PhUSE 2017

Paper AS03

How to fit longitudinal data with proc MCMC?

Hennion Maud, Arlenda, Mont-Saint-Guibert, Belgium
Rozet Eric, Arlenda, Mont-Saint-Guibert, Belgium

ABSTRACT

Longitudinal data are widely used in clinical trials and observational studies to allow an evaluation of within-sample change over time and measurement of the duration of events. While responses between subjects are independent, correlation of the measures within patient need to be modeled. As the structure of the variance-covariance matrix fully explains the correlation within repeated measures of a patient, the choice of this structure is crucial.
Popular procedures as MIXED and NLMIXED in SAS/STAT software allow to fit longitudinal model. In this presentation, we will remind what longitudinal data are and why ordinary statistical methodologies cannot be used to model the structure of the within-subject covariance. Then, a general Bayesian approach will be presented with the MCMC procedure to fit longitudinal data for any kind of correlation structure. A concrete example will be presented in order to show how this approach can be used.

INTRODUCTION

Longitudinal data are widely encountered in health science (clinical development, observational studies). They play a key role in epidemiology, clinical research and therapeutic evaluation. Due to continuous monitoring, patients included in a clinical trial are followed throughout duration of the study and repeated measures of different endpoints are recorded (treatment effects, biomarker…). The primary advantage of longitudinal data is the evaluation of within-sample change over time and the measurement of the duration of events. A common example of longitudinal study is a randomized trial with two groups, a treatment group and a control group. By measuring an output at several time points for subjects belonging to one of the treatment groups, a longitudinal model can be used to characterize the evolution against time of both groups.

In the following sections, a study will be said longitudinal when a response is observed on each subject/unit repeatedly over time. Reflecting this definition, longitudinal models are also called repeated models. Although continuous responses are more common, the responses do not need to be continuous but may be binomial or multinomial. Time interval between measures of a subject should be constant across the study but extension of the model exists for variable time intervals.

In case of continuous monitoring, the responses between subjects are independent while the repeated measures of one subject are not. Due to the correlation, usual statistical tools as Ordinary Least Square (OLS) model cannot be used because they assume independent measurements and are therefore not appropriate to fit those data.

A common way to model correlation between data of a patient is to use a multivariate regression model with a structure on the variance-covariance of the residuals. The structure of the variance-covariance matrix explains all the correlation within repeated measures of a patient. Different kind of structure can be used: unstructured matrix, autoregressive of order 1, compound symmetry. The autoregressive of order 1 structure, called AR[1], is widely used as it gives a good representation of the measures dependence over time with a limited number of parameters. The AR[1] means that values closers in time tends to fluctuate similarly, i.e. the closer in time, the higher the correlation is.

While longitudinal data are widely used in clinical trials or observational studies, they are not always analyzed in the right way to capture all information. In the following sections, we will show how to analyze efficiently repeated measures with the PROC MCMC in SAS/STAT.

NOTATION

A linear mixed model is written as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}$$

$$\boldsymbol{b} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$$
$$\boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R})$$

where $\boldsymbol{\beta}$ is the vector of fixed effects and $\boldsymbol{b}$ the vector of random effects. The matrix $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrix and the matrix $\boldsymbol{G}$ and $\boldsymbol{R}$ are the variance-covariance matrix of the random effects and residuals respectively.

As explained before, in case of longitudinal data, the correlation within subjects is explained by the structure of the variance-covariance matrix R. If a mixed model with repeated measures (MMRM) is used, the complete structure of variance-covariance can be written as

$$\boldsymbol{Var(Y) = ZGZ' + R}$$

where the ZGZ' part express the between-subject correlation and the R part express the within-subject correlation.

Here are the specifications of the variance-covariance matrix R for different structure of the within-subject variance. The matrix is shown for an example of 4 repeated measures by subject.

- Unspecified (10 parameters to estimate)

$$\begin{pmatrix} \sigma_1^2 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & \sigma_2^2 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & \sigma_3^2 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & \sigma_4^2 \end{pmatrix}$$

- Autoregressive of order 1 (2 parameters to estimate)

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \rho^3\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^3\sigma^2 & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

- Compound Symmetry (2 parameters to estimate)

$$\begin{pmatrix} \phi+\sigma & \sigma & \sigma & \sigma \\ \sigma & \phi+\sigma & \sigma & \sigma \\ \sigma & \sigma & \phi+\sigma & \sigma \\ \sigma & \sigma & \sigma & \phi+\sigma \end{pmatrix}$$

LONGITUDINAL MODEL WITH PROC MIXED

A popular procedure in SAS/STAT software to model longitudinal data is the PROC MIXED. To illustrate the use of proc MIXED for longitudinal data, the orthodontic growth data (Pinheiro and Bates, 2000) will be used. The dataset contains the growth measurements of 27 children for 11 girls and 16 boys. Every two years, the distance between pituitary and the pterygomaxillary fissure was recorded at ages 8, 10, 12 and 14.

