Hasselt University
Censtat

Longitudinal Data Analysis

HW1: Renal Transplant study

by

Banerjee Soutrik
Kishtammogari Manjula
Rosius Wannes

Diepenbeek, February 7th, 2006

Contents

List of Figures

Introduction

The current dataset contains information on 1160 post-renal transplant patients, who were followed-up for a maximum period of 10 years. The patients were assigned an ID number with following predictor variables recorded (in the dataset provided) : age at transplantation (continuous), gender (male = 1, female = 0), cardio-vascular problems before the transplant (binary - yes / no), rejection symptoms experienced in the first 3 months after the transplant (binary - yes / no). The response variable measured was the blood haematocrit value at baseline (hc0), and then at 6 months (hc06), 1 year (hc1), 2 years (hc2) and in the same manner, until 10 years (hc10), which is considered as the endpoint. Therefore in total, there were a maximum of 12 measurements for a patient. However, as typically one would find in a longitudinal dataset, there were incomplete data for several patients due to drop-outs (unbalanced data) or absents in a particular follow-up schedule. These drop-outs might have been related, at least in part, to the outcome of the patient condition such as rejection of the graft and return to dialysis or may be due to death of the patient. However, in the given dataset, the causes were not mentioned in order that one can only postulate about the causes of missing data, whether at random or not.

The normal kidneys produce erythropoietin, which is a substance necessary to form the red blood corpuscles (RBC). The lack of erythropoietin production in patients with chronic renal failure (CRF) causes a gradual diminution of the blood RBC count and consequently, the haematocrit value is lowered. In addition, there is haemodilution due to fluid retention in CRF patients, further lowering the blood haematocrit. The primary objective of this study was to follow the changes in haematocrit value after a renal transplant in repeated measures setting for each patient. Particularly of interest was to see the manner of restitution of the haematocrit level in the post-graft patients depending on the time of measurement as well as other cofactors, such as gender, cardio-vascular problems, age and rejection symptoms.

In the given dataset, there is a lack of information on the treatment received by the post-graft patients. This information is crucial, since the restitution of anti-rejection therapy with an immuno-suppressive regimen can affect the haematocrit level (or the response variable) itself either by bone marrow suppression or by renal function deterioration due to the lack of effect of the anti-rejection therapy or by the direct cytotoxic effect of the drug on the kidneys. Therefore in the present setting, it is not possible to correct for the treatment effect on the patients, which might have been related to the

outcome variable.
In this report, we first explore graphically the mean and variance structures of haematocrit values with respect to time as a continuous variable as well as individual profiles by taking into account the known cofactors as mentioned above. Following, we use some summary statistics to analyse the data and discuss the advantages and disadvantages of such techniques in brief. In the next step of longitudinal data analysis, we first fit a multivariate model and then try to reduce it to a parsimonious model of lesser parameters to define the mean structure. Then we use a two-stage model by combining a first-stage subject-specific linear regression model explaining the within subject variability part and a second-stage multivariate regression model explaining the between subject variability part. Finally, we try to fit a random-effects model to conclude our analysis. We discuss in brief our results and compare them in the conclusion.
The main purpose of this report is to present the results in a palatable way to a medical or a non-technical person, and at the same time bearing in mind that no important information is lost to describe the results to a biostatistician. The statistical softwares used to complete this report were SAS, Statable.

1 Exploratory data analysis

1.1 Description of patient characteristics

The table below shows the patient missing drop-out statistics.

Time	No. of Patients	No. of Missing patients	Percentage of missing patients
Baseline	1159	1	0,862
6 months	1160	0	0
1 year	1159	1	0,862
2 years	1073	87	7,5
3 years	955	205	17,67
4 years	846	314	27,06
5 years	742	418	36,03
6 years	652	508	43,79
7 years	565	595	51,29
8 years	488	672	57,93
9 years	411	749	64,56
10 years	348	812	70

The table below shows patient characteristics at baseline by cofactors. The mean (SD) of age was 46.43 (13.31) at presentation.

Patients characteristics at baseline	No. of patients
Male	494
Female	666
Cardio-vascular problems	207
No Cardio-vascular problems	953
Rejection symptoms	367
No Rejection symptoms	973

1.2 Mean structure

The evolution of mean haematocrit for all patients against months is shown in the Figure 1. One can see a steep rise of average haematocrit level (95%

