



Using SAS PROC MIXED to Fit Individual Growth Models

December 20, 2006

Charlie Hallahan

Overview

This presentation is based on the paper “**Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models**” by Judith Singer.

(Journal of Educational and Behavioral Statistics, Winter 1998, Vol. 24, No. 4, pp. 323-355)

The paper (and links to the data) can be found at <http://gseweb.harvard.edu/~faculty/singer/>.

This demo will cover the **Individual Growth Model example**. Last month covered the **Multilevel Model example**.

Two choices in **PROC MIXED** for modeling the error covariance structure:

1. **RANDOM** statement for modeling variation in the Level 1 parameters
2. **REPEATED** statement to directly model the residual covariance

We'll start with the **RANDOM** statement to see parallels with the Multilevel or Hierarchical models discussed last month.

An Unconditional Linear Growth Model

Start with a simple two-level model:

Level-1: linear individual growth model (within-person)

Level-2: expresses variation in parameters from the growth model as random effects unrelated to any person-level covariates (between-person).

The level-1 parameters will be denoted as π and the level-2 parameters as β .

A 3-level model in which individuals within groups are tracked over time will be mentioned at the end.

An Unconditional Linear Growth Model

One predictor : random intercept and slope

$$\text{Level 1 model: } y_{ij} = \pi_{0j} + \pi_{1j}TIME_{ij} + r_{ij} \quad r_{ij} \sim N(0, \sigma^2)$$

$$\text{Level 2 model: } \pi_{0j} = \beta_{00} + u_{0j}$$

$$\pi_{1j} = \beta_{10} + u_{1j} \quad \text{where } \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix} \right]$$

$$\text{Reduced-form model: } y_{ij} = \underbrace{\left[\beta_{00} + \beta_{10}TIME_{ij} \right]}_{\text{fixed effects}} + \underbrace{\left[u_{0j} + u_{1j}TIME_{ij} + r_{ij} \right]}_{\text{random effects}}$$

y_{ij} represents a measurement for individual j taken at a time indexed by i .

This model is similar to the models discussed last month in that measurements over time are nested within individuals (instead of students nested within schools).

An Unconditional Linear Growth Model

A dataset from a 1988 paper by **J. B. Willett** is used. The dependent variable, y , measures growth in an “opposite naming task” on four occasions for 35 individuals, **TIME** is coded 0,1,2,3 (so the intercept estimates the (true) value of opposite-naming skill at occasion 0 – initial status), and the slope estimates the rate of change in (true) opposite-naming skill across occasions.

Coding **TIME** this way allows a natural interpretation for the intercept.

Obs	id	y	time	covar	ccovar	mcovar
1	1	205	0	137	23.5429	113.457
2	1	217	1	137	23.5429	113.457
3	1	268	2	137	23.5429	113.457
4	1	302	3	137	23.5429	113.457
5	2	219	0	123	9.5429	113.457
6	2	243	1	123	9.5429	113.457
7	2	279	2	123	9.5429	113.457
8	2	302	3	123	9.5429	113.457
9	3	142	0	129	15.5429	113.457
10	3	212	1	129	15.5429	113.457

An Unconditional Linear Growth Model

```
title2 "Unconditional Linear Growth Model";  
proc mixed data=singer.willett noclprint covtest;  
  class id;  
  model y = time /solution ddfm=bw notest;  
  random intercept time / type=un sub=id;  
run;
```

1. **sub=id** identifies the clusters to be defined by the variable `id`
2. **random intercept time** specifies the intercept and slope as random variables
3. **type=un** specifies an unstructured 2x2 covariance matrix, i.e., no restrictions

By default, there is always one random effect in each model, r_{ij} , representing variation within persons. In this example, we are adding two more sources of variation, the *intercepts* and *slopes* of TIME.

An Unconditional Linear Growth Model

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	1387.72627343	
1	1	1266.82273974	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	id	1198.78	318.38	3.77	<.0001
UN(2,1)	id	-179.26	88.9634	-2.01	0.0439
UN(2,2)	id	132.40	40.2107	3.29	0.0005
Residual		159.48	26.9566	5.92	<.0001

Fit Statistics

-2 Res Log Likelihood	1266.8
AIC (smaller is better)	1274.8
AICC (smaller is better)	1275.1
BIC (smaller is better)	1281.0

An Unconditional Linear Growth Model

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	164.37	6.1188	34	26.86	<.0001
time	26.9600	2.1666	104	12.44	<.0001

Interpreting the output from an unconditional individual growth model

1. Note convergence in just two iterations. Not uncommon for a simple balanced dataset.
2. $\beta_{00} = 164.37$ is the estimate of the average initial value of \mathbf{y} across persons.
3. $\beta_{10} = 26.96$ says the average person begins with a score of 164.37 and gains 27 points per testing occasion.
4. Both fixed effects are significant.

