estimated values from the MI-procedure with the observed values (true-value) which we have deleted. Subjects from group 2 were only missing values from MAD_{25} and subjects from group 7 were missing values from BMI_{37}.

From Table 9, we see that the estimated values (measured in mm) from the imputation method for MAD are very close to the true values, the biases (observed value - imputed value) are varying in the interval $[-2.7770, 5.2460]$. That is not so far from the truth. The simulated values for BMI (measured in $\frac{\text{kg}}{m^2}$) are surprising close to the true value compared to the imputed values for MAD. The differences are in the interval $[-1.8240, 0.8760]$. Since the imputation model is fitted for a model for MAD, we may expected that we would have obtained a more narrow interval of the differences between true and imputed values for MAD than for BMI. But neither of the imputed values for the two different groups are far from the true value, and we know that our choice of imputation model is acceptable.

Group	Obs. value	Imputed values										
		m=1	*m=2*	*m=3*	*m=4*	*m=5*	*m=6*	*m=7*	*m=8*	*m=9*	*m=10*	$\overline{\text{Diff}}$
2	72	69.40	73.50	67.10	68.60	70.30	70.40	68.10	69.70	72.50	68.50	-2.1900
2	62	64.61	63.27	62.52	63.62	61.89	63.59	66.28	61.20	65.24	61.06	1.3280
2	71	68.06	70.66	69.54	66.63	69.04	63.58	65.67	68.18	69.81	71.06	-2.7770
2	61	62.64	62.25	62.74	65.91	60.51	60.07	61.64	64.30	59.75	64.10	1.3910
2	61	64.28	65.84	63.16	61.85	63.00	66.93	64.32	64.67	62.63	62.13	2.8819
2	63	66.33	65.97	64.50	70.87	65.99	61.93	67.87	66.59	64.19	59.80	2.4040
2	72	71.01	70.01	70.61	67.84	70.48	67.73	71.74	70.33	67.96	69.79	-2.2500
2	62	63.35	66.04	68.97	66.59	67.12	66.09	71.61	66.17	69.62	66.90	5.2460
2	60	64.49	63.80	62.76	62.45	66.21	62.45	66.60	63.77	63.47	62.22	3.8220
2	61	62.96	61.71	62.72	65.38	65.22	62.38	60.47	60.48	59.56	63.48	1.4360
2	56	58.22	54.94	61.04	56.52	62.14	62.05	65.61	59.45	61.20	56.71	3.7880
2	64	60.28	61.85	61.85	61.53	61.84	62.77	57.57	61.54	61.05	57.87	-3.1850
2	63	57.12	61.28	59.45	60.18	59.68	63.79	61.10	62.51	60.48	59.80	-2.4610
2	67	66.82	67.75	68.89	66.37	64.25	67.50	66.81	67.33	63.31	68.19	-0.2780
2	63	68.79	66.37	64.15	69.15	71.28	67.28	70.04	68.42	66.39	67.87	4.9740
2	67	66.27	70.48	67.44	66.44	72.26	72.96	68.10	68.74	71.64	69.03	2.3360
2	68	68.94	66.12	66.18	67.38	65.90	67.20	71.92	67.52	65.67	67.92	-0.4773
7	23.18	22.33	23.89	25.59	25.29	22.41	21.98	23.72	23.70	25.79	20.89	0.3790
7	23.62	26.27	23.34	24.63	23.57	24.19	24.48	23.60	25.12	24.25	25.51	0.8760
7	25.34	26.83	25.17	25.68	26.47	23.47	24.83	25.55	24.91	26.09	25.92	0.1520
7	24.38	23.50	23.55	21.96	22.77	23.04	22.96	23.54	24.57	23.90	23.60	-1.0410
7	23.79	22.27	23.48	21.43	24.30	23.79	25.70	24.17	21.61	25.32	22.31	-0.3520
7	30.48	28.76	29.63	29.28	30.74	29.59	28.46	29.66	29.04	29.18	29.98	-1.0480
7	24.91	22.10	23.16	25.88	23.87	23.81	27.70	24.60	21.30	24.37	23.88	-0.8430
7	23.45	21.77	22.76	22.44	21.04	21.59	22.56	19.84	20.54	19.87	23.85	-1.8240
7	27.77	29.11	27.84	27.90	27.89	26.61	29.07	28.16	26.63	28.03	28.78	0.2320

Table 9: Estimated values from the imputation model for two of the eight possible groups from the MCAR-simulation (Table 7). $\overline{\text{Diff}}$ is the average difference between each of the observed values and the 10 imputed values. Subjects from group 2 are missing MAD$_{17}$-observations and subjects from group 7 are missing BMI$_{37}$-observations.

In Table 10 we see the estimated fixed effects of the parameters when we analyze each of the 10 imputed data sets with the the growth model for MAD (perform the multilevel analysis with a complete set of data) method.

	Imputed data set									
Parameter	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
β_0	65.631	67.379	67.749	66.648	66.980	66.143	66.693	66.334	67.770	66.747
Std.dev$_{\beta_0}$	1.5617	1.5839	1.5613	1.6083	1.5789	1.5651	1.6139	1.6217	1.5865	1.6007
β_1	2.4738	2.5432	2.6299	2.5101	2.5824	2.5439	2.5108	2.5117	2.6473	2.5368
Std.dev$_{\beta_1}$	0.1506	0.1511	0.1508	0.1563	0.1519	0.1497	0.1525	0.1555	0.1527	0.1512
β_2	-0.0119	-0.0112	-0.0119	-0.0113	-0.0127	-0.0110	-0.0122	-0.0127	-0.0111	-0.0110
Std.dev$_{\beta_2}$	0.0017	0.0017	0.0017	0.0018	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
β_3	0.1365	0.0219	-0.0228	0.0659	0.0612	0.0357	0.1540	0.0486	0.0083	0.0225
Std.dev$_{\beta_3}$	0.2216	0.2213	0.2214	0.2308	0.2230	0.2212	0.2243	0.2237	0.2216	0.2237
β_4	0.1717	0.0997	0.1024	0.1433	0.1089	0.1739	0.1203	0.1164	0.1214	0.1111
Std.dev$_{\beta_4}$	0.0412	0.0420	0.0413	0.0426	0.0419	0.0415	0.0429	0.0433	0.0420	0.0426
β_5	0.2220	0.2303	0.2183	0.2064	0.2399	0.1958	0.2351	0.2584	0.1872	0.2422
Std.dev$_{\beta_5}$	0.0438	0.0442	0.0437	0.0450	0.0439	0.0436	0.0448	0.0447	0.0441	0.0444
β_6	-1.4646	-1.5728	-1.4696	-1.0711	-1.2008	-1.5663	-1.4229	-1.2142	-1.5546	-1.4242
Std.dev$_{\beta_6}$	0.5660	0.5773	0.5670	0.5845	0.5755	0.5692	0.5888	0.5945	0.5773	0.5853
β_7	0.0127	0.0080	0.0074	0.0103	0.0078	0.0137	0.0109	0.0088	0.0091	0.0090
Std.dev$_{\beta_7}$	0.0038	0.0038	0.0038	0.0040	0.0039	0.0038	0.0039	0.0040	0.0039	0.0039
β_8	0.0221	0.0244	0.0217	0.0230	0.0228	0.0178	0.0220	0.0245	0.0187	0.0235
Std.dev$_{\beta_8}$	0.0042	0.0042	0.0042	0.0043	0.0042	0.0042	0.0042	0.0043	0.0042	0.0042
β_9	-0.1032	-0.1152	-0.1172	-0.0563	-0.0768	-0.1277	-0.0773	-0.0895	-0.1173	-0.1054
Std.dev$_{\beta_9}$	0.0523	0.0528	0.0525	0.0545	0.0532	0.0522	0.0534	0.0546	0.0532	0.0530

Table 10: Estimated fixed effects and standard deviation of the parameters for the 320 subjects using complete-data method analysis from each of the 10 imputed data sets.

We see that there is only small variations between the estimates and the standard deviations of the parameters for each of the imputed data sets.

6.2.2 Combining the Results

When we use Rubin's rules to combine the results of the estimated parameters in the model for *MAD* from Table 10 we get the results of Table 11

Coefficient	*Mean*	*Average Var.*	*Between-imputation Var.*	*Total Var.*	*Standard deviation*
β_0	66.8073	2.5230e+00	4.7348e-01	3.0437e+00	1.7446
β_1	2.5490	2.3180e-02	3.0617e-03	2.6547e-02	0.1629
β_2	-0.0117	2.9933e-06	4.6081e-07	3.5002e-06	0.0019
β_3	0.0532	4.9846e-02	3.0397e-03	5.3190e-02	0.2306
β_4	0.1269	1.7754e-03	7.3153e-04	2.5801e-03	0.0508
β_5	0.2235	1.9565e-03	4.9189e-04	2.4976e-03	0.0500
β_6	-1.3961	3.3479e-01	3.0418e-02	3.682e-01	0.6068
β_7	0.0098	1.4969e-05	4.4669e-06	1.9883e-05	0.0045
β_8	0.02205	1.77590e-05	4.9041e-06	2.3153e-05	0.0048
β_9	-0.0986	2.8117e-03	5.2054e-04	3.3989e-03	0.0583

Table 11: Estimates of the fixed effects combined by Rubins rules, Equation (13-16), for the 320 subjects in the imputed data sets.

From Table 11 we see that the estimates for the different parameters are almost equal to the ones from the complete-data analysis in Table 18. The standard deviation of the parameters in Table 11 are larger than the standard deviations in Table 18 as expected.

To investigate this more throughly, we need to perform more than only one repetition of the *MCAR*-procedure. Therefore we repeated the *MCAR*-procedure 10 times as described in Figure 20. Each time we calculated:

$$\begin{aligned}\text{Bias}_\beta &= \text{mean}_{\beta_{\text{Complete-case}}} - \text{mean}_{\beta_{\text{MI-method}}},\\ \text{Fraction}_{\text{Std.dev}_\beta} &= \frac{\text{Std.dev}_{\beta_{\text{Complete-case}}}}{\text{Std.dev}_{\beta_{\text{MI-method}}}}.\end{aligned}$$

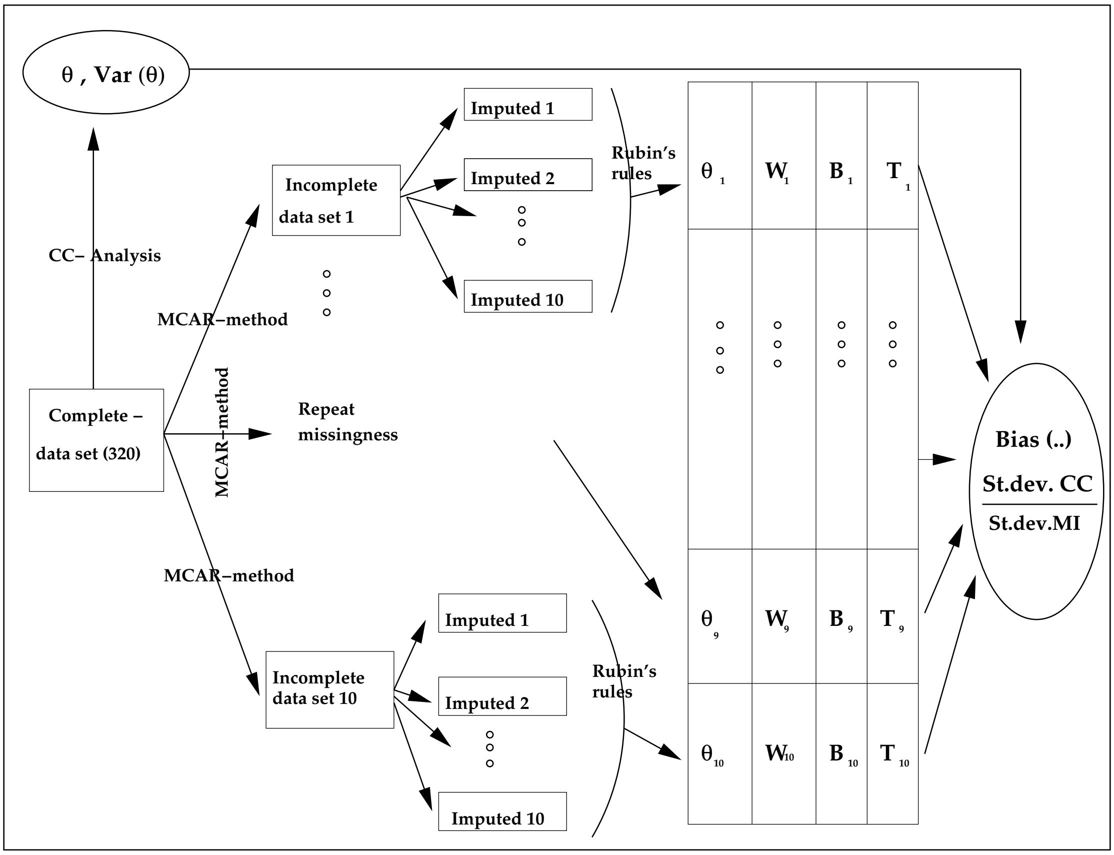

Figure 20: Simulation process to verify our imputation method, where CC is the complete case analysis of subjects with a complete set of measurements from the SGA-data set (320). θ_i is the parameter estimate of the imputed data set combined with Rubin's rules. The uncertainty in θ_i has two parts; W_i is the average within each imputation variance, B_i is the between-imputations variance and T_i is the total variability associated with θ_i. Estimated bias of the fixed effects for each parameter is the difference between $\text{mean}_{\beta_{\text{Complete-case}}}$ and $\text{mean}_{\beta_{\text{MI-method}}}$. Estimated standard deviation of the fixed effects is measured as a fraction between $\text{Std.dev}_{\beta_{\text{Complete-case}}}$ and the $\text{Std.dev}_{\beta_{\text{MI-method}}}$.

The Complete case data set is the original SGA Data set (561) with a complete set of measurements (320), and MI-method contains the data sets generated by the MCAR-procedure and the (MI-procedure + lme analysis). The result from this MCAR-simulation (the 10 repetitions of the MCAR procedure) are given in Table 12. Note that estimated parameters from the MI-method used in the estimated biases and fraction of the standard deviations from rep_i are combined with usage of Rubin's rules from the lme analysis with the 10 imputed data sets from the MCAR-procedure in each of the 10 repetitions as showed in Figure 20.

