De transnationale Universiteit Limburg is een uniek samenwerkingsverband van twee universiteiten in twee landen:
de Universiteit Hasselt en Maastricht University

UNIVERSITEIT VAN DE TOEKOMST

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

Grip strength of severely malnourished children during nutritional rehabilitation in the Jimma hospital of Ethiopia

Promotor :
dr. Herbert THIJS

Promotor :
Prof.dr. MARITA GRANITZER

Sylvanus Fonguh

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

Grip strength of severely malnourished children during nutritional rehabilitation in the Jimma hospital of Ethiopia

Promotor :
dr. Herbert THIJS

Promotor :
Prof.dr. MARITA GRANITZER

Sylvanus Fonguh

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

A Simulation Study Comparing Some Methods for Missing at Random under Missing Not at Random Settings, and the Effects of Correlation on these Methods on Longitudinal Gaussian Data.

By

Fonguh Sylvanus F.

Supervisor:

Prof. Dr. Herbert THIJS

Thesis submitted in partial fulfilment for the completion of Masters in Statistics: Epidemiology and Public Health Methodology

August 2011

CERTIFICATION

This is to certify that this project was carried out by **FONGUH Sylvanus Foryam** under supervision.

FONGUH Sylvanus Foryam

…………………………………………………………………

Signature of student

Prof. Dr. Herbert THIJS

……………………………………………………………

Signature of supervisor

Acknowledgements

“I can do everything through Christ who gives me strength.” Philippians 4:13. This research would not have been possible without the support of God and many people. I wish to express gratitude to Prof. Dr. Herbert Thijs and Prof. Dr. Marita Granitzer for their invaluable assistance, support and guidance.

Special thanks to the general staff of CENSTAT for the knowledge imparted in me and most especially my group members for sharing the literature and invaluable assistance. I will be forever grateful. I also wish to express my love and gratitude to the people of KAI, my beloved family, and most especially my wife for her understanding, unconditional endless love, and the time not spent with her during my study period abroad.

Dedication

This thesis is dedicated to Nicky, Mami Mbom and Yankee. “Miyaka Nonoh.”

List of Tables

List of Figures

Appendix

List of Tables

List of Figures

Abstract

The frequent occurence of missing data in scientific studies is not uncommon. In longitudinal clinical studies for instance, one would expect, in a "dream land" scenario, a complete profile for each study subject. Unfortunately, this is rarely the case. Plagued therefore with the issue of missingness, it becomes vital to understand the mechanism that led to the missing data to be able to perform analyses that will lead to valid inference. Under the missing at random (MAR) assumption, likelihood based methods which admit ignoring the missing process as well as Multiple Imputation are valid. On the other hand the non-ignorable *missing not at random* (MNAR) process necessitates the need for models that explicitly incorporate the dropout mechanism. Unfortunately these models are usually based on very strong and unverifiable assumptions which are difficult to implement, and sensitive to misspecifications. The objective of this thesis was to compare the effectiveness of recommended methods for MAR scenarios like Direct Likelihood, and Multiple Imputation under MNAR scenarios, and also study the possible impact of correlation on the different methods based on simulations. Results provide evidence that ignorable analysis produce reasonably stable results even when the assumption of MAR is violated. There was also no significant effect of correlation on the results.

Key words: Missing at random, missing not at random, direct likelihood multiple imputation.

Table of Contents

1 INTRODUCTION

Generally, most scientific studies rely on data-based research, and as such are faced with issues of missing data which are often not intended, and are beyond the control of the investigator. In a data matrix where the rows represent the units, cases or subjects, and the columns represent variables measured for each unit; we are faced with the issue of missing data if some of the entries in the data matrix are not observed. According to Molenberghs and Kenward (2007), missing data often arises in studies performed on human subjects such as in clinical trials, epidemiological studies, sample surveys, psychometry and econometrics, where not all planned measurements are achieved for various reasons. In surveys and epidemiological studies, incompleteness could result from reasons such as refusal by respondents to answer certain items on a questionnaire (item nonresponse) or refusal to participate (total nonresponse) failure to reach the selected subjects and/or ask all questions Wang and Fan (2004). In clinical trials, even with the best design and monitoring to yield complete data, the loss of an experimental unit is possible through dropout, or failure to show up for a planned visit (Sotto, 2009).

The pattern in which missing data occurs defines which values in the data set are observed and which are missing. A monotone pattern of missingness arises when there are no measurements for a study unit at a specific time and thereafter (dropout or attrition). On the other hand, non monotone missingness refers to missing values with no defined arrangement, but rather occurring intermittently within the set of variables. This study will focus on monotone missingness.

The missing data mechanism on the other hand concerns the reasons for missingness, and whether these reasons are related to the values observed or unobserved in the data set. These reasons are often unknown or outside the control of the investigator, hence assumptions about the process generating them need to be made. Rubin (1976), introduced an important classification scheme for missing data mechanisms. It consists of;

a. Missing completely at random (MCAR) occurring when the missingness is independent of both unobserved and observed data. For instance, a patient who fails to show up for a planned visit because he/she is travelling for non-health reasons could be considered to be MCAR.

b. Missing at random (MAR) occurs when the missingness depends on observed data, and not on the unobserved measurements. For example, a patient could be MAR if after three consecutive visits with good outcome, he/she lapses at the forth visit and as a result drops out at the fifth visit due to their poor health condition.

c. Missing not at random (MNAR) where the probability of a measurement being missing depends on the unobserved data. For example, it may happen that a patient improves till the third visit, then worsens between the third and the fourth visit and consequently drops out at the fourth visit. He/she is MNAR because the reason for missingness is related to the yet-to-be observed worsening condition (sotto, 2009).

Schafer and Graham (2002) indicated that missing data creates difficulty in scientific research, and induces additional complexity in data analysis, because most data analysis was not designed for them. It is worth mentioning that the manner in which missingness is handled can have substantial implications on the conclusions because failing to cater for missingness may lead to biased and unreliable conclusions. Based on assumptions on the missing data mechanism by Rubin (1976), simple ad-hoc methods like complete case analysis (CC) and last observation carried forward (LOCF) which edit the data to lend an appearance of completeness are often implemented. These ad-hoc edits unfortunately, usually do more harm than good, producing solutions that are often bias, lacking in power, and unreliable. However research has paved the way for more flexible and reliable "state of the art" methods such as likelihood-based, Bayesian methodologies, Weighted Generalized Estimating Equations, and Multiple Imputation (Baraldi and Enders, 2010).

Generally the likelihood and Bayesian approaches admit ignoring the missing process under MAR. On the other hand, the missingness process cannot be ignored under a MNAR scenario, which is often referred to as the non-ignorable situation. With these results, certainly the scope of viable alternatives for analysis is narrowed down, but because these are based on unverifiable assumptions of the missing data mechanisms, it is difficult to determine which missing data mechanism is in play.

