Hasselt University
Censtat

Longitudinal Data Analysis

HW2: Discrete Longitudinal Data
Renal Transplant study

by

Banerjee Soutrik
Kishtammagari Manjularani
Rosius Wannes

Diepenbeek, March 10th, 2006

Contents

1 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 and the following predictor variables were recorded (in the dataset provided) : age at transplantation (continuous variable *age*), gender (binary variable *male*), cardio-vascular problems before the transplant (binary variable *cardio*), rejection symptoms experienced in the first trimester post-transplant (binary variable *reject*). The response variable measured are the blood haematocrit values at baseline (hc_0), and then at six months (hc_{06}), one year (hc_1), two years (hc_2) and in the same manner, until 10 years (hc_{10}), which is considered as the endpoint in the present study. 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 absent patients in a particular follow-up schedule. These drop-outs might have been related, at least in part, to the outcome of the patient's 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 diminution of erythropoietin production in patients with chronic renal failure (CRF) causes a gradual decrease 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. In this study, 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 the effects of other cofactors, such as gender, cardio-vascular problems, age and rejection symptoms. Successful kidney transplantation leads to the correction of renal anaemia over an 8-10 week period [2].

In this report, we dichotomised the response variable, in order to make an analysis of the probability of a 'gain' with respect to a 'loss' and study its evolution with time. There are several ways to do this. Let's take an example - you can say that 37% is a good haematocrit value, and everything above is a desirable outcome, while everything below is an unfavourable outcome. Since, we would like to see the effect of the transplant corrected for baseline differences in respective individuals, we chose not to use this dichotomisation by directly taking a cut-off on the absolute value of the haematocrit level, but by taking the difference between the haematocrit values from the baseline haematocrit value, (*i.e.*, $hc_i - hc_0$) for all individuals. Next, we would still need a cut-off point on this new response variable for the purpose of our study. For a cut-off point (obtained by taking the average of baseline and post-transplant haematocrit levels), we used a difference of 3.5% from the baseline haematocrit value. This means that an increase of (at least) 3.5% post-transplant is a desirable effect and an increase (or even further decrease)

of less then 3.5% is an unfavourable effect of the transplant on the blood haematocrit. For the dichotomisation we then used the variable *hcco*, with $hcc0 = 1$, if the difference of the haematocrit compared to the baseline is more than or equal to 3.5%, while $hcco = 0$, if the difference is less then 3.5%.

Here, we want to model the probability that $hcco = 1$ with respect to $hcco = 0$. That is the probability of a desirable effect of the transplant with respect to the probability of an unfavourable effect.

$$\pi_{ij} := P(hcco_{ij} = 1) \qquad \text{or} \qquad 1 - \pi_{ij} := P(hcco_{ij} = 0)$$

In the previous *Longitudinal Data Analysis* assignment, we have seen that all interaction terms between the covariates $Male$, Age, $Cardio$ and $Reject$ are non-significant. That is why, in this report, we will not include any such interaction terms.

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 directly by bone marrow suppression (Azathioprine) or indirectly by renal function deterioration (Cyclosporine). Therefore in the present setting, it is not possible to correct for the treatment effects on the patients, which might have been related to the outcome variable. In addition, a condition called Post-Transplant Erythrocytosis (PTE) causes not only restitution of anaemia, but even polycythaemia in certain patients. The causes are often idiopathic, but Theophylline and Angiotensin Converting Enzyme Inhibitors (ACEI) are also thought to play a role. In this condition, there is persistently elevated haematocrit level ($> 51\%$). Since this condition is prevalent more in males than females, this may partially be responsible for 'pulling up' the mean haematocrit in males in the present study. Since this condition is also not given in the current dataset, it is not possible to ascertain, if there is difference in response associated with gender purely due to gender itself and/or due to PTE.

Our approach to analyse this dataset is illustrated in different models in the following sections. SAS is used as the principal software to treat the data. We don't extensively present the underlying theory behind the different model building in these sections, but present in as much non-mathematical way as possible the interpretation of the models and results obtained to a non-technical person.

2 Generalized Estimating Equations

2.1 Introduction

In this section we are only interested in first-order marginal mean parameters. For this we will use the Generalized Estimating Equations (GEE), which is based on non-likelihood estimation methods (quasi-likelihood). GEE require only a specification of the univariate marginal model [8]. This means that are willing to adopt *working assumptions*. The main advantage of GEE is that the association between observations on the same subject can be specified through this working correlation matrix, but the form of this one is not important, since a misspecification of it doesn't affect the consistency. Of course, when we have found our parsimonious GEE-model, we again have to check if these working assumptions are correct or not.

