“Mixed Reviews”: An Introduction to Proc Mixed

ANNMARIA DE MARS, PH.D.

THE JULIA GROUP

SANTA MONICA, CA

Obligatory naked mole rat slide

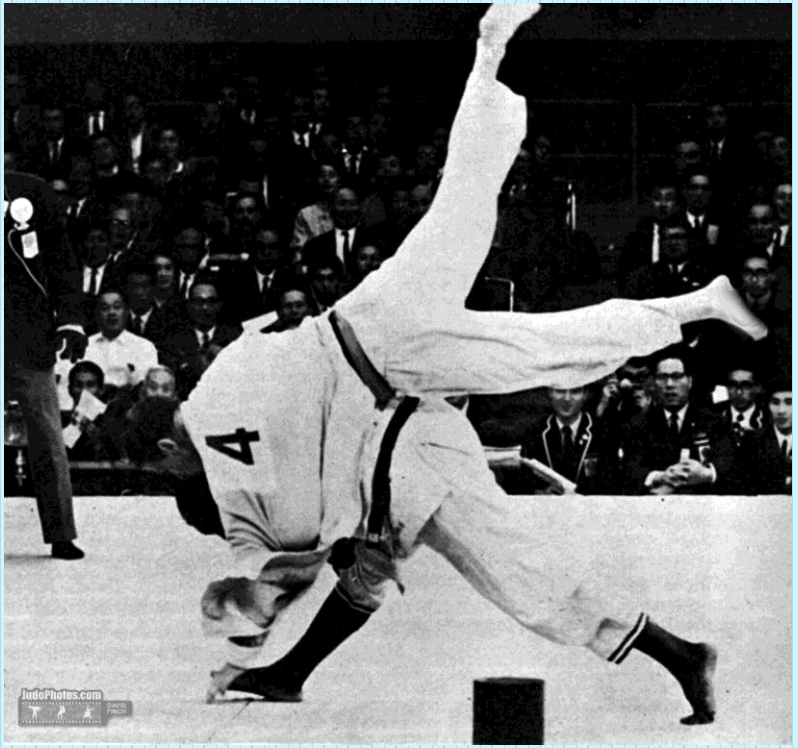
- How to do PROC MIXED, syntax using SAS 9.2 and SAS Enterprise Guide,
- Interpretation of PROC MIXED results,
- Differences between PROC MIXED and PROC GLM,
- Common error messages

Statisticians – feel free to check your text messages

"I see a significant R-squared,
can I leave now?"

It’s not as hard as it looks

Proc mixed, that is



In a nutshell

- For the vast majority of practical cases, PROC MIXED and PROC GLM will give you the same results

- If you aren't familiar with PROC GLM, the previous statement was of no help whatsoever

Baby Steps

GLM = General Linear Model

- Regression
- Analysis of Variance

PROC MIXED and PROC GLM **both**

Are used to predict the values of a numeric dependent variable

Assume the dependent variable is normally distributed

PROC GLM may be identical to PROC MIXED

MIXED MODELS include

- A mix of FIXED effects and RANDOM effects

- Seriously, this is key to understanding the whole thing

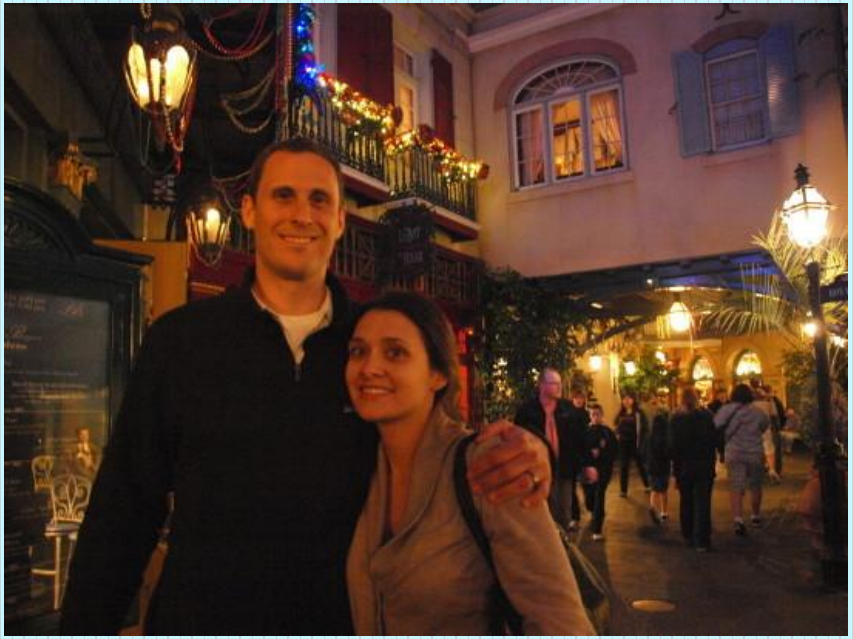
“The fixed-effects parameters are associated with known explanatory variables, as in the standard linear model.

