Hasselt University
Censtat

Longitudinal Data Analysis

Homework Assignment 3: Incomplete Longitudinal Data

by

Banerjee Soutrik
Kishtammagari Manjularani
Rosius Wannes

Diepenbeek, April 2nd, 2006

Contents

1 Introduction

In the previous two homeworks, we presented two different reports on the longitudinal analyses of the renal dataset. In the first analysis, marginal, two-stage and random-effects models were fitted where the response variable (haematocrit value post-transplant) was treated as continuous. In the second analysis, generalised estimating equations (GEE) were used to fit a marginal model and generalised linear mixed-model (GLMM) framework was employed to fit a hierarchical model to the same dataset, but on this occasion the outcome variable was dichotomised (binary). In both of these projects, we didn't *explicitly* consider missingness of the dataset to be an issue. In this project however, our main focus will be the missing data, in which we try to gain an insight into the possible factors related to the missingness.

We first explore the conditions under which the missingness mechanism can be analysed. It may be 'simple' to consider the assumption of missingness completely at random (MCAR). Under this assumption, the missingness is independent of the observed and unobserved data. However, it is unlikely in general that such a strong assumption hold true in practice. Therefore, the next possible (less strong) assumption is to consider missingness at random (MAR), in which the missingness is independent of unobserved data, conditional on observed data. Under the likelihood or Bayesian inferences, both of these conditions are sufficient for *ignorability*. The latter term means that the missingness process can be ignored. This is not possible where non-random missingness is thought to play a role. In the frequentist approach, however, MCAR assumption is usually a sufficient condition for ignorability. When both of these assumptions, MCAR & MAR, cannot be met, the missingness process is assumed to be non-ignorable, which is equivalent to saying missingness not at random (MNAR). In order that MAR is valid, one must assume that the 'missingness process' and the 'measurement process' parameters are distinct, *i.e.*, *separability* condition be satisfied, so that each can be maximised separately from their joint probability distribution and obtain the respective marginal probability distributions.

Missing data process can be explored principally under three frameworks : *selection models*, *pattern-mixture models* and *shared-parameter models*. The details can be seen in [1-3]. In this report only the first two frameworks are used for gaining insight into missingness mechanisms.

1.1 Methods

The methods available for making an inference on the data can be broadly classified into two mechanisms : simple and advanced methods.

Simple methods: Most of these methods need MCAR assumption. The first is Complete Case (CC) analysis. In this, all patients with any missing components are eliminated. Only the

patients with full follow-up are included for inference. This is unrealistic under MAR mechanism, when a bias may occur. Next, there are methods of single imputation of the missing values, e.g., unconditional mean imputation, conditional (Buck's) mean imputation (which is valid under certain MAR mechanisms) and Last Observation Carried Forward (LOCF), hot-deck imputation, regression imputation, etc. MCAR may not even be sufficient for LOCF, on the other hand CC analysis is valid only under MCAR. In addition, there is available case analysis (where inference is drawn on the available measurements and ignoring missingness). In case of mean imputations, another disadvantage is that the sample size is overestimated and the standard errors are underestimated. In this context, 'NORM' is freely available to be downloaded from [4], developed by JL Schafer. This software helps to test the normality assumptions, EM optimisation, data enhancement and single imputation.

Advanced methods: 'Advanced' methods doesn't necessarily imply more complicated or difficult to execute methods than the previously described simple methods. The word advanced implies that these methods are valid under MAR assumption in general and hence pose a less stringent condition like MCAR. The methods include Weighted GEE (since GEE, a frequentist approach, is valid only under MCAR mechanism), Direct Likelihood, Multiple Imputation (MI) and Expectation-Maximisation (EM) algorithm. Weighted GEE (WGEE) and direct likelihood methods are less complicated to execute, whereas EM and especially MI can necessitate a lot of data manipulation.

We used SAS© version 9.1 for the analysis. GEE, WGEE with proc GENMOD; Linear mixed-model (LMM) with proc MIXED; GLMM with proc NLMIXED, proc GLIMMIX; MI with proc MI and proc MIANALYZE.

We begin our analysis with some simple data exploration techniques with graphics that can always point towards a plausible cause of missingness.

1.2 Overview of relevant previous results

In previous homeworks, we built models for the evolution of the (continuous) haematocrit values as well as for the dichotomised (binary) effect of the renal transplant.

For the continuous study, we fitted an LMM. For the binary study, we fitted the data with GEE and GLMM. The results of these 3 studies can be found in Table 1. The results of the GLMM are based on a non adaptive Gaussian quadrature with number of quadrature points equal to 3 and random effects for the intercept and for the time effect (respectively, d_{11} and d_{22}). In our previous homework, we studied more advanced models, but, due to time and computational limitations, in this paper, we will only use this model in MI procedures.