An Unconditional Linear Growth Model

Random effects :

$$\begin{pmatrix} \hat{\tau}_{00} & \hat{\tau}_{01} \\ \hat{\tau}_{01} & \hat{\tau}_{11} \end{pmatrix} = \begin{pmatrix} 1198.78 & -179.26 \\ -179.26 & 132.40 \end{pmatrix} \text{ and } \hat{\sigma}^2 = 159.48 \text{ and all are significant.}$$

Thus there is still sufficient variation left in the intercepts and slopes that might possibly be explained by a level 2 (person-level) covariate.

The next step is to add a "mystery" covariate called *COVAR*.

An Linear Growth Model With a Person-Level Covariate

One fixed predictor : random intercept and slope which vary with a covariate

$$\text{Level 1 model: } y_{ij} = \pi_{0j} + \pi_{1j}TIME_{ij} + r_{ij} \quad r_{ij} \sim N(0, \sigma^2)$$

$$\text{Level 2 model: } \pi_{0j} = \beta_{00} + \beta_{01}COVAR_j + u_{0j}$$

$$\pi_{1j} = \beta_{10} + \beta_{11}COVAR_j + u_{1j} \quad \text{where } \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix} \right]$$

In order to enhance interpretation, we center $COVAR$ at its grand mean.

Reduced - form model :

$$y_{ij} = \left[\beta_{00} + \beta_{10}TIME_{ij} + \beta_{01}(COVAR_j - CO\bar{V}AR) + \beta_{11}(COVAR_j - CO\bar{V}AR)TIME_{ij} \right] \\ + \left[u_{0j} + u_{1j}TIME_{ij} + r_{ij} \right] = \text{[fixed effects]} + \text{[random effects]}$$

An Linear Growth Model With a Person-Level Covariate

Now translate the **reduced model** into **PROC MIXED** code:

```
* Center covar at its grand mean;
proc means data=singer.willett mean;
    var covar;
    output out=meancovar mean=mcovar;
run;

data willett;
    if _n_ = 1 then set meancovar;
    set singer.willett;
    ccovar = covar - mcovar;
run;

title2 "Linear Growth Model with a person-level covariate";
proc mixed data=willett noclprint covtest;
    class id;
    model y = time ccovar ccovar*time /solution ddfm=bw notest;
    random intercept time / type=un sub=id gcorr;
run;
```

An Linear Growth Model With a Person-Level Covariate

Estimated G Correlation Matrix

Row	Effect	id	Col1	Col2
1	Intercept	1	1.0000	-0.4895
2	time	1	-0.4895	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	id	1236.41	332.40	3.72	<.0001
UN(2,1)	id	-178.23	85.4298	-2.09	0.0370
UN(2,2)	id	107.25	34.6767	3.09	0.0010
Residual		159.48	26.9566	5.92	<.0001

Fit Statistics

-2 Res Log Likelihood	1260.3
AIC (smaller is better)	1268.3
AICC (smaller is better)	1268.6
BIC (smaller is better)	1274.5

An Linear Growth Model With a Person-Level Covariate

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	164.37	6.2061	33	26.49	<.0001
time	26.9600	1.9939	103	13.52	<.0001
ccovar	-0.1136	0.5040	33	-0.23	0.8231
time*ccovar	0.4329	0.1619	103	2.67	0.0087

Interpreting the output from a linear growth model with a person-level covariate

Fixed effects:

1. Since **CCOVAR** and **TIME** are uncorrelated, the intercept and slope estimates for **INTERCEPT** and **TIME** are exactly the same as before.
2. The estimate for **CCOVAR** (-0.11) captures the relationship between the covariate and initial status and is insignificant.
3. **CCOVAR**, however, does have a significant effect on growth rates. Individuals who differ by 1.0 with respect to the covariate have growth rates that differ by 0.43 .

An Linear Growth Model With a Person-Level Covariate

Random effects :

$$\begin{pmatrix} \hat{\tau}_{00} & \hat{\tau}_{01} \\ \hat{\tau}_{01} & \hat{\tau}_{11} \end{pmatrix} = \begin{pmatrix} 1236.41 & -178.23 \\ -178.23 & 107.25 \end{pmatrix}$$

and $\hat{\sigma}^2 = 159.48$ (the same as before) and all are significant.

The estimated variance, $\hat{\tau}_{00}$, has even increased with the addition of a level-2 covariate.

However, $\hat{\tau}_{11}=107.25$, is decrease from 132.40. $(132.40 - 107.25)/132.40 = .19$ says that the covariate accounts for 19% of the explainable variation in growth rates.