Coefficient	*rep* 1	*rep* 2	*rep* 3	*rep* 4	*rep* 5	*rep* 6	*rep* 7	*rep* 8	*rep* 9	*rep* 10
Bias$_{\beta_0}$	-1.1816	-0.7709	-0.2128	0.5242	-0.8013	-0.4752	0.3348	-0.3197	0.0503	-1.4100
Fraction$_{\beta_0}$	0.8756	0.9457	0.9326	0.9504	0.9523	0.9207	0.9519	0.9644	0.8420	0.9228
Bias$_{\beta_1}$	-0.0857	-0.0680	-0.0097	0.0725	-0.0854	-0.0361	0.0599	-0.0116	0.0556	-0.1070
Fraction$_{\beta_1}$	0.8600	0.9407	0.9428	0.8759	0.9279	0.9187	0.9195	0.9454	0.7950	0.9499
Bias$_{\beta_2}$	0.0000	0.0004	0.0004	0.0007	0.0001	-0.0007	-0.0008	-0.0002	0.0003	0.0005
Fraction$_{\beta_2}$	0.9029	0.9777	0.9087	0.9277	0.8437	0.8947	0.9381	0.91063	0.8320	0.8928
Bias$_{\beta_3}$	0.0576	-0.0041	0.0003	-0.0208	-0.0401	0.0546	-0.0561	-0.0215	-0.0207	0.0469
Fraction$_{\beta_3}$	0.9755	0.9621	1.0008	1.0096	0.9716	0.9821	0.9645	0.9820	1.0040	0.9730
Bias$_{\beta_4}$	0.0154	0.0213	0.0101	0.0002	0.0156	0.0148	0.0297	0.0243	0.0094	0.0456
Fraction$_{\beta_4}$	0.8711	0.9682	0.9174	0.9536	0.9493	0.8912	0.9607	0.9494	0.9482	0.8763
Bias$_{\beta_5}$	0.0295	0.0077	-0.0088	-0.0254	0.0138	0.0039	-0.0416	-0.0131	-0.0133	0.0045
Fraction$_{\beta_5}$	0.8750	0.8942	0.9591	0.9004	0.9162	0.8874	0.8980	0.9687	0.7743	0.9455
Bias$_{\beta_6}$	-0.2140	0.0280	-0.0441	-0.0969	-0.3905	-0.4419	-0.3786	-0.1386	-0.3383	-0.3974
Fraction$_{\beta_6}$	0.9409	0.9671	0.9347	0.9523	0.8955	0.9161	0.8962	0.9361	0.9640	0.8980
Bias$_{\beta_7}$	0.0003	0.0027	0.0004	-0.0007	0.0008	0.0003	0.0028	0.0010	0.0007	0.0041
Fraction$_{\beta_7}$	0.8662	0.9536	0.8974	0.9061	0.9297	0.8693	0.9439	0.9366	0.8963	0.9050
Bias$_{\beta_8}$	0.0033	-0.0003	-0.0004	-0.0022	0.0026	0.0011	-0.0056	-0.0004	-0.003	-0.0001
Fraction$_{\beta_8}$	0.8770	0.8862	0.9352	0.8488	0.9126	0.9447	0.8792	0.9587	0.7659	0.9511
Bias$_{\beta_9}$	-0.0264	0.0068	-0.0014	-0.0090	-0.0346	-0.0272	-0.0371	-0.0125	-0.0381	-0.0473
Fraction$_{\beta_9}$	0.8876	0.9609	0.9251	0.9191	0.8745	0.8971	0.8695	0.9073	0.9264	0.9000

Table 12: Estimated bias of the fixed effects for each parameter $\beta_0, ..., \beta_9$ between mean$_{\beta_{\text{Complete-case}}}$ and mean$_{\beta_{\text{MI-method}}}$, and estimated standard deviation of the fixed effects for each parameter measured as a fraction between Std.dev$_{\beta_{\text{Complete-case}}}$ and the Std.dev$_{\beta_{\text{MI-method}}}$ from the 320 subjects in the complete-case analysis.

In Table 13 we see the calculated average of the biases and average fraction of the standard deviation for each parameter

Variable	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
CC_{par}	66.677	2.5305	-0.0122	0.0277	0.1495	0.2088	-1.9599	0.0117	0.0210	-0.1682
$\overline{\text{Bias}}_\beta$	-0.5350	-0.0357	0.0001	-0.0067	0.0200	-0.0013	-0.2719	0.0015	-0.0002	-0.0259
CC_{dev}	1.5559	0.1468	0.0018	0.2269	0.0409	0.0440	0.5615	0.0037	0.0041	0.0507
$\overline{\text{Frac.}}_\beta$	0.9242	0.9066	0.9060	0.9827	0.9305	0.9020	0.9274	0.9141	0.8983	0.9058

Table 13: $\overline{\text{Bias}}_\beta$ are the calculated average bias ($\text{mean}_{\beta_{\text{Complete-case}}} - \text{mean}_{\beta_{\text{MI-method}}}$) and $\overline{\text{Fraction}}_\beta$ are average fraction of the standard deviation ($\frac{\text{Std.dev}_{\beta_{\text{Complete-case}}}}{\text{Std.dev}_{\beta_{\text{MI-method}}}}$) for each parameter $\beta_0, ..., \beta_9$. While CC_{par} and CC_{dev} are the estimates of the parameters and the standard deviations from the complete-case analysis (320).

From Table 13 we see that most of the calculated bias of the estimated parameter values from the MCAR-procedure are small and negative. The bias of β_0 is on average reduced by -0.5350 from the estimated value of the complete-data method with 320 subjects, and β_6 is on average reduced by -0.2719. The rest of the parameter estimates are approximately unchanged. The standard deviations of all the parameters have increased after the missing values have been imputed from the MI-method as expected.

6.3 Conclusions

Following the MCAR-procedure defined in Section 6.2 we have found the MI-method gives us reasonable imputed values for the missing values. Table 9 shows that there is not a large difference between the observed values and the imputed ones, when we used the complete-data method for analyzing the imputed data sets and combining the results with Rubin's rules we see from Table 11 that the estimates for the different parameters are almost equal to the ones from the complete-data method in Table 18. When we repeated the MCAR-procedure ten times we found out that there were a negative trend in the differences of the estimated parameter values and all the standard deviations of the parameters increased compared with complete case analysis.

The simulation study we used to evaluate the multiple imputation method worked very well. From Table 9 it is easy to see if the imputed values are reasonable. The results from Table 13 are as we expected, there are very little difference between the parameter estimates and the standard deviations of the parameter generated from the MCAR-procedure is larger than the estimates from the Complete case analysis with the subjects with a complete set of measurements (320) from the original SGA Data set (561).

7 Comparison of ML and MI on the SGA-Data set using a Simulation Study

In this chapter we compare the result of a maximum likelihood (ML)-procedure which uses a single model applied to $\mathrm{Y}_{\mathrm{obs}}$ alone to an analysis based on an multiple imputation (MI)-procedure (which use two models). Collins et al. (2001) assumes that the model for the complete-data population $P(Y_{\mathrm{com}};\theta)$ used in the ML analysis is the same model used to obtain the estimates and standard errors $(\hat{\theta},\ U_j)$, $j = 1, ..., m$ after performing the MI-method. Without this assumption there is no guarantee that the same population parameters are estimated under the two methods and the missing values are MAR. MAR (defined in Section 4.3) means that the probability that Y is missing may depend on the other variables X_i but not on Y itself. Collins et al. (2001) discusses three different propositions, but we will only use the first of these:

Proposition 1. *If the user of the ML procedure and the imputer use the same set of input data (same variables and observed units); if their models apply equivalent distributional assumptions to the variables and the relationships among them; if the sample size is large and if the number of imputations is sufficiently large; then the results from the ML and MI procedures will be essentially identical.*

Under this conditions the MI procedure will approximately perform a Bayesian analysis under the same model used in the ML-procedure. The asymptotic equivalence between the Bayesian and likelihood-based procedures is well known (Gelman et al. (1995b)). Note that with large samples the effect of a diffuse prior distribution will be diminished which will cause the MI (Bayesian) and ML analyses to produce similar results.

7.1 Comparing ECME and the LME-Method using the Complete Data

In the SGA-data set values are missing both for the variable MAD (that is the response variable in the lme model) and for the BMI variable (which is a covariate in the lme model for MAD). The MI-method based on the multivariate lme-method of Schafer (implemented in the PAN package) can handle this situation. This is however not the case for the EM-method of Schafer implemented in the PAN package, but this is theoretical possible to accomplish. BMI should have been used as a predictor for MAD. Therefore to compare the performance of the MI-Method with the ML-Method we use the original data set with a complete set of measurements (320) and will only delete values in the MAD variable.

The ML procedure applied to Y_{obs} uses a single model, and this *ecme* procedure should give us approximate equal results as with the multilevel analysis which is performed with the *lme* procedure. The results from this two methods is given in Table 14.

Parameter	*ECME*	*LME$_{REML}$*	*LME$_{ML}$*	*Difference$_{REML}$*	*Difference$_{ML}$*
β_0	68.5946	66.6775	66.6692	1.9171	1.9254
$\beta_{0_{\text{Std.Error}}}$	1.7482	1.5559	1.5460	0.1923	0.2022
β_1	2.6969	2.5305	2.5298	0.1664	0.1671
$\beta_{1_{\text{Std.Error}}}$	0.1605	0.1468	0.1461	0.0137	0.0144
β_2	-0.0122	-0.0122	-0.0122	0.0000	0.0000
$\beta_{2_{\text{Std.Error}}}$	0.0018	0.0018	0.0018	0.0000	0.0000
β_3	0.0277	0.0277	0.0285	0.0000	0.0008
$\beta_{3_{\text{Std.Error}}}$	0.2245	0.2269	0.2258	0.0024	0.0013
β_4	0.1495	0.1495	0.1495	0.0000	0.0000
$\beta_{4_{\text{Std.Error}}}$	0.0406	0.0409	0.0406	0.0003	0.0000
β_5	0.2094	0.2088	0.2091	0.0006	0.0003
$\beta_{5_{\text{Std.Error}}}$	0.0438	0.0440	0.0438	0.0002	0.0000
β_6	-1.9588	-1.9599	-1.9595	0.0011	0.0007
$\beta_{6_{\text{Std.Error}}}$	0.5579	0.5615	0.5577	0.0036	0.0002
β_7	0.0117	0.0117	0.0117	0.0000	0.0000
$\beta_{7_{\text{Std.Error}}}$	0.0037	0.0037	0.0037	0.0000	0.0000
β_8	0.0211	0.0210	0.0210	0.0001	0.0001
$\beta_{8_{\text{Std.Error}}}$	0.0041	0.0041	0.0041	0.0000	0.0000
β_9	-0.1682	-0.1682	-0.1682	0.0000	0.0000
$\beta_{9_{\text{Std.Error}}}$	0.0504	0.0507	0.0505	0.0003	0.0001

Table 14: Analyzing the complete-case (320) by the ecme and the lme-method, where Difference$_{\text{REML}}$ is the difference between the estimates of the ECME and LME$_{\text{REML}}$ method and Difference$_{\text{ML}}$ is the difference between the estimates of the ECME and LME$_{\text{ML}}$ method.

We emphasize that this was done to verify that the ML-procedure and the growth model for MAD (perform the multilevel analysis with a complete set of data) method give approximately equal results. The estimated parameters of the *ecme*-procedure are almost equal to the estimates from the *lme*-procedure, except the estimate of β_0 and β_1 where the difference is approximately 1 standard deviation. The 95% confidence interval for β_0 and β_1 with the *lme* procedure is: $\beta_0 \in \{63.5657, 69.7893\}$ and $\beta_1 \in \{2.2369, 2.8241\}$. The estimate of β_0 and β_1 by the *ecme* procedure is within this interval and therefore there is not a significant difference between the estimates of the two methods.

7.2 Simulation Study

We will now generate data set with different types if missing mechanisms, when we generated the different data sets with the different missingness-procedures we wanted to keep the rate of missingness as in the SGA Data set (561). From Table 2 we see that the rate of missingness of MAD by the four different time of gestations during the pregnancy are 3%, 23%, 21% and 24%, which means that from the complete data set (320), subjects with a complete set of measurements (320) from the original SGA

Data set, approximately 10 observations will be missing from the first antenatal visit, 74 observations will be missing from the second antenatal visit, 67 observations will be missing from the third antenatal visit and 77 observations will be missing from the fourth and last of the antenatal visits.

7.2.1 MCAR

Following a MCAR procedure we know that the missingness does not depend on the observed or unobserved data. We therefore randomly drew MAD observations to be missing such that the missing rates at each of the four occasions were fulfilled.

7.2.2 MAR

Following a MAR procedure we know that the missingness of the response variable may depend on the covariates, but not on the value of the response. To construct the simulation method using the MAR mechanism simple, we used the observed values of BMI of the mother at each of the four occasions as predictor variable of missingness by the MAD variable.

If the values of $\text{BMI}_i > \text{Constant}_i \quad i = \{1, 2, 3, 4\}$ it will cause MAD_i to be missing. The different values of the Constant which are selected to generate similar missing rates of MAD as in the full data set are

$$\text{Constant}_i = \begin{cases} 28.5 & \text{at the first occasion.} \\ 25.6 & \text{at the second occasion.} \\ 27.2 & \text{at the third occasion.} \\ 28.0 & \text{at the fourth occasion.} \end{cases}$$

A woman who has an $\text{BMI} > 25$ is classified as over-weighted.

7.2.3 MNAR_1

Following a MNAR procedure we know that the missingness of the response variable may depend on the covariates and on the missing value itself. Again to make the simulation-method using the MNAR mechanism simple we omit the current value of MAD if $\text{MAD}_i > \text{Constant}_i \quad i = \{1, 2, 3, 4\}$. The different values of the Constants which are selected to generate similar missing rates of MAD as in the full data set are

$$\text{Constant}_{\text{MAD}_i} = \begin{cases} 46.00 & \text{at the first occasion.} \\ 67.50 & \text{at the second occasion.} \\ 96.50 & \text{at the third occasion.} \\ 105.50 & \text{at the fourth occasion.} \end{cases}$$

By executing this procedure only one of the subjects got all four repeated measurements removed, and by performing an lme analysis with the complete case-data we only have a complete set of data from 187 of the 320 subjects.