More so the non ignorable MNAR, missingness cannot be fully ruled out based on the observed data, thus necessitating the need for models that explicitly incorporate the dropout mechanism, like the selection models and pattern mixture models, see Molenberghs and Verbeke (2005). Unfortunately these models are usually based on very strong and unverifiable assumptions that maybe difficult to implement, and sensitive to misspecifications (Jansen et al., 2006). Plagued therefore with this difficulty, the researcher might turn to sensitivity analysis which considers different methods under varying missing data scenarios, and compare results.

The objective of this thesis is to compare the effectiveness of recommended methods for MAR scenarios like Direct Likelihood, and Multiple Imputation under MNAR scenarios, and also study the possible impact of correlation on the different methods based on simulations. The following paragraph gives an overview of the organization of this thesis.

Focusing on missingness in the response, this thesis first presents in section 2, a brief perspective of incomplete longitudinal data, and some fundamental concepts of missing data, and methods for handling missingness under the MAR assumption. Section 3 elaborates on the simulation study, data generation and analysis. The results of the simulation analysis are presented in section 4. Finally discussion conclusion and recommendations are presented in section 5.

2 INCOMPLETE LONGITUDINAL DATA

Generally, longitudinal studies often entail monitoring the effect of an experimental treatment via measurements of a particular variable repeatedly over time, with time being a covariate of interest. The duration and number of measurements are usually preplanned and are equal for all subjects (balanced data). The entire subjects' profiles are often of importance in evaluating the effectiveness of an experimental treatment. In a “dream land” scenario we would expect a complete profile for each study subject; completers. Unfortunately, this is unsurprisingly rarely the case. Plagued therefore with the issue of missingness, the approach to cater for incompleteness and the choice of statistical method may have important implications on the resulting conclusion. Consequently, this section considers some fundamental concepts and commonly used methodologies in the area of incomplete data.

2.1 LAY OUT

To better appreciate the notion of the various modeling frameworks, let's consider a longitudinal study where there is incomplete data. A full data density is as below:

$$f(\boldsymbol{y}_{i\,,}\boldsymbol{r}_i \backslash \boldsymbol{\theta}, \boldsymbol{\psi}) \tag{1}$$

$\boldsymbol{y}_i$ and $\boldsymbol{r}_i$ respectively represent vectors for the joint distribution of the ith subject's outcomes and missingness indicators. $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are the respective parameter vectors describing the response and the missingness processes respectively. The choice of factorization of (1) characterizes the various modeling frameworks.

Under a selection model (SeM) framework, (1) is factorized into a marginal model for the measurements and a conditional model for the non-response given the measurements (Rubin, 1976; Little and Rubin, 1987), that is,

$$f(\boldsymbol{y}_i\ \boldsymbol{r}_i \backslash \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i \backslash \boldsymbol{\theta}) f(\boldsymbol{r}_i \backslash \boldsymbol{y}_i\ , \boldsymbol{\psi}) \tag{2}$$

The selection models are an obvious choice for clinicians, who often are interested in the marginal effect ($\boldsymbol{\theta}$) of the independent variables on the response.

The reverse factorization of (1) into a marginal model for non-response and a conditional model for the measurements given the non response,

$$f(\boldsymbol{y}_i\ \boldsymbol{r}_i \backslash \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i \backslash \boldsymbol{r}_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i \backslash \boldsymbol{\psi}) \tag{3}$$

characterizes the Pattern-mixture models (PMM). In contrast with SeM, the parameter $\boldsymbol{\theta}$ in a PMM denotes pattern specific effects of independent variables on the response.

Finally, if (1) is factorized such that the measurements and the missing data processes are considered independent, conditioned on a set of latent variables or random effects (b_i),

$$f(\boldsymbol{y_i}\ \boldsymbol{r_i}\backslash\boldsymbol{\theta},\boldsymbol{\psi}) = f(\boldsymbol{y_i}\ \backslash\boldsymbol{b_i},\boldsymbol{\theta})f(\boldsymbol{r_i}\backslash b_i\ ,\boldsymbol{\psi}) \qquad (4)$$

Then a shared parameter model (SPM) results. Here, $\boldsymbol{\theta}$ represents the effect of independent variables, given the random effects.

2.2 METHODS OF HANDLING MISSING DATA

The intent of any analysis is to make valid inferences regarding a population of interest. Missing data threatens this goal, if the missing data creates a biased sample. Therefore, it is important to respond to a missing data problem in a manner which reflects the scientific question. Significant developments have been made in recent years regarding methodologies which handle responses to these problems and biases. Unfortunately, these methodologies are often not available to many researchers for reasons like the lack of familiarity and computational challenges, thus researchers often resort to ad-hoc approaches which may ultimately do more harm than good (Little & Rubin, 1987; Schafer and Graham, 2002). These ad-hoc approaches include complete case analysis (CC), last observation carried forward (LOCF), simple forms of imputation (conditional or unconditional mean imputation). However more reliable and “state of the art” methods such as direct likelihood, Bayesian analyses, multiple imputation, weighted generalized estimating equations and expectation maximization algorithms have been recommended by several authors (Schafer and Graham, 2002; Molenberghs and Kenward, 2007; Rubin *et.al.,* 2007). Bearing in mind that under certain assumptions the missingness process can be ignored, a sensible choice on the modeling process is important.

Since the main focus of this thesis is to compare the effectiveness of Direct Likelihood and Multiple Imputation under MNAR scenario, the following sections will present a general perspective of MAR and Ignorability.

2.3 MISSING AT RANDOM AND IGNORABILITY

This notion of non-response introduced by Rubin (1976) helps clarify under what conditions it could be possible to ignore the missing data process and still make valid inferences. In this section the concepts of Ignorability, direct likelihood and multiple imputation will be presented.

2.3.1 IGNORABILITY

The Likelihood-based methods commonly involve maximization of full data likelihood. However, faced with missingness, inference must be based on what is observed. Rubin (1976) stated that under precise assumptions, likelihood-based inference is valid when the missing data mechanism is ignored. Molenberghs and Verbeke (2005) formally showed that, once appropriate account is taken of what has been observed, there remains no dependence on unobserved data (at least in terms of probability models). Thus, as stated by Kenward and Molenberghs (1998), if the parameter describing the measurement process (θ) is functionally independent of the parameter describing the missingness process (ψ), then a separability or parameter distinctness condition is satisfied (within the likelihood framework), and under such a condition MCAR and MAR are ignorable (while an MNAR is non-ignorable). This implies that θ can be estimated directly from the observed data while ignoring the missingness process (unless the missingness process is of scientific interest). This is a task easily done by standard software procedures that allow for missing values. Worthy to mention is that contrary to Likelihood and Bayesian inference, the frequentist inference is ignorable only under MCAR (Molenberghs and Verbeke, 2005).