		Continuous		Binary			
Effect		LMM		GEE(AR(1))		GLMM	
		est	s.e.	est	s.e.	est	s.e.
Int		-	-	1.911	0.283	3.496	0.497
age		0.060	0.008	-0.016	0.005	-0.042	0.008
$male$	0	31.990	0.440	-0.325	0.104	0.000	0.000
	1	33.900	0.440	0.000	0.000	0.972	0.188
$cardio$	0	-	-	-0.301	0.141	-0.838	0.192
	1	-	-	0.000	0.000	0.000	0.000
$reject$	0	-	-	-	-	-0.630	0.194
	1	-	-	-	-	0.000	0.000
$year$		-	-	-0.68	0.035	-	-
$year * male$	0	1.870	0.074	-	-	-	-
	1	1.450	0.082	-	-	-	-
$age * year$		-	-	0.0019	0.0008	-	-
$cardio * year$	0	-	-	-	-	0.308	0.066
	1	-	-	-	-	0.000	0.000
$year_2 * male$	0	-0.190	0.008	-	-	-	-
	1	-0.140	0.009	-	-	-	-
$cardio * year_2$	0	-	-	-	-	-0.033	0.0064
	1	-	-	-	-	0.000	0.000
d_{11}		-	-	-	-	4.008	0.293
d_{22}		-	-	-	-	0.556	0.077

Table 1: Results of previous studies.

2 Exploratory Data Analysis

2.1 Description of missingness

In Table 2, we see the different patterns of missingness of the response variable in the given dataset. As we can see, there are 20 different patterns of response missingness in the dataset. These patterns, will divide the subjects into 20 different groups, each group with their own missing pattern. The 20 groups can be divided into three categories.

1. Completers: There is one group of completers. This group consists of 339 subjects, which is equal to 29.2% of all subjects.

Group	HC_0	HC_{06}	HC_1	HC_2	HC_3	HC_4	HC_5	HC_6	HC_7	HC_8	HC_9	HC_{10}	#
completers													
1	X	X	X	X	X	X	X	X	X	X	X	X	339
dropouts													
2	X	X	X	X	X	X	X	X	X	X	X	-	68
3	X	X	X	X	X	X	X	X	X	X	-	-	73
4	X	X	X	X	X	X	X	X	X	-	-	-	76
5	X	X	X	X	X	X	X	X	-	-	-	-	86
6	X	X	X	X	X	X	X	-	-	-	-	-	89
7	X	X	X	X	X	X	-	-	-	-	-	-	103
8	X	X	X	X	X	-	-	-	-	-	-	-	109
9	X	X	X	X	-	-	-	-	-	-	-	-	117
10	X	X	X	-	-	-	-	-	-	-	-	-	86
non-monotone missingness													
11	X	X	X	X	X	X	X	X	X	X	-	X	4
12	X	X	X	X	X	X	X	X	-	-	X	X	2
13	X	X	X	X	X	X	X	-	X	X	X	X	1
14	-	X	X	X	X	X	X	X	X	X	X	X	1
15	X	X	X	X	X	X	-	-	X	X	-	-	1
16	X	X	X	X	-	-	X	X	X	X	-	-	1
17	X	X	-	X	X	-	X	X	X	-	-	X	1
18	X	X	X	X	X	-	X	X	-	-	-	-	1
19	X	X	-	X	X	X	X	-	-	-	-	-	1
20	X	X	X	X	-	X	-	-	-	-	-	-	1

Table 2: Missing data patterns of the continuous dataset.

2. Dropouts: This consists of subjects having a missingness pattern of the form

$$(\underbrace{X, ..., X}_{\text{i times}}, -, ..., -).$$

For this example we say that the subject drops out at time $i + 1$.

The dataset consists of 807 dropouts, distributed throughout nine groups. This coincides with 69.5% of all subjects, and 98.3% of the subjects with missing response variables.
How the dropouts are divided over the nine groups, compared to the full dataset, through the 10 study years, is summarised in Figure 1. Here we can see that the dropout rate is not very high, and can be seen almost constant through the years (between 5.9% and 10.0% per year).

3. Non-monotone: There are 10 groups of non-monotone missingness subjects. Only 14 subjects are distributed among these groups. Which coincides with only 1.20% of all study subjects.