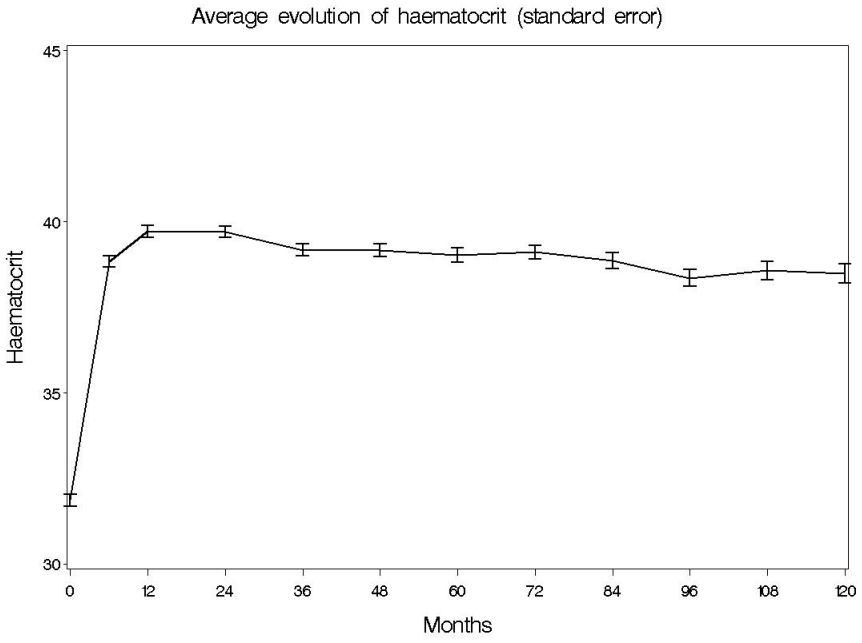

Figure 1: Average evolution of haematocrit

CI) during the first 6 months after transplant and followed by a smaller rise in the next 6 months. During the period from the 1st year until the 10th year, the level remains a trifle below 40 and decreases slowly over time. The increasing CI is most likely due to patient drop-outs in later years (and does not *per se* indicate a non-constant variance function). The Figures in figure 2 show the evolution of mean haematocrit level against months for the covariates gender, cardio-vascular problems agegroup and rejection symptoms respectively. We divided the age variable into two class variables based on the median (median split) as $age < 48 and age \geq 48$ We note that the evolution is apparently different (although parallel) for the two genders, but not so for the cardio-vascular problems. For the rejection symptoms, it is not evident from the figure if the two evolutions are significantly different (although they are also parallel in evolution with respect to time). Hence among the three cofactors, gender appears to have most pronounced effect.

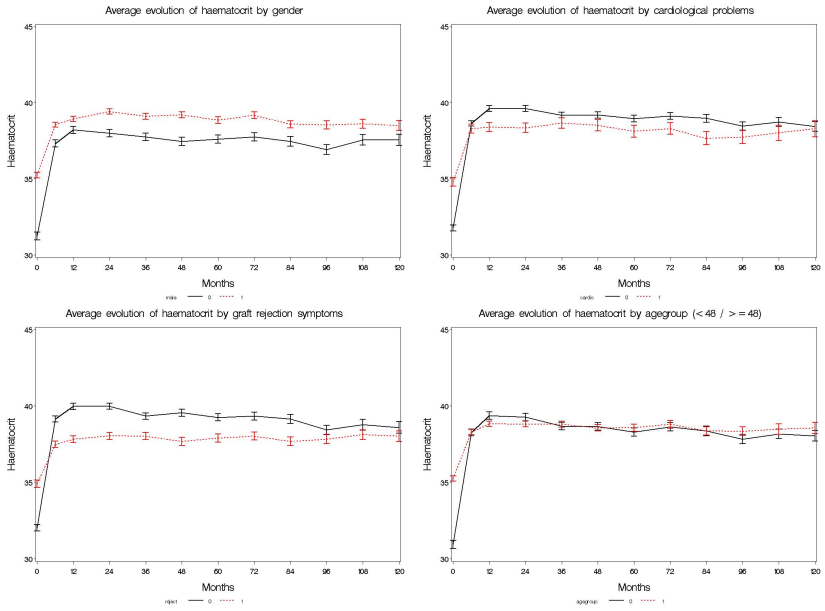

Figure 2: Evolution of haematocrit for different factors

1.3 Variance structure

Figure 3 below shows the evolution of variance with time (month), hence called the 'variance function'. The scatter plot in the figure are the squared residuals. Although the variance function 'looks' very straight, but on regressing the squared residuals on the time variable, it gives a significant negative slope indicating that it is not constant with time. This means that we are not able to use a random intercept model later on in our dataset, neither a semi-variogram would be appropriate in this case. Since our data is unbalanced, we are also not able to assess the correlation structure with respect to time for the subjects.

1.4 Individual profiles

Since the number of patients were quite large, we plotted 8 different graphs from the 3 binary variables (cardio-vascular problems, gender, rejection symptoms) to evaluate visually the within- and between-variability part in

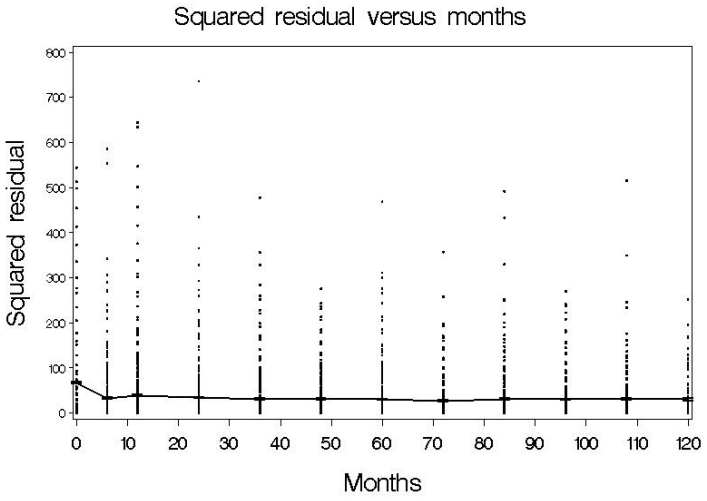

Figure 3: Variance function

each of these 8 graphs and also to observe if any particular combination of the binary variables leads to a difference in evolution of haematocrit level in that particular sub-group. The analysis is nevertheless subjective, although its role cannot be overemphasized in that it forms a baseline to built a good mean function to explain the variability with respect to time.
There is a wide baseline between-variability in all the 8 sub-groups. The within-variability part looks less pronounced, however the individual evolutions doesn't appear to follow, in general, a linear course, but probably could be explained by a higher order (e.g., quadratic) polynomial function of the time variable to take into account the initial 'bent' in the evolutions of the patient profiles.