7.2.4 MNAR$_2$

Another way of simulate a data set with missing values with the *MNAR* missing mechanism is to use the observed values of BMI of the mother at each of the four occasions as predictor variable in addition to the MAD variable itself to be predictors of missingness by the MAD variable. If the values of $\text{BMI}_i > \text{Constant}_i$ and $\text{MAD}_i > \text{Constant}_i \ \ i = \{1, 2, 3, 4\}$ it will cause the MAD variable$_i$ to be missing. The different values of the Constants to generate similar missing rates of MAD as in the full data set are:

$$\text{Const.}_{\text{BMI}_i} = \begin{cases} 22.04 & \text{at the first occasion.} \\ 23.88 & \text{at the second occasion.} \\ 25.24 & \text{at the third occasion.} \\ 26.42 & \text{at the fourth occasion.} \end{cases} \quad \text{Const.}_{\text{MAD}_i} = \begin{cases} 46.00 & \text{at the first occasion.} \\ 67.50 & \text{at the second occasion.} \\ 96.50 & \text{at the third occasion.} \\ 105.50 & \text{at the fourth occasion.} \end{cases}$$

All the Constants used by the BMI-variable are equal to the average values at each of the different occasion, and those observed MAD measurements which are larger that the constant used by the MNAR procedure represent the $3 - 20\%$ largest measurements at each occasion during the pregnancy. By executing this procedure none of the subjects had all four repeated measurements removed, but 15 subjects had the last three measurements during the study removed, and by performing an lme analysis with the complete-data set we only have a complete set of data from 251 of the 320 subjects.

7.2.5 Comparing the Results of the Different Missingness Mechanisms

The convergence behavior of the log-likelihood using the *ECME-procedure* on the complete data set, data set with missing data generated by the *MCAR-procedure*, data set with missing data generated by the *MAR-procedure* and the data set with missing data generated by the *MNAR*$_1$ and *MNAR*$_2$*-procedures* are shown in Figure 21.

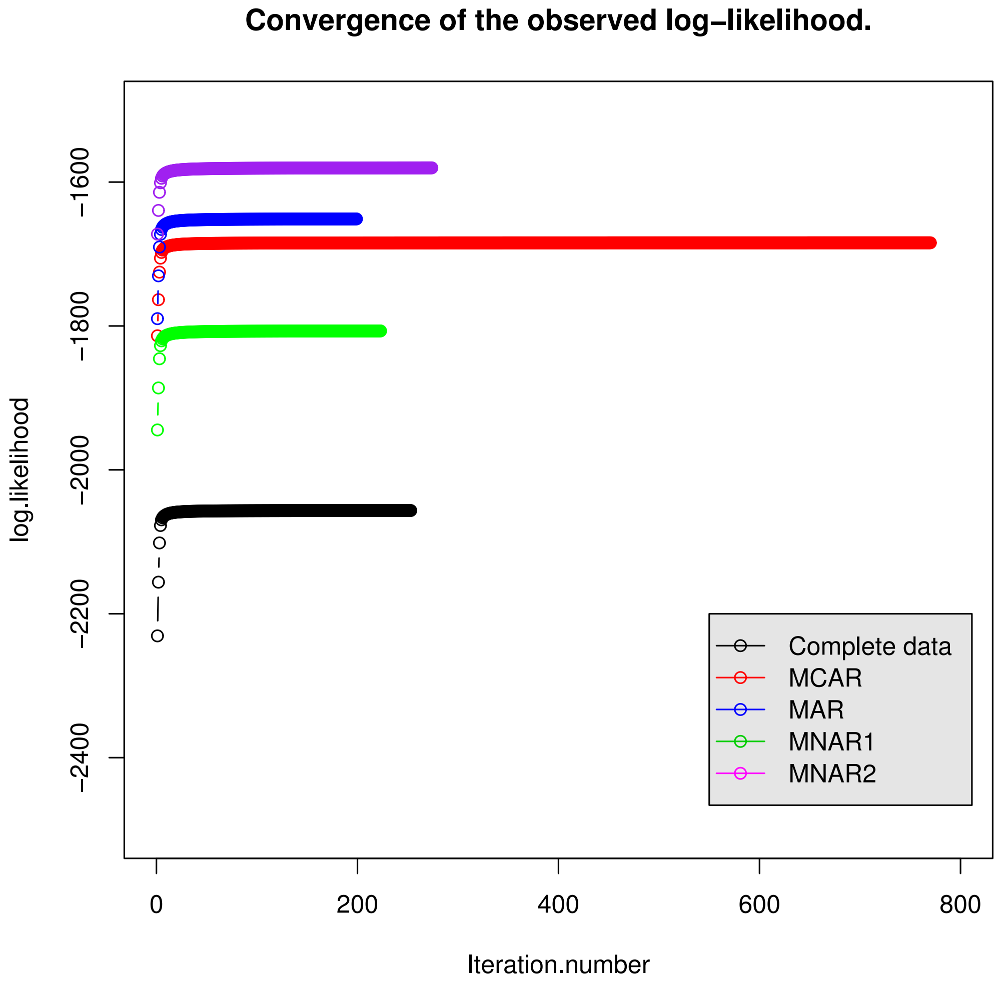

Figure 21: Convergence behavior of the log-likelihood by the different missingness-procedures. MNAR1 is based on the values of MAD and MNAR2 is based on the BMI values in addition to the MAD values.

We see that the log-likelihood of the *ECME*-procedure used on different data set with different missing mechanisms converged at different steps. The log-likelihood of the complete data set converged at 253 iterations and had the lowest value (−2056.505). The log-likelihood of the data set generated by the *MCAR* procedure converged at 770 iterations which was slowest at all the procedures, and it had the third highest value (−1684.405). The log-likelihood of the data set generated by the *MAR* procedure converged fastest (199 iterations) and had the second highest value with −1651.240. The log-likelihood of the data set generated by the $MNAR_1$ procedure which is based upon the MAD values itself in addition to the BMI values converged at 223 iterations, which was almost as fast as the *MAR* procedure and with almost the same value of the log-likelihood (-1806.901). The log-likelihood of the data set generated by the $MNAR_2$ procedure which is based upon only the MAD values itself converged at 274 iterations and had the highest value with -1580.245.

The result of executing the ECME, MI and lme method for the data sets with the three different missing mechanisms, is summarized in Table 16, together with the results on the complete data set for each situation.

Parameter-	MCAR			MAR			MNAR_1			MNAR_2			non-missing
Estimate	ECME	MI	CC	ECME	MI	CC	ECME	MI	CC	ECME	MI	CC	$\text{lme}^{(*)}$
β_0	68.3829	67.0163	68.5653	68.4573	69.4967	65.8628	**72.8140**	**71.1523**	**70.7421**	**70.7779**	**69.9675**	67.4455	66.6775
$\beta_{0_{\text{Std.Error}}}$	1.8096	1.6565	2.3663	2.2358	1.8841	2.3437	1.7786	1.6817	1.8001	1.6553	1.4890	1.6777	1.5559
β_1	2.6353	2.5796	2.4880	2.5654	**2.8543**	2.3753	**3.0966**	**2.9926**	**2.9588**	**2.9833**	**2.9550**	**2.8744**	2.5305
$\beta_{1_{\text{Std.Error}}}$	0.1699	0.1555	0.2129	0.2084	0.1812	0.2218	0.1634	0.1628	0.1680	0.1512	0.1511	0.1589	0.1468
β_2	-0.0111	-0.0102	-0.0125	-0.0132	-0.0099	-0.0141	-0.0100	-0.0098	-0.0095	-0.0095	-0.0089	**-0.0082**	-0.0122
$\beta_{2_{\text{Std.Error}}}$	0.0020	0.0019	0.0025	0.0021	0.0020	0.0025	0.0018	0.0017	0.0019	0.0018	0.0018	0.0021	0.0018
β_3	0.0022	0.0215	0.2651	-0.0677	-0.0839	-0.0225	-0.0590	-0.0575	-0.1336	-0.0665	-0.0632	-0.2864	0.0277
$\beta_{3_{\text{Std.Error}}}$	0.2268	0.2341	0.3315	0.2333	0.2378	0.2681	0.2213	0.2237	0.2507	0.2155	0.2203	0.2718	0.2269
β_4	0.1545	0.1474	0.1209	0.1375	0.1237	0.1231	0.0985	0.0952	0.1293	0.0861	0.0775	0.1239	0.1495
$\beta_{4_{\text{Std.Error}}}$	0.0420	0.0441	0.0592	0.0449	0.0501	0.0494	0.0408	0.0417	0.0459	0.0381	0.0387	0.0443	0.0409
β_5	0.2115	0.1914	0.1605	0.2581	**0.1160**	0.2771	**0.0896**	**0.0729**	**0.0429**	0.1500	0.1217	0.1465	0.2088
$\beta_{5_{\text{Std.Error}}}$	0.0459	0.0471	0.0682	0.0707	0.0746	0.0859	0.0447	0.0464	0.0513	0.0416	0.0415	0.0466	0.0440
β_6	-2.0228	-1.9669	-2.8266	-2.4900	-2.3240	-2.2483	-2.0997	-2.0686	-2.0215	-1.6490	-1.5989	-1.2603	-1.9599
$\beta_{6_{\text{Std.Error}}}$	0.5699	0.5695	0.7297	0.5919	0.6026	0.6604	0.5377	0.5399	0.5813	0.4892	0.4920	0.5138	0.5615
β_7	0.0131	0.0119	0.0099	0.0099	0.0085	0.0085	0.0063	0.0058	0.0073	0.0048	**0.0039**	**0.0028**	0.0117
$\beta_{7_{\text{Std.Error}}}$	0.0038	0.0040	0.0050	0.0039	0.0046	0.0042	0.0037	0.0038	0.0041	0.0034	0.0036	0.0040	0.0037
β_8	0.0222	0.0189	0.0253	**0.0300**	**0.0112**	**0.0309**	**0.0103**	**0.0083**	**0.0078**	0.0139	**0.0107**	0.0135	0.0210
$\beta_{8_{\text{Std.Error}}}$	0.0044	0.0045	0.0060	0.0066	0.0055	0.0081	0.0042	0.0044	0.0048	0.0039	0.0042	0.0044	0.0041
β_9	-0.1771	-0.1677	-0.2508	-0.2022	-0.1810	-0.1663	-0.1659	-0.1607	-0.1542	-0.1368	-0.1300	-0.1195	-0.1682
$\beta_{9_{\text{Std.Error}}}$	0.0522	0.0549	0.0621	0.0516	0.0538	0.0564	0.0476	0.0484	0.0517	0.0431	0.0450	0.0467	0.0507

Table 15: Results of analysis performed by the ECME and MI procedures with data set generated by the MCAR, MAR, MNAR_1 and MNAR_2 methods, and at each case a lme procedure is performed with a complete set of data, which is named CC. $\text{lme}^{(*)}$ is the lme procedure performed with the complete SGA Data Set (320). The MNAR_1 is the missing mechanism based upon the MAD values itself and in MNAR_2 is the missing mechanism based upon the MAD values and the BMI values. In the MI-method we have used 10 different imputed data sets and merged the estimates using Rubins rules. Use of boldface type in the table indicates the parameter-estimates which is not within the 95% $(\overline{\beta_i} \pm 2\sqrt{T})$ confidence interval from Table 21.

A rule of thumb that Schafer and Graham (2002) have found useful is that bias becomes problematic if its absolute size is greater than $|\text{bias}|_{\hat{\beta}} > \frac{1}{2}\text{standard deviation}_{\hat{\beta}}$. We make the following general observations of the missing values, which were imposed by the three different missing-mechanisms:

- MCAR – all the estimated parameter values from the ECME, MI and the lme method with a complete data set lie within the 95% - confidence intervals, and the estimated standard deviation from all the three methods are greater than the standard deviations estimated from the lme$^{(*)}$ method (defined in the caption in Figure 16). *Note* that the lme method with a complete data set have a larger standard deviation for each parameter than the ECME and MI methods.

- MAR – all the estimated parameter values from the ECME, the lme with a complete data set and almost all the parameter estimates from the MI method lies within the 95% - confidence intervals. The absolute values of the bias from the estimates of β_1, β_5 and β_8 from the MI method are greater than the rule of thumb. The estimated standard deviations from the ECME and MI methods are greater than the standard deviations from the MCAR simulation, but the standard deviations from the *lme* method with a complete data set have been reduced on six of the ten parameters, unchanged in one parameter (β_2) and increased on three parameters (β_1, β_5 and β_8).

- MNAR$_1$ – the parameter values estimated by this type of missing-mechanism are biased and their standard deviations are smaller than the standard deviations estimated from MAR and MCAR with both the ECME , MI and the lme method with a complete set of data. The absolute size of the biases of $\beta_0, \beta_1, \beta_5$ and β_8 are greater than the rule of thumb and will cause a larger residual in the model which describes MAD during the pregnancy . The estimated standard deviations from the MNAR simulation are smaller than the standard deviations estimated from the MCAR and MAR simulations, and smaller than the estimated standard deviations estimated by the lme$^{(*)}$ method except for the CC method. ($\text{Std.dev}_{\text{MAR}} > \text{Std.dev}_{\text{MCAR}} > \text{Std.dev}_{\text{lme}^{(*)}} > \text{Std.dev}_{\text{MNAR}_1}$)

- MNAR$_2$ – the parameter values estimated by this type of missing-mechanism are biased, but not so much as the other MNAR option and their standard deviations are smaller than the standard deviations estimated from MAR, MCAR and MNAR$_1$ with both the ECME, MI and the lme method with a complete set of data. We also see that the estimated values of β_5 and β_7 are not significantly biased compared with the MNAR$_1$ method. The fraction of the standard deviations of the CC analysis of the different missing mechanisms are: $\text{CC}_{\text{MCAR}} > \text{CC}_{\text{MAR}} > \text{CC}_{\text{MNAR}_1} > \text{CC}_{\text{MNAR}_2}$.

7.2.6 Further Investigations under the MAR Assumption

When we used the BMI variable in the MAR missing mechanism to generate data set containing missing values, and used both ECME and MI method to give us a complete-case analysis. It resulted in that the parameter estimates of the variable used to create the missing values in the data set were significantly biased compared to the parameter estimates from the lme$^{(*)}$ method. Checking this assumption further, we wanted to use AGE and LBW to generate two different data set with missing values using the MAR mechanism, and performed the ECME and MI method to generate complete cases, trying to assess the same trend.

MAR$_1$.

We use AGE as a predictor for MAD in the MAR missing mechanism. We divided the data set into four different groups such that

1. The first group contained only women whose AGE $\leq$ 22. In the full data set (561) there are 28 (4.99%) women who fulfill the age criteria, and the missing rates of MAD at the different time of gestation are 14.29%, 42.86%, 32.14% and 42.86%.
2. The second group contained women whose age is: 23 $\geq$ AGE $\leq$ 29. There were 304 (54.19%) women who fulfilled the age criteria for being in the second group, the missing rates of MAD at the different gestational time are 3.62%, 20.07%, 18.75% and 23.36%.
3. The third group contained women whose 23 $\geq$ AGE $\leq$ 35. There were 194 (34.58%) women who fulfilled the age criteria for being in the third group, the missing rates of MAD at the different gestational time are 1.03%, 24.23%, 21.65% and 21.65%.
4. The last group contained women whose AGE $>$ 35. There were 35 (6.24%) women who fulfilled the age criteria for being in the last group, the missing rates of MAD at the different gestational time are 0%, 28.60%, 25.71% and 28.57%.