In addition to these missing response variables, we also have one missing value in the explanatory

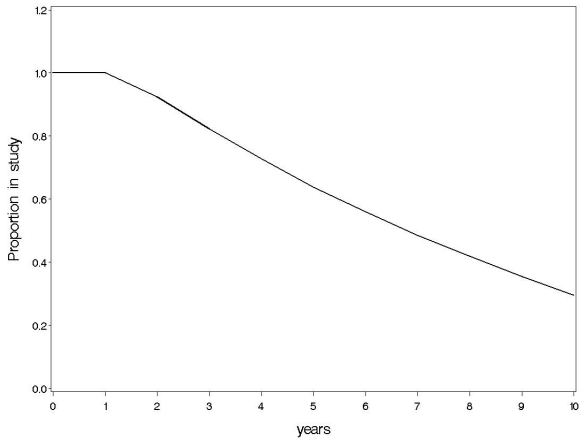

Figure 1: Representation of dropouts

variables. One subject misses the covariate *age*. This subject belongs to group six.

2.2 Description of the groups

In this section, we will study a possible link between the pattern of the missingness and the response variable. For this we look at the individual profiles of the 20 groups separately (with a 95% CI), given in Figure 7, Appendix A.

The individual profiles for the first 10 groups look very dense, as we expected, since the first 10 groups almost covers 99% of the study subjects. The other 10 profiles are not dense at all, because of the small group-sizes. However, there is no trend to see in the last 10 groups.

For the first 10 groups (completers and dropouts) we have to look for something else to study the profiles. For this, we give the mean profiles of the first 10 groups, given in Figure 2. In this plot, we can see a possible link: the earlier groups (the subjects with long stay period in the study) seems to have lower baseline values then the other dropout groups.

2.3 Reasons for dropouts

In this section, we will look if we can obtain an insight on which covariates have an influence on missingness. We will only look at the 1146 subjects being completers or dropouts. We can then

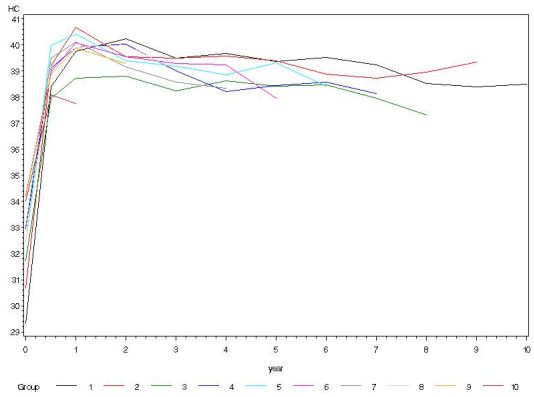

Figure 2: Mean profiles, for complete and monotone groups given in table 2.

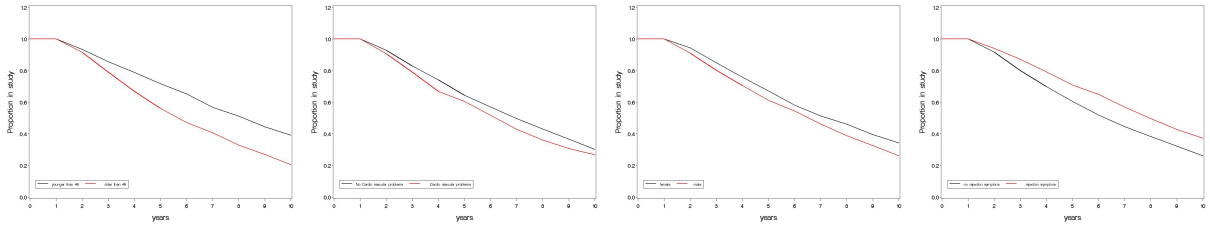

Figure 3: Influence of covariates on the observed dropouts (from left to right: Age-cardio-male-reject).

give separated plots for every category to see which covariates influence the dropout. The plots are given in Figure 3.

In the plots, we can see that the dropout rate for older people is higher than for younger. Also for the covariate *reject* there seems to be a larger rate of dropout than that for the non-rejections. This is maybe a thing we should take into account in our study.

Another way of seeing these conclusions is the following: If we insert the covariate D_{ij} for which

$$D_{ij} = \begin{cases} 1 & \text{if} \quad \text{dropout of subject } i \text{ at time } j \\ 0 & \text{if} \quad \text{no dropout of subject } i \text{ at time } j \end{cases}$$

Then we can try to model the probability of D_{ij} as a function of these covariate. We than obtain the following model

$$\text{Logit}(Pr(D_{ij})) = -4.55 + 0.024 * age - 0.209 * (1 - male) + 0.26 * (1 - reject) + 0.26 * j,$$

Here we can see what we also suspected from looking at Figure 3. We obtain significant results for $year$, $reject$ and age. Also the covariate $male$ is significant, but at a lower level than the others ($p = .01$).

Now, we can extend this method, and also look at the interactions between the covariates. If we do that we obtain the final model

$$\text{Logit}(Pr(D_{ij})) = -4.46 + 0.021 * age + -0.0044 * age * (1 - male) + 0.006 * age * (1 - reject) + 0.26 * j,$$

Now we can see that the covariates $male$ and $reject$ are only significant with an interaction with age, Again $male$ at a lower significance level than $reject$. These interactions can off course not be seen in the plots.