A linear growth curve model will be fitted on the distance response with the gender as categorical variable (0/1). To incorporate the correlation of the four measures (8-10-12-14) from the same subject, a structure of variance within subject is added to the model. The data are assumed to be Gaussian. The PROC MIXED to fit a growth curve model with a structure on the variance-covariance is shown on the following code:

```
proc mixed data=Orthodont;
       class Subject Sex(ref="0") Age(ref="8");
       model Dist = Sex Age Age*Sex / solution;
       repeated Age / Subject=Subject type= AR(1) R;
run;
```

where the repeated statement specifies the R covariance matrix, which model dependency between observations. In this case, the AR[1] structure for the variance-covariance is requested (`type=AR(1)`). The R matrix is block diagonal with 27 blocks, one by subject, where each block consisting of identical 4x4 autoregressive matrix (see Table 3).

Table 1 PROC MIXED Parameter estimates

Solution for Fixed Effects

Effect	Sex	age	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			22.8750	0.5681	25	40.26	<.0001
Sex	1		-1.6932	0.8901	25	-1.90	0.0687
Sex	0		0	.	.	.	.
age		10	0.9375	0.7170	75	1.31	0.1950
age		12	2.8437	0.6497	75	4.38	<.0001
age		14	4.5937	0.5156	75	8.91	<.0001
age		8	0	.	.	.	.
Sex*age	1	10	0.1080	1.1234	75	0.10	0.9237
Sex*age	1	12	-0.9347	1.0179	75	-0.92	0.3615
Sex*age	1	14	-1.6847	0.8077	75	-2.09	0.0404
Sex*age	1	8	0	.	.	.	.
Sex*age	0	10	0	.	.	.	.
Sex*age	0	12	0	.	.	.	.
Sex*age	0	14	0	.	.	.	.
Sex*age	0	8	0	.	.	.	.

Table 2 PROC MIXED Parameters of Sample Variance

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
AR(1)	Subject	0.5882
Residual		5.1641

Table 3 Subject's block of the block diagonal matrix R

Estimated R Matrix for Subject 1

Row	Col1	Col2	Col3	Col4
1	5.1641	1.0512	1.7870	3.0378
2	1.0512	5.1641	3.0378	1.7870
3	1.7870	3.0378	5.1641	3.0378
4	3.0378	1.7870	3.0378	5.1641

LONGITUDINAL MODEL WITH PROC MCMC

In a classical approach, longitudinal models are analyzed in a frequentist way where fixed-effect parameters are considered fixed with an unknown mean and the random effects are considered as unobserved latent variables. In a frequentist approach, parameters of the model are estimated by maximizing the likelihood and are summarized by a point estimate and a standard deviation. Moreover, the asymptotic normality is assumed most of the time in inference.

In a Bayesian approach, all the unknown quantities (fixed effects and random effects) in a statistical model are considered as random variable. Therefore, a complete distribution is available for any parameter of interest. Bayesian statistics are based on the Bayesian paradigm which combine different sources of information such as the prior information and the likelihood (the data from the study as in frequentist). Bayesian solutions are recommended for data with complex structure as there is no modelling limitations.

PROC MCMC STRUCTURE

The structure of the PROC MCMC is similar to the PROC NLMIXED structure. A PROC MCMC code can be divided into four blocks:

- **Parameters block**: In the first block, all the parameters of the model have to be listed with their optional initial values. Multiple PARMS statements are allowed and if several parameters are listed in the same PARMS statement, the Metropolis algorithm will update the list of parameters simultaneously.
  If a vector of parameters is used, an array has to be defined before with the ARRAY statement.
- **Prior block**: Prior distributions have to be defined for each model parameter defined in the parameters block with the PRIOR statement. A PRIOR statement is composed of a single (or a list of) parameter(s) followed by a tilde (~) and the prior distribution with its parameters. Multiple PRIOR statements are allowed.
- **Program Statements**: As in PROC NLMIXED, programming statements can be added to define new parameters or for computation of the likelihood
- **Model Statement**: The MODEL statement is used to specify the conditional distribution of the data given the parameters (i.e. the likelihood). A single (or a list of) dependent variable has to be defined followed by a tilde (~) and the distribution with the parameters.

By default, PROC MCMC assumes that observations of the input data set are independent. Then, the joint log-likelihood is evaluated as the sum of individual log-likelihood functions and the log-likelihood is specified in the MODEL statement. Longitudinal model can be fitted in PROC MCMC as long as the correlation between measures of a subject are taken into account. In the following sections, two options to fit longitudinal models will be presented. The first one will be shown based on Orthodont data. The second one is a more general approach and will be developed based on a more complex example.