Then we removed as many observations completely at random from each occasion of measurements in each of the four groups to obtain the appropriate amount of missing values. This is a MAR and not $MCAR$ mechanism because the quantity of MAD observations to be removed depends of the age of the women. We see that there are different rates of missingness in each group. In group 1 there are most missing values at the second and fourth occasion of measurements, compared with the other groups we see that from the second occasion of measurement there are missing more than 40% of the values and that is twice the amount (measured in percentage) which is missing in the other groups. From the third occasion of measurements we see that there is missing almost one third of the measurements, while in the other three groups there are missing approximately one fifth of the measurements. From the last occasion of measurements we see that there is missing 40% of the results in group one and that is almost twice the amount which is missing in the other groups.

MAR$_2$.

We use *LBW* as a predictor for MAD in the MAR missing mechanism, we divided the data set into three different groups such that

1. The first group contained only women who have not given birth to a child with low birth weight in one of the earlier pregnancies. In the full data set there are 494 (88.06%) women who fulfill the this criteria, and the missing rates of MAD at the different time of gestation are 2.65%, 22.20%, 20.57% and 23.01%.

2. The second group contained women who have given birth to one low weighted child from earlier pregnancies. There were 61 (10.87%) women who fulfilled this criteria for being in the second group, the missing rates of MAD at the different gestational time are 4.48%, 28.36%, 19.40% and 29.85%.

3. The third group contained women who have given birth to two low weighted children from earlier pregnancies. Totally there were only three (0.53%) women who fulfilled this criteria for being in the third group, the missing rates of MAD at the different gestational time are 33.33%, 66.67%, 100.00% and 66.67%.

Then we removed as many observations completely at random from each occasion of measurements in each of the three groups to obtain the appropriate quantity of missing values. This is also a MAR mechanism since the quantity of MAD observations to be removed depends of the different outcome from the *LBW* variable. We see that the number of subjects in each group are very different, there are 494 subjects in the first group and only 3 subjects in the third group. There are missing approximately the equal amount of missing values in the first and second group. The results of the complete-data method and complete-case analysis are given in Table 16.

Parameter-	MAR_1			MAR_2			
Estimate	ECME	MI	CC	ECME	MI	CC	$\text{lme}^{(*)}$
β_0	68.2614	66.9566	66.8450	67.8971	66.8012	68.0726	66.6775
$\beta_{0_{\text{Std.Error}}}$	1.8355	1.7091	2.4940	1.8551	1.7545	2.4501	1.5559
β_1	2.7119	2.6380	2.5310	2.6671	2.5669	2.6421	2.5305
$\beta_{1_{\text{Std.Error}}}$	0.1742	0.1612	0.2292	0.1707	0.1669	0.2399	0.1468
β_2	-0.0114	-0.0111	-0.0129	-0.0115	-0.0110	-0.0130	-0.0122
$\beta_{2_{\text{Std.Error}}}$	0.0020	0.0020	0.0026	0.0020	0.0019	0.0029	0.0018
β_3	0.0654	0.0822	0.3814	0.0289	0.0156	0.0618	0.0277
$\beta_{3_{\text{Std.Error}}}$	0.2343	0.2335	0.3568	0.2330	0.2397	0.3677	0.2269
β_4	0.1486	0.1447	0.1536	0.1605	0.1513	0.0988	0.1495
$\beta_{4_{\text{Std.Error}}}$	0.0422	0.0436	0.0673	0.0424	0.0438	0.0655	0.0409
β_5	0.2191	0.2026	0.1842	0.2281	0.2052	0.2180	0.2088
$\beta_{5_{\text{Std.Error}}}$	0.0466	0.0471	0.0647	0.0476	0.0510	0.0703	0.0440
β_6	-1.9175	-1.9447	-1.6932	-1.9741	-1.9036	-1.7673	-1.9599
$\beta_{6_{\text{Std.Error}}}$	0.5680	0.5765	0.7203	0.5825	0.5835	1.0778	0.5615
β_7	0.0109	0.0103	0.0125	0.0141	0.0134	0.0118	0.0117
$\beta_{7_{\text{Std.Error}}}$	0.0039	0.0040	0.0059	0.0038	0.0040	0.0061	0.0037
β_8	0.0214	0.0186	0.0201	0.0207	0.0181	0.0170	0.0210
$\beta_{8_{\text{Std.Error}}}$	0.0045	0.0046	0.0060	0.0045	0.0047	0.0069	0.0041
β_9	-0.1631	-0.1665	-0.1473	-0.1880	-0.1617	-0.1859	-0.1682
$\beta_{9_{\text{Std.Error}}}$	0.0528	0.0540	0.0631	0.0537	0.0539	0.1002	0.0507

Table 16: MAR_1 is the missing mechanism based upon the AGE of the mother and MAR_2 is the missing mechanism based upon the LBW (low birth weighted children from earlier pregnancies) variable. Use of boldface type in the table indicates the parameter-estimates which is not within the 95% confidence interval from Table 21. In each case a lme procedure is performed with a complete set of data, which is named CC.

We see that none of the estimated values of the parameters used generating missing values with the *MAR*-assumption are significantly different from the $\text{lme}^{(*)}$ estimates. One reason for this might be that the dependence (correlation) of the variables AGE and LBW are smaller on to MAD than the BMI variable. Since BMI is such a strong predictor for MAD a missing-procedure for BMI will also act as a missing-procedure for MAD.

7.3 Conclusion

We have verified that the ML-procedure and the growth model for MAD (perform the multilevel analysis with a complete set of data) gives approximately equal results. From our simulation study we have verified *proposition* 1 from Collins et al. (2001), the MI-procedure and the ML-procedure gives us approximately equal parameter estimates for the generated data set with different types of missing mechanisms. Both MNAR-procedures produced biased parameter estimates and the standard deviations were reduced. We believe that the covariates used in the MAR-procedure approximately will act as a MNAR-procedure if the correlation between the covariates and the response variable are large enough.

Another interesting observation in the simulation study is that estimates of the coefficients for variables used to generate the MAR and MNAR missing mechanism are "suffering" because they tend to be more biased compared to the values from the lme procedure performed on the complete SGA Data Set (320) than the other variables. According to the *MAR* assumption such a procedure should give unbiased parameter estimates.

8 Analyzing MAD Growth in the SGA Data set

8.0.1 Introduction

In this chapter we

(a) describe the selection procedure of variables included into the growth model for MAD.

(b) perform a complete data analysis on the SGA Data set (561) and compare the results to the complete case analysis with the subjects with a complete set of measurements (320) from the original SGA Data set.

8.1 Variable Selection Based on a Complete-Case SGA Data Set

In Section 2.2 the SGA Data Set was presented. Here we repeat the work based on backward elimination to arrive at a set of explanatory variables to be based in a linear mixed effects model with MAD as response variable.

We first fit a full set of covariates. From the estimates of the fixed effects from Table 17 we see from the *p-values* that there are many covariates which were not significant ($p > 0.05$). The covariates were omitted based upon the *p-values.* We omitted the covariate which had the biggest *p-value*, and re-estimated new coefficients until we only had significant variables left. This procedure is called *backwards elimination.*

Coefficient	*Value*	*Std.dev*	*DF*	*t-value*	*p-value*
Intercept	69.03488	2.5301035	555	27.285396	0.0000
Time*	2.66509	0.2495621	555	10.679081	0.0000
Time^{*2}	-0.00846	0.0023226	555	-3.642911	0.0003
Smoking	0.37239	0.4557491	184	0.817091	0.4149
Cigarettes	-0.02785	0.0412494	555	-0.675143	0.4999
Age	0.14434	0.0548158	184	2.633186	0.0092
Parity	0.76006	0.4629415	184	1.641810	0.1023
Weight	0.00585	0.0405019	555	0.144332	0.8853
BMI	0.17594	0.1318462	555	1.334435	0.1826
HB	-0.25660	0.1290669	555	-1.988134	0.0473
Gender	0.07923	0.4037541	184	0.196225	0.8447
LBW	-2.32386	0.7334272	184	-3.168496	0.0018
Time*:Smoking	0.10099	0.0406742	555	2.482940	0.0133
Time*:Cigarettes	-0.00471	0.0038111	555	-1.236099	0.2169
Time*:Age	0.00938	0.0048808	555	1.920991	0.0552
Time*:Parity	0.02336	0.0410673	555	0.568813	0.5697
Time*:Weight	0.00279	0.0035146	555	0.795007	0.4269
Time*:BMI	0.00905	0.0117005	555	0.773093	0.4398
Time*:HB	-0.00441	0.0145524	555	-0.302803	0.7622
Time*:Gender	0.02042	0.0358550	555	0.569413	0.5693
Time*:LBW	-0.17193	0.0648776	555	-2.650088	0.0083

Table 17: Estimates of the fixed effects from $(561-369)$ 192 subjects who had a complete set of observed measurements for the relevant variables.

We omitted Cigarettes, Smoking, Weight, Parity and HB from the model. Now we only have "significant" covariates left, because all the covariates are significant except *gender*, but we know that *gender* has an important effect. The growth of fetuses are different between the gender, girls are smaller and weigh less than boys. Concerning the MAD-variable this will show itself in from approximately 18 weeks of gestation, and the differences will increase towards the delivery. This have been confirmed by Goldenber et al. (1993). Their primary objective were to determine the importance of several maternal risk factors and fetal sex on specific fetal anthropometric measurements assessed by ultrasonograhy. Serial ultrasonographic examinations were performed on 1205 fetuses of multiparous women who ultimately gave birth at term. They measured among others femur length and abdominal circumference (MAD) at mean gestational ages of $18, 25, 31$ and 36 and they estimated a fetal weight. They used regression analyses to determine the effects on each measurement of maternal race, results from the study were that acting through their effect on head circumference, abdominal circumference and fetal length, each of the risk factors together with female sex were shown to have a negative effect on fetal weight.

The new model with the significant covariates for MAD is

$$\begin{aligned} y_{ij} = \beta_{0_j} + \beta_{1_j} x^*_{ij_{Time}} + \beta_{2_j} x^{*^2}_{ij_{Time}} + \beta_3 x_{ij_{Gender}} + \beta_4 x_{ij_{AGE}} + \beta_5 x_{ij_{BMI}} + \beta_6 x_{ij_{LBW}} \\ + \beta_7 x_{ij_{AGE}} x^*_{ij_{Time}} + \beta_8 x_{ij_{BMI}} x^*_{ij_{Time}} + \beta_9 x_{ij_{LBW}} x^*_{ij_{Time}} + \epsilon_{ij}. \end{aligned} \tag{38}$$

Complete data set

The estimates of the fixed effects from the complete data set are given in Table 18.

Covariate	*Coefficient*	*Value*	*Std.dev*	*DF*	*t-value*	*p-value*
Constant	β_0	66.6775	1.5559	954	42.8549	0.0000
Time*	β_1	2.5305	0.1468	954	17.2356	0.0000
Time^{*2}	β_2	-0.0122	0.0018	954	-6.9294	0.0000
Gender	β_3	0.0277	0.2269	316	0.1219	0.9031
AGE	β_4	0.1495	0.0409	316	3.6575	0.0003
BMI	β_5	0.2088	0.0440	954	4.7443	0.0000
LBW	β_6	-1.9599	0.5615	316	-3.4905	0.0006
Time*:Age	β_7	0.0117	0.0037	954	3.1757	0.0015
Time*:BMI	β_8	0.0210	0.0041	954	5.0877	0.0000
Time*:LBW	β_9	-0.1682	0.0507	954	-3.3147	0.0010

Table 18: Estimates of the fixed effects from the 320 subjects who had a complete set of values.

8.2 Complete Case (561)

Since we have assessed that the MI-method gives us reasonable imputed values for the missing values in Section 6 we used the MI-method (defined in Section 5.4.1) on the full data set based upon the SGA Data Set (561). In Table 19 we see the estimated values of the parameters from the *complete-data method* when we analyzed each of the 10 imputed data sets.

Parameter	*m=1*	*m=2*	*m=3*	*m=4*	*m=5*	*m=6*	*m=7*	*m=8*	*m=9*	*m=10*
$\hat{\beta}_0$	68.423	68.166	66.969	68.978	67.806	68.509	67.833	67.728	68.637	68.664
$\beta_{0_{\text{Std.Error}}}$	1.1508	1.1537	1.1827	1.1385	1.1329	1.1610	1.1543	1.1502	1.1748	1.1395
$\hat{\beta}_1$	2.6406	2.5915	2.5052	2.6545	2.5906	2.6158	2.5563	2.5503	2.6343	2.6640
$\beta_{1_{\text{Std.Error}}}$	0.1149	0.1133	0.1164	0.1120	0.1118	0.1148	0.1147	0.1131	0.1132	0.1161
$\hat{\beta}_2$	-0.0107	-0.0104	-0.0115	-0.0104	-0.0117	-0.0124	-0.0116	-0.0110	-0.0102	-0.0118
$\beta_{2_{\text{Std.Error}}}$	0.0013	0.0013	0.0014	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0014
$\hat{\beta}_3$	0.0503	0.0698	0.0912	0.0273	0.1345	-0.0476	-0.0219	0.0939	-0.1032	0.0902
$\beta_{3_{\text{Std.Error}}}$	0.1663	0.1695	0.1691	0.1673	0.1682	0.1685	0.1658	0.1696	0.1710	0.1657
$\hat{\beta}_4$	0.0732	0.0706	0.1070	0.0582	0.0746	0.0602	0.0650	0.0807	0.0635	0.0640
$\beta_{4_{\text{Std.Error}}}$	0.0282	0.0282	0.0290	0.0278	0.0276	0.0284	0.0283	0.0281	0.0288	0.0279
$\hat{\beta}_5$	0.2319	0.2377	0.2480	0.2236	0.2483	0.2441	0.2648	0.2456	0.2257	0.2308
$\beta_{5_{\text{Std.Error}}}$	0.0340	0.0342	0.0350	0.0339	0.0338	0.0346	0.0343	0.0341	0.0348	0.0340
$\hat{\beta}_6$	-1.5268	-1.2546	-1.1252	-1.2656	-1.2660	-1.2847	-1.2941	-1.5400	-1.2196	-1.5993
$\beta_{6_{\text{Std.Error}}}$	0.3613	0.3614	0.3720	0.3568	0.3538	0.3645	0.3627	0.3602	0.3693	0.3581
$\hat{\beta}_7$	0.0060	0.0082	0.0088	0.0072	0.0065	0.0057	0.0072	0.0082	0.0068	0.0063
$\beta_{7_{\text{Std.Error}}}$	0.0027	0.0026	0.0027	0.0026	0.0026	0.0027	0.0027	0.0026	0.0026	0.0027
$\hat{\beta}_8$	0.0242	0.0227	0.0257	0.0217	0.0247	0.0245	0.0253	0.0246	0.0226	0.0219
$\beta_{8_{\text{Std.Error}}}$	0.0033	0.0033	0.0034	0.0033	0.0033	0.0034	0.0033	0.0033	0.0033	0.0034
$\hat{\beta}_9$	-0.1437	-0.1151	-0.1034	-0.1279	-0.1168	-0.1114	-0.1098	-0.1461	-0.1160	-0.1383
$\beta_{9_{\text{Std.Error}}}$	0.0345	0.0338	0.0348	0.0335	0.0333	0.0344	0.0343	0.0339	0.0339	0.0348

Table 19: Estimates of the fixed effects and standard deviation of the parameters from the 561 subjects from each of the M imputed data sets.