These variables can be either qualitative (as in the traditional analysis of variance) or quantitative (as in standard linear regression).”

SAS 9.2 SAS/STAT Users Guide

Fixed effect

Not a random sample of genders



Random effects - schools

Well, either type of school, really

Why age is not a random effect

It's all about the (error)variance

Predicting height of a new species

- Family is a random effect
- Age is not

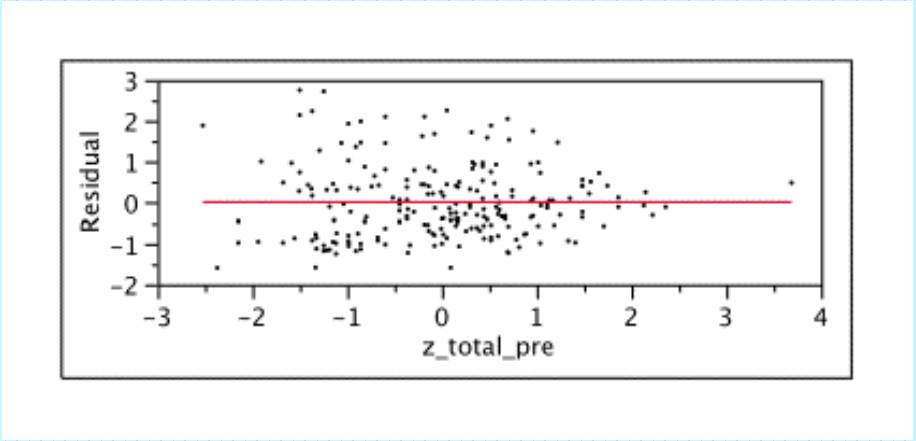
Why PROC MIXED may not matter

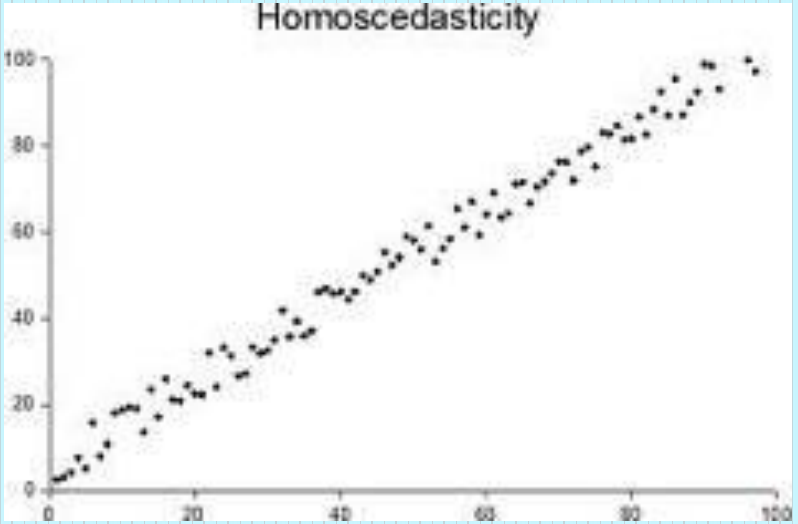
It may be that there is not much covariance within schools, sites, etc.

How Mixed Models differ

$$Y = ß1 * FAM\_INC + ß2 * SCHOOL + ß3 * IQ + e$$

IID* and homoscedasticity





* Independent, identically distributed

GLM assumes errors are uncorrelated, homoscedastic and normally distributed

In PROC MIXED, errors are not assumed to be uncorrelated

That's IT ?!

You went through this whole thing to say that in PROC GLM errors are assumed to be uncorrelated and in PROC MIXED they're not?