In order to build towards a parsimonious model for the mean structure, we used the PROC GENMOD procedure in SAS. We started by considering a model having an intercept part, a linear function of time and also included a quadratic function of time part for all the continuous and discrete class variables that we are provided with. We did not include interaction terms as was mentioned in the introduction.

It is also possible to do a second order extension of the estimating equations (GEE2). This parts also include the marginal pairwise association. However, in this project, we will not go into this.

2.2 Fitting the marginal model

Several choices are available for adopting the working correlation matrix in a GEE-model. The most commonly used structures are Unstructured, Independence, Exchangeable, Auto-Regressive(1). But the choice of the working correlation structure is not very important in this context, since a non-appropriate working correlation structure still provides consistent mean regression parameter estimates as the cluster size becomes large, however at the cost of loss of precision or efficiency.

We start with the full model:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= Int + M * male + A * age + C * Cardio + R * Reject \\ &\quad +(Y + M1 * male + A1 * age + C1 * Cardio + R1 * Reject) * year \\ &\quad +(Y2 + M2 * male + A2 * age + C2 * Cardio + R2 * Reject) * year^2. \end{aligned} \tag{1}$$

Starting from this full model, we try to come to a parsimonous model.

The initial parameter estimates are obtained by fitting the model under the assumption of independent correlation structure by SAS. SAS later uses these initial estimates to calculate the model-based (or naive) and empirical (or robust or Sandwich) estimates for the working correlation structure. One must be cautious not to read too much from the initial estimates.

Next, we compare the empirical and model-based standard errors in order to select a suitable working correlation structure from the above four types of correlation structures. If the standard errors of the model-based estimates are not too different from those of the empirical estimates, it implies that the working correlation matrix that we used in the model was not too far from reality. Table1 shows the parameter estimates and standard errors of model-based and empirical estimates for the correlation structures : Unstructured, Exchangeable, Independent and Auto-Regressive(1).

In Table 1, it can be seen that the differences between the empirical from the model-based standard errors for the Independent and Auto-Regressive(1) types are quite different, but for Exchangeable (Compound Symmetry in SAS) and Unstructured types are not very different (only regression parameters of the final model is presented). Unstructured type has always the smallest differences between the empirical and model-based standard errors. Therefore, we decided to use the Unstructured type as our working correlation structure, although Exchangeable type could equally be adopted.

The next step is to select which regression parameters are significant in the parsimonious model. Since log-likelihood method can't be used to reduce the model, we use Type 3 option in SAS for the backward model selection method, with $\alpha = .05$. The model is as follows :

$$\begin{aligned} \log\left(\frac{\pi}{1-\pi}\right) &= 1.911 - 0.325(1-M) - 0.30(1-C) \\ &\quad -0.016Age - 0.68Year + 0.0019Age * Year \end{aligned} \tag{2}$$

type		**Un**			**Exch**			**Ind**			**AR(1)**		
Parms		Est	Emp S.E.	MB S.E.	Est	Emp S.E.	MB S.E.	Est	Emp S.E.	MB S.E.	Est	Emp S.E.	MB S.E.
Int		2.016	0.276	0.277	1.982	0.275	0.268	1.932	0.291	0.1734	1.911	0.283	0.266
$Male$	0	-0.325	0.101	0.101	-0.322	0.102	0.100	-0.372	0.109	0.049	-0.325	0.104	0.083
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$Cardio$	0	-0.303	0.136	0.141	-0.292	0.137	0.139	-0.2455	0.147	0.069	-0.301	0.141	0.117
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Age		-0.016	0.005	0.004	-0.016	0.005	0.004	-0.016	0.005	0.003	-0.016	0.005	0.005
$Year$		-0.088	0.028	0.029	-0.082	0.027	0.022	-0.070	0.039	0.031	-0.68	0.035	0.043
$Age * year$		0.0015	0.0006	0.0006	0.0013	0.0006	0.0005	0.0023	0.0023	0.0007	0.0019	0.0008	0.0009

Table 1: Estimates of GEE models

3 Random effects model

3.1 Introduction

In this section, we are interested in describing the individual evolutions of each subject separately [8].
We will try to fit the model with a Generalized Linear Mixed Model (GLMM). It assumes that the outcome variables are independent, following a generalized linear model, corrected with a subject-specific regression parameter.

In SAS, and also in other software tools, there are several ways to fit these models. They are described in [8], Chapter 15. We will mainly use PROC NLMIXED to analyse the problem, and the GLIMMIX macro to get starting values.