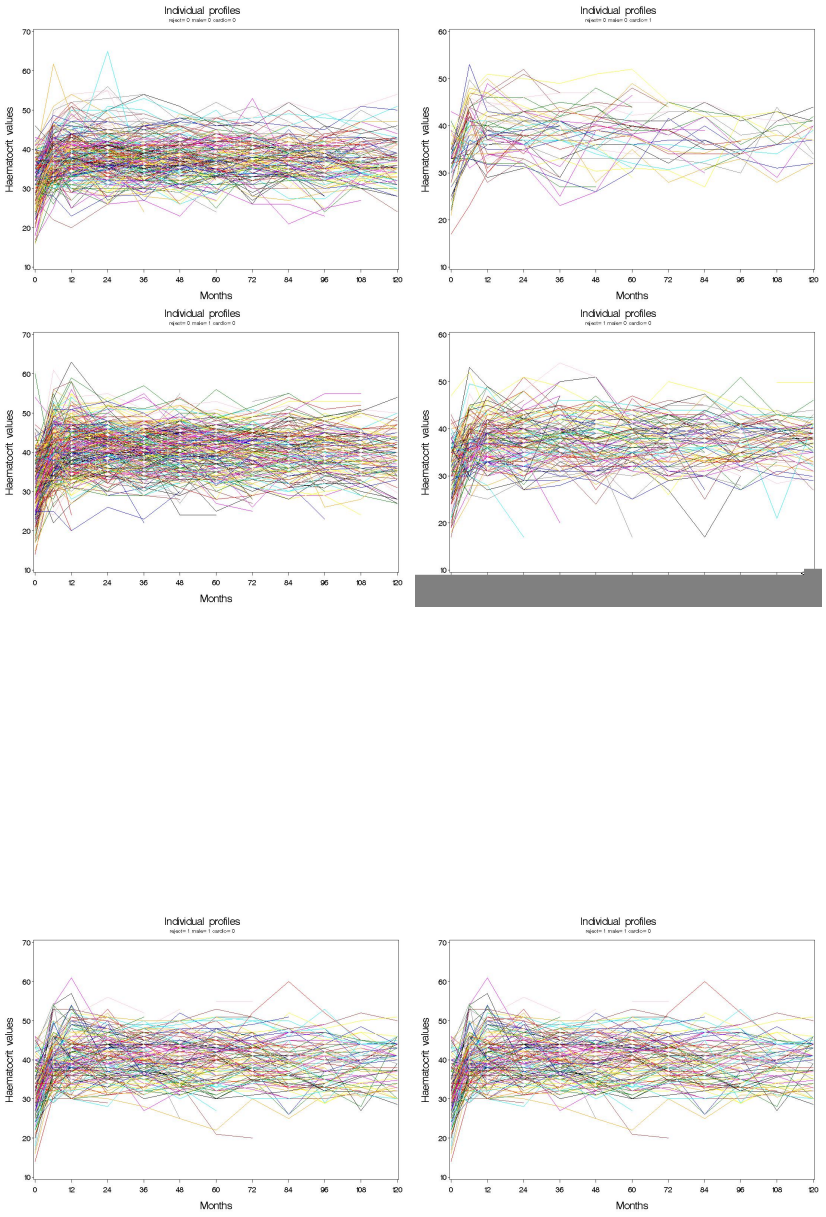

Figure 4: Individual profile in each study subgroup

2 Summary statistics

Longitudinal data are often correlated, and therefore the classical methods to analyse data often fail in this field. There are a number of different methods of using summary statistics to analyse the data, but often the longitudinal aspect of the data are lost due to data reduction and in addition it doesn't say anything about the manner of evolution of patients in different groups. Some typical examples of summary statistics widely used are :- area under the curve (AUC) analysis, which is not applicable where the profiles intersect ; analysis of increment using only partial information ; analysis of endpoints (this is particularly used in a randomized trial, where the baseline responses are supposedly equal in the groups) ; comparison of each time point simultaneously (with the correction for multiple comparisons) with the disadvantage that time ordering is lost ; paired t-test (or paired Wilcoxon's test for matched data, where the data are not Normally distributed) between the endpoints and baseline (uses partial information) ; analysis of covariance (where the baseline response is entered as a covariate) ; taking the means of the two (or more groups) and comparing them by classical parametric or non-parametric tests (using partial information) ; analysis of mid-points or peaks (also with the disadvantage of using partial information), etc.
In our report, we present the paired t-test result to compare if there were any significant difference between the haematocrit at baseline and at the end of 10th year. Therefore, only for those patients having participated till the end of the study will be taken into account, in addition, the result will not take into account the evolution of haematocrit with time for those patients. We obtained a significant difference (improvement) of means between the two haematocrit values matched for patients with an increase of 9.2% ($t = 21.9$, $df = 346$, bilateral, $P < .0001$).
We also used the ANCOVA procedure to fit a model for the haematocrit level at the end of 10th year (hc10) with age, male, reject, cardio, hc0 as predictor variables. In the final model, only age, male with intercept were individually significant at the 0.05 significance level ($P < .001$ for all the parameters), as well as the model was also significant ($P < .0001$). This model is given as follows :