Exploring the Structure of Variance-Covariance Matrix Within Persons

Covariance Matrix Structures in Mixed Models:

Recall the general structure of a mixed model:

$$y = X\beta + Z\gamma + \varepsilon \quad \text{where } \gamma \sim N(0, G), \quad \varepsilon \sim N(0, R), \quad \gamma \text{ and } \varepsilon \text{ are independent.}$$

Thus, the random part of the model ($Z\gamma + \varepsilon$) has covariance structure: $Z'GZ + R$

Structure can be imposed on the compound error term ($Z\gamma + \varepsilon$) by modeling either G and/or R .

In **PROC MIXED**, G is specified with the **RANDOM** statement and R with the **REPEATED** statement.

Exploring the Structure of Variance-Covariance Matrix Within Persons

So far, we've modeled correlations within each subject with the random part of the model via the **RANDOM** statement.

The **RANDOM** statement really says something about the behavior of the parameters in the model **across subjects**.

The **REPEATED** statement specifies how the residual term r_{ij} behaves **within each subject**.

In **Growth Models**, with the observations being ordered by time within each subject, an **autoregressive structure** might be appropriate.

Note that the last model estimated, accounted for **heteroskedasticity** by allowing the variance of the error term to vary with **TIME**.

With **PROC MIXED**, it is easy to estimate and compare several models with different covariance structures.

Exploring the Structure of Variance-Covariance Matrix Within Persons

Consider the following simpler model:

$$Y_{ij} = \pi_{0j} + \pi_{1j}TIME_{ij} + r_{ij} \quad \text{where } r_{ij} \sim N(0, \Sigma)$$

$$\pi_{0j} = \beta_{0j}$$

$$\pi_{1j} = \beta_{1j}$$

Here, the intercept and slope for *TIME* are fixed, but the covariance within each subject is modeled with the block-diagonal covariance matrix Σ .

The structure of Σ is specified with the **REPEATED** statement.

The number of time observations per subject determines the number of parameters in Σ . In our example, there are four time observations per subject.

We'll try three covariance structures: **Compound symmetry**, **AR(1)**, and **Unstructured** 17

Exploring the Structure of Variance-Covariance Matrix Within Persons

Compound symmetry: each 4x4 block is
$$\begin{pmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 \end{pmatrix}$$
 two parameters,

$$\text{where } \rho = \frac{\sigma_1}{\sigma^2 + \sigma_1} \text{ and } \sigma_1 = \text{Cov}(y_{ij}, y_{kj}) \quad i \neq k$$

For some reason, this is the way **SAS** chooses to parametrize this matrix.

It can equivalently be written as
$$\sigma^2 + \sigma_1 \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

Exploring the Structure of Variance-Covariance Matrix Within Persons

AR(1): each 4x4 block is σ^2
$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$
 two parameters

Unstructured: each 4x4 block is
$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix}$$
 ten parameters

Exploring the Structure of Variance-Covariance Matrix Within Persons

```
title2 "Linear Growth Model with compound symmetry";  
proc mixed data=willettnocprint covtest;  
  class id wave;  
  model y = time /solution notest;  
  repeated wave / type=cs sub=id r;  
run;
```

The variable `wave` on the repeated statement is the same as the variable `time`, except it is treated as a **class variable**, i.e. **MIXED** creates the appropriate dummy variables. Only a class variable is allowed on the **REPEATED** statement.

The option `type=cs` specifies the **compound symmetry** covariance structure.

The option `r` causes the estimate of the covariance matrix for the 1st subject to be printed. In this case, it's the same 4 x 4 matrix for each subject.

Exploring the Structure of Variance-Covariance Matrix Within Persons

Estimated R Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	1280.71	904.81	904.81	904.81
2	904.81	1280.71	904.81	904.81
3	904.81	904.81	1280.71	904.81
4	904.81	904.81	904.81	1280.71

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
CS	id	904.81	242.59	3.73	0.0002
Residual		375.90	52.1281	7.21	<.0001

Estimates of R: $\hat{\sigma}^2 + \hat{\sigma}_1 = 1280.71$ and $\hat{\rho} = \frac{\hat{\sigma}_1}{\hat{\sigma}^2 + \hat{\sigma}_1} = 904.81 / 1280.71 = 0.706$

Estimate of within subject residual variation: $\hat{\sigma}^2 = 1280.71 - 904.81 = 375.90$

Exploring the Structure of Variance-Covariance Matrix Within Persons

Fit Statistics

-2 Res Log Likelihood	1300.3
AIC (smaller is better)	1304.3
AICC (smaller is better)	1304.4
BIC (smaller is better)	1307.5

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	164.37	5.7766	34	28.45	<.0001
time	26.9600	1.4656	104	18.40	<.0001

All fixed and random effects are significant – although the tests for significance of the random effects (option `covtest`) requires “large” samples, usually more than we have in this example.