How Mixed Models differ

$$Y = ß1 * FAM\_INC + ß2 * IQ + \gamma Z1 + e$$

But in this case e is no longer assumed to be independent

Also, there can be (and usually are) a bunch of gamma effects

One last try …

- In the general model PROC GLM we are trying to fit means

In the mixed model…

- In addition to means, we are trying to fit covariances. We don't have to worry about that with the general linear model because we have assumed that all the observations are independent and thus there is no covariance.

One last statistical point

Yes, it is really the last one

Crossed versus nested factors

Often, in PROC MIXED you'll need to specify if your data are nested

Crossed factors: Each level of each factor appears in every other level

Marital status and employment

Nested factors

Subjects are usually NESTED within group.

A subject will be either an experimental group or a control group.

An example and syntax at last!

We want to test for differences between control and experimental groups.

We'd like to take into account our repeated measures, so when we compare our groups later we can say that any differences are due to our wonderful training.

Syntax

```
PROC MIXED DATA = mixed ;
   CLASS group name pre_post ;
   MODEL score = group pre_post group*pre_post ;
   REPEATED  pre_post / SUBJECT = name(group) TYPE = cs
   ;
   LSMEANS group group*pre_post / adjust = tukey ;
```

That's IT ?

Yes.

Now, for each statement …

PROC MIXED DATA = mixed ;

Okay, that's pretty self-evident

CLASS group name pre_post ;

Identify your categorical variables in the class statement.

This INCLUDES your subject identifier

MODEL score = group pre_post group*pre_post ;

Model dependent = FIXED EFFECTS ;

REPEATED pre_post / SUBJECT = name(group) TYPE = cs ;

Specify the variable that is repeated

Specify the variable that identifies the subject AND IF IT IS NESTED

Subject-identifier (variable it is nested within)

You don't have to specify a covariance structure type

LSMEANS group group*pre_post / adjust = tukey ;

LSMEANS requests the means for the variables and subgroups specified

/ adjust = requests test of differences between means

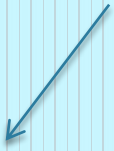
Output!

Model Information	
Data Set	WORK.MIXED
Dependent Variable	zscore
Covariance Structure	Compound Symmetry
Subject Effect	Name(Group)
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

There were no random effects

Dimensions	
Covariance Parameters	2
Columns in X	9
Columns in Z	0
Subjects	171
Max Obs Per Subject	4

In this example should be 2 * # of subjects



Number of Observations	
Number of Observations Read	342
Number of Observations Used	342
Number of Observations Not Used	0

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	2798.56848725	
1	2	2709.06943103	0.00000079
2	1	2709.06860756	0.00000000

Convergence criteria met

Covariance Parameter Estimates		
Cov Parm	**Subject**	**Estimate**
CS	Name(Group)	134.04
Residual		75.3247

Here is the estimate for the covariance due to Name (within subjects), the type of covariance matrix is compound symmetry.

Akakike Information Criterion

Fit statistics are used to compare models. If I re-ran the model without one of the variables, I could see if the resulting model was better or worse.

Fit Statistics	
-2 Res Log Likelihood	2709.1
AIC (smaller is better)	2713.1
AICC (smaller is better)	2713.1
BIC (smaller is better)	2719.4

Is your model [illegible]ng?

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
1	89.50	<.0001

Type 3 test of FIXED EFFECTS

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Group	1	169	3.17	0.0770
Pre_Post	1	169	49.86	<.0001
Group*Pre_Post	1	169	18.71	<.0001

What hypothesis is being tested?

Least Squares Means							
Effect	**Group**	**Pre/Post**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Group	CONTROL		101.06	1.3890	169	72.76	<.0001
Group	EXP		104.62	1.4460	169	72.35	<.0001
Group*Pre_Post	CONTROL	Post	102.34	1.5338	169	66.72	<.0001
Group*Pre_Post	CONTROL	Pre	99.7753	1.5338	169	65.05	<.0001
Group*Pre_Post	EXP	Post	109.95	1.5952	169	68.93	<.0001
Group*Pre_Post	EXP	Pre	99.2919	1.5952	169	62.24	<.0001

THIS IS REALLY IMPORTANT !!!

