

COURSE NOTES

# Getting Started with Mixed Models

Danny Modlin  
Senior Analytical Training Consultant

*Getting Started with Mixed Models Course Notes* was developed by Chris Daman, Danny Modlin, and Jill Tao. Additional contributions were made by John Amrhein, David Dickey, Ramon Littell, Bob Lucas, Stephen Mistler, Danny Modlin, Mike Patetta, Chris Riddiough, Catherine Truxillo, and Russ Wolfinger. Instructional design, editing, and production support was provided by the Learning Design and Development team.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

### **Getting Started with Mixed Models Course Notes**

Copyright © 2024 SAS Institute Inc. Cary, NC, USA. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

# Table of Contents

<b>Lesson 1</b>	<b>Introduction to Mixed Models .....</b>	<b>1-1</b>
1.1	Defining Mixed Models .....	1-3
1.2	Introduction to Mixed Models and Theory .....	1-9
	Demonstration: Fitting a Mixed Model Using the Mixed Models Task.....	1-12
1.3	Lesson Summary .....	1-29
<b>Lesson 2</b>	<b>Examples of Mixed Models .....</b>	<b>2-1</b>
2.1	Two-Way Mixed Models .....	2-3
	Demonstration: Plotting the Data.....	2-6
	Demonstration: Fitting the Two-Way Mixed Model .....	2-8
2.2	Analysis of Covariance with Random Effects .....	2-13
	Demonstration: Fitting an ANCOVA Model .....	2-18
2.3	Introduction to Repeated Measures Analysis.....	2-24
	Demonstration: Producing Profile Plots for the Three Drugs .....	2-29
<b>Lesson 3</b>	<b>LMIXED Procedure in SAS Viya.....</b>	<b>3-1</b>
3.1	LMIXED Procedure in SAS Viya .....	3-3
	Demonstration: Using PROC LMIXED within CAS .....	3-11
<b>Appendix A</b>	<b>References .....</b>	<b>A-1</b>
A.1	References .....	A-3

To learn more...



For information about other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to [training@sas.com](mailto:training@sas.com). You can also find this information on the web at <http://support.sas.com/training/> as well as in the Training Course Catalog.

For a list of SAS books (including e-books) that relate to the topics covered in this course notes, visit <https://www.sas.com/sas/books.html> or call 1-800-727-0025. US customers receive free shipping to US addresses.

# Lesson 1      Introduction to Mixed Models

<b>1.1</b>	<b>Defining Mixed Models .....</b>	<b>1-3</b>
<b>1.2</b>	<b>Introduction to Mixed Models and Theory .....</b>	<b>1-9</b>
	Demonstration: Fitting a Mixed Model Using the Mixed Models Task .....	1-12
<b>1.3</b>	<b>Lesson Summary .....</b>	<b>1-29</b>



# 1.1 Defining Mixed Models

---

## Objectives

- Define fixed and random effects.
- Define and explain a mixed model.

3

Copyright © SAS Institute Inc. All rights reserved.



## Where Do Data Come From?

- designed experiments
- sample surveys
- observational studies

4

Copyright © SAS Institute Inc. All rights reserved.



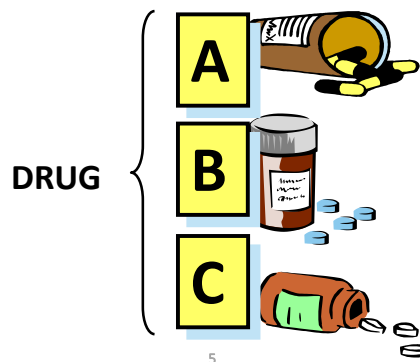
Almost any data set is produced by one of these three sources.

- In *designed experiments*, some form of treatment is applied to experimental units, and responses are observed.
- In *sample surveys*, data are collected on units according to a plan, called a *survey design*, but treatments are not applied to units. The units, typically people, already contain certain attributes such as age and gender. A value, such as annual income, is determined for each unit.
- In *observational studies*, data are collected on units that are available, rather than on units that are chosen according to a plan. For example, in a study to evaluate first-graders' reading skills, you collect data from several schools in a school district. The data include the students' ages, genders, attending schools, teachers, and the scores on some reading skill tests.