We used the equations from Rubin (12,13, 14, 15) when we merged the results for $\beta_0, ..., \beta_9$ from each of the $M = 10$ data sets. In Table 20 we see the combined results for the parameters in the model for *MAD*.

Coefficient	*Mean*	*Average Var.*	*Between-imputation Var.*	*Total Var.*	*Std.deviation*
β_0	68.1720	1.3316e+00	3.5351e-01	1.7204e+00	1.3117
β_1	2.6003	1.3007e-02	2.6307e-03	1.5901e-02	0.1261
β_2	-0.0112	1.7959e-06	5.1933e-07	2.3672e-06	0.0015
β_3	0.0384	2.8263e-02	5.5689e-03	3.4389e-02	0.1854
β_4	0.0717	7.9753e-04	2.0314e-04	1.0210e-03	0.0320
β_5	0.2400	1.1742e-03	1.5867e-04	1.3487e-03	0.0367
β_6	-1.3376	1.3109e-01	2.5129e-02	1.5873e-01	0.3984
β_7	0.0071	7.0434e-06	1.0847e-06	8.2365e-06	0.0029
β_8	0.0238	1.1060e-05	2.0179e-06	1.3280e-05	0.0036
β_9	-0.1229	1.1639e-03	2.29167e-04	1.4160e-03	0.0376

Table 20: Estimates of the fixed effects from all 561 subjects from the SGA Data Set.

The square root of total variance is the overall standard error associated within $\overline{\beta}$. Note that if the were no missing data then $\hat{\beta}_{j_1}, \hat{\beta}_{j_2}, ..., \hat{\beta}_{j_M}$ would be identical, and B, Equation (14), would be zero and T, Equation (15), would simply be equal to W, Equation (13). The size of B relative to $\overline{W}$ reflects how much information is contained in the missing part of the data relative to the observed part. A rough 95% confidence interval can be obtained as $\overline{\beta_i} \pm 2\sqrt{T}$, but in general is is better to calculate intervals using the approximation $\overline{\beta_i} \pm t_{\mathrm{df}}\sqrt{T}$ where t_{df} denotes a quantile of the Students t-distribution with degrees of freedom equal to Equation (18). If the originally β-values lie in this confidence interval, then there is not a significant change in the estimated parameter values.

We see that the estimated value of β_4 from the complete case analysis in Table 21 have been reduced from 0.1495 to 0.0717, and does not lie in the 95% confidence interval from the MI-method. That means that there is a significant change in the estimated parameter value. The β_0 parameter have increased from 66.6775 to 68.1720, but it is not a significant change. The remaining parameters estimates are also not significant changed. Note that all the standard deviations of the parameters from the complete data method (561) analysis are reduced compared to the SGA Data Set complete case analysis (320).

	Analysis model (320)			Imputation model (561)		
Coefficients	Estimate	Std.dev	95% - confidence interval	Estimate	Std.dev	95% - confidence interval
β_0	66.6775	1.5559	$\{63.6279, 69.7271\}$	68.1720	1.3117	$\{65.5486, 70.7954\}$
β_1	2.5305	0.1468	$\{2.2428, 2.8182\}$	2.6003	0.1261	$\{2.3481, 2.8525\}$
β_2	-0.0122	0.0018	$\{-0.0157, -0.0087\}$	-0.0112	0.0015	$\{-0.0142, -0.0082\}$
β_3	0.0277	0.2269	$\{-0.4261, 0.4875\}$	0.0384	0.1854	$\{-0.3324, 0.4092\}$
β_4	0.1495	0.0409	$\{0.0693, 0.2297\}$	0.0717	0.0320	$\{0.0077, 0.1357\}$
β_5	0.2088	0.0440	$\{0.1226, 0.2950\}$	0.2400	0.0367	$\{0.1666, 0.3134\}$
β_6	-1.9599	0.5615	$\{-3.0604, -0.8594\}$	-1.3376	0.3984	$\{-2.1344, -0.5408\}$
β_7	0.0117	0.0037	$\{0.0044, 0.0190\}$	0.0071	0.0029	$\{0.0013, 0.0129\}$
β_8	0.0210	0.0041	$\{0.0130, 0.0290\}$	0.0238	0.0036	$\{0.0166, 0.0310\}$
β_9	-0.1682	0.0507	$\{-0.2676, -0.0688\}$	-0.1229	0.0376	$\{-0.1981, -0.0477\}$

Table 21: Estimate of the coefficients of the parameters from the model for MAD for the complete-case (320) and the MI complete-data (561).

8.3 Conclusion

In Section 8.1 we have found using complete case analysis that the growth model for MAD may be expressed by Equation (38). From the masters thesis of Eilertsen (2006) we know that during the first 20 weeks of gestation there is little individual variation in fetal growth. This is because the fetal genome is the major determinant of growth in early pregnancy. Later in pregnancy will environmental, nutritional and hormonal influences become increasingly more important. Thus growth differences and disorders become more evident in the second half of the pregnancy. Therefore, variables such as smoking and age, will affect the growth of the fetus such that it will not reach its growth potential.

Since MAD is the most important predictor of the fetal weight (Smith et al. (1994), Manning (1995) and Snijders and Nicolaides (1994))[6], we made a linear mixed effects model with MAD as response variable. In a growth model it is important to include both genetical variables of the mother and lifestyle variables such as smoking and overweight. We included the following variables, where the ones in italic were found to be significant in the growth model for MAD: *Gender, LWB, BMI, AGE, Time*, Parity, Smoking, Cigarettes and HB. It is surprising that Smoking did not became significant. In the SGA Data Set there are 200 (35.65%) subjects who were smoking, 359 (63.99%) were not smoking at the time of the conception and 2 (0.36%) subjects did not answer. Of the 200 subjects who were smoking, 50 (8.91%) smoked $1-9$ cigarettes per day and 150 (26.74%) smoked over 10 cigarettes each day. Throughout the pregnancy an increasing numbers quitted smoking and the average number of cigarettes smoked each day felled. Another interesting observation is that there were more and more missing values due to pregnancy length, at approximately 33 weeks of gestation 134 (23.89%) of the subjects reported smoking. Of the 134 subjects who were smoking, 57 (10.16%) of these smoked $1-9$ cigarettes per day and 77 (13.73%) smoked over 10 cigarettes each day. About time

[6] The references says that abdominal circumference (and not MAD) are most correlated to birth weight, but ultrasound linked to abdominal circumference is most linked to MAD.

of birth (47 (8.38%) and (72 (12.38%) of the 561 subjects reported respectively $1-9$ and over 10 cigarettes per day. It is important to be aware of that the growth model for MAD is based upon subjects from the SGA Data Set with a complete measurements on all the initial variables and not the full SGA Data set (561). Of the 200 (35.65%) subjects who were smoking at the time of conception only 94 (16.76%) subjects smoked during the pregnancy. Of the 192 subjects 131 (68.23%) did not smoke, and that may cause the effects of smoking during the pregnancy to have an non-significant effect in the model.

The parity of the women should also have been a significant variable in the growth model for MAD, a reason that it became non-significant is that every women who participated into the *SGA-study* have one or two children from earlier pregnancy. It is possible that we would have seen an effect of parity in the growth model between women who are pregnant for the first time and women who have earlier given birth, but we have no possibility to verify this hypothesis. We have observed that there is a non-significant effect between women who have given birth to one or two children from earlier pregnancies.

When we used the MI-method on the full SGA Data Set with missing values it resulted in that all the standard deviations of the parameter have been reduced compared with the complete case analysis. There were not a significant change in the parameter estimates except the coefficient of the age of the mother. From this results we see the benefits of using the MI-method compared to the ordinary complete case analysis.

9 Discussion and Conclusions

The analysis of data sets with missing values is one area is statistical science where real advances recently have been made. Modern missing data techniques which substantially improve upon old ad hoc methods are finally becoming available. Among these new techniques multiple imputation is especially powerful because of its generality. Unlike the ad hoc methods, MI solves the missing data problem in a statistically reasonable manner and incorporates missing-data uncertainty into all summary statistics. We now look into some issues that is of importance when analyzing missing data.

What is the Relationship Between the Model Used for Imputation and the Model Used for Analysis? - An imputation model should be chosen to be (at least almost) compatible with the analyses that subsequent will be performed on the imputed data sets. The imputation model should be good enough to preserve the relationships among the variables that will be focus of later investigation. An example from Collins et al. (2001), suppose that a variable Y is imputed under a normal model that includes the variable X_1. After imputation, the analyst then uses linear regression to predict Y from X_1 and another variable X_2 which was not in the imputation model. The estimated coefficient for X_2 from this regression would tend to be biased toward zero, because Y has been imputated without regard for its possible relationship with X_2. This means that in general any association that may prove important in subsequent analyses should be present in the imputation model, but he converse of this rule is not necessary. If Y has been imputed under a model that includes X_2, there is no need to include X_2 in future analyses involving Y unless its relationship to Y is of substantive interest. In our imputation model we have excluded the *gender* of the fetus and *age* of the mother, but they are included into the lme-method for predicting *MAD* because they are significant. Results appropriate to MAD are not biased by inclusion of extra variables in the imputation phase, because we have a rich imputation model that preserves a large number of associations. This is very desirable because it allows us to use the SGA Data Set (561) for a variety of post-imputation analyses.

What If the Missing Data Are Not "Missing at Random" - Most of the techniques presently available for creating multiple imputations assume that the missing values are missing at random (MAR), which means that missing values in the data set carry no information about probabilities of missingness (as in Figure 5). This assumption is mathematically convenient because it allows one to avoid an explicit probability of nonresponse. In some applications Schafer (2003) thinks that ignorability may seem implausible but with attrition in a longitudinal study such as in our case, it is possible that subjects drop out for reasons related to current data values. It is therefore important to note that the MI-model does not require or assume that nonresponse is ignorable. Imputations may in principle be created under any kind of assumptions/model for the missing-data mechanism and the resulting inferences will be valid under that mechanism.

Use of Growth Curves in the Clinic - During the pregnancy the clinicians use ultrasound to measure BPD (Biparietal diameter), MAD and FL (Femur length). To evaluate fetal growth, these measurements are compared to size-charts especially developed for the particular population. Studies have shown that customized size-charts, where the fetal growth is adjusted for pre pregnancy characteristics, can reduce the false-positive rate of growth restriction diagnosis in a normal populations Mongenelli and Gardosi (1996). Thus better revealing the fetuses who is pathological small (IGUR) from those are genetically small. The statistical methods in this thesis can be used to make customized size chart based on pre-pregnancy characteristics of the fetus and mother, but we need data set with children who is not IGUR, to estimate the coefficients in the growth model.

Missing Gestational Age - The imputation model where gestational age was included into the matrix with response variables from Section 5.4.3 were not useful compared with the imputation model where the missing values in the gestational Time-variable were replaced with the average gestational age at each of the 4 antenatal visits. Therefore we use the MI-procedure where the gestational age variable is included as covariates in the imputation model. This model has much smaller variance and the ACF-plot of the parameters in the imputation-model clearly shows that the estimates of the parameters are much more uncorrelated and the mixing is much better.

Results on the SGA Data Set - We have found a procedure to verify if the MI-method gives us reasonable imputed values for the missing values by following the MCAR-procedure defined in Section 6.2. Table 9 shows that there is not a large difference between the observed values and the imputed ones, when we used the complete-data method for analyzing the imputed data sets and combining the results with Rubin's rules we see from Table 11 that the estimates for the different parameters are almost equal to the ones from the complete-data method in Table 18.

Simulation Study - From the simulation study in Section 7.2 we have verified that the ML-procedure and the growth model for MAD gives approximately equal results. From our simulation study we have verified *proposition* 1 from Collins et al. (2001), the MI-procedure and the ML-procedure gives us approximately equal parameter estimates for the generated data set with different types of missing mechanisms. Both MNAR-procedures produced biased parameter estimates and the standard deviations were reduced. We believe that the covariates used in the MAR-procedure approximately will act as a MNAR-procedure if the correlation between the covariates and the response variable are large enough. Another interesting observation in the simulation study is that estimates of the coefficients for variables used to generate the MAR and MNAR missing mechanism are "suffering" because they tend to be more biased compared to the values from the lme procedure performed on the complete SGA Data Set (320) than the other variables. According to the *MAR* assumption such a procedure should give unbiased parameter estimates. As we discovered during the simulation study the implemented PAN ML-method can only handle data set with missing values in the response variable compared with the MI-method which can handle missing values in both the response variable and the covariates.

Predicting Fetal Weight - In Section 8.1 we have found using complete case analysis that the growth model for MAD may be modeled using by Equation (38). Since MAD is the most important predictor of the fetal weight we developed a model for this variable. In a growth model there is important to include both genetical variables of the mother and lifestyle variables such as smoking and overweight. We included the following variables, where the italic ones became significant in the growth model for MAD: *Gender, LWB, BMI, AGE, Time*, Parity, Smoking, Cigarettes and HB. It is surprising that Smoking did not became significant. A reason may be that there is such a small amount in the SGA Data Set who are smoking during the whole pregnancy and will cause this variable to have an non-significant effect in the model. The parity of the subjects should also have been a significant variable in the growth model for MAD, a reason that it became non-significant is that every women who participated into the *SGA-study* have one or two children from earlier pregnancy. It is possible that we would have seen an effect of parity in the growth model between women who are pregnant for the first time and women who have earlier given birth, but we have no possibility to verify this hypothesis. We have observed that there is a non-significant effect between women who have given birth to one or two children from earlier pregnancies. The growth model for MAD is

$$\begin{aligned} y_{ij} &= \beta_{0_j} + \beta_{1_j} x^{*}_{ij_{Time}} + \beta_{2_j} x^{2*}_{ij_{Time}} + \beta_3 x_{ij_{Gender}} + \beta_4 x_{ij_{AGE}} + \beta_5 x_{ij_{BMI}} + \beta_6 x_{ij_{LBW}} \\ &\quad + \beta_7 x_{ij_{AGE}} x^{*}_{ij_{Time}} + \beta_8 x_{ij_{BMI}} x^{*}_{ij_{Time}} + \beta_9 x_{ij_{LBW}} x^{*}_{ij_{Time}} + \epsilon_{ij}. \end{aligned}$$

When we used the MI-method on the SGA Data Set (561) with missing values it resulted in that all the standard deviations of the parameter have been reduced compared with the complete case analysis. There were not a significant change in the parameter estimates except for the coefficient for the age of the women. From this results we see the benefits of using the MI-method compared to the ordinary complete case analysis.