$$hc_{10} = 34.29 + 0.08 * age + 2.01 * male$$

3 Multivariate model

In order to build towards a parsimonious model for the mean structure, we used the PROC MIXED procedure in SAS. We started by considering a model having an intercept part, a linear function of time part and also included a quadratic function of time part for all the continuous and categorical variables that we are provided with. We also included important interactions that may be present based on clinical relevance. They were interaction between 'age*cardio' (since cardio-vascular problems could increase with age), interaction between 'male*cardio' (since males have more cardio-vascular problems than females), interactions between 'male*reject' & 'age*reject' (because of some unknown reasons, age and/or gender might be involved with renal graft rejection). We didn't consider interactions between 'male*age' and 'reject*cardio' in our full model, since these interactions were not clinically important to consider. We corrected for age in the intercept part, but didn't include age in the linear or quadratic functions of time variable due to the same reason.
We start with our 'baseline full model' as having the four interactions as mentioned in the previous paragraph. The interactions were not significant by -2loglikelihood method, which follows G^2 distribution (asymptotically X^2 distribution). The stepwise model reduction procedure using Maximum Likelihood is shown in table below. Finally, we get a model (number 9 in the table), which could not be reduced further. All the parameters were individually significant at the 0.05 significance level ($P < .0001$ for all the parameters), as well as the full model was also significant ($P < .0001$).
The parsimonious model is shown in the form of equation below :

$$\text{Haematocrit level} = \text{age male male*month male*month2}$$

We can therefore see that from a full model of 31 parameters in the beginning, we could finally obtain a more parsimonious model for the mean structure, where all the number of parameters is reduced to 7.
In the model building process, we used the method of -2loglikelihood. It was also prossible to use AIC, BIC, Akaike-Schwarz criteria for model selection, which we didn't use in our case.

nr	Model	Par	-2l	Ref	G^2	diff df	p-value
1	cardio reject male age cardio*age male*cardio male*reject age*reject cardio*month reject*month male*month cardio*month2 reject*month2 male*month2	31	54648.4				
2	cardio reject male age cardio*age male*cardio age*reject cardio*month reject*month male*month cardio*month2 reject*month2 male*month2	27	54648.4	1	0	4	1
3	cardio reject male age cardio*age male*cardio cardio*month reject*month male*month cardio*month2 reject*month2 male*month2	25	54648,5	2	0.1	2	0.95
4	cardio reject male age male*cardio cardio*month reject*month male*month cardio*month2 reject*month2 male*month2	23	54649	3	0.5	2	0.78
5	cardio reject male age cardio*month reject*month male*month cardio*month2 reject*month2 male*month2	19	54649,5	4	0.5	4	0.97
6	reject male age cardio*month reject*month male*month cardio*month2 male*month2 reject*month2	17	54649,7	5	0.2	2	0.90
7	reject male age cardio*month male*month cardio*month2 male*month2	13	54652,3	6	2.6	4	0.63
8	male age cardio*month male*month cardio*month2 male*month2	11	54652,9	7	0.6	2	0.74
9	male age male*month male*month2	7	54655,9	8	3	4	0.56
10	age male*month male*month2	5	54706,6	9	50.7	2	0

Regarding the covariance structure, we didn't reduce it by any special structure with the options for special covariance structures given in SAS, because our data were unbalanced. In the unstructured covariance matrix, there were 78 parameters (12 for variances and 66 for covariances). The parameter estimators can be found in the following table:

Effect	Estimate	Standard Error
Age effect	0.067	0.008
Gender		
Male	33.30	0.414
Female	31.27	0.423
Month Effect		
Male	0.11	0.007
Female	0.09	0.008
Month^2 Effect		
Male	-0.00083	0.000057
Female	-0.00065	0.000062

4 2-Stage analyses

4.1 Introduction

The aim of the two-stage analysis is first to fit a linear regression model for each subject separately, which gives the subject-specific intercepts and slopes. This model describes the observed variability within subject. Then we explain the variability in the subject-specific regression coefficients using known covariates. This is the analysis of these intercepts and slopes, which allows to study the between-subject variability.

4.2 Marginal model

As a first stage model we look at the polynomial $HC_{ij} = \sum_{k=0}^{n} \alpha_k t_{ij}^k$. With an F_{Meta}-test we find the degree of polynomial of this first model. The results of these tests can be found in the following table:

Model	R^2_{Meta}	model comparison	F_{Meta}	p-value
quadratic	0.2388	quadratic to linear	2.2416	0.000
cubic	0.4669	cubic to quadratic	2.2372	0.000
4th	0.6398	4th to cubic	2.3389	0.000
5th	0.7681	5th to 4th	2.3005	0.000
6th	0.8559	6th to 5th	1.8997	0.000
7th	0.9085	7th to 6th	0.4354	0.6967

In this table we see that every model (upto the 6th one) is rejected compared to a model with a higher degree time polynomial. This means that we should look at a first stage model as a sixth order time polynomial. However, due to overparametrization and because it is difficult to interpret, we will only look at a second order time polynomial. We used the parsimonious model obtained in the multivariate model to build a 2-stage model.