This course is concerned primarily with data from *designed experiments*, although the concepts can be easily extended to sample surveys and observational studies.

## Fixed Effects

- All levels of interest are selected by a nonrandom process and are included in the study.
- Inferences are to be made only about those levels that are included in the study.



Copyright © SAS Institute Inc. All rights reserved.



*Fixed effects* are those factors whose levels are selected by a nonrandom process or whose levels consist of the entire population of possible levels. All levels of interest are in your data set. The researcher is interested in comparing the effects of the factors on the response variable only for those levels included in the study.

For example, in a drug study, the researcher wants to compare the effect of three drugs (A, B, and C). She is interested only in the comparison of these three drugs, and she knows what they are before she conducts the experiment. The variable **drug** is a fixed effect.

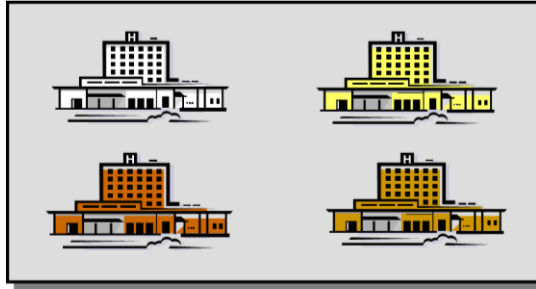
A model containing only fixed effects is called a *fixed effects model*.



## Random Effects

- Levels consist of a random sample of levels from a population of possible levels.
- Inference is about the population of levels, not only the subset of levels that are included in the study.

### CLINICS



6

Copyright © SAS Institute Inc. All rights reserved.



In some situations, a factor might have several levels and the researcher or data analyst selects a subset of the levels to include in the study. The inference from the data analysis is about the population of levels and not only the subset of levels included in the study. Effects like these are *random effects*. For example, in the same drug study, if four clinics are randomly selected from a population of clinics in a region and the researcher wants to make an inference about the drug effects across the population of clinics, not only the ones included in the study, then **clinic** is a random effect.

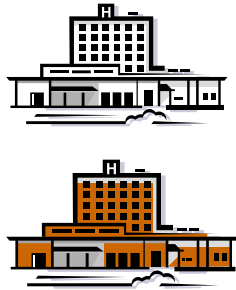
Models in which all effects are random are called *random effects models*. Variances associated with random effects are known as *variance components*.

**Note:** When deciding whether effects are fixed or random, you should ask yourself whether inferences are going to be drawn for only the levels of factors included in the experiment or across the population of possible levels.

## Mixed Models

Models in which some factors are fixed effects and other factors are random effects are called *mixed models*.

### CLINIC – random



### DRUG – fixed

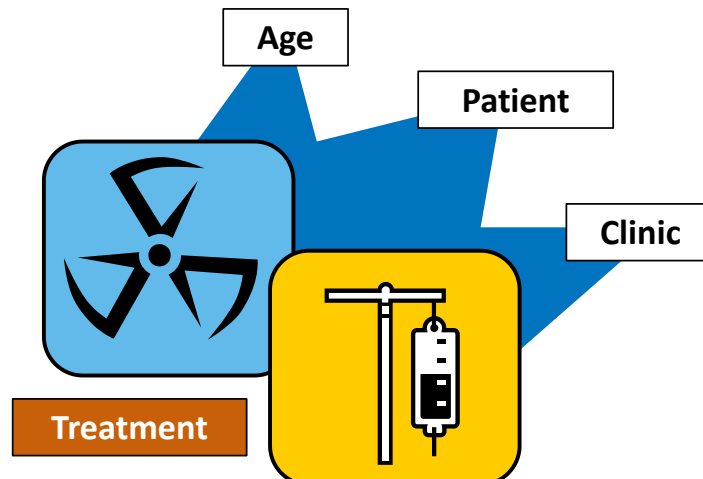


7

Copyright © SAS Institute Inc. All rights reserved.