2.3.2 DIRECT LIKELIHOOD

It is worth mentioning that likelihood based methods (such as the Generalized Linear Mixed Models) can be applied to incomplete data after prior treatment of the missing values via CC or LOCF. Here, since the missing values are no longer present, the likelihood approach is based on the full-data likelihood. In contrast to this view, in this section for incomplete longitudinal data (with no manipulations), any method within the likelihood framework would require working with the observed-data-likelihood (Sotto, 2009). Recall that under ignorability (when the separability condition is met), under a MAR assumption for instance, only available cases can be used (observed-data-likelihood) for analysis. Thus, Verbeke and Molenberghs (2000) and Sotto (2009), pointed out that if the focus of inference lies on the response process parameter, estimation of the (conditional) non-response model (given the observed measurements) can altogether be ignored. Moreover, standard software procedures are available that allow for incomplete observations, hence, a fairly simple approach that permits the amount of information in the data to be maintained, leading to more valid (efficient) inference (Verbeke and Molenberghs, 2000). However, in scenarios where missingness (for instance, dropout) needs to be addressed, either by means of a dropout model for WGEE or by an imputation model for multiple-imputation-based GEE, then the missing data mechanism is not ignorable (Sotto, 2009).

2.3.3 MULTIPLE IMPUTATION

Introduced by Rubin (1978), and stated by Molenberghs and Kenward (2007), multiple imputation (MI) has become an important approach for dealing with statistical analysis of incomplete data. According to Molenberghs and Kenward (2007), the key idea is to replace each missing value with a set of M plausible values (Bayesian draw) from the conditional distribution of the unobserved values, given the observed ones such that the imputed values properly represents the information about the missing value that is contained in the observed data for the chosen model. This will result to a set of M complete datasets, which are then analyzed using standard complete data methods, and the results from the M analysis have to be combined into a single inference. The model used to perform the imputation is the imputation model, while that used to analyze the complete datasets is the substantive model. MI in its basic form requires the missing mechanism to be MAR, though the technique has been applied in MNAR settings as well (Molenberghs, Kenward and Laseffre 1997).

Given a vector of repeated measures $\boldsymbol{Y}_i = (\boldsymbol{Y}_i{}^o, \boldsymbol{Y}_i{}^m)$ with $\boldsymbol{Y}_i{}^o$ being the observed and $\boldsymbol{Y}_i{}^m$ the missing components, described by the parameter vector $\boldsymbol{\theta}$, during imputation, we aim at filling the missing data with draws from the conditional distribution$f(\boldsymbol{y}_i{}^m/\boldsymbol{y}_i{}^o, \boldsymbol{\theta})$. Since $\boldsymbol{\theta}$ is unknown, an estimate for it say $\widehat{\boldsymbol{\theta}}$, is first obtained from the data, after which we use $f(\boldsymbol{y}_i{}^m/\boldsymbol{y}_i{}^o, \widehat{\boldsymbol{\theta}})$ to impute the missing data. This means that we are generating samples from the distribution of $\widehat{\boldsymbol{\theta}}$, thus sampling uncertainty is catered for. With the imputed values, incomplete data are augmented to complete data, which are then used to obtain estimates of $\boldsymbol{\theta}$ and its variance $V_\theta = \widehat{Var}(\widehat{\boldsymbol{\theta}})$. These steps are repeated multiple times, say M times, producing $\widehat{\boldsymbol{\theta}}^m$, and $V_\theta{}^m$, for $m = 1, \ldots, M$. In pooling the results of the analysis of the M complete data sets into a single inference, an average of the estimates are taken, and are given by:

$$\overline{\widehat{\boldsymbol{\theta}}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{\theta}}^m \quad and\ estimated\ variance\ given\ by \quad \boldsymbol{V} = \boldsymbol{W} + \left(\frac{M+1}{M}\right)\boldsymbol{B}.$$

Where

$$\boldsymbol{W} = \sum_{m=1}^{M} \frac{V_\theta{}^m}{M} \quad and\ \boldsymbol{B} = \sum_{m=1}^{M} \frac{\left(\widehat{\boldsymbol{\theta}}^m - \overline{\widehat{\boldsymbol{\theta}}}\right)\left(\widehat{\boldsymbol{\theta}}^m - \overline{\widehat{\boldsymbol{\theta}}}\right)'}{M-1}$$

with $\boldsymbol{W}$ being the average within imputation variance, and **B** the between imputation variance (Rubin, 1986) cited by (Wayman, 2003; Molenberghs and Verbeke, 2005; Sotto, 2009). MI is an attractive method because of its ease of use, available software, and also its high efficiency even for small values of M, with 3-5 imputations sufficient to obtain excellent results (Molenberghs and Verbeke, 2005).

3 THE SIMULATION STUDY

A simulation is an imitation of some real process. The act of simulating something generally entails representing certain key characteristics or behaviors of a selected physical or abstract system. In this study we define two stages. Below the data generation mechanism is presented.

3.1 DATA GENERATION

Using the SAS IML procedure, 5000 simulations were run and data generated based on a data generating model which consists of a measurement model on one hand, and a dropout model given the measurement model on the other hand. For the measurement process, since we have a continuous outcome we assume a multivariate normal, a special case of the linear mixed effect model (Verbeke and Molenberghs 2000, cited by Janssen 2005) :

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + \varepsilon_i \tag{5}$$

Where $\boldsymbol{Y_i}$ is a n-dimensional response vector for subject $i, 1 \leq i \leq N$, N is the number of subjects, X_i is an (n x p) known design matrix, $\boldsymbol{\beta}$ is the p-dimensional vector containing the fixed effects, $\varepsilon_i \sim N(0, V_i)$ with V_i a covariance matrix having a AR(1) covariance structure. The Diggle-Kenward logistic regression model for the dropout process was used. (Diggle and Kenward 1994, cited by Molenberghs and Kenward 2007). This model allows the conditional probability for the dropout at occasion j given that the subject was still observed at the previous occasion to depend on the history ($\boldsymbol{h_{ij}}$), and possibly the unobserved current outcome, but not on future outcome $y_{ik}, k > j$. The model is represented as below:

$$logit\left[P\left(D_i = j \backslash D_i \geq j, \boldsymbol{h_{ij}}, y_{ij}, \boldsymbol{\varphi}\right)\right] = \varphi_o + \varphi_1 y_{ij-1} + \varphi_2 y_{ij}. \tag{6}$$

From the above equation, D_i is the occasion at which dropout occurs, and it is greater than 1, thus all first measurements were observed. Setting $\varphi_1 = 0$ in equation above resulted in a MNAR dropout process, and $\varphi_2 = 0$ resulted in a MAR dropout process. Based on the hypothetical data (Table 1), starting values for the parameters and their standard deviations were defined, with the parameter of interest (Treatment by time interaction) having an estimate of 12 with a standard error of 5. Data was generated from a normal distribution. The correlation (ρ) between outcome measurements was defined as low (0.1), medium (0.5), and high (0.8), in order to investigate the effect of correlation on the applied methods.