$$HC_{ij} = \alpha_{1i} + \alpha_{2i} * t_{ij} + \alpha_{3i} * t_{ij}^2 + \epsilon_{ij}$$

For a subject i at a certain time t_{ij}. With the assumption that $\epsilon_{ij} \sim N(\mathbf{0}, \Sigma_i)$ We think however that these $\alpha_{.i}$ are not the same for every subject, so there

is also some variance between the subjects. We also try to capture this into a model. We will try to write these as a model of age (A_i) and gender (M_i) because this are the covariates we have found in the parsimonious multivariate model. Normally we should do the same analysis as we did for our multivariate model. This means starting with the full model, and deleting the non-significant terms in the model. However, due to a lack of time, we only looked at the same model we found in the multivariate part. This means we have the following second stage model:

$$\begin{cases} \alpha_{1i} &= \beta_0 * A_i + \beta_1 * (M_i) + \beta_2 * (1 - M_i) + b_{1i} \\ \alpha_{2i} &= \beta_3 * (M_i) + \beta_4 * (1 - M_i) + b_{2i} \\ \alpha_{3i} &= \beta_5 * (M_i) + \beta_6 * (1 - M_i) + b_{3i} \end{cases}$$

Here A and M_i respectively represents the age and gender of the i-th subject, and $b_i \sim N(\mathbf{0}, D)$.

If we combine these 2 models we become the following model:

$$HC_{ij} = \begin{cases} \beta_0 A_i + \beta_1 + \beta_3 * t_{ij} + \beta_5 * t_{ij}^2 + b_{1i} + b_{2i} * t_{ij} + b_{3i} * t_{ij}^2 + \epsilon_{ij} \\ \qquad\qquad \text{if } M_i = 1 \text{ i.e. i-th subject is male} \\ \\ \beta_0 A_i + \beta_2 + \beta_4 * t_{ij} + \beta_6 * t_{ij}^2 + b_{1i} + b_{2i} * t_{ij} + b_{3i} * t_{ij}^2 + \epsilon_{ij} \\ \qquad\qquad \text{if } M_i = 1 \text{ i.e. i-th subject is female} \end{cases}$$

Now we try to estimates these parameters with SAS, however, it noted convergency problems in calculating the parameters because the variance of b_{3i} was converging to 0.
To solve these problems we transformed our time variable (which was originally expressed in months) to the number of years. Now the procedure do converges, and it gives the following model for expected HC-value on a time t (after the transplant) of a A year old person.

$$HC(t, A) = \begin{cases} 33.90 + 0.06A + 1.85 * t - 0.18 * t^2 & \text{if male} \\ \\ 31.99 + 0.06A + 1.43 * t - 0.13 * t^2 & \text{if female} \end{cases}$$

With the following estimators for D

$$D \approx \begin{pmatrix} 11.68 & -0.52 & -0.03 \\ -0.52 & 1.26 & -0.09 \\ -0.03 & -0.09 & 0.01 \end{pmatrix}$$

And $\Sigma = \sigma^2 \approx 18.24 I_{n_i}$.
The parameter estimators can be found in the following table:

Effect	Estimate	Standard Error
Age effect	0.059	0.008
Gender		
Male	33.98	0.435
Female	31.98	0.444
Year Effect		
Male	1.87	0.074
Female	1.45	0.082
Year^2 Effect		
Male	-0.19	0.008
Female	-0.14	0.009

Note that if you compare these table with the table in Chapter 3, you should divide the parameters found in the year and year^2 effect divide by respectively 12 and 12^2, this in order to maintain the month$\rightarrow$ year transformation. The transformation of the standard errors are not so easy.
In this model, again all mean parameters turns out to be significant.

5 Random effect model

In the previous 2-stage model formation, we used the model obtained from the multivariate model by using Restricted Maximum Likelihood Method (REML). In our final model building stage, we try to build a model by exploring serial correlation in our data. We used two most commonly used serial correlation function, viz., the exponential and the Gaussian functions. The full model described in the beginning of our multivariate model formulation was also considered to be our building block in this stage. We reduced the model stepwise as shown in the following table.

nr	Model	Parameters	Random effects	Cov. Par.	-2l	Ref	G^2	diff df	p-value
1	FULL	31	int year year2	9	57712.2				
2	-male*reject	27	int year year2	9	57712.6	1	0.4	4	0.98
3	-age*reject	25	int year year2	9	57706.4	2	-6.2	2	1
4	-age*cardio	23	int year year2	9	57701.9	3	-4.5	2	1
5	-cardio*male	19	int year year2	9	57703.1	4	1.2	4	0.88
6	-reject	17	int year year2	9	57702.9	5	0.2	2	0.91
7	-reject*year2	15	int year year2	9	57702.6	6	-0.3	2	1
8	-reject*year	13	int year year2	9	57698.5	7	-4.1	2	1
9	-ALL cardio's	7	int year year2	9	57690	8	8.5	6	0.20
10		7	int year	6	57796.1	9	106.1	2:3	$< .0001$

The parameter estimates with standard errors are provided below:

Effect	Estimate	Standard Error
Age effect	0.06	0.008
Gender		
Male	33.90	0.44
Female	31.99	0.44
Year Effect		
Male	1.87	0.074
Female	1.45	0.082
Year^2 Effect		
Male	-0.19	0.008
Female	-0.14	0.009

We find no difference in -2loglikelihood by using exponential or Gaussian serial correlation functions. We were able to reduce the full model to a par-

simonious model of 7 parameters, which were same as found in the multivariate stage. These parameters were all significant at the 0.05 level ($P < .001$ for all parameters and $P < .0001$ for the model). We also note that the parameter estimates are very similar in random effects and the 2-stage models (both being built using REML).