## Cancer Example



8

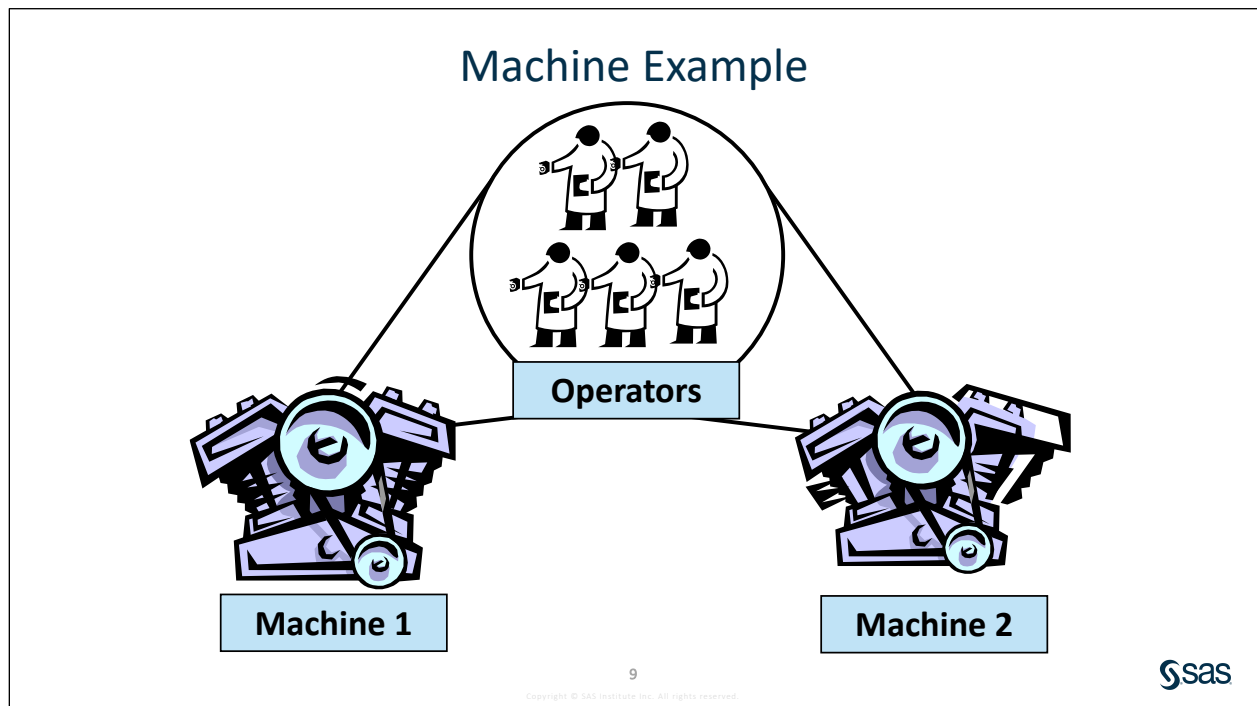
Copyright © SAS Institute Inc. All rights reserved.



A physician studies the effect of a chemotherapy treatment and a radiotherapy treatment on a certain form of cancer. The physician also wants to determine which treatment is more effective with children. Ten adults and 10 children are selected from each of five cancer treatment clinics. The reduction in the size of the tumor is measured after a two-month treatment period.

The experiment consists of the following effects:

- treatment** is a **fixed** effect (**chemotherapy** and **radiotherapy**) because the physician wants to compare only these two treatments and does not attempt to draw an inference beyond these two treatments.
- age** is a **fixed** effect (**child** and **adult**).
- clinic** might be a **fixed** or a **random** effect depending on the physician's interest. If the physician is interested only in the variability associated with these five clinics, then **clinic** is a **fixed** effect. However, if these clinics were randomly selected from the population of clinics, and the physician wants to make an inference about the population variation of clinics, then **clinic** is a **random** effect.
- patient** is a **random** effect because patients are a sample from a population of patients with this form of cancer.