Test of Mean Differences

Differences of Least Squares Means					
Effect	**Group**	**Pre/Post**	**Group**	**Pre/Post**	**Adj P**
Group	CONTROL		EXP		0.0770
Group*Pre_Post	CONTROL	Post	CONTROL	Pre	0.2039
Group*Pre_Post	CONTROL	Post	EXP	Post	0.0040
Group*Pre_Post	CONTROL	Post	EXP	Pre	0.5161
Group*Pre_Post	CONTROL	Pre	EXP	Post	<.0001
Group*Pre_Post	CONTROL	Pre	EXP	Pre	0.9963
Group*Pre_Post	EXP	Post	EXP	Pre	<.0001

Compared to PROC GLM

NOTE: Dataset is structured differently

PROC MIXED multiple observations per subject

PROC GLM one observation per subject, with multiple fields for test score

MIXED

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Group	1	169	3.17	0.0770
Pre_Post	1	169	49.86	<.0001
Group*Pre_Post	1	169	18.71	<.0001

GLM

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Group	1	1098.62572	1098.62572	3.21	0.0748
Error	169	58116.52254	341.86190		

Source	DF	Type III SS	Mean Square	F Value	Pr > F
prepost	1	3755.47423	3755.47423	49.74	<.0001
prepost*Group	1	1409.19516	1409.19516	18.66	<.0001
Error(prepost)	169	12835.23216	75.50137		

The less than exciting point

It is not a very huge difference whether you use

PROC GLM or

PROC MIXED

How about RANDOM effects ?

Maybe that will be better?

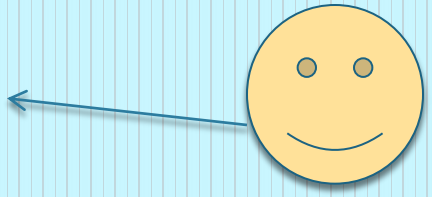
Curriculum tested at three sites with pretest and posttest

RANDOM effect = school

REPEATED = trial

At last! More syntax!

```
proc mixed data =  mixedR   ratio ;
    class   site trial case  ;
    model score =   trial   ;
    Random site / subject = case(site);
    Repeated trial ;
    lsmeans trial / adjust = tukey   ;
```



NOTE

Model score = trial ;

Only fixed effects on model statement

Site is a random effect

Random site / subject = case(site);

Repeated trial ;

We are happy

Covariance Parameter Estimates			
Cov Parm	**Subject**	**Ratio**	**Estimate**
type	Name(type)	2.0383	78.4081
Trial		1.0000	38.4679

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Trial	1	201	52.12	<.0001

Differences of Least Squares Means									
Effect	Trial	_Trial	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
Trial	1	2	-4.4554	0.6171	201	-7.22	<.0001	Tukey-Kramer	<.0001

Seriously, what difference does it make?

- Sometimes you can use RANDOM or REPEATED
- Sometimes RANDOM doesn't matter

Random & Repeated

Are not the same – except when they are.

Certain overspecified models …" can be specified by using a **random** or **repeated** statement alone. Unfortunately, one such model is the commonly encountered repeated measures with compound symmetry. "

http://www.jerrydallal.com/LHSP/mixedq.htm

Random may not matter

Conceptually, you have a random effect if it is sampled from the population of individuals, machines, schools, etc.

Statistically, a random effects explains some of the covariance. If there is not any difference among the families/ schools / sites in your sample, the RANDOM statement won't matter.

Repeated Measures ANOVA using Proc Mixed

With no programming

Pre- Post Test Experiment

A typical experimental design – subjects were either an experimental group or control group.