3.2 Model fitting

Here again, we will start with a full model:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= Int + M * male + A * age + C * Cardio + R * Reject + b_1 \\ &\quad + (Y + M * male + A1 * age + C1 * Cardio + R1 * Reject + b_2) * year \\ &\quad + (Y2 + M * male + A2 * age + C2 * Cardio + R2 * Reject + b_3) * year^2. \end{aligned} \tag{3}$$

Where b_1, b_2 and b_3 represents a random effect in the intercept term, the slope term, and the $year^2$ effect, respectively.

We will fit four different starting models. The first model is where we only look at a random intercept term. In a second model, we include a linear independent random slope effect. As a third model we add an independent random $year^2$ effect. Then we end with these 3 random effects, but also let them be correlated. At the end the covariance matrix of (b_1, b_2, b_3) is given by

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{12} & d_{22} & d_{23} \\ d_{13} & d_{23} & d_{33} \end{pmatrix}.$$

With the last model, we can try to reduce the random effects with a χ^2 test to look for significant difference in log likelihood. However, all variances and covariance between the random effects turns out to be significant.

In order to fasten the computing time of the NLMIXED procedure in SAS, we already need initial estimates for the parameters in the full model. These initial estimates, we get from the GLIMMIX macro.

Now, we can start with reducing the models. For faster computation time, we will first uses the Gaussian quadrature of order three. This does not give a very accurate estimation, however it can give a good indication of the significance of parameters estimates.

With this we come to four different models, the parameter estimates of which can be seen in the table in Appendix A.1. As you can notice, in almost every case, the full model (3) reduces to the model:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= Int + M * male + A * age + C * Cardio + R * Reject + b_1 \\ &\quad +(C1 * Cardio + b_2) * year \\ &\quad +(C2 * Cardio + b_3) * year^2. \end{aligned} \tag{4}$$

Now, in order to have better estimates, we can try to run this model with a non adaptive gaussian quadrature of order 15 (or even in an adaptive Gaussian way) However, this last step, we will not do, because if we take a large order of quadrature points, the non adaptive estimates will come very close to the adaptive estimates.

The final model becomes then:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= 4.042 + 0.91 * male - 0.037 * age - 1.02 * Cardio - 0.48 * Reject + b_1 \\ &\quad +(0.27 * Cardio + b_2) * year \\ &\quad +(-0.027 * Cardio + b_3) * year^2. \end{aligned} \tag{5}$$

with the following variance-covariance matrix of the error terms

$$\begin{pmatrix} 13.45 & -0.86 & 0.094 \\ -0.86 & 1.1229 & -0.092 \\ 0.094 & -0.092 & 0.0084 \end{pmatrix}.$$

3.3 Empirical Bayes Estimation

In this part, we are interested in estimating the random effects. A way to estimate these effect, is described in [8], Chapter 14.2.4. With this method we obtained estimates are called the *Emperical Bayes estimates* (EB Estimates). In practise, EB estimates are usually used for diagnostic purposes, such as detection of outliers [7]. Normally, this is done with histograms and normal quantile plots.

We will start from the final random-effects model (4)[1]. As we can see, in the histograms and QQ-plots in appendix B.1, for all three parameters, the distributions do not present a normal shape.

[1]We have not used (5) because these results were only obtained just before the hand-in time of this report. We used the model obtained with three quadrature points, and three dependent random effects.

The distribution of the random intercept has mean -0.24 with variance 2.24. On the contrary the random slope is very concentrated around 0. The slope has mean -0.03 with a deviation of 0.43. The $year^2$ effect has also a mean close to zero (0.0001), and a standard deviation of 0.047.

Hereby, we would like to add that, in general, the EB-estimates show less variability than actually present in the population. This phenomenon is called *shrinkage*.

On the scatter plot we can see that there are no outliers observed.

4 Transition Model

4.1 Introduction

In this section we will write a regression model for a subject's outcome value, given the previous outcome value(s) of that subject. In our case, we will only look back two outcomes variables. We then have two extra covariates, which we call $previous_1$ and $previous_2$. We generate these variables with the SAS macro Dropout, given in [8], Chapter 32.5.

4.2 Model fitting

The same full model with the covariates $previous_1$ and $previous_2$ of the marginal model was considered as the outset. The model reduction was done in the same way as before. The parameters estimates are given in Table 2.