6 Conclusion

We proceeded with analysing our data graphically in the beginning, followed by summary techniques. Then we used PROC MIXED (GLM) in SAS to explore the mean structure by developing a parsimonious model of 7 parameters for the mean structure (however we couldn't reduce the covariance structure as it was not balanced data or by heterogeneous compound symmetry option).
In the next step, we built a 2-stage model and report our estimates for a marginal model. In the end we tried a plausible random effects model, which is by far the best of all the methods used, using exponential function for the serial correlation part. We obtain a model with 7 parameters for the mean structure as we obtained with the multivariate model. The main difference between the multivariate and the random effects model is the parameter estimates, which is not surprising, since these two methods use different procedures, ML and REML respectively. Both of the models show a significant time and time squared effect of the gender.

References

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

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

7 Appendix

7.1 SAS-codes

```
libname in "C:\Documents and Settings\soutrik.banerjee\My Documents\LDA_hw1";
proc contents data = in.Renal;
run;

/*adding an agegroup to the dataset, (younger or older then 48)*/

data renal1;
set in.renal;
agegroup = 1;
if age < 48 then agegroup = 0;
run;

/*transposing the horizontal data to vertical*/

data renal (keep = id hc months age male cardio reject agegroup);
set Renal1;
  id =_n_;
  array time  {12} hc0 hc06 hc1 hc2 hc3 hc4 hc5 hc6 hc7 hc8 hc9 hc10;
  do i = 1 to 12;
        months = i;
        hc = time{i};
        output renal;
  end;
run;

data renal2;
set renal;
if months = 1 then month = 0;
if months = 2 then month = 6;
if months = 3 then month = 12;
if months = 4 then month = 24;
if months = 5 then month = 36;
if months = 6 then month = 48;
if months = 7 then month = 60;
if months = 8 then month = 72;
if months = 9 then month = 84;
if months = 10 then month = 96;
if months = 11 then month = 108;
if months = 12 then month = 120;
run;

/*exploring the mean structure*/

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = renal2;
plot hc*month / haxis = axis1 vaxis = axis2;
symbol1 c = black i = std1mjt l = 1 w = 2 mode =  include; /*l-line type, w-thickness*/
axis1 label = (h = 2 'Months') value = (h = 1.5) order = (0 to 120 by 12) minor = none;
axis2 label = (h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by 5) minor = none;
title h = 2 'Average evolution of haematocrit (standard error)';
run;
quit;

/*exploring the mean structure by gender*/

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = renal2;
plot hc*month=male / haxis = axis1 vaxis = axis2;
symbol1 c = black i = std1mjt l = 1 w = 2 mode =  include; /*l-line type, w-thickness*/
symbol2 c = red i = stdm1jt l = 2 w = 2;
axis1 label = (h = 2 'Months') value = (h = 1.5) order = (0 to 120 by 12) minor = none;
axis2 label = (h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by 5) minor = none;
title h = 2 'Average evolution of haematocrit by gender';
run;
quit;

/*exploring the mean structure by cardio*/
```

```
goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = renal2;
plot hc*month=cardio / haxis = axis1 vaxis = axis2;
symbol1 c = black i = std1mjt l = 1 w = 2 mode =  include; /*l-line type, w-thickness*/
symbol2 c = red i = stdm1jt l = 2 w = 2;
axis1 label = (h = 2 'Months') value = (h = 1.5) order = (0 to 120 by 12) minor = none;
axis2 label = (h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by 5) minor = none;
title h = 2 'Average evolution of haematocrit by cardiological problems';
run;
quit;

/*exploring the mean structure by reject*/

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = renal2;
plot hc*month=reject / haxis = axis1 vaxis = axis2;
symbol1 c = black i = std1mjt l = 1 w = 2 mode =  include; /*l-line type, w-thickness*/
symbol2 c = red i = stdm1jt l = 2 w = 2;
axis1 label = (h = 2 'Months') value = (h = 1.5) order = (0 to 120 by 12) minor = none;
axis2 label = (h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by 5) minor = none;
title h = 2 'Average evolution of haematocrit by graft rejection symptoms';
run;
quit;

/*exploring the mean structure by agegroup*/

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = renal2;
plot hc*month = agegroup / haxis = axis1 vaxis = axis2;
symbol1 c = black i = std1mjt l = 1 w = 2 mode = include; /*l-line type, w-thickness*/
symbol2 c = red i = stdm1jt l = 2 w = 2;
axis1 label = (h = 2 'Months') value = (h = 1.5) order = (0 to 120 by 12) minor = none;
axis2 label = (h = 2 A = 90 'Haematocrit') value = (h = 1.5) order = (30 to 45 by 5) minor = none;