OPTION 1: ORTHODONT DATA

In order to illustrate the use of PROC MCMC for longitudinal data, the example on the Orthodontic growth data shown before will be applied to the PROC MCMC. To fit a repeated measurement model with PROC MIXED, the dataset needs to be rolled out i.e. each row is an observation. On the contrary, as PROC MCMC assumes, by default, that all observations are independent, the input data set should store all observations from one subject in one row in order to model the within-subject covariance structure. Therefore, a PROC TRANSPOSE should be applied to the input data set before fitting the model.

```
proc transpose data=Orthodont out=Orthodont_tr prefix=Dist;
by Subject Sex;
var dist;

RUN;
```

The following SAS code fit a Bayesian longitudinal model on the Orthodontic growth data with a AR[1] structure for the variance-covariance matrix. The distribution used in the model statement is the mvnar distribution which correspond to a multivariate normal distribution with an AR[1] structure where *S2* is the variance and *Rho* the correlation.

```
proc mcmc data=Orthodont_tr outpost=outpost nbi=5000 nmc=25000 thin=15 plots=ALL;
array mean[4];
array a_age[3]      a_age10 a_age12 a_age14;
array a_sex_age[3]  a_sex_age10  a_sex_age12  a_sex_age14;
array dist[4]       dist1 dist2 dist3 dist4;
```

```
parms intercept 0.1
      a_Sex 0.1
      a_age: 0
      a_sex_age: 0;
parms S2 0.0011;
parms Rho;

prior intercept   ~ normal(0,var=1e6);
prior a_Sex       ~ normal(0,var=1e6);
prior a_age:      ~ normal(0,var=1e6);
prior  a_sex_age: ~ normal(0,var=1e6);
prior S2          ~ igamma(shape=0.01, scale=0.01);
prior Rho         ~ uniform(-1,1);

mean[1] = intercept + a_Sex * (Sex=1)                                    ;/*time = 8*/
mean[2] = intercept + a_Sex * (Sex=1)+ a_age10 +  a_sex_age10 * (Sex=1);/*time = 10*/
mean[3] = intercept + a_Sex * (Sex=1)+ a_age12 +  a_sex_age12 * (Sex=1);/*time = 12*/
mean[4] = intercept + a_Sex * (Sex=1)+ a_age14 +  a_sex_age14 * (Sex=1);/*time = 14*/

model dist ~ mvnar(mean, var=S2, rho=Rho);

run;
```

To check that the above model fit the same model than with PROC MIXED, the output from the two procedures will be compared. From the PROC MIXED, the 'Solution For Fixed Effects' table and the 'Covariance Parameter Estimates' table (see Table 1 and Table 2) is compared to the 'Posterior Summaries and Intervals' table from the PROC MCMC (see Table 4).

As can be seen, the point estimate of the fixed effects and covariance parameters obtained with PROC MIXED are similar to the mean of the posterior distributions of the model parameters from PROC MCMC.

Table 4 PROC MCMC Posterior distributions

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
intercept	1667	22.9087	0.5805	21.8206	24.1410
a_Sex	1667	-1.7067	0.9156	-3.4139	0.1431
a_age10	1667	0.9110	0.5405	-0.2294	1.8702
a_age12	1667	2.8195	0.6663	1.4767	4.1156
a_age14	1667	4.5687	0.7312	3.1774	6.0638
a_sex_age10	1667	0.1243	0.8306	-1.4092	1.8347
a_sex_age12	1667	-0.9233	1.0679	-2.9652	1.2060
a_sex_age14	1667	-1.6841	1.1582	-3.7653	0.7570
S2	1667	5.4673	1.0603	3.5980	7.5686
Rho	1667	0.6042	0.0833	0.4419	0.7690

DIFFERENT COVARIANCE STRUCTURES

In previous example, the structure of the within-subject covariance was fixed to an auto regressive of order 1 because the appropriate distribution is available in PROC MCMC. Of course, the same longitudinal model could be fitted on the Orthodont data with different structure for the within-subject covariance. In the following code, we propose to fit the same model with a compound symmetry structure as no built-in distributions are provided by PROC MCMC. The

Compound Symmetry covariance matrix has two parameters, σ and ϕ, which are specified as below, in case of four measures:

$$R = \begin{pmatrix} \phi+\sigma & \sigma & \sigma & \sigma \\ \sigma & \phi+\sigma & \sigma & \sigma \\ \sigma & \sigma & \phi+\sigma & \sigma \\ \sigma & \sigma & \sigma & \phi+\sigma \end{pmatrix}.$$

As no built-in distribution is provided, the mvn distribution will be used and the variance-covariance matrix will be specified with the program statements. The elements of the matrix will be filled according to the covariance specification. The BEGINNODATA and ENDNODATA statements are used to fill the covariance matrix because statements between those statements are computed once.

```
beginnodata;
do i = 1 to 4;
       do j = 1 to 4;
              Cov[i,j] = sigma + phi * (i=j);
       end;
end;
endnodata;
```