Parms		Est.	S.E.
Int		-1.7929	0.0829
$Male$	0	0.1760	0.0778
$Male$	1	0.000	0.000
$previous_1$		2.6061	0.0861
$previous_2$		1.7817	0.0874

Table 2: Estimates for the Transition Model

5 Discussion

In this report, we have presented the results of the 3 different types of models. The Marginal model (GEE) provides 'population-based' average of the evolution of the probability of a desirable to an unfavourable outcome with respect to time. These estimates don't include individual random effects, hence the models are presumably more 'crude' than the random-effects or hierachical models (this can be seen visually from the model-fits in the figures provided in appendix B.2), which includes the 'subject-specific' random effects. The latter may be useful in detecting outliers or special evolution patterns in a given subject of a group of subjects.

We chose to use the 'difference in haematocrit level' method in order to correct for the baseline differences in respective individuals given there were 4 known cofactors as well as there might have been more latent cofactors to correct for, which were not provided in the dataset. This could result in a 'classification bias' associated with known or unkwown cofactors if we used a direct cut-off value on the absolute haematocrit values. That is why, we tried to minimize the classification bias by correcting for the baseline differences in all subjects. Next, the bibliography didn't indicate a single cut-off value, due to which reason, we used the midpoint between the post-transplant haematocrit and baseline haematocrit.

Since patients with PTE or other causes might have been persistently associated with high haematocrit value, which is unfavourable as well, a multinomial logit model of three or more levels might be plausible in this or further controlled studies in order not to mix patients with polycythaemia and normal haematocrit levels. In addition, if we are interested for different outcomes (based on more than one cut-off values), a multinomial model is also warranted.

The transition model is different from these two models in that it gives the probability of same outcome given the previous 2 outcomes. We found from our estimates that both previous outcomes were significant in predicting the next outcome. In other words this model signifies most patients who had $hcco = 1$, stayed in the same category. (The same goes for $hcco = 0$.) It translates into a state of stability in the haematocrit level for most of the patients.

It is also possible to get a 'marginalised' mean evolution for a random-effects model by sample-based numerical integration, but we haven't included it in this report. Since a logistic function was used to model the outcome, the 'marginalised' and marginal (GEE) models are not equivalent in regression parameter estimates, even if they had the same number of significant covariates in the respective final models.

In our hierarchical model building, we used a non-adaptive Gaussian quadrature to economise the time, however under optimal conditions, it would be ideal to use between 20 - 50 quadrature points by adaptive Gaussian quadrature. However, the log likelihoods cannot be compared between the different quadrature points models.

We finally observe that in the final 3 models the 'significant' parameters were different, but in the first home-work report, the 3 models (marginal, two-stage and random-effects models), they were same and the parameter estimates were quite close as well.

References

[1] Gaston RS, Julian BA, Curtis JJ. Posttransplant erythrocytosis: an enigma revisited. *Am J Kidney Dis* 1994 Jul;**24**(1):1-11.

[2] Kessler M. Erythropoietin and erythropoiesis in renal transplantation. *Nephrol Dial Transplant* 1995;**10**(Suppl 6):114-116.

[3] Lee DB. Interrelationship between erythropoietin and erythropoiesis: insights from renal transplantation. *Am J Kidney Dis* 1991 Oct;**18**(4 Suppl 1):54-56.

[4] *Oxford Handbook of Clinical Medicine* (reprint 2002), 5th Edition, Oxford University Press, USA.

[5] Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using The SAS® System,* Second Edition. Cary, NC:SAS Institute Inc, 2000.

[6] Turkowski-Duhem A, Kamar N, Cointault O, Lavayssiere L, Ribes D, Esposito L, Fillola G, Durand D, Rostaing L. Predictive factors of anemia within the first year post renal transplant *Transplantation* 2005 Oct 15;**80**(7):903-909.