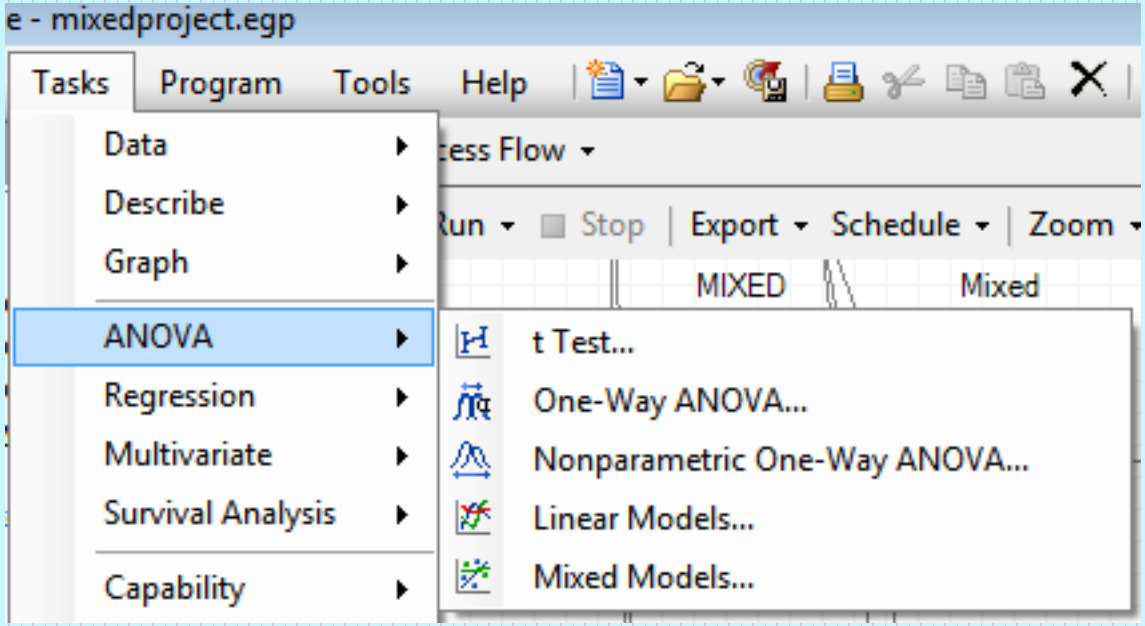
Both groups were given a pre-test and a post-test.

You want to test for significance of interaction between group and test. Your hypothesis is that such an interaction exists and the experimental group improved more.

What to do & how to do it

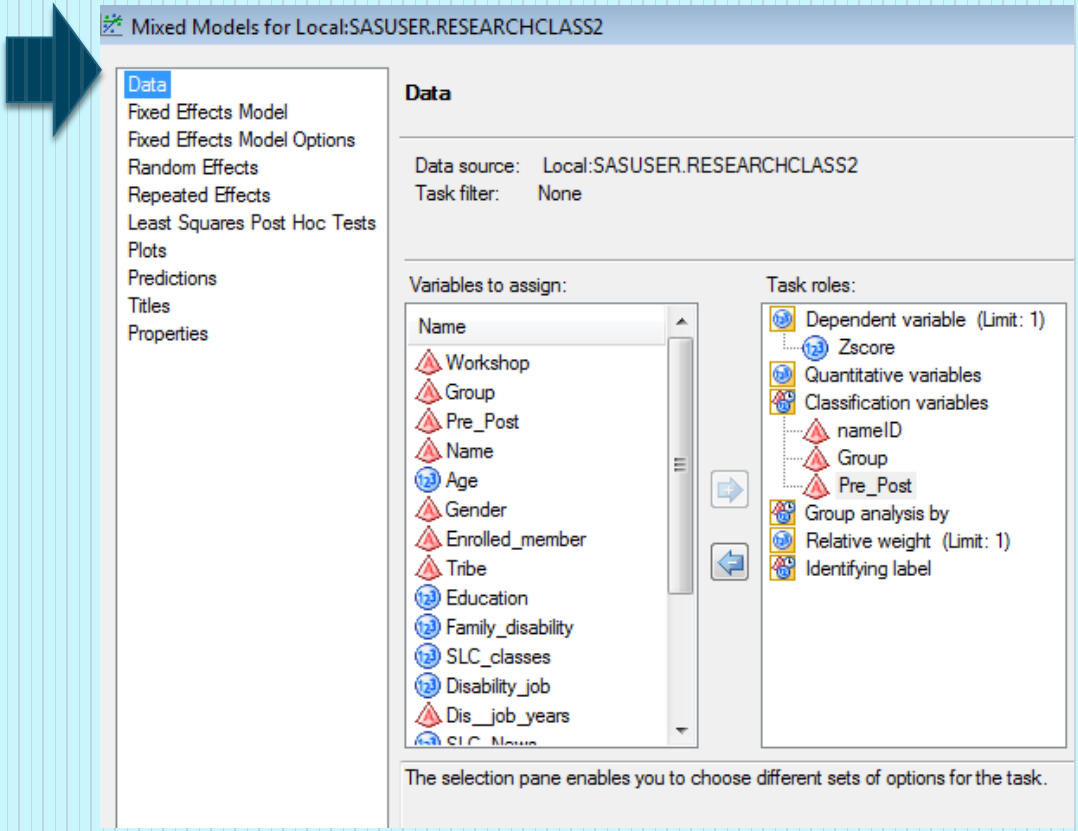
- You could do a mixed model ANOVA
- It is called mixed because it has two types of effects, fixed and random
- Your data should be in the format of one record for EACH measurement for each person, i.e., multiple records per person.