The following code will be used to fit the model on the Orthodontic growth data:

```
proc mcmc data=Orthodont_tr outpost=outpost nbi=10000 nmc=50000 thin=15 plots=ALL;
array mean[4];
array a_age[3] a_age10 a_age12 a_age14;
array a_sex_age[3]  a_sex_age10  a_sex_age12  a_sex_age14;
array dist[4] dist1 dist2 dist3 dist4;
array Cov[4,4];

parms intercept 0.1
        a_Sex 0.1
        a_age: 0
        a_sex_age: 0;
parms sigma 0.1;
parms phi 0.1;

prior intercept  ~ normal(0,var=1e6);
prior a_Sex ~ normal(0,var=1e6);
prior a_age:  ~ normal(0,var=1e6);
prior a_sex_age: ~ normal(0,var=1e6);
prior sigma ~ general(0,lower=0);
prior phi ~ general(0,lower=0);

beginnodata;
do i = 1 to 4;
       do j = 1 to 4;
              Cov[i,j] = sigma + phi * (i=j);
       end;
end;
endnodata;


mean[1] = intercept + a_Sex * (Sex=1)                                    ;/*time = 8*/
mean[2] = intercept + a_Sex * (Sex=1)+ a_age10 +  a_sex_age10 * (Sex=1);/*time = 10*/
mean[3] = intercept + a_Sex * (Sex=1)+ a_age12 +  a_sex_age12 * (Sex=1);/*time = 12*/
mean[4] = intercept + a_Sex * (Sex=1)+ a_age14 +  a_sex_age14 * (Sex=1);/*time = 14*/

model dist ~ mvn(mean, Cov) ;

run;
```

OPTION 2: A GLOBAL METHODOLOGY

In this section, a general methodology is developed to fit longitudinal model with any kind of within-subject covariance structure. To illustrate the methodology, a concrete example from the pharmaceutical industry will be used.

In this example, three treatments are tested on 35 subjects. For each subject, three measures are taken during the follow-up for the three treatments. Thus, each subject has 9 measures divided in three blocks of three repeated measures. A log transformation is applied on the response and the log response is assumed to be Gaussian. Figure 1 shows the data after log transformation by treatment group and across time. The first few records of the data set are shown on Table 5.

Table 5 The first few records of the data set (dataa)

SUBJID	Time	Time_	Treatment	Log_data	Log_baseline
1	7	A	Treat_1	0.7433525738	-2.701838946
1	10	B	Treat_1	1.7931685928	-2.701838946
1	13	C	Treat_1	1.8373370166	-2.701838946
1	7	A	Treat_2	1.207974746	-1.287853505
1	10	B	Treat_2	1.7665760577	-1.287853505
1	13	C	Treat_2	1.5722355967	-1.287853505
1	7	A	Treat_3	1.1845374172	-1.227868137
1	10	B	Treat_3	1.6798559863	-1.227868137
1	13	C	Treat_3	1.8015436668	-1.227868137
2	7	A	Treat_2	0.4935423239	-1.065999397
2	10	B	Treat_2	1.8086245895	-1.065999397
2	13	C	Treat_2	1.6359807559	-1.065999397
2	7	A	Treat_1	0.8491363131	-0.527489125
2	10	B	Treat_1	1.7202170603	-0.527489125
2	13	C	Treat_1	1.7899980368	-0.527489125
2	7	A	Treat_3	0.8854636559	-1.062219455
2	10	B	Treat_3	1.7287352946	-1.062219455
2	13	C	Treat_3	1.7995826685	-1.062219455
3	7	A	Treat_3	1.0618368576	-0.535595847
3	10	B	Treat_3	1.9057520419	-0.535595847
3	13	C	Treat_3	1.6867314905	-0.535595847
3	7	A	Treat_2	0.8548678168	-0.455030091
3	10	B	Treat_2	1.7791205078	-0.455030091
3	13	C	Treat_2	1.6033793542	-0.455030091
3	7	A	Treat_1	0.772940122	-0.550858551
3	10	B	Treat_1	1.6940256749	-0.550858551
3	13	C	Treat_1	1.7432377954	-0.550858551

A longitudinal mixed model is applied on the data and adjusted for baseline value (log transformed). The treatment, the time and the treatment-time interaction will also be evaluated in the model. The within-subject covariance will be taken into account in the model. As the three treatments are tested on each patient, the repeated measures of one treatment are correlated and a AR[1] structure is assumed. It is assumed that, within-subject, there is no correlation between treatment. A random effect on the subject is added. The model and the variance-covariance matrix for a subject are as follows