3 Multiple Imputation

The key idea of Multiple Imputation (MI) is to replace the missing value with M different plausible values, where M is the number of imputations. We then have M complete datasets, which can be analysed. The results from these M analyses are then captured in a single analysis.

An advantage of MI is that it can be very efficient for a small number of imputations M. In this paper, we will use 5 imputations. the grade of efficiency is then given by

$$\left(1 + \frac{\gamma}{5}\right)^{-1},$$

where γ is the fraction of missing information for the quantity being estimated. In our case we have $\gamma = 0.313$, which leads to an efficiency of 94.1%.

We will first do an MI analysis on the continuous data, and afterwards we will go further on a dichotomised dataset.

3.1 MI for the continuous LMM study

For calculating the imputations, we will use the Markov chain Monte Carlo (MCMC) method. This method can handle arbitrary missing data. (The regression method can only work with dropouts, so we should have to delete 14 subjects.) With this MCMC method, we will try to estimate a missing value, depending on its previous value(s).

Now we are ready for doing the analysis. We choose to do the linear mixed model, discussed in our homework 1.

$$HC \quad = \quad male \quad age \quad male * year \quad male * year^2. \tag{1}$$

	Estimations		test if Est= μ_0		
Parameter	Est	s.e.	μ_0	t	p
Int	32.606	0.489	31.99	1.26	0.2247
$male$	2.112	0.272	1.91	0.74	0.4647
age	0.052	0.009	0.06	-0.87	0.3972
$year$	1.148	0.073	1.87	-9.92	<.0001
$male * year$	0.131	0.086	-0.42	6.41	<.0001
$year^2$	-0.103	0.008	-0.19	10.74	<.0001
$male * year^2$	-0.019	0.009	-0.05	-7.68	<.0001

Table 3: Test for LMM model.

Where *male* is a binary class variable.

After the analysis of the dataset is done for each of the 5 imputations, we want to combine these analyses to a single analysis, using proc MIANALYZE. We want to *compare* this analysis with the LMM study of previous homeworks. So we will test if the results of MIANALYZE are equal to the results in Table 1. We will test this with the *mu0* command in proc MIANALYZE

However, in PROC MIANALYZE, we cannot use a class statement together with a test function. For this we will change our model (1) to the following model

$$HC = Int \quad male \quad age \quad year \quad male * year \quad year^2 \quad male * year^2. \tag{2}$$

Where *male* is not seen as a class variable. The results of these tests can be seen in Table 3.

In this table, we can see a comparison between the study in Homework 1, and the same study with Multiple Imputations. In the table, the estimates of the model are shown, together with their standard errors. Then these are tested whether there is a significant difference between the new estimates and the estimates found in the previous homework. We can see that there are no significant differences for the Int, *male* and *age* variables of model (2) (i.e. the *male* and *age* variable in model (1)) . All the other variables (*i.e.* all the variables with a year effect) have significant differences.

This is a logical result, because, what we see in the models, the number of patients at baseline is the same in the two studies. This is because there is only one missing value for the baseline rate. For the time effect, there are much more missing values due to attrition. Because of this, in the time effect, the 2 studies show a significant difference.

3.2 MI for the discrete GEE study

As in Homework 2, we dichotomise our dataset in a binary dataset, with the following dichotomisation rule:

$$HCd_i = \begin{cases} 1 & \text{if} \quad HC_i - HC_0 > 3.5 \\ 0 & \text{if} \quad HC_i - HC_0 \leq 3.5 \end{cases}$$

The results of the previous studies are summarised in Table 1.

To keep a binary dataset, we first do Multiple Imputation, of the continuous data, and then do the dichotomisation. After that, we study it in the same way as we did before.

In the last homework, we derived a final GEE-model:

$$HCd \quad = \quad male \quad cardio \quad age \quad year \quad age * year. \tag{3}$$

With *male* and *cardio* binary class variables. But we again want to compare this to the MI-study, so we should not have class variables in our model, therefore we change (3) to

$$HCd \quad = \quad Int \quad female \quad ncardio \quad age \quad year \quad age * year. \tag{4}$$

where we defined the variables $female := 1 - male$ and $ncardio := 1 - cardio$. Results from this study and comparison with the model in homework 2, are summarised in Table 4.

	Estimations		test if Est= μ_0		
Parameter	Est	s.e.	μ_0	t	p
Int	1.802	0.159	1.911	-0.69	0.498
$female$	-0.234	0.044	-0.325	2.08	0.054
$ncardio$	-0.218	0.066	-0.301	1.26	0.235
age	-0.013	0.003	-0.016	0.94	0.358
$year$	-0.070	0.029	-0.680	20.87	<.0001
$age * year$	0.0006	0.0007	0.002	-2.02	0.083

Table 4: Test for GEE model.