title h = 2 'Average evolution of haematocrit by agegroup (<48 / >=48)';
run;
quit;

/*exploring individual profiles*/

proc sort data=renal2;
by reject male cardio;
run;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape i = join;
proc gplot data = renal2;
plot hc*month=id /nolegend skipmiss haxis = axis1 vaxis = axis2;
axis1 label = (h = 2 'Months') value = (h = 1.5) order = (0 to 120 by 12) minor = none;
axis2 label = (h = 2 A = 90 'Haematocrit values') value = (h = 1.5) minor = none;
title h = 2 'Individual profiles';
by reject male cardio;
run;
quit;

/*difference between baseline and endpoint*/

proc ttest data = in.Renal;
paired hc0*hc10;
title;
run;

/*Starting ANCOVA model*/

proc glm data = in.Renal;
model hc10 = age male cardio reject hc0;
run;

/*Final ANCOVA model*/

proc glm data = in.Renal;
model hc10 = age male;
run;

/*normality testing for hc0 (marginal deviation from normality)*/

proc univariate normaltest plot data = in.Renal;
var hc0;
```

```
output;
run;

/*normality testing for hc10*/

proc univariate normaltest plot data = in.Renal;
var hc10;
output;
run;

/*exploring the variance structure */

/*plotting the variance function*/
title;
proc glm data = renal2;
model hc = month;
output out = out r = residual;
run;

proc sort data = out;
by month;
run;

data out2;
set out;
residual2 = residual**2;
run;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = out2;
plot residual2*month = 1 residual2*month = 2 / overlay skipmiss haxis = axis1 vaxis = axis2;
symbol1 c = black i = std1mjt w = 2 mode = include ;
symbol2 c = black v = dot h = 0.2 mode = include ;
axis1 label = (h = 2 'Months') value = (h = 1.5) minor = none;
axis2 label = (h = 2 A = 90 'Squared residual') minor = none;
title h = 2 'Squared residual versus months';
run;
quit;

title;
proc glm data = out2;
model residual2 = month;
output;
run;

/*plotting the semivariogram*/
/*(opgelet !! proc variogram bis zum proc gplot ist sehr lange !!!)*/

proc variogram data = out outpair = out;
coordinates xc = month yc = id;
compute robust novariogram;
var residual;
run;

data variogram;
set out;
if y1 = y2;
vario = (v1 - v2)**2/2;
run;

data variance;
set out;
if y1 < y2;
vario = (v1 - v2)**2/2;
run;

proc means data = variance mean;
var vario;
output;
run;

proc loess data = variogram;
ods output scoreresults = out;
model vario = distance;
score data = variogram;
run;
```

```
proc sort data = out;
by distance;
run;

goptions reset = all ftext = swiss gsfmode = replace rotate = landscape;
proc gplot data = out;
plot vario*distance = 1 p_vario*distance = 2 / overlay haxis = axis1 vaxis = axis2 vref = 40.49 lvref = 3;
symbol1 c = red v = dot h = 0.2 mode = include;
symbol2 c = black i = join w = 2 mode = include;
axis1 label = (h = 2 'Time lag') value = (h = 1.5) minor = none;
axis2 label = (h = 2 A = 90 'v(u)') value = (h = 1.5);
title h = 3 'Semi-variogram';
run;quit;

/*Question 3*/
/*Adding a variable monthcls to the dataset which is just a copy of month*/

/*Deleting the rows which has missing hc-values, inserting the squared time variable and inserting an intercept 1*/
data renal3;
set  renal2;
where hc ^= .;
monthcls = month;
month2 = month**2;
run;

/*Likelihood ratio testing of different multivariate models*/

title "first model";
proc mixed data = renal3 method = ml;
class id cardio reject male monthcls;
model hc = cardio reject male age cardio*age male*cardio male*reject age*reject cardio*month reject*month male*month cardio*month2 reject*month2 male*mon
repeated monthcls / type = un subject = id;
run;

title "nineth model";
proc mixed data = renal3 method = ml;
class id cardio reject male monthcls;
model hc = male age male*month male*month2 / noint s;
repeated monthcls / type = un subject = id;
run;

title "tenth model";
proc mixed data = renal3 method = ml;
class id cardio reject male monthcls;
model hc = male age male*month male*month2 / noint s;
repeated monthcls / type = un subject = id;
run;

proc sort data = renal2;
by id;
run;

data new;
set Renal2;
month2 = month**2;
int = 1;
run;

data macro;
set new;
where hc ^= .;
run;

/*Macro*/

%macro gof(INDATA =  ,
           OUTDATA = ,
           Y =  ,
           ID = ,
           X1 =  ,
           X2 =  );

proc sort data = &indata;
by &id;
run;
```

```
proc freq data = &indata;
tables &id / out=uit  noprint ;
run;
proc iml;
use uit;
read all var {count} into aantal;
close uit;
use &indata;
labelx1 = {&X1};
labelx2 = {&X2};
labely = {&Y};
labelid = {&id};
read all var labelid into id;
read all var labely into y;
read all var labelx1 into x1;
read all var labelx2 into x2;
close &indata;

p=ncol(x1)+ncol(x2);
ftel1 = 0;
fnoem1 = 0;
dftel1 = 0;
dfnoem1 = 0;
begin = 1;
do i = 1 to nrow(aantal);
   ni = aantal[i,];