Exploring the Structure of Variance-Covariance Matrix Within Persons

With **PROC MIXED**, we can fit several competing covariance structures and use fit statistics to select a “best” model.

Assumption	N parameters	AIC	BIC	-2RLL
Compound Symmetry	2	1304.3	1307.5	1300.3
AR(1)	2	1277.5	1280.6	1273.5
Unstructured	10	1283.4	1299.0	1263.4

The various options on the **REPEATED** statement are:

```
type=cs  
type=ar(1)  
type=un
```

Exploring the Structure of Variance-Covariance Matrix Within Persons

For **type = un**, we get $\hat{\Sigma} = \begin{pmatrix} 1308 & 977 & 921 & 564 \\ 977 & 1120 & 1018 & 856 \\ 921 & 1018 & 1289 & 1081 \\ 564 & 856 & 1081 & 1415 \end{pmatrix}$

Note that the diagonal elements are approximately the same and the covariances appear to decrease over time. This suggests that **type = ar(1)** may be appropriate with only two free parameters in Σ instead of 10.

For **type = ar(1)**, we get $\hat{\Sigma} = \begin{pmatrix} 1324 & 1092 & 901 & 743 \\ 1092 & 1324 & 1092 & 901 \\ 901 & 1092 & 1324 & 1092 \\ 743 & 901 & 1092 & 1324 \end{pmatrix}$

Exploring the Structure of Variance-Covariance Matrix Within Persons

Based on the above results, the **AR(1)** structure is the preferred model.

We can now see what happens if we also allow the intercept and slope of **TIME** to vary randomly according to the covariate **COVAR**.

```
title2 "Linear Growth Model with a person-level covariate";  
proc mixed data=willettnoc1print covtest;  
  class id wave;  
  model y = time ccovar ccovar*time /solution ddfm=bw notest;  
  random intercept time / type=un sub=id g;  
  repeated wave / type=ar(1) sub=id r;  
run;
```

Exploring the Structure of Variance-Covariance Matrix Within Persons

Estimated R Matrix for id 1

Row	Col1	Col2	Col3	Col4
1	141.37	-19.3631	2.6522	-0.3633
2	-19.3631	141.37	-19.3631	2.6522
3	2.6522	-19.3631	141.37	-19.3631
4	-0.3633	2.6522	-19.3631	141.37

Estimated G Matrix

Row	Effect	id	Col1	Col2
1	Intercept	1	1258.10	-182.41
2	time	1	-182.41	110.94

Exploring the Structure of Variance-Covariance Matrix Within Persons

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	id	1258.10	333.25	3.78	<.0001
UN(2,1)	id	-182.41	84.5520	-2.16	0.0310
UN(2,2)	id	110.94	34.5299	3.21	0.0007
AR(1)	id	-0.1370	0.2589	-0.53	0.5968
Residual		141.37	36.3449	3.89	<.0001

Fit Statistics

-2 Res Log Likelihood	1260.0
AIC (smaller is better)	1270.0
AICC (smaller is better)	1270.5
BIC (smaller is better)	1277.8

Exploring the Structure of Variance-Covariance Matrix Within Persons

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	164.42	6.1990	33	26.52	<.0001
time	26.9082	1.9775	103	13.61	<.0001
ccovar	-0.1234	0.5034	33	-0.25	0.8079
time*ccovar	0.4357	0.1606	103	2.71	0.0078

Once we've allowed variation with the **RANDOM** statement, the estimate of the off-diagonal element of Σ is now insignificant, showing that we don't gain anything by allowing for an AR(1) structure with the **REPEATED** statement.

This conclusion is validated by the AIC, BIC, and -2RLL fit statistics.

Other models could still be tried – perhaps including an AR(1) structure along with just random intercepts.

Conclusion

1. **Three-level models** (or higher) use multiple **RANDOM** statements with appropriate nesting specification given in the **SUB=** option. For example, we could nest students within teachers and teachers within schools.

```
proc mixed noclprint covtest;  
  class teacher school;  
  model mathach= / solution;  
  random intercept / sub=school;  
  random intercept / sub=teacher(school);  
run;
```

2. Can combine the **Multilevel** model considered last month with the **Growth** model.

```
proc mixed noclprint covtest;  
  class student teacher;  
  model mathach = time / solution ddfm = bw;  
  random intercept time / type=un sub=teacher;  
  random intercept time / type=un sub=student(teacher);  
run;
```

Conclusion

3. **Heteroskedasticity** in the error covariance matrix can be modeled with the **GROUP** option on the **RANDOM** statement.
4. Sampling-based **Bayesian** analysis can be conducted using a **PRIOR** statement that permits a variety of distributional specifications for the variance components parameters prior density.
5. PROCs **GLIMMIX** and **NLINMIX** can be used to fit generalized linear mixed models and nonlinear mixed models.