Suggestions to further work - The MI-method based on the multivariate lme-method of Schafer implemented in the PAN package can handle situations where missing values occur in both the response variable and the covariates. Since this situation is not possible to execute for the EM-method of Schafer (also implemented in the PAN package), a natural extension is to write a program allowing the EM-method handling the same situation.

References

Barnard, J. and Rubin, D. B. “Small-sample degrees of freedom with multiple imputation.” *Biometrika Trust*, 86(4):948–955 (1999).

Bråthen, E. W. “Multilevel analysis by growth of fetuses.” (2005). NTNU.

Brockwell, P. J. and Davis, R. A. *Introduction to Time Series and Forecasting*. Springer, second edition (2002).

Casella, G. and Berger, R. L. *Statistical Inference*. Duxbury, second edition (2001).

Collins, L. M., Schafer, J. L., and Kam, C.-M. “A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures.” *Psychological Methods*, 6(4):330–351 (2001).

Dempster, A. P., Laird, N. M., and Rubin, D. B. “Maximum likelihood from incomplete data via the EM algorithm.” *Journal of the Royal statistical society*, 39(1):1–38 (1977).

Dempster, A. P. and Rubin, D. B. *Incomplete Data in Sample Surveys*, volume 2. New York: Academic Press (1983). Pp 3-10.

Eilertsen, S. F. “Intrauterine growth patterns and birth outcomes.” Master’s thesis, NTNU (2006).

Gelman, A., Roberts, G. O., and Gilks, W. R. *Bayesian Statistics*, volume 5. Oxford University Press (1995a).

Gelman, A., Rubin, D. B., Carlin, J. B., and Stern, H. S. *Bayesian Data Analysis*. Chapman and Hall, London, first edition (1995b).

Goldenber, R. L., Davis, R. O., Cliver, S. P., Cutter, G. R., Hoffmann, H. J., Dubard, M. B., and Copper, R. L. “Maternal risk factors and their influence on fetal anthropometric measurements.” *American Journal of Obstetrics and Gynecology*, 168(4):1197–2003 (1993).

Goldstein, H. *Multilevel statistical models*. Oxford University Press Inc.,New York, third edition (2003).

Kenward, M. G. and Molenberghs, G. “Likelihood based frequentist inference when data are missssing at random.” *Statistical Science*, 13(3):236–247 (1998).

Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc. Hoboken, New Jersey, second edition (2002).

Liu, C. and Rubin, D. B. “ML estimation of the t distribution using EM and its extensions, ECM, ECME.” *Statistica Sinica*, 5:19–39 (1995).

Manning, F. A. *Fetal Medicine, Principles and practice*, chapter Intrauterine growth retardation, 317. Norwalk (1995).

Meng, X.-L. and Rubin, D. B. "Maximum likelihood estimation via the ECM algorithm: A general framework." *Biometrika*, 80:267–278 (1993).

Mongenelli, M. and Gardosi, J. "Reduction of false-positive diagnosis of fetal growth restriction by application of customized fetal growth standards." *Obstet Gynechol*, 88(5):8–844 (1996).

Pinheiro, J. C. and Bates, D. M. *Mixed-Effects Models on S and in S-PLUS*. Springer-Verlag New York,Inc, first edition (2000).

Rubin, D. B. "Inference and missing data." *Biometrika*, 63(3):581–592 (1976).

—. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, first edition (1987).

Schafer, J. L. *Analysis of Incomplete Multivariate data*. St. Edmundsbury Press, Bury St Edmunds, Suffolk, first edition (1997).

—. "Multiple imputation: a primer." *Statistical Methods in Medical Research*, 8(1):3–15 (1999).

—. "Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ." *Statistica Neerlandica*, 57(1):19–35 (2003).

Schafer, J. L. and Graham, J. W. "Missing data: Our View of the State of the Art." *Psychological Methods*, 7:147–177 (2002).

Schafer, J. L. and Yucel, R. M. "Computional Strategies for Multivariate Linear Mixed-Effects Models with Missing values." *Journal of Computational and Graphical Statistics*, 11(2):437–457 (2002).

SGA-Scandinavia. "SGA: Successive Small-for-Gestational-Age Births, a Longitudinal Study on Fetal Growth and Perinatal Outcome." (1986-1988). Small-for-Gestational-Age; The study design, the population and the variable list.

Smith, G. C. S., Smith, M. F. S., and McNay, M. B. "The relation between fetal abdominal circumference and birth weigth: Findings in 3512 pregnancies." *Br. J Obstet Gynecol*, 104–186 (1994).

Snijders, R. J. M. and Nicolaides, K. J. "Fetal biomery at 14 to 40 weeks gestation." *Ultrasound Obstet Gynecol*, 4(34) (1994).

A R documentation

A.1 LME (Linear Mixed-Effects) - Model Formulation

Description:

This generic function fits a linear mixed-effects model in the formulation described in Laird and Ware (1982) but allowing for nested random effects. The within-group errors are allowed to be correlated and/or have unequal variances.

Usage:

$lme(fixed, data, random, method, correlation, weights, subset, method, na.action,$
$control, contrasts = NULL)$

Variance functions are specified in the *lme* method using the *weights* argument. Weights equal to NULL corresponds to a homoscedastic variance model for the within-group errors. Variance models can be specified in *weights* either as a one-sided formula, in which case it is passed as the single argument to the *varFixed* constructor or as a *varFunc* object. This may be created using the standard constructors are described in §5.2.1 from *Mixed-Effects Models in **S** and **S-PLUS*** (Pinheiro and Bates, 2000).

Note:
The *lme-model* used in the lme fit always assume that there is a within-group error term being added to the response.

A.2 Description of The R routines used in Pan

This is a reproduction of the auxiliary file for pan and the ecme methods defined in Section 4.

A.3 MI-method

Gibbs sampler for the multivariate linear model with incomplete data. This function will typically be used to produce multiple imputations of missing data values in multivariate panel or clustered data. The underlying model is the same as in equation (19),where

- $y_i \ = (n_i * r)$ – matrix of incomplete multivariate data for subject or cluster i;
- $X_i = (n_i * p)$ – matrix of covariates;
- $z_i \ = (n_i * q)$ – matrix of covariates;
- $\beta \ = (p * q)$ – matrix of coefficients common to the population (fixed effects);
- $b_i \ = (q * r)$ – matrix of coefficients specific to subject or cluster i (random effects);

- $\epsilon_i \; = (n_i * r)$ – matrix of residual errors.

The matrix b_i, when stacked into a single column is assumed to be normally distributed with mean zero and unstructured covariance matrix Ψ, and the rows of ϵ_i are assumed to be independently normal with mean zero and unstructured covariance Σ. Missing values may appear in y_i in any pattern.

In most applications of this model, the first columns of X_i and Z_i will be constant (one) and Z_i will contain a subset of the columns of X_i.

usage:

pan(y, subj, pred, xcol, zcol, prior, seed, iter, start)

Arguments:

- y – matrix of responses. Each column of y corresponds to a response variable. Each row of y corresponds to a single subject-occasion. Missing values (NA) may occur in any pattern.

- subj –vector of length nrow(y), giving the subject indicators i for the rows of y.

- pred – matrix of covariates used to predict y. This should have the same number of rows as y. The first column will typically be constant (one), and the remaining columns correspond to other variables appearing in X_i and Z_i.

- xcol – a vector of integers indicating which columns of pred will be used in X_i.

- zcol – a vector of integers indicating which columns of pred will be used in Z_i.

- prior – a list with four components specifying the hyper-parameters of the prior distributions for Ψ and Σ.

- seed – integer seed for initializing pan()'s internal random number generator.

- iter – total number of iterations or cycles of the Gibbs sampler to be carried out.

- start – an optional list of quantities to specify the initial state of the Gibbs sampler. If "start" is omitted then pan() chooses its own initial state.

Note

This function assumes that the rows of y (and thus the rows of subj and pred) have been sorted by subject number. That is, we assume that subj=sort(subj),y=y[order(subj),], and pred=pred[order(subj),]. If the matrix y is created by stacking y_i, $i = 1, ..., m$ then this will automatically be the case.

A.4 ECME-method

Performs maximum-likelihood estimation for generalized linear models. This function will typically be used to produce multiple imputations of missing data values in multivariate panel or clustered data. The underlying model is the same as in equation (19),where

- $y_i \ = (n_i * 1)$ – vector of incomplete multivariate data for subject or cluster i
- $X_i = (n_i * p)$ – matrix of covariates;
- $z_i \ = (n_i * q)$ – matrix of covariates;
- $\beta \ = (p * 1)$ – vector of coefficients common to the population (fixed effects);
- $b_i \ = (q * 1)$ – vector of coefficients specific to subject or cluster i (random effects);
- $\epsilon_i \ = (n_i * 1)$ – vector of residual errors.

The vector b_i is assumed to be normally distributed with mean zero and unstructured covariance matrix psi, $b_i \sim N(0, \Psi)$ independently for $i = 1, ..., m$. The residual vector e_i is assumed to be $e_i \sim N(0, sigma^2 V_i)$, where V_i is a known $(n_i * n_i)$ matrix. In most applications is V_i equal to the identity matrix.

usage:

ecme(y, subj, occ, pred, xcol, zcol, vmax, start, maxits=1000, eps=0.0001, random.effects=F)

Arguments:

- y – vector of responses. This is simply the individual y_j vectors stacked upon one another. Each element of y represents the observed response for a particular subject-occasion or for a particular unit within a cluster.
- subj – vector of length nrow(y), giving the subject indicators i for the rows of y.
- occ – vector of same length as y indicating the "occasions" for elements of y. In a longitudinal data set where each individual is measured on at most $n_{\max}$ distinct occasions, each element of y corresponds to one subject-occasion and elements of occ should be coded as $1, 2, ..., n_{\max}$ to indicate these occasion labels.
- pred – matrix of covariates used to predict y. This should have the same number of rows as y. The first column will typically be constant (one), and the remaining columns correspond to other variables appearing in X_i and Z_i.
- xcol – a vector of integers indicating which columns of pred will be used in X_i.
- zcol – a vector of integers indicating which columns of pred will be used in Z_i, if zcol=NULL then the model is assumed to have no random effects.

- vmax – optional matrix of dimension c(max(occ),max(occ)) from which the V_i matrices will be extracted. In a longitudinal data set $v_{\max}$ would represent the V_i matrix for an individual with responses at all possible occasions $1, 2, ..., n_{\max} =$ max(occ). For individuals with responses at only a subset of these occasions, the V_i will be obtained by extracting the rows and columns of $v_{\max}$ for those occasions. If no $v_{\max}$ is specified will an identity matrix be used ($v_{\max}$ = identity).

- start – optional starting values of the parameters. If this arguments is not given then ecme() chooses its own starting values. This argument should be a list of three elements named β, Ψ and σ^2. *Note* that β should be a vector of the same length as "xcol", ψ should be a matrix of dimension c(length(zcol),length(zcol)) and σ^2 should be a scalar. This arguments has no effect if zcol=NULL.

- maxits – maximum number of cycles of ECME to be performed. The algorithm runs to convergence or until "maxits" iterations, whichever comes first.

- eps – convergence criterion. The algorithm is considered to have if the relative differences in all parameters from one iteration to the next are less than eps - that is if all($|\text{eps}_{\text{new}} - \text{eps}_{\text{old}}| < \text{eps} * |\text{eps}_{\text{old}}|$).

- random.effects – if *TRUE* it returns empirical Bayes estimates of all the random effects b_i, $i = 1, 2, ..., m$ and their estimated covariance matrices.