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

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

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

A SAS outcome

A.1 Estimates for the Random Effect Parameters

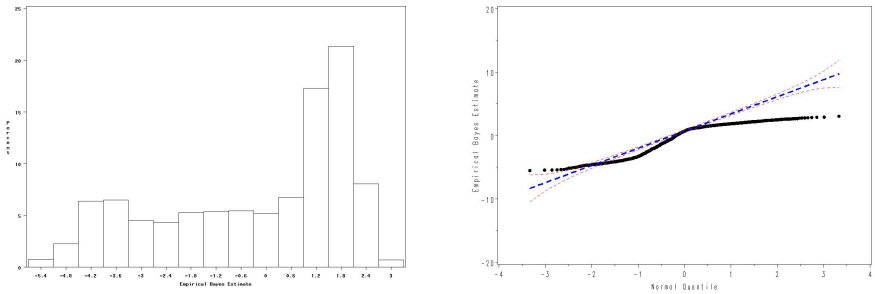
No. of Quadrature Points		3		3		3		3		15	
		1 indep. R.E.		**2 indep. R.E.**		**3 indep. R.E.**		**3 dep. R.E.**		**3 dep. R.E.**	
Parms		Est	S.E.	Est	S.E.	Est	S.E.	Est	S.E.	Est	S.E.
Int		3.8942	0.4985	3.4961	0.4973	4.1641	0.4821	4.0194	0.4529	4.0419	0.7152
$Male$	1	1.0472	0.1594	0.9720	0.1882	0.6638	0.1712	0.7753	0.1663	0.9106	0.2610
	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Age		-0.04745	0.007554	-0.04238	0.007894	-0.04349	0.007100	-0.04384	0.007304	-0.03708	0.01053
$Cardio$	0	-0.7940	0.2288	-0.8384	0.1923	-1.0249	0.2263	-0.8700	0.2267	-1.0244	0.3792
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$Reject$	1	-0.6121	0.1846	-0.6303	0.1942	-0.7932	0.1922	-0.9127	0.1745	-0.4825	0.2858
	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$Year$		-0.2198	0.06799	—		—		—		—	
$Age * Year$		0.002798	0.001135	—		—		—		—	
$Cardio * Year$	0	0.2116	0.06075	0.3083	0.06647	0.2836	0.06524	0.3616	0.07648	0.2657	0.09602
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$Cardio * Year^2$	0	-0.01530	0.005247	-0.03308	0.006418	-0.02406	0.007413	-0.03687	0.007783	-0.02714	0.009899
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
d_{11}		4.2678	0.2165	4.0084	0.2929	7.4351	0.6174	12.9995	2.1538	13.4571	1.9483
d_{22}		—		0.5566	0.07660	0.1592	0.02204	1.2060	0.2075	1.1229	0.2112
d_{33}		—		—		0.000540	0.000126	0.01605	0.002996	0.008422	0.001988
d_{12}		—		—		—		-2.0288	0.6482	-0.8589	0.5594
d_{13}		—		—		—		0.2748	0.07161	0.09427	0.05985
d_{23}		—		—		—		-0.1248	0.02360	-0.09167	0.01981
Log Likelihood		- 3528.29634		-3403.26861		-3340.0426		-3300.25054		-3274.07834	

In this table, the first column represents the model with an random intercept. The second column represents the model of a random effect of time, as well as year, with independent random effects. The next column denotes the model with three independent random effects (Intercept, year and year^2). The fourth column represents the model with three correlated random effects. We also indicate in the last column, the regression parameter estimates by use of non-adaptive Gaussian quadrature, 15 quadrature points. Since this result was nearly at the end of this report, we just present it in a comparative way, rather than altering our report and figures

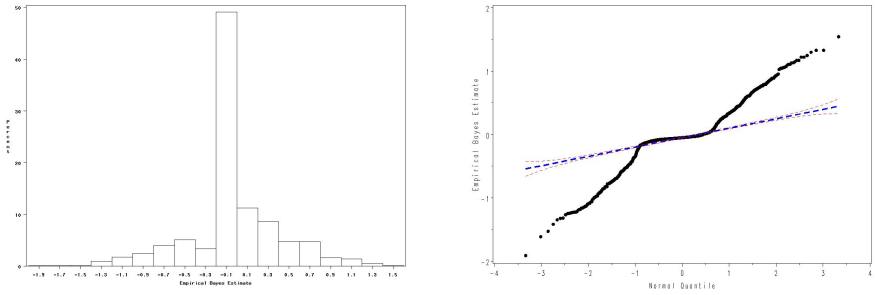
B Plots

B.1 Empirical Bayes Estimates

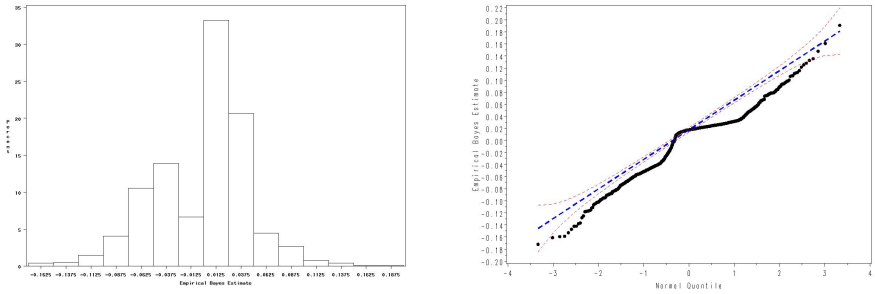
For the QQ plots we used a macro *nqplot*, written by Michael Friendly, which can be found on http://www.math.yorku.ca/SCS/sssg/nqplot.html