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\,\boldsymbol{b_i} + \boldsymbol{\epsilon_i}$$

(Equation 1)

where $\boldsymbol{Y_i}$ is the vector of response for patient $\boldsymbol{i}$, $\boldsymbol{\beta}$ is the vector of the fixed-effects, $\boldsymbol{X_i}$and $\boldsymbol{Z_i}$ are the design matrix for patient i, $\boldsymbol{b_i}$ is the patient random effect ($\boldsymbol{b_i} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_g^2})$) and $\boldsymbol{\epsilon_i}$ are the residuals ($\boldsymbol{\epsilon_i} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R_i})$). As explained before, the variance of $\boldsymbol{Y_i}$ is a combination of the between-subject variance and the matrix $\boldsymbol{R}$:

$$\boldsymbol{V}(\boldsymbol{Y_i}) = \begin{pmatrix} \boldsymbol{1} \\ \vdots \\ \boldsymbol{1} \end{pmatrix} \boldsymbol{\sigma_g^2} (\boldsymbol{1} \quad \ldots \quad \boldsymbol{1}) + \boldsymbol{R_i}$$

$$= \begin{pmatrix} \sigma_g^2 & \sigma_g^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 & \sigma_g^2 \end{pmatrix} + \begin{pmatrix} R_{i1} & 0 & 0 \\ 0 & R_{i2} & 0 \\ 0 & 0 & R_{i3} \end{pmatrix}$$

(Equation 2)

where each element is 3x3 matrix and R_{ij} is a AR[1] matrix.

The model can be fitted in a frequentist way with a PROC MIXED as shown below. The results are displayed in Table 6 (fixed effects estimates and standard errors) and in Table 7 (variance estimate) for the longitudinal mixed model.

```
proc mixed data=dataa;
class time_(ref="C") treatment subjid;
model log_data = treatment time_ treatment*time_ log_baseline/ solution;
random int /subject=subjid;
repeated time_ / subject=treatment(subjid) type=AR(1) R Rcorr;
run;
```

By default, the PROC MCMC assumes that the input observations are independent. The joint log-likelihood is hence the sum of the individual log-likelihood. In case of dependence between observations, this method produces an incorrect log-likelihood.

In case of longitudinal data, the repeated measures of a subject are correlated. Then, all measures of a subject should be stored in one row and a multivariate distribution should be used. If needed, the variance-covariance matrix can be also specified as shown previously.

A second option is to create the design matrix X and to store all relevant variables (responses and model covariates) in arrays. In this way, the joint log-likelihood can be constructed for the entire data set and take into account correlation between observations. As before, any structure of the variance-covariance matrix can be computed and a multivariate distribution will be used. This methodology will be explained in more details in the following sections. Firstly, it will be shown how to construct easily the design matrix X. Then, the longitudinal Bayesian model with PROC MCMC will be presented. The SAS code is shown in the Appendix