3.2 HYPOTHETICAL DATA

The simulated dataset consisted of two groups of patients, treated and untreated, with a group size of 50. Four measurements were taken for 4 weeks. The variable of main interest was the treatment by time interaction. The hypothetical data for 2 patients in the treatment arm, where response=1 for observed and 0 for missing, is presented on table 1.

Table 1: Hypothetical Dataset

ID	WEEKS	TREATMENT	RESPONSE
1	1	1	1
1	2	1	1
1	3	1	0
1	4	1	0
2	1	1	1
2	2	1	1
2	3	1	1
2	4	1	0

3.3 DATA ANALYSIS

The generated datasets were first explored to see the percentage missing, and the evolution of the subject profiles. In accordance with Verbeke and Molenberghs (2000), Molenberghs *et al.,* (2004), Janssen *et al.*, (2006), Molenberghs and Kenward (2007), the SAS procedure MIXED, can be used for missing data analysis without the need for any data manipulation when the separability condition is met, thus the MIXED procedure was used to fit a mixed model first to the complete data, then the data resulting from the MAR and MNAR processes. Multiple imputation with 3 imputations was used to complete the datasets, and the completed data sets were then analyzed with the MIXED procedure. The AR (1) covariance structure that was used for the data generation was maintained at the analysis stage. The MIANALYSE procedure was used to combine inferences from the imputed datasets, into a single inference, according to Rubin's formula. All analysis was repeated three times under three different correlations. Results were compared based on the mean of the parameter estimates, and the standard errors around the mean. Scatter and box plots were also used to better visualize results. In the sections that follow, results from the analysis are presented.

4 RESULTS

4.1 EXPLORATORY DATA ANALYSIS

The generated data sets were explored before analysis, to see the evolution of the subjects in the study. Figure 1 presents the evolution of individual profiles under the complete data, MAR and MNAR scenarios, for simulation number 2000. A histogram of the response under MNAR and MAR scenarios are shown in Fig 1 and 2 of the appendix respectively, they show that there is no departure from normality in the datasets; this was further confirmed by the Shapiro-Wilk test which showed no departure from normality.

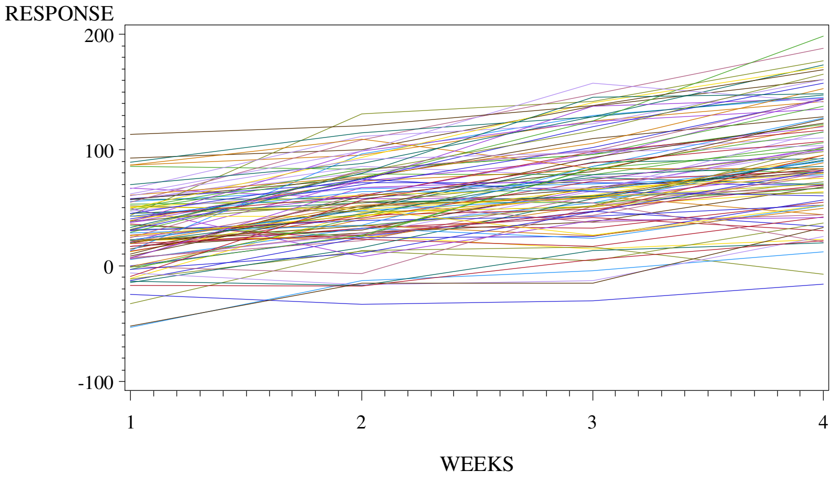

Figure 1A: Individual profile for the simulated complete dataset.

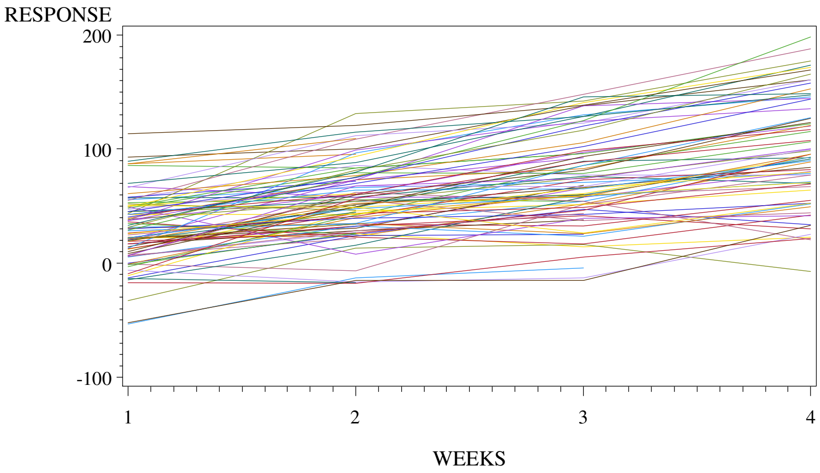

Figure 1b: Individual profile for the MAR scenario.

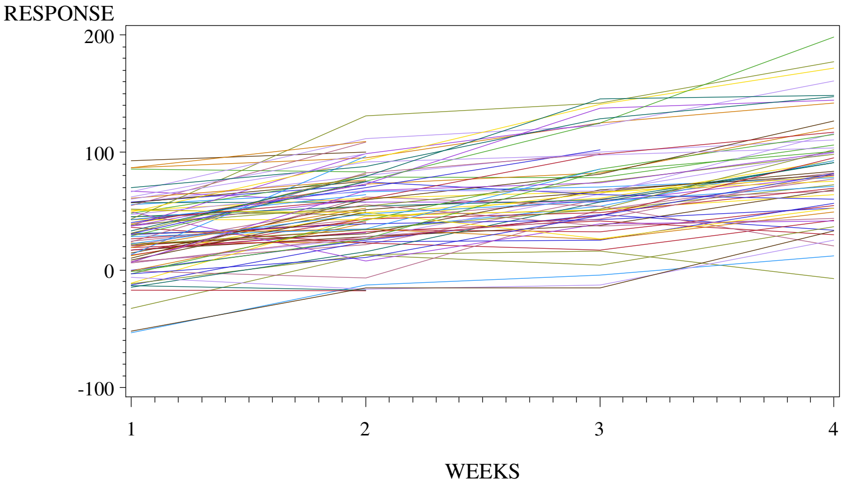

Figure 1c: Individual profile for the MNAR scenario.