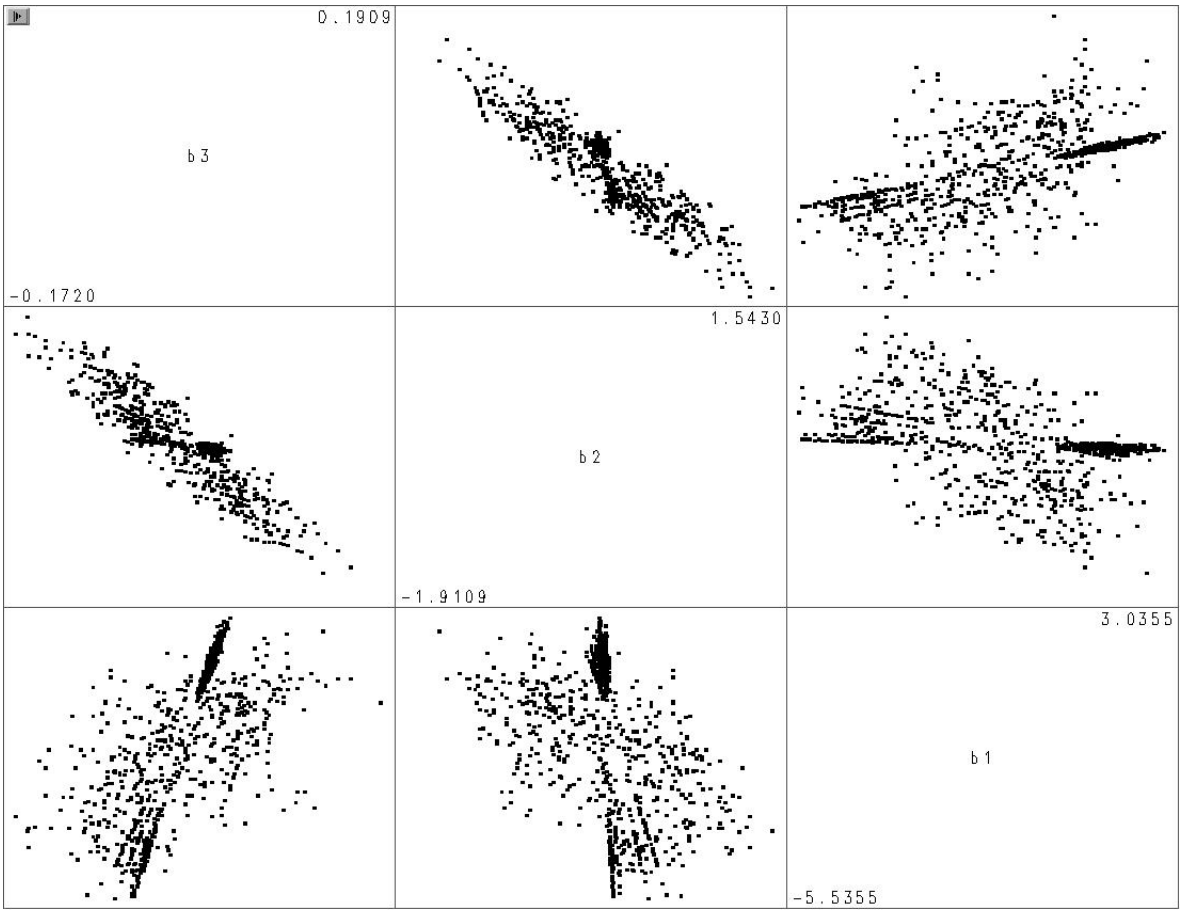
Histogram and QQ-plot of the intercept random effect



Histogram and QQ-plot of the year random effect



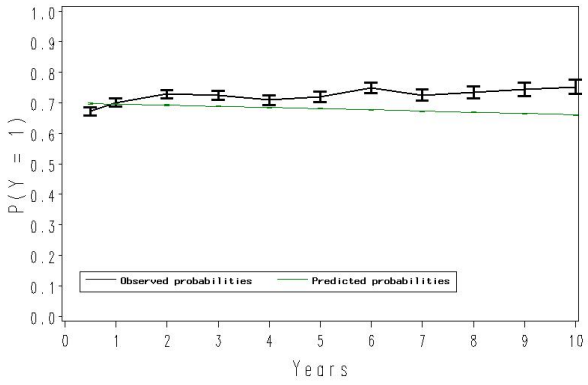
Histogram and QQ-plot of the year^2 random effect