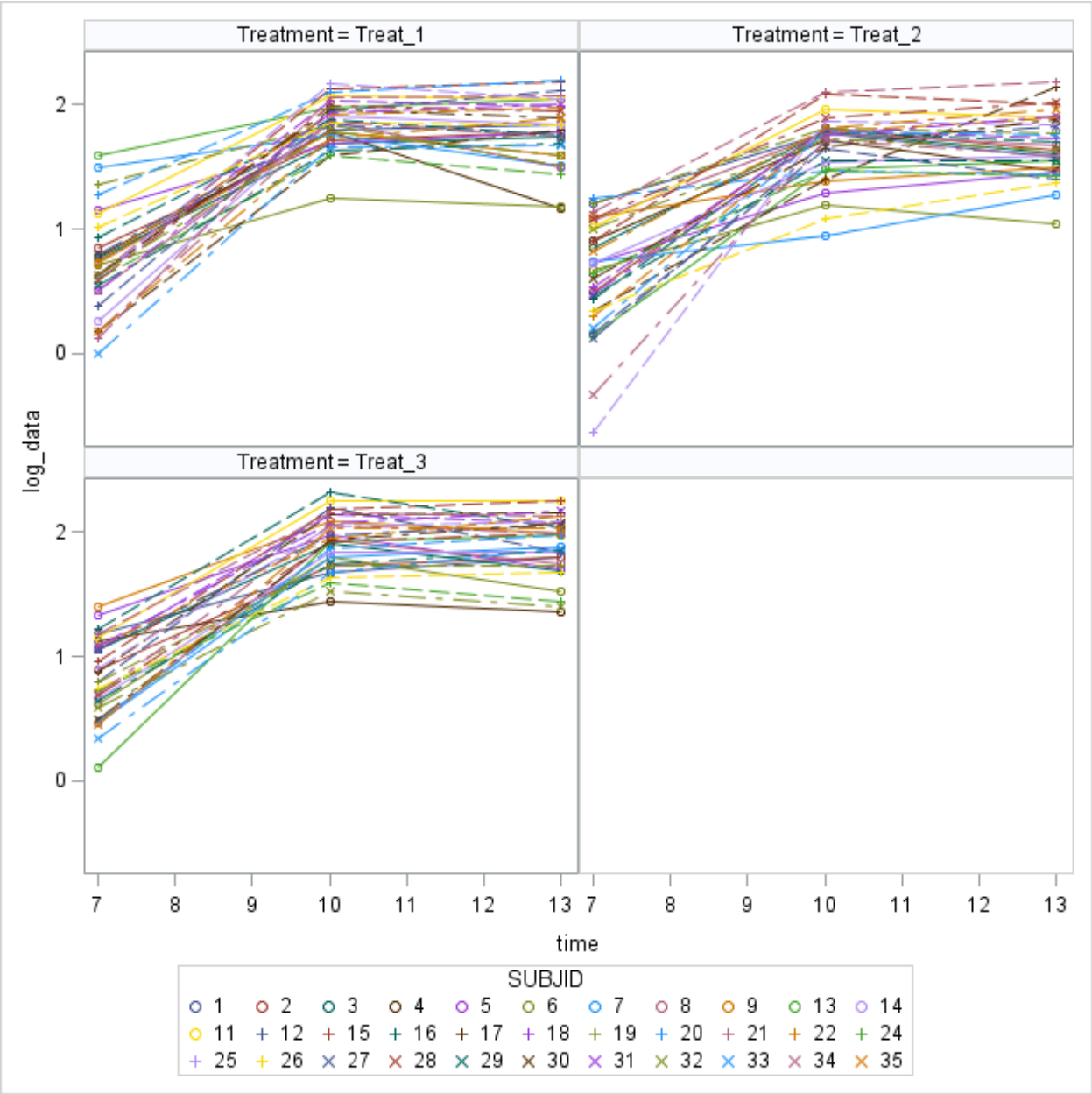

Figure 1 Plot of the logarithm of the response across time. Data are presented by treatment for all patients

Table 6 PROC MIXED Parameter Estimates

Solution for Fixed Effects

Effect	time_	Treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			1.9134	0.06096	32	31.39	<.0001
Treatment		Treat_1	-0.1022	0.07029	255	-1.45	0.1473
Treatment		Treat_2	-0.2205	0.07027	255	-3.14	0.0019
Treatment		Treat_3	0	.	.	.	.
time_	A		-1.0657	0.06624	255	-16.09	<.0001
time_	B		0.02408	0.05744	255	0.42	0.6754
time_	C		0	.	.	.	.
time_*Treatment	A	Treat_1	-0.02286	0.09368	255	-0.24	0.8074
time_*Treatment	A	Treat_2	0.01933	0.09368	255	0.21	0.8367
time_*Treatment	A	Treat_3	0	.	.	.	.
time_*Treatment	B	Treat_1	0.02572	0.08123	255	0.32	0.7518
time_*Treatment	B	Treat_2	-0.04690	0.08123	255	-0.58	0.5642
time_*Treatment	B	Treat_3	0	.	.	.	.
time_*Treatment	C	Treat_1	0	.	.	.	.
time_*Treatment	C	Treat_2	0	.	.	.	.
time_*Treatment	C	Treat_3	0	.	.	.	.
log_baseline			0.02316	0.03607	255	0.64	0.5214

Table 7 PROC MIXED Parameters of Sample Variance

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	SUBJID	0.007631
AR(1)	Treatment(SUBJID)	0.3300
Residual		0.08125

STEP 1: MODEL MATRIX

Before fitting the longitudinal model on the data (shown on Table 5), the model matrix should be computed in order to simplify the writing of the model in PROC MCMC. For this purpose, PROC TRANSREG allows to code the variable for the design matrix. The model statement creates new dummy variables (1/0) where the reference level is the last level of the categorical variable. The variable listed in the ID statement will be copied in the output data set called model_matrix_X.

```
proc transreg data=dataa design;
       model class(treatment treatment*time time / zero=last);
       id subjid log_data log_baseline;
       output out=model_matrix_X(drop=time _TYPE_ _NAME_);
RUN;
```

Then the data set covariates will contain all the variables evaluated in the model (intercept, log baseline, treatment, time and treatment-time) with dummy variables for the categorical variables. The data set covariates is the model matrix X in equation 1.

```
data covariates (keep = intercept &_trgind log_baseline);
   set model_matrix_X;
RUN;
```