From the figures above we can see that all subjects enter the study from the first week and dropout only occurs thereafter. From the plots it is evident that one can hardly distinguish MAR from MNAR, thus the need for a sensitivity analysis as advocated by Molenberghs and Kenward (2007). The MAR data had 50.24% missingness, while the MNAR data had 54.66% missingness.

4.2 STATISTICAL ANALYSIS

4.2.1 DIRECT LIKELIHOOD

The SAS procedure proc MIXED was used to analyze the generated data, since we had continuous repeated measurements and a Gaussian data according to Molenberghs and Kenward (2007). Table 2 below shows the mean of estimates, and standard errors around the mean in square brackets under the three different scenarios, using data generated with a correlation of 0.8

Table 2: Mean estimates (standard errors) from Direct Likelihood analysis of the complete, MAR and MNAR datasets.

Parameter	**Completers**	**MAR**	**MNAR**
Intercept	9.96 (5.04)	9.94 (5.11)	10.17 (5.11)
Time	13.99 (1.58)*	14.00 (1.77)*	13.78 (1.79)*
Treatment	-1.95 (7.19)	-2.02 (7.29)	-1.90 (7.29)
Treatment*Time	12.00 (2.33)*	12.06 (2.57)*	11.95 (2.65)*

**Significant at 5% level*

From Table 2 above, we see that results are slightly different under the different conditions. A closer look at the treatment by time interaction, we see that there is a gradual increase in standard error as we move from the completers through MAR to MNAR, and a decrease in the estimates. Also the difference between the MAR and MNAR is not alarming. One may attribute this slight difference to the differences in the simulated datasets, and maybe the differences in percentage missingness.

Fig 2 below shows no clear deviation from each other as points fall almost on the same line.

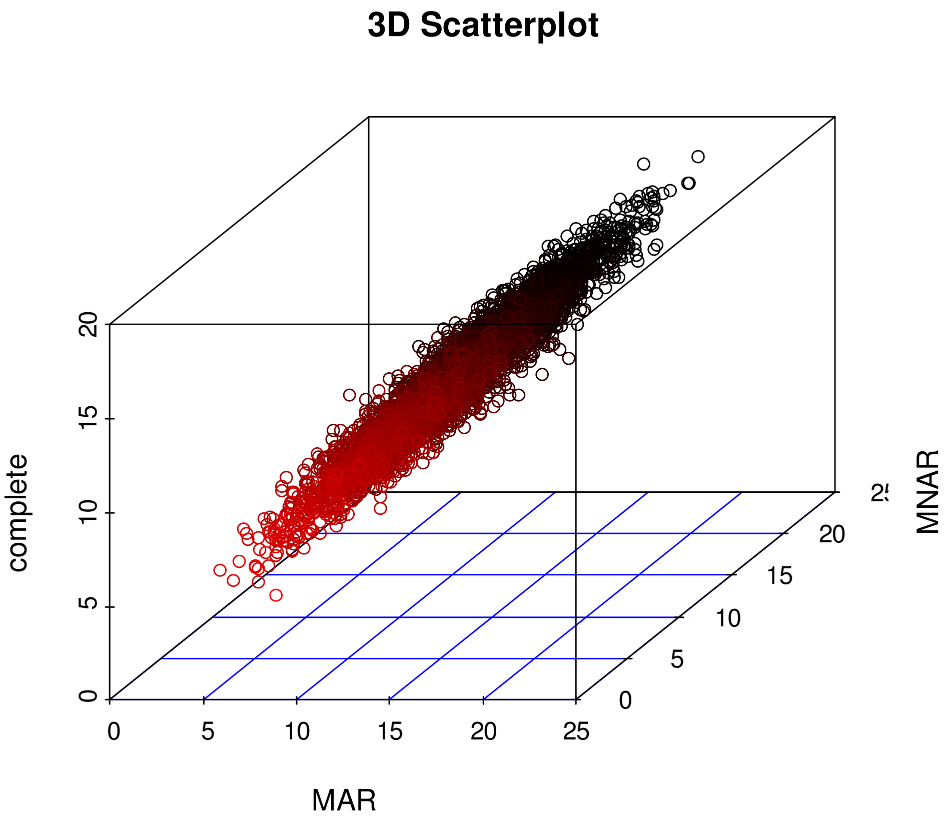

Figure 2: Three dimensional Plot of mean estimates of Treatment by time interaction for Completers, MAR and MNAR.

4.2.2MULTIPLE IMPUTATION

The missing values in the data were first completed by multiple imputation which caters for the variability around the missing values, and thus do not suffer from the issues of overestimating the precision by treating imputed and observed data on equal footing. The completed datasets were then analyzed with the MIXED procedure, and results combined with the MIANALYSE procedure. Table 3 below presents these results.

Table 3: Mean estimates (standard errors) from Direct Likelihood analysis of the complete data, MAR and MNAR datasets after Multiple Imputation.

Parameter	Completers	MAR	MNAR
Intercept	9.96 (5.04)	9.96 (5.00)	10.18 (5.10)
Time	13.99 (1.58)*	13.98 (1.45)*	13.78 (1.45)*
Treatment	-1.95 (7.19)	-1.96 (7.13)	-1.91 (7.10)
Treatment*Time	12.00 (2.33)*	12.01 (2.19)*	11.96 (2.15)*

**Significant at 5% level*

Results from multiple imputation under MAR and MNAR were also similar as can be seen from Table 3, and they are also very close to results from the completers, thus indicating that MI can be used under both scenarios, though not totally ruling out sensitivity analysis as a tool to gain more power and assurance when making inferences. As mentioned earlier, the very slight differences in the values could be attributed to differences in the simulated data and the percentage missing values.

4.3 DIRECT LIKELIHOOD AND MULTIPLE IMPUTATION UNDER VARYING CORELATION.

In an attempt to examine the effect of correlation on the different methods, the data were generated under varying correlations, and the methods were implemented. Since data were generated at every occasion with a different correlation, it is obvious that totally different values will be generated, thus for a better visualization of the results, box plots of the interaction between treatment and time will be presented. See tables 1-5 of the appendix for mean estimate and standard errors.