Select MIXED MODELS task



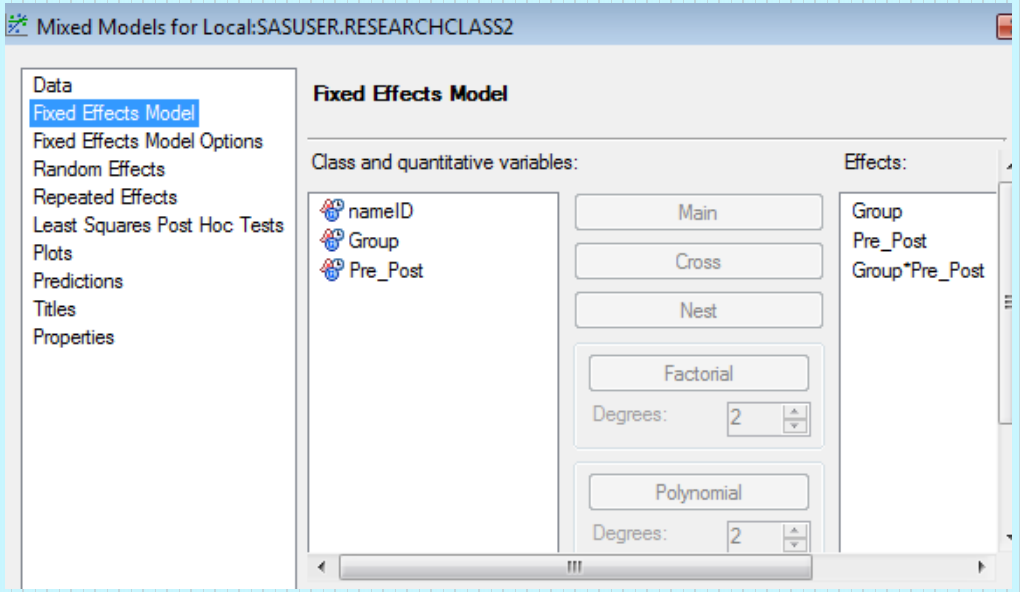
Mixed Model

1. Click on Data
2. Drag ZSCORE under dependent variable
3. Drag NAMEID PRE_POST & GROUP under classification



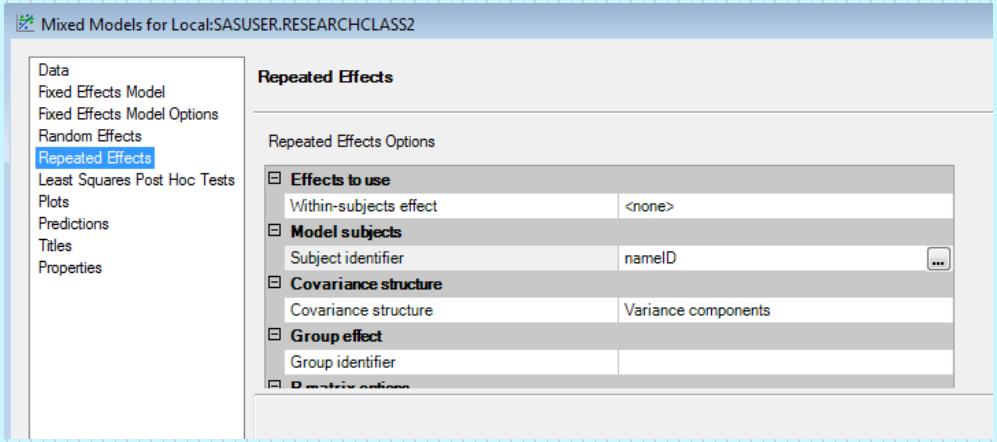
Mixed Model: Continued

1. Select FIXED EFFECTS MODEL
2. Select GROUP & PRE_POST and click on the **Main** button
3. Select GROUP & PRE_POST <u>at the same time</u> by holding down the shift key and click on the **Cross** button