3.3 MI for the discrete GLMM study

Before we start with this analysis, we should mention that the GLMM study in homework 2 is done on a much smaller dataset then we do here. Because PROC NLMIXED can not work with

	Estimations		test if Est= μ_0		
Parameter	Est	s.e.	μ_0	t	p
$Intercept$	4.218	0.751	3.496	0.96	0.376
$male(=1)$	0.542	0.256	0.972	-1.68	0.152
age	-0.047	0.013	-0.042	-0.37	0.723
$cardio(=0)$	-0.841	0.264	-0.838	-0.01	0.992
$reject(=0)$	-0.837	0.191	-0.630	-1.09	0.297
$cardio(=0) * year$	0.131	0.067	0.308	-2.65	0.039
$cardio(=0) * year^2$	-0.017	0.006	-0.033	2.82	0.025
d_{11}	5.964	0.411	4.008	4.76	0.0001
d_{22}	0.061	0.006	0.556	-82.35	<.0001

Table 5: Test for GLMM model.

missing values, and we did not had tools to cover for these missing values, we had to delete all the subject with missingness to study the GLMM.

Now, however, we don't have missing values, because we did a Multiple Imputation procedure on the dataset. We again will do the same thing as we did before. This means testing the estimates we obtained in our previous homework (see Table 1), to the values we obtained by analysing the dataset with Multiple Imputations.

In Table 5, we again see a a non-significant difference between for baseline values' parameters. In addition, a significant difference on the time and time^2 effects. These are also the same things we saw in previous studies.

To remark in this table, is the significant difference of random effects, for the intercept, as well as for the time effects. This difference can be explained by what was mentioned above. In the previous homework, we only could work with the completers (=339 subjects) of the study. While now, we can work with the entire dataset. This means that there will be a change in both of the random effects, also for the intercept.

4 CC and LOCF analyses

In this section, a Complete Case analysis and Last Observation Carried Forward analysis are done for comparative purposes only. Under MAR assumption, they may produce biased inference. Here we used the same full model that was used in the first homework in order to build a

		Multivariate		CC		LOCF	
Parameter		Est	s.e.	Est	s.e.	Est	s.e.
				ML			
male	0	31.374	0.423	30.270	0.732	31.696	0.571
	1	33.300	0.414	32.250	0.701	33.753	0.564
cardio	0	-	-	-	-	-0.097	0.342
	1	-	-	-	-	0.000	0.000
age		0.067	0.008	0.076	0.015	0.068	0.008
$year * male$	0	1.053	0.092	1.306	0.105	-	-
	1	1.276	0.083	1.197	0.126	-	-
$year * cardio$	0	-	-	-	-	0.633	0.051
	1	-	-	-	-	0.460	0.110
$year^2 * male$	0	-0.094	0.009	-0.110	0.010	-	-
	1	-0.119	0.008	-0.096	0.011	-	-
$year^2 * cardio$	0	-	-	-	-	-0.048	0.004
	1	-	-	-	-	-0.033	0.008
				REML			
male	0	31.374	0.424	30.270	0.735	31.696	0.572
	1	33.300	0.415	33.753	0.565	33.753	0.566
cardio	0	-	-	-	-	-0.100	0.343
	1	-	-	-	-	0.000	0.000
age		0.067	0.008	0.076	0.015	0.068	0.009
$year * male$	0	1.054	0.092	1.306	0.106	-	-
	1	1.277	0.083	1.197	0.127	-	-
$year * cardio$	0	-	-	-	-	0.633	0.051
	1	-	-	-	-	0.460	0.109
$year^2 * male$	0	-0.094	0.009	-0.110	0.010	-	-
	1	-0.1194	0.008	-0.097	0.019	-	-
$year^2 * cardio$	0	-	-	-	-	-0.048	0.003
	1	-	-	-	-	-0.033	0.008

Table 6: Results of CC and LOCF study compared with Multivariate model (Continuous data).

parsimonious reduced model. Loglikelihood method was used to reduce CC/LOCF/Multivariate models.

The results are shown in Table 6. The graphical comparison of CC, LOCF, Multivariate and observed values is shown in Figure 4.

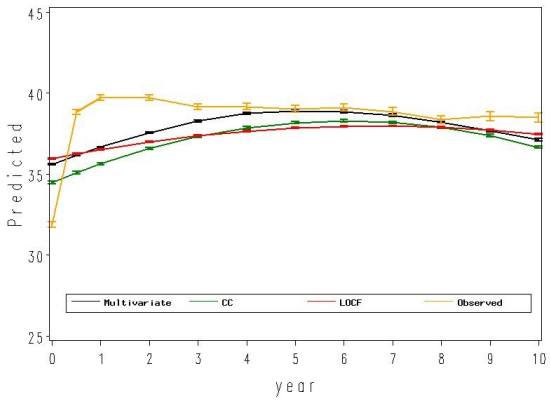

Figure 4: Graphical comparison between CC, LOCF, Multivariate and observed.