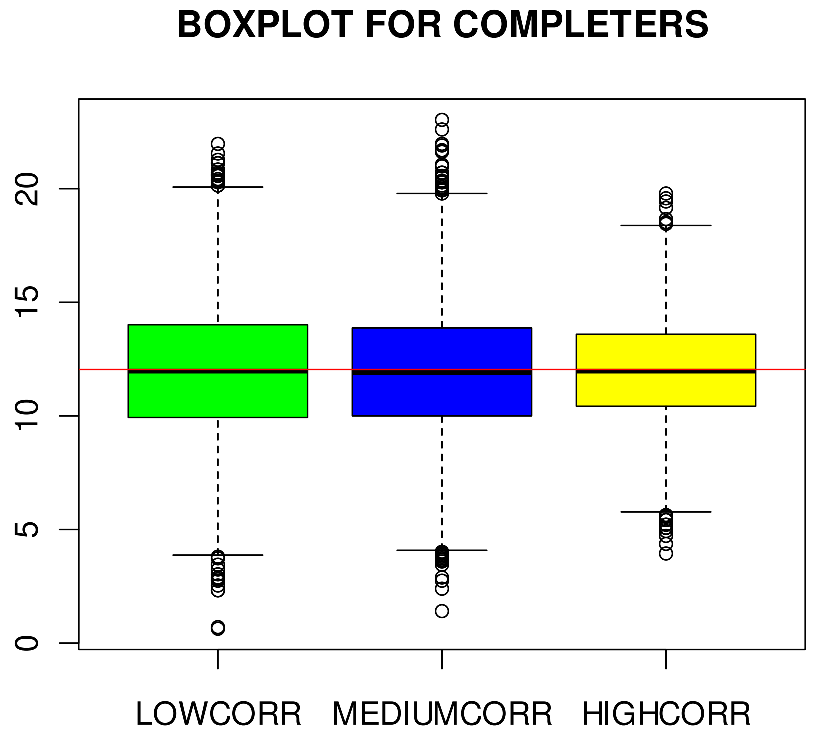

Figure 3: Box plot of the mean estimates of treatment by time interaction for the complete data under different correlations.

From figure 3 above, we see that the correlation has no major effect on the direct likelihood method, as all the estimates are approximately around 12 (red horizontal line) which is the expected, and do not deviate from each other. It is worth mentioning that though the estimates do not deviate from each other, the outlying observations gradually decrease as we move from the low to high correlation and variability decreases at high correlation.

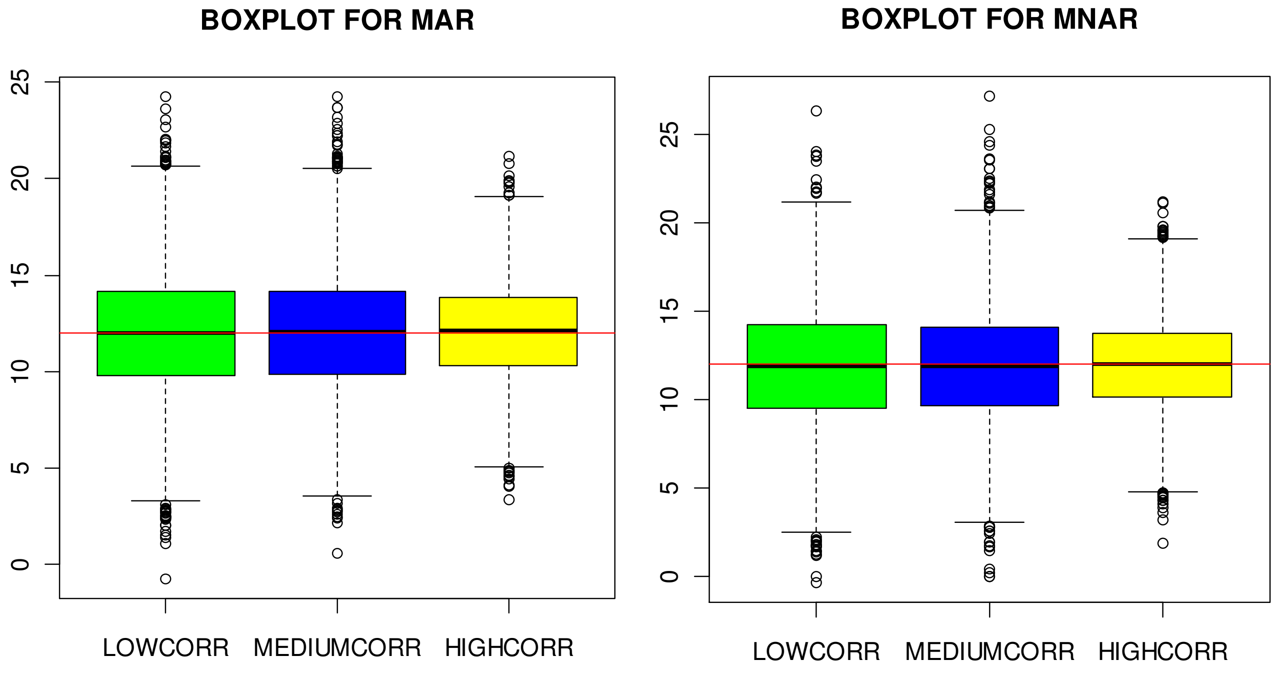

Figure 4: Box plots of the mean estimates for MAR and MNAR data under different correlations.

As for the complete case, we see from Figure 4 above, the same trend as we move from low to high correlation. Also we still see that the mean estimates are pretty stable and not deviating very much from each other.

For the case of Multiple Imputation, as can be seen from Figure 5, there are no deviations, and the outlying observations gradually reduce as we move towards a high correlation with an increase in precision of the estimates seen with a decrease in variability as depicted by the size of the boxes as we move from low to high correlations.