B R code:

Listing 1: Data file

```
############### loading the packages ###############

library(nlme)
library(foreign)
library(pan)
library(stats)

###############  Reading the data file ###############

datasett <-  read.spss("eystein␣ekstrakt␣7␣sept␣2005.sav")

#Make a new data set, with the random sample (561)

data <- is.na(datasett$V0053A)
data1 <- datasett$V0053A
index <- matrix(0,561,1)
j <- 0
for(i in 1:2072){
if (data[i] != "TRUE"){
  j <- j + 1
  index[j] <- i
}
}
datasett <- as.data.frame(datasett)
datasettnew <- datasett[index,]

#All the variables from datasettnew which are used during the analysis have
     been accumulated in Miss.Dat

Miss.Dat <- dget("Missing.dat")

#Using the imputation procedure with a complete set of time measurements,
    fetus and gegstational time are being used as random effects in the MI–
    method

data.set <- Miss.Dat

nc <- dim(data.set)[2]
nr <- dim(data.set)[1]
bmi <- data.set[,c(8,11,14,17)]/(data.set[,4]^2)
MAD <- data.set[,c(6,9,12,15)]
ymat <- cbind(c(t(MAD)),c(t(bmi)))
realtime <- data.set[,c(7,10,13,16)]
mastertime <- cbind(rep(16.76,nr),rep(24.62,nr),rep(32.47,nr),rep(36.63,nr)
    )
time <- mastertime
time[!is.na(realtime)] <- realtime[!is.na(realtime)]
time <- c(t(time))
```

```
#Centering the gestational age

time1 <- time - 27.62
time2 <- time1^2

xmat <- cbind(rep(1,nr*4),rep(data.set[,3]==2,each=4),time1,time2)
fetus <- rep(data.set[,1],each=4)

res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a=2,
    Binv=diag(2),c=4,Dinv=diag(4)),seed=123,iter=10000)

# Plot showing convergence of the parameter values

svec=1000:10000

# sigma

par(mfrow=c(2,2))
plot(svec,(res$sigma[1,1,svec]),type="l",main="MAD")
plot(svec,(res$sigma[1,2,svec]),type="l")
plot(svec,(res$sigma[2,1,svec]),type="l")
plot(svec,(res$sigma[2,2,svec]),type="l",main="BMI")
par(mfrow=c(1,1))

plot(density(res$sigma[1,2,svec]),type="l")
quantile(res$sigma[2,1,svec],c(0.15, 0.05,0.01))

par(mfrow=c(2,2))
acf((res$sigma[1,1,svec]),lag.max=100)
acf((res$sigma[1,2,svec]),lag.max=100)
acf((res$sigma[2,1,svec]),lag.max=100)
acf((res$sigma[2,2,svec]),lag.max=100)

#psi

par(mfrow=c(4,4),mar=c(3,3,3,1))
for (j in 1:4){
    for (i in 1:4){
        plot(svec,(res$psi[j,i,svec]),type="l")
    }
}

par(mfrow=c(4,4),mar=c(2.5,2.5,2.5,1))
for (j in 1:4){
    for (i in 1:4){
        acf(res$psi[j,i,svec],lag.max=100,main="")
    }
}

#beta

par(mfrow=c(2,4),mar=c(3,3,3,1))
for (i in 1:4){
    plot(res$beta[i,1,svec],type="l",main="MAD")

}

for (i in 1:4){

    plot(res$beta[i,2,svec],type="l",main="BMI")

}

par(mfrow=c(2,4),mar=c(3,3,3,1))
for (i in 1:4){
    acf(res$beta[i,1,svec],lag.max=100,main="MAD")
}
for (i in 1:4){
    acf(res$beta[i,2,svec],lag.max=100,main="BMI")
}

##### Executes the MCAR-method 10 times to assess the accuracy of the
    estimates of the MI-method #####

X <- is.na(Miss.Dat)
missing.pattern <- as.data.frame(table(as.data.frame(X)))
Missing.pattern <- missing.pattern[missing.pattern$Freq >0,]
Missing.pattern <- missing.pattern[missing.pattern$Freq >=10,]

Gruppe <- seq(1:dim(Missing.pattern)[1])
Missing.patnew <- cbind(Gruppe,Missing.pattern)

prob <- as.vector(Missing.patnew[,19])/(sum(Missing.patnew[,19]))

Mean.Par <- NULL
Tot.Par <-  NULL

j <- 0

while(j <= 10){
  Miss.Com <- na.omit(Miss.Dat)
  Miss.com <- na.omit(Miss.Dat)
  MCAR <- sample(1:length(Gruppe),320,replace = TRUE,prob)

  for(i in 1:length(MCAR)){

    if(MCAR[i] == 2){
      Miss.com[i,9] <- NA
    }

    if(MCAR[i] == 3){
      Miss.com[i,11] <- NA
    }

    if(MCAR[i] == 4){
      Miss.com[i,14] <- NA
    }

    if(MCAR[i] == 5){
      Miss.com[i,c(9,15)] <- NA
    }

    if(MCAR[i] == 6){
      Miss.com[i,c(9,12,15)] <- NA
    }

    if(MCAR[i] == 7){
      Miss.com[i,17] <- NA
    }

    if(MCAR[i] == 8){
      Miss.com[i,c(15,17)] <- NA
    }

  }

 nc <- dim(Miss.com)[2]
 nr <- dim(Miss.com)[1]
 bmi <- Miss.com[,c(8,11,14,17)]/(Miss.com[,4]^2)
 MAD <- Miss.com[,c(6,9,12,15)]
 time <- Miss.com[,c(7,10,13,16)]
 time <- c(t(time))

#Centering the gestational age

time1 <- time - 27.62
time2 <- time1^2

#Enter the response data into a matrix, one column for each variable

ymat <- cbind(c(t(MAD)),c(t(bmi)))

#Specifying the model to be used for imputation

 xmat <- cbind(rep(1,nr*4),rep(Miss.com[,3]==2,each=4),time1,time2)
 fetus <- rep(Miss.com[,1],each=4)

#Now we are ready to run pan().
#First we do a preliminary run of 10000 iterations.

 res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
    =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=123,iter=10000)

#Impute the missing data m=10 times taking 1000 steps between the
    imputations

  y1 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=9565,iter=1000,start=res$last)
  y2 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=6047,iter=1000,start=res$last)
  y3 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=3955,iter=1000,start=res$last)
  y4 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=4761,iter=1000,start=res$last)
  y5 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=9188,iter=1000,start=res$last)
  y6 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=9029,iter=1000,start=res$last)
  y7 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=4343,iter=1000,start=res$last)
  y8 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=2372,iter=1000,start=res$last)
  y9 <- res$y
  res <- pan(y=ymat,subj=fetus,pred=xmat,xcol=1:4,zcol=c(1,3),prior=list(a
     =2,Binv=diag(2),c=4,Dinv=diag(4)),seed=7081,iter=1000,start=res$last)
  y10 <- res$y

#MAD.lme are being executed 10 times
#creating data set for lme-method

Gender <- rep(Miss.Com[,2],each=4)
LWB <- rep(Miss.Com[,3],each=4)
AGE <- rep(Miss.Com[,5],each=4)
fetus <- rep(Miss.com[,1],each=4)
temp.Mad <- cbind(y1[,1],y2[,1],y3[,1],y4[,1],y5[,1],y6[,1],y7[,1],y8[,1],
    y9[,1],y10[,1])
temp.Bmi <- cbind(y1[,2],y2[,2],y3[,2],y4[,2],y5[,2],y6[,2],y7[,2],y8[,2],
    y9[,2],y10[,2])

B0 <- NULL
B1 <- NULL
B2 <- NULL
B3 <- NULL
B4 <- NULL
B5 <- NULL
B6 <- NULL
B7 <- NULL
B8 <- NULL
B9 <- NULL

stdB0 <- NULL
stdB1 <- NULL
stdB2 <- NULL
stdB3 <- NULL
```

```
stdB4 <- NULL
stdB5 <- NULL
stdB6 <- NULL
stdB7 <- NULL
stdB8 <- NULL
stdB9 <- NULL

for(i in 1:10){

  MAD <- temp.Mad[,i]
  BMI <- temp.Bmi[,i]
  MI.dat <- as.data.frame(cbind(fetus,MAD,time1,time2,AGE,BMI,Gender,LWB))

  MAD.lme <- lme(MAD ~ time1 + time2 + as.factor(Gender) + (AGE + BMI + as.
      factor(LWB) )*time1,random= ~time1|fetus,data=MI.dat)

  Std.Error <- data.frame(sqrt(diag(MAD.lme$"varFix")))
  Value <- data.frame(MAD.lme$coef[1])

  B0 <- c(B0, Value[1,1])
  B1 <- c(B1, Value[2,1])
  B2 <- c(B2, Value[3,1])
  B3 <- c(B3, Value[4,1])
  B4 <- c(B4, Value[5,1])
  B5 <- c(B5, Value[6,1])
  B6 <- c(B6, Value[7,1])
  B7 <- c(B7, Value[8,1])
  B8 <- c(B8, Value[9,1])
  B9 <- c(B9, Value[10,1])

  stdB0 <- c(stdB0,Std.Error[1,1])
  stdB1 <- c(stdB1,Std.Error[2,1])
  stdB2 <- c(stdB2,Std.Error[3,1])
  stdB3 <- c(stdB3,Std.Error[4,1])
  stdB4 <- c(stdB4,Std.Error[5,1])
  stdB5 <- c(stdB5,Std.Error[6,1])
  stdB6 <- c(stdB6,Std.Error[7,1])
  stdB7 <- c(stdB7,Std.Error[8,1])
  stdB8 <- c(stdB8,Std.Error[9,1])
  stdB9 <- c(stdB9,Std.Error[10,1])

}

#Calculate the parameter estimates by  Rubin's rules

param <- cbind(B0,B1,B2,B3,B4,B5,B6,B7,B8,B9)
std.param <- cbind(stdB0,stdB1,stdB2,stdB3,stdB4,stdB5,stdB6,stdB7,stdB8,
    stdB9)
var.param <-  std.param^2
average.par <- NULL
var.par <- NULL
mean.param <- NULL
tot.var <- NULL
```

```
for(i in 1:dim(var.param)[1]){

  mean.param <- c(mean.param,mean(param[,i]))
  average.par <- c(average.par, mean(var.param[,i]))
  var.par <- c(var.par, var(param[,i]))
  tot.var <- c(tot.var,(average.par[i] + ((1/dim(param)[1])+1)*var.par[i]))
}
  Mean.Par <- c(Mean.Par,mean.param)
  Tot.Par <-  c(Tot.Par,sqrt(tot.var))
  j <- j + 1
}

#Calculates the bias of the parameters and the fraction of the standard
    deviations

#Initial values from the complete case analysis (320)

Miss.Dat <- dget("Missing.dat")
Miss.Com <- na.omit(Miss.Dat)
bmi <- c(t(Miss.Com[,c(8,11,14,17)]/(Miss.Com[,4]^2)))
MAD <- c(t(Miss.Com[,c(6,9,12,15)]))
time <- Miss.Com[,c(7,10,13,16)]
time <- c(t(time))

#centered gestational age

time1 <- time - 27.62
time2 <- time1^2

Gender <- rep(Miss.Com[,2],each=4)
LWB <- rep(Miss.Com[,3],each=4)
AGE <- rep(Miss.Com[,5],each=4)
fetus <- rep(Miss.Com[,1],each=4)
MI.dat <- as.data.frame(cbind(fetus,MAD,time1,time2,AGE,bmi,Gender,LWB))

MAD.lme <- lme(MAD ~ time1 + time2 + as.factor(Gender) + (AGE + bmi + as.
    factor(LWB) )*time1,method="ML",random= ~time1|fetus,data=MI.dat)

Std.Error <- round(data.frame(sqrt(diag(MAD.lme$"varFix"))),4)
Value <- round(data.frame(MAD.lme$coef[1]),4)


Init.beta <- c(66.67750291,2.53052347,-0.01221834,0.02764627,0.14953252,
               0.20877076,-1.95990456,0.01172218,0.02099182,-0.16820022)

Init.Std <- c(1.555889215,0.146819685,0.001763264,0.226877126,0.040884245,
              0.044004905,0.561504001,0.003691179,0.004125998,0.050743426)

#Estimates from the MCAR-method

Simulation.ResBeta <- cbind(Mean.Par[1:10],Mean.Par[11:20],Mean.Par[21:30],
    Mean.Par[31:40],Mean.Par[41:50],Mean.Par[51:60],Mean.Par[61:70],Mean.
```

```
   Par[71:80],Mean.Par[81:90],Mean.Par[91:100],Mean.Par[101:110])

Simulation.ResStd <- cbind(Tot.Par[1:10],Tot.Par[11:20],Tot.Par[21:30],Tot.
   Par[31:40],Tot.Par[41:50],Tot.Par[51:60],Tot.Par[61:70],Tot.Par[71:80],
   Tot.Par[81:90],Tot.Par[91:100],Tot.Par[101:110])

#Calculates the results

for(i in 1:10){
  Simulation.ResBeta[i,] <-  Init.beta[i] - Simulation.ResBeta[i,]
  Simulation.ResStd[i,] <-  Init.Std[i]/Simulation.ResStd[i,]
}

totalt <- rbind(Simulation.ResBeta,Simulation.ResStd)

mean.beta <- NULL
mean.Std <- NULL
for(i in 1:10){
  mean.beta <- c(mean.beta,mean(Simulation.ResBeta[i,]))
  mean.Std <- c(mean.Std,mean(Simulation.ResStd[i,]))
}

################ End of MCAR-method ################

################ Simulation Study ################

#Complete case analysis

Miss.Com <- na.omit(dget("Missing.dat"))

Miss.Com[,6]<- MAD[,1]
Miss.Com[,9]<- MAD[,2]
Miss.Com[,12]<- MAD[,3]
Miss.Com[,15]<- MAD[,4]

Complete.MAR <- na.omit(Miss.Com)

bmi <- c(t(Complete.MAR[,c(8,11,14,17)]/(Complete.MAR[,4]^2)))
MAD <- c(t(Complete.MAR[,c(6,9,12,15)]))
time <- Complete.MAR[,c(7,10,13,16)]
time <- c(t(time))

#centered gestational age

time1 <- time - 27.62
time2 <- time1^2
Gender <- rep(Complete.MAR[,2],each=4)
LWB <- rep(Complete.MAR[,3],each=4)
AGE <- rep(Complete.MAR[,5],each=4)
fetus <- rep(Complete.MAR[,1],each=4)
MI.dat <- as.data.frame(cbind(fetus,MAD,time1,time2,AGE,bmi,Gender,LWB))
```

```
MAD.lme <- lme(MAD ~ time1 + time2 + as.factor(Gender) + (AGE + bmi + as.
    factor(LWB) )*time1,random= ~time1|fetus,data=MI.dat)

Std.Error <- round(data.frame(sqrt(diag(MAD.lme$"varFix"))),4)
Value <- round(data.frame(MAD.lme$coef[1]),4)

res <- cbind(Value,Std.Error)