Finally, two new variables are created (last and first) to enclose the measures from one patient (the three measures for the three treatments). The total number of measures is recorded in the macro variable nrec and the number of measures by patient is recorded in the macro variable nrep.

```
data recodedsplit1(drop=counter intercept &_trgind);
  set model_matrix_X nobs=_n;
  call symputx('nrec', _n);
  by subjid;
  if first.subjid then counter=0;
  counter+1;
  first = first.subjid;
  last = last.subjid;
  if _n_ = _n then
  do;
       call symputx('nrep',counter);
  end;
run;
```

STEP 2: PROC MCMC

As the design matrix was created, it is not needed to import a data set with the response and the variables in the PROC MCMC statement. Instead, response and variables will be stored in arrays thanks to the BEGINCNST and ENDCNST statements. Those statements define a block within which PROC MCMC processes the programming statements only during the setup stage of the simulation. The read_array function store variables (ex: log_data) from a dataset (ex: recodedsplit1) into an array (ex: data). Those arrays have to be defined before with a length of 1 and the option nosymbols because they must be dynamic. The read_array function resizes dynamically the array to match with the dimension of the input data set.

```
array covar[1] /nosymbols ;
array data[1]/nosymbols;
array first1[1]/nosymbols;
array last1[1]/nosymbols;

begincnst;
 rc = read_array("recodedsplit1",data,"log_data");
 rc = read_array("recodedsplit1",first1,"first");
 rc = read_array("recodedsplit1",last1,"last");
 rc = read_array("covariates",covar);
endcnst;
```

As for the Compound Symmetry example, the variance-covariance matrix will be specified between the beginnodata and endnodata statements. First, the matrix VCV is filled up with the variance parameter sig2g which is the between-subject variance.

```
call Fillmatrix(VCV,sig2g);
```

Then, the matrix is filled up to be a block diagonal matrix with a AR[1] structure in the blocks as specified in Equation 2.

```
do k= 0 to 2;
       do i = 1 to 3;
              do j= 1 to 3;
                     VCV[i+(k*3), j+(k*3)] = sig2e * rho**abs(j-i) + sig2g;
              end;
       end;
end;
```

Non-informative priors are defined for the model parameters and variance parameters in the same block.

```
prior beta: ~ normal(0,var=1e6);
prior sig2g ~ general(0,lower=0);
prior sig2e ~ general(0,lower=0);
prior rho   ~ general(0,lower=0,upper=1);
```

Finally, the fixed part of the model ($X\beta$) is computed with the `call mult(covar,beta,mu);` statement.

In the last part of the PROC MCMC, the model will be specified by incrementing of the log-likelihood patient by patient. The JOINTMODEL options has to be added in the proc mcmc statement to indicate that the function used in the model statement calculates the log-likelihood for the entire data set rather than just for one observation.
The variables 'first' and 'last' allow to limit the repeated measures of a patient (3 measures for the 3 treatments). The repeated measures of the three treatments for a patient follow a multivariate normal distribution with the variance-covariance matrix VCV specified before. The log-likelihood is incremented patient by patient.

```
ljointpdf = 0;

 do irec = 1 to &nrec;
       if (first1[irec] = 1) then counter=0;
          counter = counter + 1;
          ytemp[counter] = data[irec];
          mutemp[counter] = mu[irec];
       if (last1[irec] = 1) then  ljointpdf = ljointpdf + lpdfmvn(ytemp, mutemp, VCV);
  end;


  model general(ljointpdf);
```

The results of the PROC MCMC are shown on Table 8

Table 8. The mean of the posterior distribution of the beta parameters (Table 8) correspond to the estimate of the fixed effects from the PROC MIXED (Table 6). The mean of the posterior distributions of the variance parameters, sig2e, rho and sig2g (Table 8), is equal to the estimate of the variance parameters (Table 7).

Table 8 PROC MCMC Posterior distributions

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
sig2e	4000	0.0830	0.00880	0.0671	0.1017
rho	4000	0.3230	0.0839	0.1558	0.4808
sig2g	4000	0.0107	0.00678	0.000074	0.0233
beta1	4000	1.9149	0.0623	1.7933	2.0357
beta2	4000	-0.1042	0.0702	-0.2465	0.0287
beta3	4000	-0.2194	0.0709	-0.3710	-0.0865
beta4	4000	-1.0649	0.0669	-1.1989	-0.9351
beta5	4000	0.0247	0.0572	-0.0818	0.1399
beta6	4000	-0.0239	0.0930	-0.2084	0.1537
beta7	4000	0.0166	0.0951	-0.1720	0.2013
beta8	4000	0.0255	0.0816	-0.1315	0.1909
beta9	4000	-0.0464	0.0814	-0.2185	0.1067
beta10	4000	0.0259	0.0364	-0.0441	0.0985

There are many advantages to use this methodology to fit longitudinal model in a Bayesian way. Firstly, as the complete variance matrix is specified, any structure can be used to fit to the data. In this case, the same variance parameters are used in the AR[1] block but it would be possible to specify different variances depending on the treatment.
Moreover, by writing explicitly the variance matrix, prior distributions have to be defined on variance and correlation parameters and not on a variance matrix. The usual prior distribution for variance matrix is the inverse Wishart. This distribution is not easy to define and less intuitive than inverse gamma distribution which are used for variance parameters.
Finally, as the design matrix is defined and used in the proc MCMC, it is not necessary to write explicitly the equation of the model which can be complex in some cases.