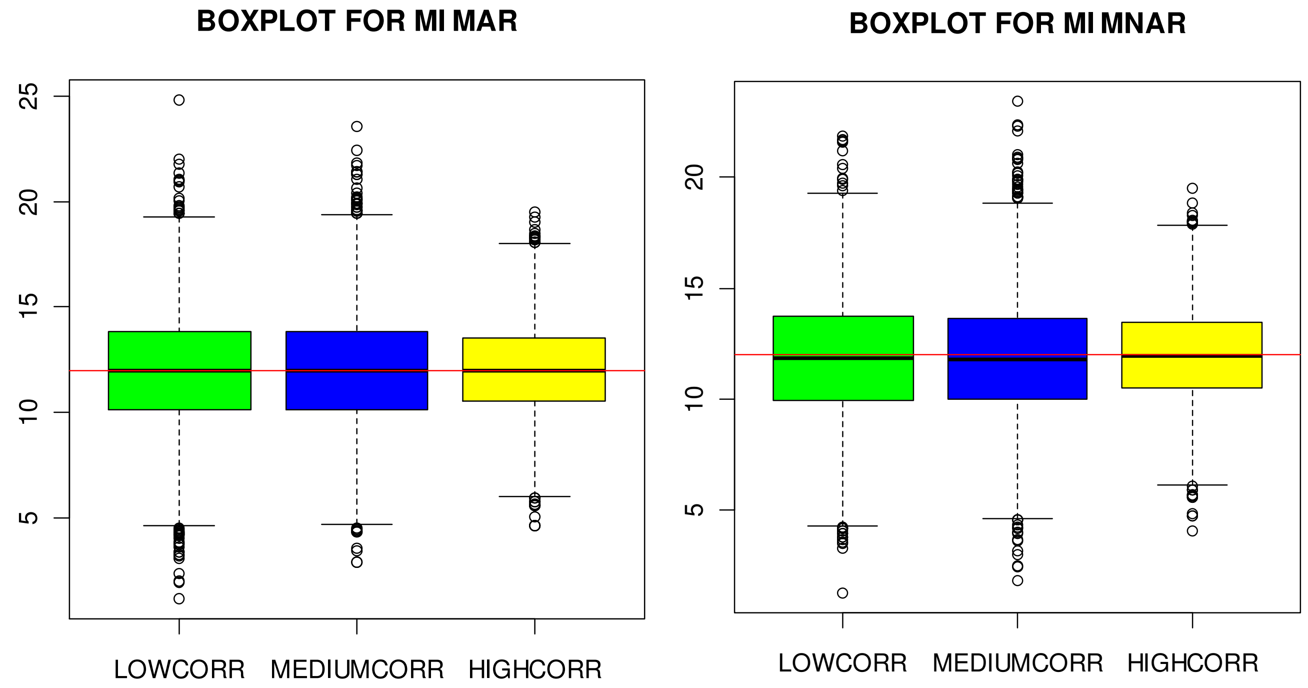

Figure: 5 Box Plots of the mean estimates for MAR and MNAR after analysis with Multiple Imputation.

5 DISCUSSIONS AND CONCLUSIONS

Missing data are ubiquitous in quantitative research studies, and often occurs in situations which are beyond the control of the researcher. Due to its pervasive nature, some methodologists have described missing data as "one of the most important statistical and design problems in research" (Azar, 2002). In longitudinal clinical studies for instance, a special case of missingness occurs when there is a loss in subjects' measurements at a specific point in time and thereafter, commonly termed *dropout* or attrition. Consequently, it becomes vital to understand the mechanism that led to the missing data to be able to perform analyses that will lead to proper inference. Thus, the decision to use a particular method can be based on a researcher's specific questions and/or preference for certain analytic techniques. Rubin (1976), and Little and Rubin (2002) showed that under precise assumptions, likelihood-based inference is valid when the missing data mechanism is ignored. Molenberghs and Verbeke (2005) formally showed that, once appropriate account is taken of what has been observed, there remains no dependence on unobserved data (at least in terms of probability models). Furthermore Kenward and Molenberghs (1998) showed that it is better to use the observed information matrix rather than the expected. They further advocated for the use of methods such as likelihood based and multiple imputation as they offer a general framework from which valid inferences could be developed under MAR. These methods do not only enjoy the much wider validity, but also have the advantage that they are easy to implement without any data manipulation given the availability of appropriate software.

On the other hand departures from MAR should be considered, in any analysis, and the possible consequence of such departures on the conclusions reached, since it is usually difficult to justify beforehand the assumption of MAR. This leads to more general MNAR models which explicitly incorporate the dropout mechanism. Unfortunately the inferences they produce are typically highly dependent on the un-testable and often implicit built-in assumptions regarding the distribution of the unobserved measurements given the observed ones Janssen *et al.* (2006). Considering the fact that likelihood based methods and multiple imputation are valid and easy to implement under the assumption of MAR, this thesis through a simulation study focused on the effectiveness of these methods under the MNAR scenario given the fact that they require modeling processes that are difficult and sensitive to misspecifications.

Results from the ignorable analysis under non ignorable conditions were quite stable as presented in table 2. Though simulations are necessarily limited, the results fall in line with those of Molenberghs, Kenward and Laseffre (1997) who applied multiple imputation in MNAR settings and found stable results, and Janssen et al (2006) who argued that even when the MAR assumption are violated, results are stable under ignorable analyses because these analysis constrain the behavior of the unseen data to be similar to that of the observed data. There was also no significant effect of correlation on the results as the box plots show no deviation in the distribution of the mean estimates from all analysis.

It is worth mentioning that though results are stable, once data has been missing, no modeling method whether MAR or MNAR can fully recover the lack of information that occurs due to incompleteness of the data. As a recommendation, it may be of interest to extend this study under varying percentage of missingness, and study the impact on inference.

In conclusion, the use of ignorable likelihood methods is attractive in analyzing incomplete data, and might be used for primary analysis purpose since according to Rubin, Stern and Vehovar (1995), the assumption of MAR is often to be regarded as a realistic one in well-conducted experiments, while MNAR models should be used in a sensitivity analysis to explore the impact of deviations from the MAR assumption on the conclusions.

REFERENCES:

Azar, B. (2002) Finding a solution for missing data. Monitor on Psychology, **33**, 70.

Baraldi, A. N. and Enders, C. K. (2010) An introduction to modern missing data analyses. *Journal of School Psychology*, **48**, 5-7.

Diggle, P. J. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Applied statistics*, **43, 49-93**

Janssen, I. (2005) *Flexible Model Strategies and Sensitivity Analysis Tools for Non-Monotone Incomplete Categorical Data.* Censtat, Universiteit Hasselt

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., Mallinckrodt, C. (2006). Analyzing incomplete binary longitudinal clinical trial data. Statistical Science **21**, 52–69.

Kenward, M.G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random. Statistical Science. **13**, 236-247

Little, R. and Rubin, D. (1987) *Statistical Analysis with Missing Data.* New York: John Wiley & Sons, Inc.

Little, R. and Rubin, D. (2002) *Statistical Analysis with Missing Data (2nd Edition).* New York: John Wiley & Sons, Inc.

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

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C. and Carroll, R. J. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics,* **5** 445–464.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G and Kenward, M.G. (2007) *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons, Inc.

Rubin, D. (1976) Inference and missing data. *Biometrika,* **63**, 581-92.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, MR0899519.

Rubin, D. B., Stern H. S., and Vehovar V. (2005) Handling 'don't know' survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, **90**, 822-828.

Rubin, L., Witkiewitz, K., Andre, J. and Reilly, S. (2007) *The Journal of Undergraduate Neuroscience Education* ,**5**, 71-77.

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

Sotto, C. (2009) *Topics in Analysis and Sensitivity Analysis for Incomplete Longitudinal Data*. Censtat, Universiteit Hasselt.

Wang, L. and Fan, X. (2004) Missing data in disguise and implication for survey data analysis. *Field Methods*, **3**, 332-351.

Verebeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data. New York: Springer.*

Wayman, J. C. (2003) Multiple Imputation for Missing Data: What is it and how can I use it? Paper presented at the 2003 Annual Meeting of the American Educational Research Association.

Appendix.