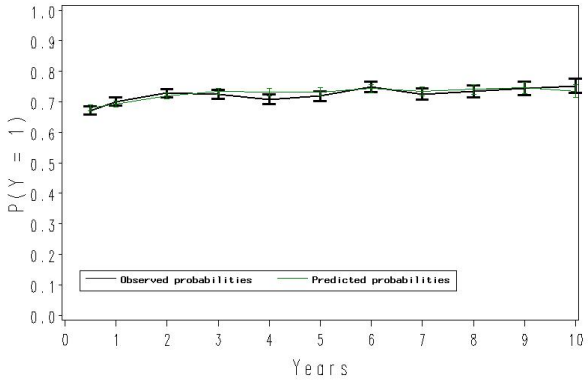
The scatterplot of the 3 random effect EB estimates

B.2 Observed Vs. Predicted

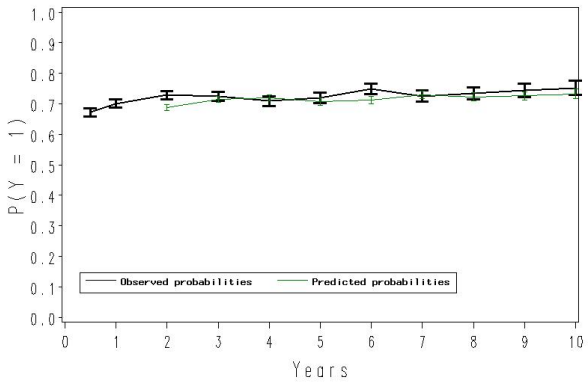
Marginal model: observed vs predicted



Hierarchical model: observed vs predicted



Transition model: observed vs predicted



C SAS code

```
libname Mit "C:\Documents and Settings\soutrik.banerjee\My Documents\Datasets";

/*Loading the Macros*/

%include "E:\My Documents\My SAS Files\macros\dropout.sas";
%include "E:\My Documents\My SAS Files\macros\GLIMMIX.sas";
%include "E:\My Documents\My SAS Files\macros\nqplot.sas";

/*transposing the horizontal data to vertical*/

data renalv;
set Mit.renal;
array zeit (11) HC06 HC1 HC2 HC3 HC4 HC5 HC6 HC7 HC8 HC9 HC10;
do j = 1 to 11;
HCdiff = zeit (j) - HC0;
year = j;
if j = 1 then year = 0.5;  if j = 2 then year = 1;  if j = 3 then year = 2;
if j = 4 then year = 3; if j = 5 then year = 4; if j = 6 then year = 5;
if j = 7 then year = 6; if j = 8 then year = 7; if j = 9 then year = 8;
if j = 10 then year = 9; if j = 11 then year = 10;
output;
end;
run;

data renal2;
set renalv;
yearclss = year;
year2 = year**2;
drop j HC0 HC06 HC1-HC10;
run;

/*sort the dataset*/

proc sort data = renal2;
by id;
run;

/*with cut-off (co) at hc = 3.5*/

data renalco;
set renal2;
if hcdiff = . then hcco = .;
else
if hcdiff > 3.5 then hcco = 1;
else hcco = 0;
run;

/*making a variable with the previous measurement*/

%dropout(data=renalco,id=id,time=year,response=hcco,out=previous1);
data previous1;
set previous1;
prev1 = prev;
drop prev;
run;

/*making a variable with the second previous measurement*/

%dropout(data=previous1,id=id,time=year,response=prev1,out=previous2);
data previous2;
set previous2;
prev2 = prev;
drop prev;
run;

/*transition model reduction*/
/*full model*/

proc genmod data = previous2 descending;
class id yearclss male cardio reject;
model hcco = male cardio age reject year year2
             male*year cardio*year age*year reject*year
 male*year2 cardio*year2 age*year2 reject*year2
```

```
            prev1 prev2 / link = logit dist = binomial type3;
repeated subject = id / withinsubject = yearclss type = un corrw modelse;
run;

/*final model*/

proc genmod data = previous2 descending;
class id yearclss male cardio reject;
model hcco = male prev1 prev2 / link = logit dist = binomial type3;
repeated subject = id / withinsubject = yearclss type = un corrw modelse;
output out = sortie predicted = pred_probabilities;
run;

/*GEE model building*/
/*full model*/

proc genmod data = renalco descending;
class id yearclss male cardio reject;
model hcco = male cardio age reject year year2
            male*year cardio*year age*year reject*year
 male*year2 cardio*year2 age*year2 reject*year2
            / link = logit dist = binomial type3;
repeated subject = id / withinsubject = yearclss type = un corrw modelse;
run;

/*final model*/

proc genmod data = renalco descending;
class id yearclss male cardio reject;
model hcco = male cardio age year age*year / link = logit dist = binomial
type3;
repeated subject = id / withinsubject = yearclss type = un corrw modelse;
output out = sortie2 predicted = pred_probabilities;
run;

/*graphical plotting of the transition model*/

proc gplot data = sortie;
plot hcco*year = 1 pred_probabilities*year = 2 / overlay haxis = axis1 vaxis = axis2 legend = legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside) offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor = none;