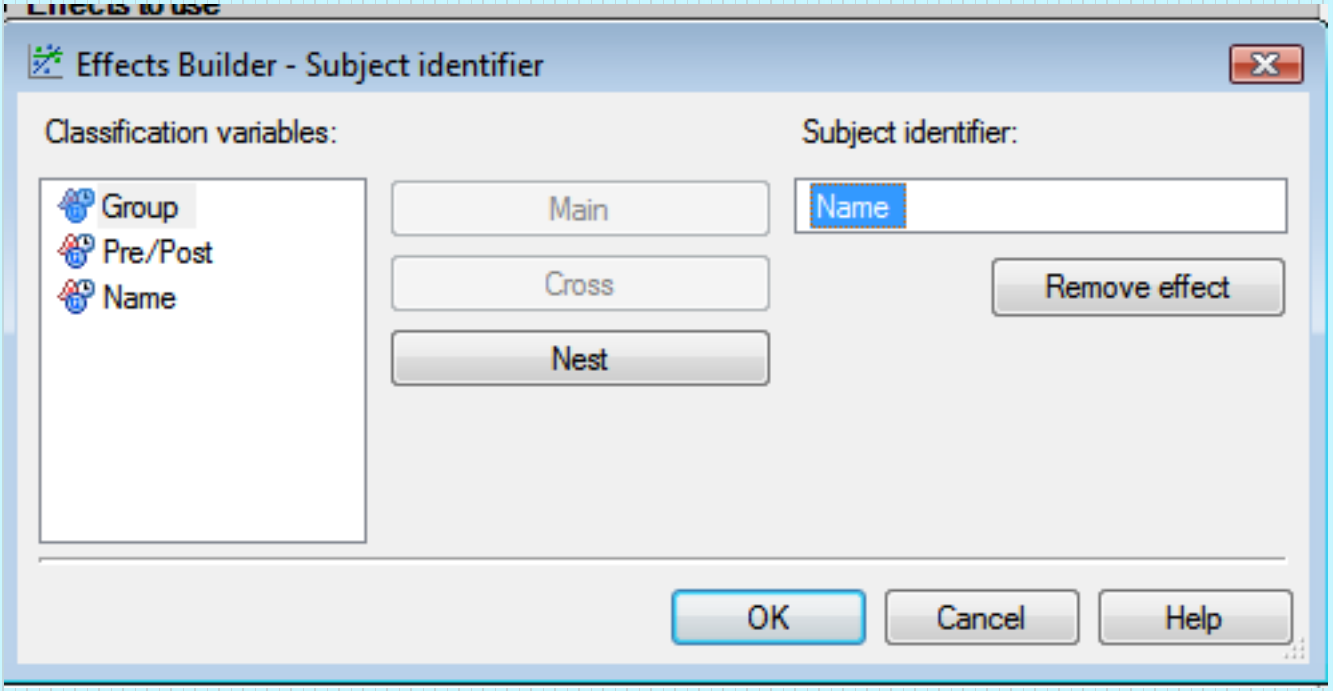
Mixed Model: Repeated

1. Click on REPEATED EFFECTS
2. Click on the … next to Subject Identifier
3. Select NAMEID as the Subject Identifier.



Mixed Model: Repeated

1. Click on Name, the NEST button will no longer be grayed out
2. Click on Group
3. Click NEST



Mistakes not to make

1. Even though it makes perfect sense to think of the subject identifier as a random effect (which it is) do NOT identify it as a random effect. The random effect is for random effects that are not repeated. In this example, there were no such random effects.
2. Know the difference between crossed & nested effects. Here we have both crossed and nested effects
   http://support.sas.com/learn/statlibrary/statlib_eg4.2/eg_anova_4.htm

My point !

Sorry, but ...

- Whether you use REPEATED vs RANDOM, the type of covariance, whether you use PROC GLM vs PROC MIXED. None of it matters a great deal unless your model is borderline.
- What does matter is if your model is completely WRONG, that is if you leave out the repeated effects, don't realize that subjects are nested within schools

So, Chris Rock was wrong. You need to know why*