   einde = begin + ni - 1;
   if ni >= p then do;
      x1i = x1[(begin:einde),];
      x2i = x2[(begin:einde),];
      yi = y[(begin:einde),];
      xi = x1i||x2i;
      ri = yi-xi*(inv(xi`*xi))*xi`*yi;
      rHi = yi - x1i*(inv(x1i`*x1i))*x1i`*yi;
      rssi = ri`*ri;
      rssHi = rHi`*rHi;
      ftel1 = ftel1 + (rssHi - rssi);
      fnoem1 = fnoem1 + rssi;
      dftel1 = dftel1 + ncol(x2);
      dfnoem1 = dfnoem1 + (ni-p);
   end;
   begin = einde + 1;
end;
f = (ftel1/dftel1)/(fnoem1/dfnoem1);
c = {"F"  "ndf" "ddf" "p-value"};
F_test = (F||dftel1||dfnoem1||(1-probf(f,dftel1,dfnoem1)));
print F_test[colname=c format=10.4];


p=ncol(x1);
R = 0||0||0;
ssto = 0;
ssr = 0;
begin = 1;
do i = 1 to nrow(aantal);
   ni = aantal[i,];
   einde = begin + ni - 1;
   if ni >= p then do;
      idi = id[begin,];
      x1i = x1[(begin:einde),];
      yi = y[(begin:einde),];
      ri = yi - x1i*(inv(x1i`*x1i))*x1i`*yi;
      ssei = ri`*ri;
      sstoi = (yi-(j(1,ni)*yi/ni))`*(yi-(j(1,ni)*yi/ni));
      ssri = sstoi-ssei;
      ssto = ssto+sstoi;
      ssr = ssr+ssri;
      if sstoi > 0 then do;
         r = r//(idi||(ssri/sstoi)||ni);
      end;
      else do;
         r = r//(idi||1||ni);
      end;
   end;
   else do;
      idi = id[begin,];
```

```
      yi = y[(begin:einde),];
      sstoi = (yi-(j(1,ni)*yi/ni))'*(yi-(j(1,ni)*yi/ni));
      ssto = ssto+sstoi;
      ssr = ssr+sstoi;
      r = r//(idi||1||ni);
   end;
   begin = einde + 1;
end;
Tot_R2=ssr/ssto;
c = {"R2"};
print Tot_R2[colname=c format=10.4];
r=r[2:nrow(r),];
naam = labelid||"R2"||"ni";
create &outdata var naam;
append from r;
quit;
run;
%mend;

%gof(indata=macro,
outdata=rsquares,
Y=hc,
ID=id,
X1=int month month2,
X2=month3);
run;

title "Ri_square_meta & F_meta";
goptions reset = all ftext = swiss device = psepsf gsfmode = replace rotate = landscape;
proc gplot;
plot R2*ni / overlay vref = 0.4669 lvref = 1;
symbol v = dot h = 0.3;
run;

/*2 stage model*/

title "REMLE";
proc mixed data = renal3 method = reml;
class id monthcls male;
model hc = age male male*month male*month2 / noint solution ddfm = satterth;
random intercept month month2 / type = un subject = id g gcorr v vcorr;
repeated monthcls / type = simple subject = id r rcorr;
run;

/*This procedure doesn't converge, So we try to change the time expression in years instead of months*/

data renal4;
set renal3;
year = month / 12;
year2 = month2 / 12**2;
yearcls = monthcls / 12;
run;

title "REMLE";
proc mixed data = renal4 method = reml;
class id yearcls male;
model hc = age male male*year male*year2 / noint solution ddfm = kr;
random intercept year year2 / type = un subject = id g gcorr v vcorr;
repeated yearcls / type = simple subject = id r rcorr;
run;

title "MLE";
proc mixed data = renal4 method = ml;
class id yearcls male;
model hc = age male male*year male*year2 / noint solution ddfm = kr;
random intercept year year2 / type = un subject = id g gcorr v vcorr;
repeated yearcls / type = simple subject = id r rcorr;
run;

/*Likelihood ratio testing of different random effects model*/

title "first model";
proc mixed data = renal4 method = reml;
class id cardio reject male yearcls;
model hc = cardio reject male age cardio*age male*cardio male*reject age*reject cardio*year reject*year male*year cardio*year2 reject*year2 male*year2 / s
random intercept year year2 / type = un subject = id g gcorr v vcorr;
```

```
repeated yearcls / type = sp(exp)(year) local subject = id r rcorr;
run;

title "eighth model";
proc mixed data = renal4 method = reml;
class id cardio reject male yearcls;
model hc = cardio male age cardio*year male*year cardio*year2 male*year2 / noint solution ddfm = kr;
random intercept year year2 / type = un subject = id g gcorr v vcorr;
repeated yearcls / type = sp(exp)(year) local subject = id r rcorr;
run;

title "nineth model";
proc mixed data = renal4 method = reml;
class id cardio reject male yearcls;
model hc = male age male*year male*year2 / noint solution ddfm = kr;
random intercept year year2 / type = un subject = id g gcorr v vcorr;
repeated yearcls / type = sp(exp)(year) local subject = id r rcorr;
run;

title "nineth model -year2";
proc mixed data = renal4 method = reml;
class id cardio reject male yearcls;
model hc = male age male*year male*year2 / noint solution ddfm = kr;
random intercept year / type = un subject = id g gcorr v vcorr;
repeated yearcls / type = sp(exp)(year) local subject = id r rcorr;
run;

title "tenth model";
proc mixed data = renal4 method = reml;
class id cardio reject male yearcls;
model hc = age male*year male*year2 / noint solution ddfm = kr;
random intercept year year2 / type = un subject = id g gcorr v vcorr;
repeated yearcls / type = sp(exp)(year) local subject = id r rcorr;
run;
```