CONCLUSION

Longitudinal data are widely used in clinical development to evaluate a within-sample change over time and the measurement of the duration of events. To make a valid inference, the special features of longitudinal has to be taken into account as the correlation between the repeated measures of a subject. In a frequentist approach, popular procedures as PROC MIXED and PROC NLMIXED are used to fit longitudinal model with the appropriate structure of the variance-covariance matrix.

Two options to use PROC MCMC to do a Bayesian analysis of longitudinal data exists. In the first one, all data of a subjects should be store in one row and a multivariate distribution should be used. If needed, the variance-covariance matrix can be also specified. The second is a general Bayesian approach to fit longitudinal data with any structure of within-subject covariance. The methodology was illustrated by a concrete case where repeated measures of 3 treatments are recorded on every subject. In this approach, a model matrix dataset is needed which allow to not write explicitly the model.

REFERENCES

Pinheiro, J.C. and Bates, D. (2000). Mixed-Effects Models in S and S-PLUS. Springer, New York.

Robert J. Tempelman (2012). Workshop Materials for KSU Conference on Applied Statistics in Agriculture, Michigan State University

Fang Chen, Gordon Brown and Maura Strokes, SAS Institute Inc. Fitting your favorite mixed models with PROC MCMC. (Paper SAS5601-2016)

APPENDIX

PROC MCMC CODE

```
proc transreg data=dataa design;
       model class(treatment time time*treatment / zero=last);
       id subjid log_data log_baseline;
       output out=model_matrix_X(drop=time _TYPE_ _NAME_);
run;

Proc Sort data=model_matrix_X;
by subjid;
Run;

data covariates (keep = intercept &_trgind log_baseline);
   set model_matrix_X;
run;

data recodedsplit1(drop=counter intercept &_trgind);
  set model_matrix_X nobs=_n;
  call symputx('nrec', _n);
  by subjid;
  if first.subjid then counter=0;
  counter+1;
  first = first.subjid;
  last = last.subjid;
  if _n_ = _n then
  do;
       call symputx('nrep',counter);
  end;
run;

%put &nrep;
%put &nrec;
%let nvar=10;


data a;
run;


ODS Graphics on;
proc mcmc data=a outpost=postar1 propcov=quanew nmc=40000 thin=10  jointmodel;

  array covar[1] /nosymbols ;
  array data[1]/nosymbols;
  array first1[1]/nosymbols;
  array last1[1]/nosymbols;


  array beta[&nvar] ;
  array mu[&nrec];
  array ytemp[&nrep];
  array mutemp[&nrep];
  array VCV[&nrep,&nrep];

  begincnst;
     rc = read_array("recodedsplit1",data,"log_data");
        rc = read_array("recodedsplit1",first1,"first");
        rc = read_array("recodedsplit1",last1,"last");
        rc = read_array("covariates",covar);
  endcnst;
```

```
  parms sig2e 0.1  ;
  parms rho   0.1 ;
  parms sig2g 0.1  ;
  parms (beta1-beta&nvar) 1;


  beginnodata;
    prior beta: ~ normal(0,var=1e6);
       prior sig2g ~ general(0,lower=0);
    prior sig2e ~ general(0,lower=0);
       prior rho ~ general(0,lower=0,upper=1);


       call Fillmatrix(VCV,sig2g);

       do k= 0 to 2;
              do i = 1 to 3;
                     do j= 1 to 3;
                            VCV[i+(k*3),j+(k*3)] = sig2e * rho**abs(j-i) + sig2g ;
                     end;
              end;
    end;


       call mult(covar,beta,mu);
  endnodata;

    ljointpdf = 0;

    do irec = 1 to &nrec;
     if (first1[irec] = 1) then counter=0;
          counter = counter + 1;
          ytemp[counter] = data[irec];
          mutemp[counter] = mu[irec];
         if (last1[irec] = 1) then  ljointpdf = ljointpdf + lpdfmvn(ytemp, mutemp,
VCV);
        end;


  model general(ljointpdf);
run;
ODS graphics off;
```

CONTACT INFORMATION (HEADER 1)

Your comments and questions are valued and encouraged. Contact the author at:

Hennion Maud
Arlenda
Rue Edouard Belin, 5
1348 Mont-Saint-Guibert (Belgium)
Work Phone: +32 (0) 10 46 10 15
Email: maud.hennion@arlenda.com
Web: www.arlenda.com