#complete case (320) ecme-procedure

Complete <- na.omit(dget("Missing.dat"))

time <- c(t(Complete[,c(7,10,13,16)]))
time1 <- time - 27.62
time2 <- time1^2
MAD <- c(t(Complete[,c(6,9,12,15)]))
AGE <- rep(Complete[,5],each=4)
Gender <- rep(Complete[,2],each=4)
BMI <- c(t(Complete[,c(8,11,14,17)]/(Complete[,4]^2)))
LWB <- rep(Complete[,3],each=4)

pred <- cbind(rep(1,320),time1,time2,Gender,AGE,BMI,LWB,AGE*time1,BMI*time1
    ,LWB*time1)
occ <- rep(c(1,2,3,4),320)
subj <- rep(Complete[,1],each=4)
ymat <- MAD

result <- ecme(y=ymat,subj,occ,pred,xcol=1:10,zcol=c(1,2))

log.likelihood <- result$loglik
Iteration.number <- seq(1:length(log.likelihood))

#MCAR-procedure

mcar.dat <- na.omit(dget("Missing.dat"))

mcar17 <- sample(1:320,10)
mcar25 <- sample(1:320,74)
mcar33 <- sample(1:320,67)
mcar37 <- sample(1:320,77)

mcar.dat[mcar17,6] <- NA
mcar.dat[mcar25,9] <- NA
mcar.dat[mcar33,12] <- NA
mcar.dat[mcar37,15] <- NA

MAD <- cbind(mcar.dat[,6],mcar.dat[,9],mcar.dat[,12],mcar.dat[,15])

Miss.Com <- na.omit(dget("Missing.dat"))
Miss.Com[,6]<- MAD[,1]
Miss.Com[,9]<- MAD[,2]
Miss.Com[,12]<- MAD[,3]
Miss.Com[,15]<- MAD[,4]
```

```r
Complete.MCAR <- na.omit(Miss.Com)

bmi <- c(t(Complete.MCAR[,c(8,11,14,17)]/(Complete.MCAR[,4]^2)))
MAD <- c(t(Complete.MCAR[,c(6,9,12,15)]))
time <- Complete.MCAR[,c(7,10,13,16)]
time <- c(t(time))

#centered gestational age

time1 <- time - 27.62
time2 <- time1^2
Gender <- rep(Complete.MCAR[,2],each=4)
LWB <- rep(Complete.MCAR[,3],each=4)
AGE <- rep(Complete.MCAR[,5],each=4)
fetus <- rep(Complete.MCAR[,1],each=4)
MI.dat <- as.data.frame(cbind(fetus,MAD,time1,time2,AGE,bmi,Gender,LWB))

MAD.lme <- lme(MAD ~ time1 + time2 + as.factor(Gender) + (AGE + bmi + as.
    factor(LWB) )*time1,random= ~time1|fetus,data=MI.dat)

Std.Error <- round(data.frame(sqrt(diag(MAD.lme$"varFix"))),4)
Value <- round(data.frame(MAD.lme$coef[1]),4)

res <- cbind(Value,Std.Error)

#MAR-procedure

MAR <- na.omit(dget("Missing.dat"))
BMI <- MAR[,c(8,11,14,17)]/(MAR[,4]^2)

ant.BMI <- NULL
value.BMI <- c(28.5,25.6,27.2,28)

for(j in 1:4){
  teller.BMI <- 0;
  for(i in 1:320){
    if(BMI[i,j] > value.BMI[j]){
      teller.BMI <- teller.BMI + 1
    }
  }
  ant.BMI <- c(ant.BMI,teller.BMI)
}

for(j in 1:4){
  for(i in 1:320){
    if(BMI[i,j]> value.BMI[j]){
      MAR[i,(3+3*j)] <- NA
    }
  }
}

MAD <- MAR[,c(6,9,12,15)]
```

```
#MNMAR_1-procedure

MNAR <- na.omit(dget("Missing.dat"))
MAD <- MNAR[,c(6,9,12,15)]
ant.MAD <- NULL
value.MAD <- c(46,67.5,96.5,109.5)

for(j in 1:4){
  teller.MAD <- 0;
  for(i in 1:320){
    if(MAD[i,j] > value.MAD[j]){
      teller.MAD <- teller.MAD + 1
    }
  }
  ant.MAD <- c(ant.MAD,teller.MAD)
}

for(j in 1:4){
  for(i in 1:320){
    if(MAD[i,j] > value.MAD[j]){
      MNAR[i,(3+3*j)] <- NA
    }
  }
}

MAD <- MNAR[,c(6,9,12,15)]

#Complete case analysis (MNAR_1-procedure)

Miss.Com <- na.omit(dget("Missing.dat"))
Miss.Com[,6]<- MAD[,1]
Miss.Com[,9]<- MAD[,2]
Miss.Com[,12]<- MAD[,3]
Miss.Com[,15]<- MAD[,4]

Complete.MNAR <- na.omit(Miss.Com)

bmi <- c(t(Complete.MNAR[,c(8,11,14,17)]/(Complete.MNAR[,4]^2)))
MAD <- c(t(Complete.MNAR[,c(6,9,12,15)]))
time <- Complete.MNAR[,c(7,10,13,16)]
time <- c(t(time))

#centered gestational age

time1 <- time - 27.62
time2 <- time1^2
Gender <- rep(Complete.MNAR[,2],each=4)
LWB <- rep(Complete.MNAR[,3],each=4)
AGE <- rep(Complete.MNAR[,5],each=4)
fetus <- rep(Complete.MNAR[,1],each=4)
MI.dat <- as.data.frame(cbind(fetus,MAD,time1,time2,AGE,bmi,Gender,LWB))
```

```
MAD.lme <- lme(MAD ~ time1 + time2 + as.factor(Gender) + (AGE + bmi + as.
    factor(LWB) )*time1,random= ~time1|fetus,data=MI.dat)

Std.Error <- round(data.frame(sqrt(diag(MAD.lme$"varFix"))),4)
Value <- round(data.frame(MAD.lme$coef[1]),4)

res <- cbind(Value,Std.Error)

#MNAR_2-procedure

MNAR <- na.omit(dget("Missing.dat"))
BMI <- MNAR[,c(8,11,14,17)]/(MNAR[,4]^2)
MAD <- MNAR[,c(6,9,12,15)]
ant.BMI <- NULL
ant.MAD <- NULL
value.BMI <- c(22.04,23.88,25.24,26.42)
value.MAD <- c(46,67.5,96.5,109.5)

for(j in 1:4){
  teller.BMI <- 0;
  for(i in 1:320){
    if(BMI[i,j] > value.BMI[j]){
      teller.BMI <- teller.BMI + 1
    }
  }
  ant.BMI <- c(ant.BMI,teller.BMI)
}

for(j in 1:4){
  teller.MAD <- 0;
  for(i in 1:320){
    if(MAD[i,j] > value.MAD[j]){
      teller.MAD <- teller.MAD + 1
    }
  }
  ant.MAD <- c(ant.MAD,teller.MAD)
}

for(j in 1:4){
  for(i in 1:320){
    if((BMI[i,j] > value.BMI[j]) && (MAD[i,j] > value.MAD[j])){
      MNAR[i,(3+3*j)] <- NA
    }
  }
}

MAD <- MNAR[,c(6,9,12,15)]

#Complete case analysis (MNAR_2-procedure)

Miss.Com <- na.omit(dget("Missing.dat"))
Miss.Com[,6]<- MAD[,1]
Miss.Com[,9]<- MAD[,2]
Miss.Com[,12]<- MAD[,3]
Miss.Com[,15]<- MAD[,4]

Complete.MNAR <- na.omit(Miss.Com)

bmi <- c(t(Complete.MNAR[,c(8,11,14,17)]/(Complete.MNAR[,4]^2)))
MAD <- c(t(Complete.MNAR[,c(6,9,12,15)]))
time <- Complete.MNAR[,c(7,10,13,16)]
time <- c(t(time))

#centered gestational age

time1 <- time - 27.62
time2 <- time1^2
Gender <- rep(Complete.MNAR[,2],each=4)
LWB <- rep(Complete.MNAR[,3],each=4)
AGE <- rep(Complete.MNAR[,5],each=4)
fetus <- rep(Complete.MNAR[,1],each=4)
MI.dat <- as.data.frame(cbind(fetus,MAD,time1,time2,AGE,bmi,Gender,LWB))

MAD.lme <- lme(MAD ~ time1 + time2 + as.factor(Gender) + (AGE + bmi + as.
    factor(LWB) )*time1,random= ~time1|fetus,data=MI.dat)

Std.Error <- round(data.frame(sqrt(diag(MAD.lme$"varFix"))),4)
Value <- round(data.frame(MAD.lme$coef[1]),4)

res <- cbind(Value,Std.Error)


#ecme-procedure

Complete <- na.omit(dget("Missing.dat"))

occ.MAD <- NULL
subj.MAD <- NULL
MAD.new <- MAD
a <- is.na(MAD)

for(i in 1:320){
  for(j in 1:4){
    if(a[i,j]==FALSE){
      occ.MAD <- c(occ.MAD,j)
      subj.MAD <- c(subj.MAD,i)
      MAD.new[i,j] <- 1
    }
    else if(a[i,j]==TRUE){
      MAD.new[i,j] <- 0
    }
  }
}
```

```
time <- Complete[,c(7,10,13,16)]
BMI <- Complete[,c(8,11,14,17)]/(Complete[,4]^2)
bmi <- NULL
Time.keep <- NULL
for(i in 1:320){
  for(j in 1:4){
    if(MAD.new[i,j] > 0){
      Time.keep <- c(Time.keep,time[i,j])
      bmi <- c(bmi,BMI[i,j])
    }
  }
}

#centered gestational age

time1 <- c(t(Time.keep)) - 27.62
time2 <- time1^2

Sum <-  apply(MAD.new,1,sum)

LBW <- NULL
AGE <- NULL
Gender <- NULL
for(i in 1:320){
  LBW <- c(LBW,rep(Complete[i,3],each=Sum[i]))
  AGE <- c(AGE,rep(Complete[i,5],each=Sum[i]))
  Gender <-c(Gender,rep(Complete[i,2],each=Sum[i]))
}

ymat <- na.omit(c(t(MAD)))
ant<- length(ymat)

pred <- cbind(rep(1,ant),time1,time2,Gender,AGE,bmi,LBW,AGE*time1,bmi*time1
    ,LBW*time1)

#executing the ecme-method

resultMCAR <- ecme(y=ymat,subj=subj.MAD,occ=occ.MAD,pred,xcol=1:10,zcol
    =1:2)
log.likelihoodMCAR  <- resultMCAR$loglik
Iteration.numberMCAR <- seq(1:length(log.likelihoodMCAR))

resultMAR <- ecme(y=ymat,subj=subj.MAD,occ=occ.MAD,pred,xcol=1:10,zcol=1:2)
log.likelihoodMAR <- resultMAR$loglik
Iteration.numberMAR <- seq(1:length(log.likelihoodMAR))

resultMNAR <- ecme(y=ymat,subj=subj.MAD,occ=occ.MAD,pred,xcol=1:10,zcol
    =1:2)
log.likelihoodMNAR <- resultMNAR$loglik
Iteration.numberMNAR <- seq(1:length(log.likelihoodMNAR))

resultMNAR2 <- ecme(y=ymat,subj=subj.MAD,occ=occ.MAD,pred,xcol=1:10,zcol
    =1:2)
```

```
log.likelihoodMNAR2 <- resultMNAR2$loglik
Iteration.numberMNAR2 <- seq(1:length(log.likelihoodMNAR2))

#results

result <- c(resultMCAR,resultMAR,resultMNAR,resultMNAR2)

#PAN-procedure (only MAD-values are estimated)

Complete <- na.omit(dget("Missing.dat"))
Complete[,6]<- MAD[,1]
Complete[,9]<- MAD[,2]
Complete[,12]<- MAD[,3]
Complete[,15]<- MAD[,4]

MAD <- c(t(Complete[,c(6,9,12,15)]))
time <- c(t(Complete[,c(7,10,13,16)]))
time1 <- time - 27.62
time2 <- time1^2
AGE <- rep(Complete[,5],each=4)
Gender <- rep(Complete[,2],each=4)
BMI <- c(t(Complete[,c(8,11,14,17)]/(Complete[,4]^2)))
LBW <- rep(Complete[,3],each=4)

pred <- cbind(rep(1,320),time1,time2,Gender,BMI,AGE,LBW)

ymat <- c(t(MAD))
subj <- rep(Complete[,1],each=4)

res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
    (1),c=2,Dinv=diag(2)),seed=123,iter=10000)

  y1 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=9565,iter=1000,start=res$last)
  y2 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=6047,iter=1000,start=res$last)
  y3 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=3955,iter=1000,start=res$last)
  y4 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=4761,iter=1000,start=res$last)
  y5 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=9188,iter=1000,start=res$last)
  y6 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=9029,iter=1000,start=res$last)
  y7 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=4343,iter=1000,start=res$last)
```

```r
  y8 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=2372,iter=1000,start=res$last)
  y9 <- res$y
  res <- pan(y=ymat,subj,pred,xcol=1:7,zcol=c(1,2),prior=list(a=1,Binv=diag
      (1),c=2,Dinv=diag(2)),seed=7081,iter=1000,start=res$last)
  y10 <- res$y

#Executing lme-method 10 times

MAD.pan <- cbind(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10)

B0 <- NULL
B1 <- NULL
B2 <- NULL
B3 <- NULL
B4 <- NULL
B5 <- NULL
B6 <- NULL
B7 <- NULL
B8 <- NULL
B9 <- NULL

stdB0 <- NULL
stdB1 <- NULL
stdB2 <- NULL
stdB3 <- NULL
stdB4 <- NULL
stdB5 <- NULL
stdB6 <- NULL
stdB7 <- NULL
stdB8 <- NULL
stdB9 <- NULL

fetus <- rep(Complete[,1],each=4)

for(i in 1:10){

  MAD <- MAD.pan[,i]
  MI.dat <- as.data.frame(cbind(fetus,MAD,time1,time2,AGE,BMI,Gender,LBW))

  MAD.lme <- lme(MAD ~ time1 + time2 + as.factor(Gender) + (AGE + BMI + as.
      factor(LBW) )*time1,random= ~time1|fetus,data=MI.dat)

  Std.Error <- data.frame(sqrt(diag(MAD.lme$"varFix")))
  Value <- data.frame(MAD.lme$coef[1])

B0 <- c(B0, Value[1,1])
B1 <- c(B1, Value[2,1])
B2 <- c(B2, Value[3,1])
B3 <- c(B3, Value[4,1])
B4 <- c(B4, Value[5,1])
B5 <- c(B5, Value[6,1])
B6 <- c(B6, Value[7,1])
B7 <- c(B7, Value[8,1])
B8 <- c(B8, Value[9,1])
B9 <- c(B9, Value[10,1])

stdB0 <- c(stdB0,Std.Error[1,1])
stdB1 <- c(stdB1,Std.Error[2,1])
stdB2 <- c(stdB2,Std.Error[3,1])
stdB3 <- c(stdB3,Std.Error[4,1])
stdB4 <- c(stdB4,Std.Error[5,1])
stdB5 <- c(stdB5,Std.Error[6,1])
stdB6 <- c(stdB6,Std.Error[7,1])
stdB7 <- c(stdB7,Std.Error[8,1])
stdB8 <- c(stdB8,Std.Error[9,1])
stdB9 <- c(stdB9,Std.Error[10,1])

}

#Calculates the parameter estimates by Rubin's rules

param <- cbind(B0,B1,B2,B3,B4,B5,B6,B7,B8,B9)
std.param <- cbind(stdB0,stdB1,stdB2,stdB3,stdB4,stdB5,stdB6,stdB7,stdB8,
    stdB9)
var.param <-  std.param^2
average.par <- NULL
var.par <- NULL
mean.param <- NULL
tot.var <- NULL

for(i in 1:dim(var.param)[1]){

  mean.param <- c(mean.param,mean(param[,i]))
  average.par <- c(average.par, mean(var.param[,i]))
  var.par <- c(var.par, var(param[,i]))
  tot.var <- c(tot.var,(average.par[i] + ((1/dim(param)[1])+1)*var.par[i]))
}

Mean.Par <- mean.param
Tot.Par <-  sqrt(tot.var)
res <- round(cbind(Mean.Par,Tot.Par),4)

###############  END ###############
```