axis2 label =(h = 2 A = 90 'P(Y = 1)') value = (h = 1.5) order = (0 to 1 by 0.1) minor = none;
legend1 value = ('Observed probabilities' 'Predicted probabilities' h = 1.5) label = none frame position = (bottom left inside) offset = (3,3);
title h = 3 'Transition model: observed vs predicted';
run;quit;

/*graphical plotting of the marginal model*/

proc gplot data = sortie2;
plot hcco*year = 1 pred_probabilities*year = 2 / overlay haxis = axis1 vaxis = axis2 legend = legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside) offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor = none;
axis2 label =(h = 2 A = 90 'P(Y = 1)') value = (h = 1.5) order = (0 to 1 by 0.1) minor = none;
legend1 value = ('Observed probabilities' 'Predicted probabilities' h = 1.5) label = none frame position = (bottom left inside) offset = (3,3);
title h = 3 'Marginal model: observed vs predicted';
run;quit;

/*deleting all missing data  (only for NLMIXED!!!)*/

data renalNLM;
set renalco;
where hcco ^= . & age ^= . & reject ^= . & cardio ^= . & male ^= . & hcdiff
^= .;
drop agegroup;
run;

/*final model NLMIXED*/

proc nlmixed data = renalNLM noad qpoints = 15 maxiter = 300;
parms Int = 4.33 M = 0.8926 A = -0.05229 C = -0.9134 R = -0.3767
      CY = 0.2788 CY2= -0.0223
  d11 = 9.5 d22 = 1.05 d33 = 0.014 d12 = -1.02 d13 = 0.17 d23 = -0.096;
```

```
teta = Int + M*male + A*age + C*(1-cardio) + R*reject + b1 +
   CY*(1-cardio)*year + b2*year + CY2*(1-cardio)*year2 + b3*year2;
expteta = exp(teta);
p = expteta / (1 + expteta);
model hcco ~ binary(p);
random b1 b2 b3 ~ normal([0,0,0],[d11,d12,d22,d13,d23,d33]) subject = id out = eb;
predict p out = sortie3;
run;

/*graphical plotting of the hierarchical model*/

proc gplot data = sortie3;
plot hcco*year = 1 pred*year = 2 / overlay haxis = axis1 vaxis = axis2 legend = legend1;
legend1 value =(height = 1.5) label = none frame position = (bottom left inside) offset = (3,3);
symbol1 c = black v = none i = std1mjt w = 3 mode = include;
symbol2 c = green v = none i = std1mjt w = 1 mode = include;
axis1 label =(h = 2 'Years') value = (h = 1.5) order = (0 to 10 by 1) minor = none;
axis2 label =(h = 2 A = 90 'P(Y = 1)') value = (h = 1.5) order = (0 to 1 by 0.1) minor = none;
legend1 value = ('Observed probabilities' 'Predicted probabilities' h = 1.5) label = none frame position = (bottom left inside) offset = (3,3);
title h = 3 'Hierarchical model: observed vs predicted';
run;quit;

\*The QQ-plots, and the Histograms for the EB Estimates*\

\*3 different dataset, one for every R.E.*\

data Mit.EBEstInt;
set eb;
where Effect="b1";
run;

data Mit.EBEstY;
set eb;
where Effect="b2";
run;

data Mit.EBEstY2;
set eb;
where Effect="b3";
run;

/*3 QQ plots*/

%nqplot(data=Mit.EBEstInt, var=Estimate);
%nqplot(data=Mit.EBEstY, var=Estimate);
%nqplot(data=Mit.EBEstY2, var=Estimate);

/*3 histograms*/

proc univariate data=Mit.EBEstInt;
var Estimate;
histogram Estimate;
run;

proc univariate data=Mit.EBEstY;
var Estimate;
histogram Estimate;
run;

proc univariate data=Mit.EBEstY2;
var Estimate;
histogram Estimate;
run;

\*The Database for the sqatter plot. The plot itself is made by SAS/INSIGHT *\

data Mit.EB (keep=id b1 b2 b3);
set Mit.EBEstInt;
id= id;
b1= estimate;
set Mit.EBEstY;
b2= estimate;
set Mit.EBEstY2;
b3= estimate;
run;
```