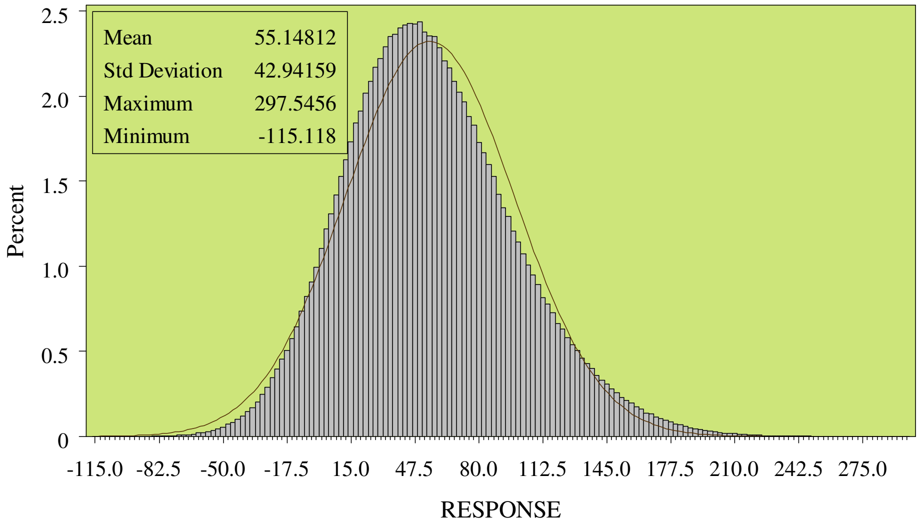

Fig 1 Histogram of response for the MNAR data

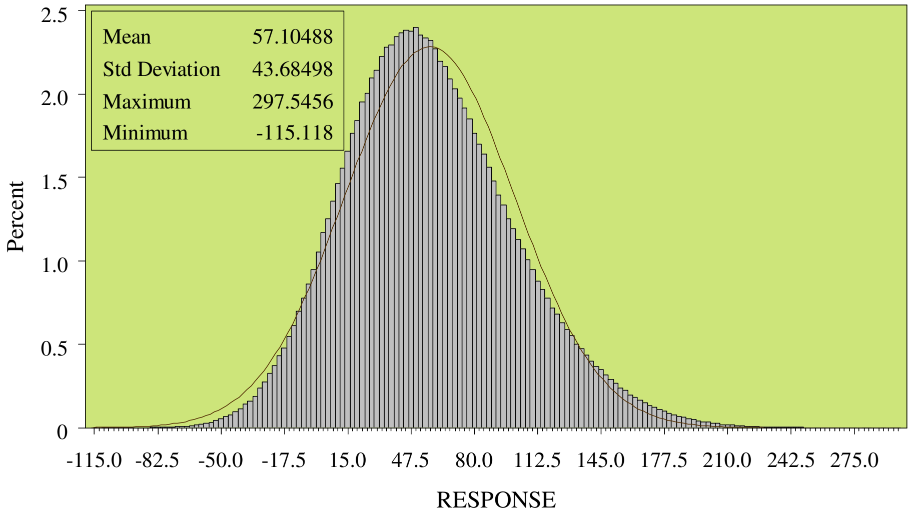

Fig 2 Histogram of response for the MAR data

Table :1 Mean estimates (standard errors) of Completers under different correlations

	$\boldsymbol{\rho = 0.1}$	$\boldsymbol{\rho = 0.5}$	$\boldsymbol{\rho = 0.8}$
Parameter	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
Intercept	9.99 (5.34)	9.93 (6.26)	9.96 (5.04)
Time	13.98 (2.10)*	14.04 (2.02)*	13.99 (1.58)*
Treatment	-1.86 (7.63)	-1.98 (8.86)	-1.95 (7.19)
Treatment*Time	11.96 (2.99)*	11.94 (2.86)*	12.00 (2.33)*

**Significant at the 5% level*

Table :2 Mean estimates (standard errors) of MAR under different correlations

	$\boldsymbol{\rho = 0.1}$	$\boldsymbol{\rho = 0.5}$	$\boldsymbol{\rho = 0.8}$
Parameter	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
Intercept	9.91 (5.60)	9.93 (5.72)	9.94 (5.70)
Time	14.03 (2.31)*	14.04 (2.20)*	14.01 (1.64)*
Treatment	-1.93 78)	-2.10 (8.09)	-2.02 (8.06)
Treatment*Time	12.03 (3.30)*	12.06 (3.20)*	12.06 (2.31)*

**Significant at the 5% level*

Table :3 Mean estimates (standard errors) of MNAR under different correlations

	$\boldsymbol{\rho = 0.1}$	$\boldsymbol{\rho = 0.5}$	$\boldsymbol{\rho = 0.8}$
Parameter	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
Intercept	10.19 (5.62)	10.20 (5.72)	10.17 (5.68)
Time	13.66 (2.36)*	13.72 (2.23)*	13.77 (1.67)*
Treatment	-1.77 (8.04)	-1.91 (8.16)	-1.89 (8.05)
Treatment*Time	11.88 (3.45)*	11.89 (3.23)*	11.95 (2.38)*

**Significant at the 5% level*

Table :4 Mean estimates (standard errors) of MAR MI under different correlations

	$\boldsymbol{\rho = 0.1}$	$\boldsymbol{\rho = 0.5}$	$\boldsymbol{\rho = 0.8}$
Parameter	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
Intercept	9.96 (5.28)	9.97 (5.55)	9.96 (5.79)
Time	14.00 (1.96)*	14.00 (1.90)*	13.95 (1.66)*
Treatment	-1.82 (9.22)	-2.03 (7.82)	-1.96 (8.20)
Treatment*Time	11.95 (2.78)*	11.99(2.76)*	12.01 (2.37)*

**Significant at the 5% level*

Table :5 Mean estimates (standard errors) of MNAR MI under different correlations

	$\boldsymbol{\rho = 0.1}$	$\boldsymbol{\rho = 0.5}$	$\boldsymbol{\rho = 0.8}$
Parameter	Estimate(s.e)	Estimate(s.e)	Estimate(s.e)
Intercept	10.23 (5.20)	10.19(5.50)	10.18 (5.81)
Time	13.65 (1.91)*	13.73(1.84)*	13.77 (1.69)*
Treatment	-1.69 (7.41)	-1.86(7.81)	-1.91 (8.22)
Treatment*Time	11.83 (2.76)*	11.86(2.70)*	11.96 (2.44)*

**Significant at the 5% level*

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Grip strength of severely malnourished children during nutritional rehabilitation in the Jimma hospital of Ethiopia

Richting: **master of Statistics-Epidemiology & Public Health Methodology**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Fonguh, Sylvanus

Datum: **12/09/2011**