5 Weighted Generalised Estimating Equations

When the given data is binary, under frequentist approach standard GEE needs MCAR assumption. Therefore, weights are to be calculated in order to fit a WGEE model. In order to calculate the weights (which are inverse probabilities that a subject drops out at a particular time or occasion), a dropout probability model (binary) needs to be first fit using some data manipulation and two macros : 'dropout' and 'dropwgt' available at the CenStat faculty.

In the dataset, only 14 patients had non-monotone missingness, which is a problem for WGEE procedure with SAS. Therefore, we decided to delete those patients and only use the rest of the patients for WGEE analysis (a better possible way would be to do MI for the non-monotone missingness and then do WGEE on the data containing monotone missingness). Here we used the same full model that was used in the second homework in order to build a parsimonious reduced model. Type 3 Wald statistics were used to reduce GEE/WGEE models. GEE/WGEE were fitted with dropout patients only.

The results of the parameter estimates with empirical standard errors are provided in Table 7. The graphical comparison of GEE, WGEE and observed probabilities are shown in Figure 5.

6 Sensitivity analysis

In this section we will only look at the completers and the dropouts. So we delete the 14 subjects not belonging to these categories

Parameter		Est	s.e.
	WGEE		
$male$	0	0.606	0.120
	1	0.794	0.103
$year * male$	0	-0.014	0.056
	1	0.102	0.038
$year^2 * male$	0	-0.001	0.006
	1	-0.013	0.004
	GEE		
$male$	0	1.691	0.279
	1	2.010	0.278
$cardio$	0	-0.329	0.142
	1	0.000	0.000
age		-0.016	0.005
$year$		-0.086	0.029
$age * year$		0.002	0.001

Table 7: Parameter for WGEE and GEE.

6.1 Pattern Mixture Models

In this section we will use model (2) for the Pattern Mixture model. We would again like to note that this model is equivalent to model (1) We will look at the parameters for this model for each of the different 10 groups. The results of this study are summarised in Table 8.

Now we will look at the average estimated values, and the average residuals for each group. They are plotted in Figure 6. In this, we can see that the mean structures, as well as the residuals are not very different for each group.

6.2 Models with group as a covariate

In this section, we will see group as a class variable. We will start with model 1, but enlarge this with the interactions of group with every other variable. We do this for the baseline as well as for the intercept effect. The result of the T3 tests of the interactions are given in Table 9.

From this table, we can see that there is a significant intercept and time effects from the group. In addition, there may be a significant interaction between the *cardio* and the group. All other effects are highly non-significant.

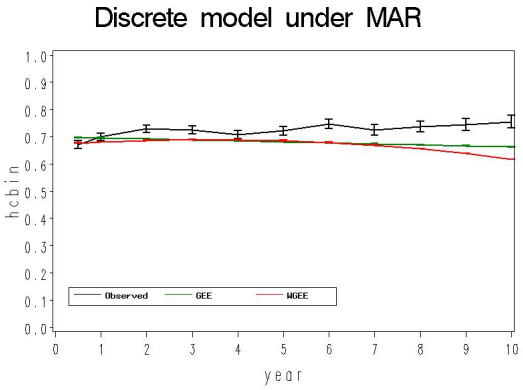

Figure 5: Graphical comparison between WGEE, GEE and observed.

7 Discussion

In order to infer about the parameter estimates under the assumption of MAR, it is assumed that the measurement and the missing processes are independent (separability or parameter distinctness condition). Although MAR assumption may not hold in general, nevertheless ignorable analyses may provide reasonably stable results, even when the assumption of MAR is violated [2].

In this report, we did not explicitly use available case (AC) analysis for inference on parameter estimates, since this was done in the previous homeworks (by frequentist approach with proc GENMOD under MCAR assumption, or by direct likelihood or Bayesian approach with proc MIXED under MAR assumption, for example). In addition, an assumption called covariate-dependent missingness can also be described, which is a stronger assumption than MAR, but a weaker assumption than MCAR. In this case, the missingness is independent of the observed and unobserved data, but given the covariates. However, we restrict our focus to only the three assumptions (MCAR, MAR, MNAR) stated in the introduction of this report.

We proceeded with simple and advanced methods to see the impact of different methods on the parameter estimates of the fitted models and hence compare them with our previous results. This would, in addition, give an insight to the mechanism of missingness process. The first important aspect to test is to verify whether this missingness process is an MAR or MCAR. For this, we fitted a dropout probability model in terms of the previous outcome and the covariates $male$, age, $reject$, $year$ and $year^2$, starting from a full model. All these parameter estimates were significant implying that there is evidence against MCAR in favour of MAR [2]. The fitted

Gr	Int	$male$	age	$male * year$	$year$	$male * year^2$	$year^2$
1	30.986	2.316	0.067	1.845	0.126	-0.159	-0.020
2	30.562	1.957	0.086	0.633	1.571	-0.056	-0.154
3	30.658	2.144	0.072	1.449	0.712	-0.167	-0.068
4	33.673	2.464	0.021	1.065	1.010	-0.143	-0.127
5	30.423	0.995	0.110	2.029	0.757	-0.276	-0.147
6	31.755	0.732	0.082	1.889	2.309	-0.364	-0.381
7	33.986	0.975	0.013	3.560	1.548	-0.767	-0.354
8	32.540	0.132	0.020	6.708	2.709	-1.697	-0.906
9	31.009	1.839	0.046	9.234	0.350	-3.551	0.035
10	29.360	1.959	0.068	11.811	1.107	-7.474	-2.004

Table 8: Pattern Mixture Models.

	Intercept				Year effect			
group	Num	Den			Num	Den		
Interacted with	DF	DF	F	p	DF	DF	F	p
$male$	9	1096	0.491	0.881	9	8225	1.210	0.283
$reject$	10	1096	1.362	0.193	10	8225	1.045	0.401
$cardio$	10	1096	2.445	0.007	10	8225	1.613	0.096
age	9	1096	1.126	0.340	10	8225	0.638	0.782
Int	9	1096	2.165	0.022	9	8225	2.976	0.001

Table 9: Interactions with group as covariate.

equation is given below :

$$\text{logit}(Pr(D)) = -5.43 + 0.67 * t_{j-1} + 2.78 * (reject - male) + 0.63 * year - 0.04 * year^2,$$

where t_{j-1} is the time or occasion of previous measurement, D = dropout ; (All the parameter estimates are significant at .05 level).

In the Exploratory data analysis, we found that the covariates age, $reject*age$ and $male*reject$ are the most significant covariates to determine the chances of dropout. Off course, this together with the time effect.

With the Multiple Imputation chapter, we saw that the result, obtained in previous homeworks, were not really valid. We found a significant difference, especially in the time effect on the haematocrit value.

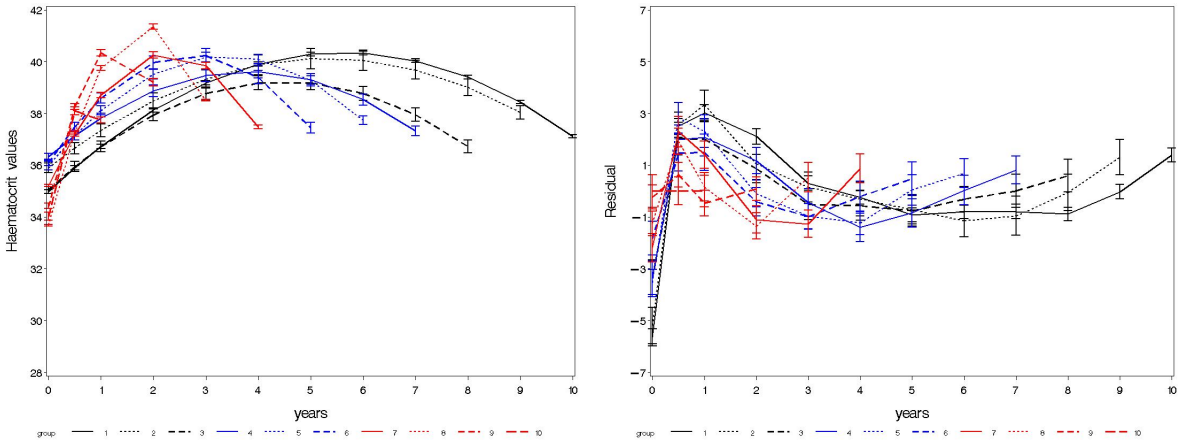

Figure 6: Pattern Mixture model (on the left the mean profiles right, and on the right, the mean residuals.

The LOCF, CC and multivariate models' parameter estimates were different. The graphical plot (Figure 4) shows that both CC, LOCF and multivariate models' predicted values follow different courses, although the graphs are close to each other. This might point towards an invalid MCAR mechanism as well, since CC and LOCF are valid under MCAR mechanism. *(A point to note that we fitted a multivariate model, and not a random-effects model due to lack of time).*

WGEE estimates correct for the weights that are the inverse probabilities that the subject drops out at a particular occasion. in Figure 5, we observe that there is a difference in the course of WGEE and GEE estimates particularly from the 7th year of follow-up. This again indicates that a possible MAR mechanism might be playing a significant role in the missingness process, although different concurrent missingness mechanisms cannot be overruled.

A Sensitivity analysis for Pattern-Mixtures models is done to observe the differences in models between the groups. We only did this study on the continuous dataset, to fit a LMM. We tried to do this in two ways. First, we fitted the model for every group separately and than, we fitted one model, with group as a class variable inserted in the model.

In conclusion, we find a significant evidence of MAR mechanism in this study. In addition, an MNAR mechanism cannot be completely ruled out. In spite of this, we can still say that this MAR mechanism is not too strong as to alter our results to a great extent.

References

[1] Molenberghs, G and Verbeke, G. (2005-2006) Longitudinal Data Analysis, courses notes, Universiteit Hasselt.

[2] Molenberghs, G and Verbeke, G. (2000) *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

[3] Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

[4] Website to download NORM.exe:
http://www.stat.psu.edu/~jls/misoftwa.html

A Individual profiles per pattern group

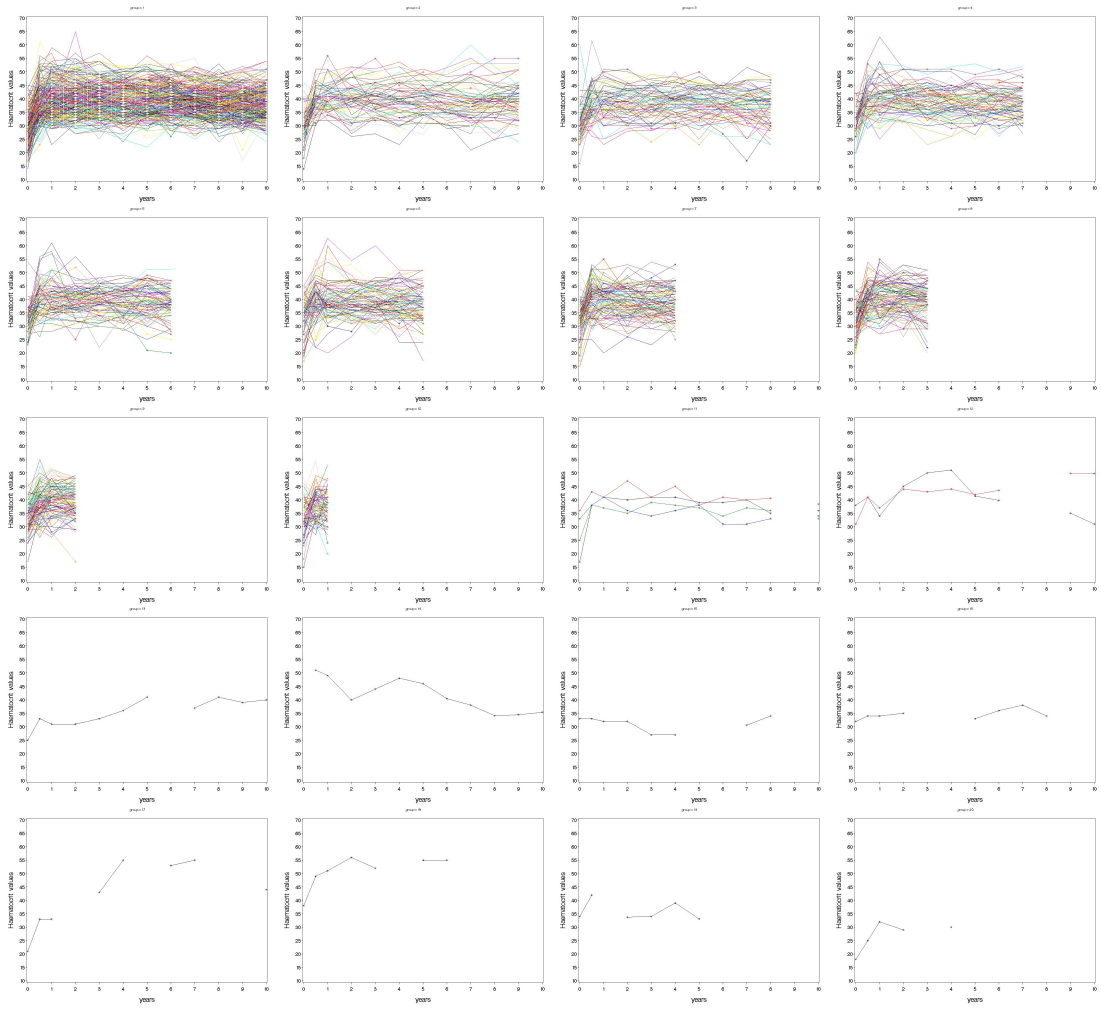

Figure 7: Individual profiles, per group given in Table 2