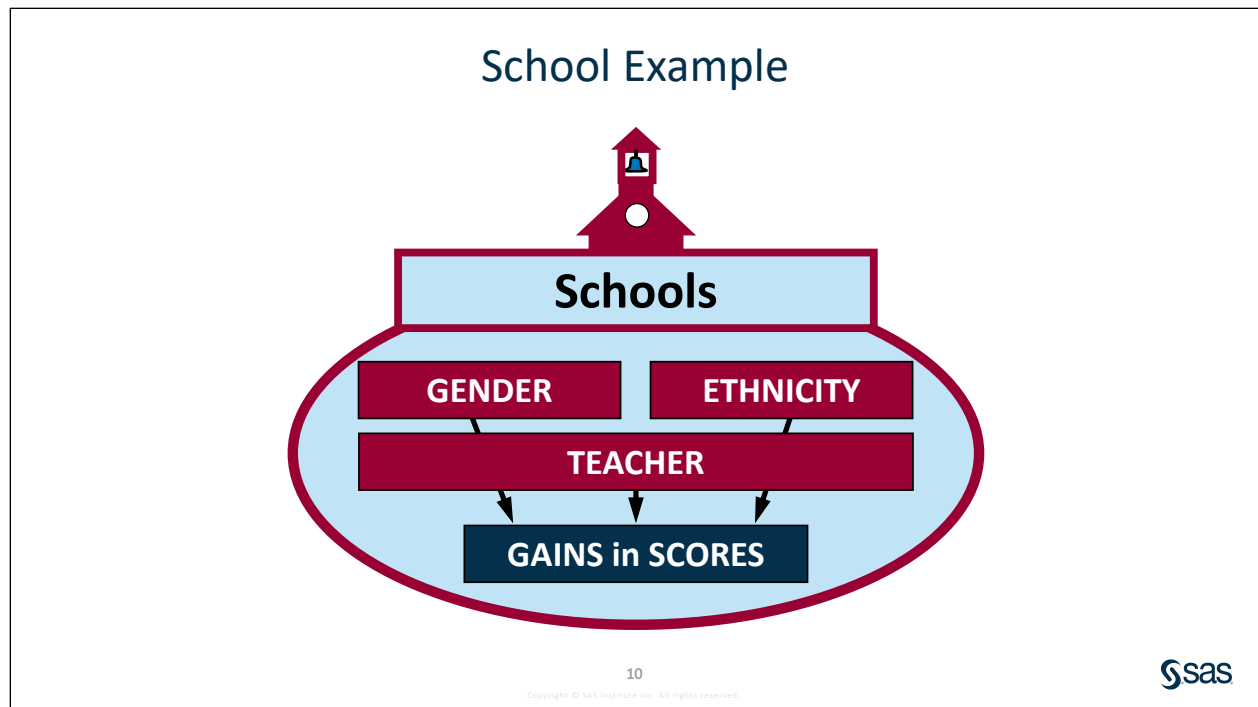


The manager of an automotive plant must replace the machines that produce a certain component that is used in automatic transmissions. Two different machines are available. The manager wants to evaluate the productivity of the machines when they are operated by the plant's employees. Five employees are randomly selected from the work force to operate each machine at three time periods.

This *two-way mixed* model consists of the following effects:

- machine** is a **fixed** effect because the manager considers only these two machines.
- operator** is a **fixed** or **random** effect. If only these five employees operate the machine, then **operator** is a **fixed** effect. However, if these five operators are a random sample of all possible operators, then **operator** is a **random** effect.
- machine\*operator** is a **fixed** effect when **operator** is fixed and a **random** effect when **operator** is random.

This is a two-way mixed model (if **operator** is random) because each selected employee's productivity is evaluated on each machine.



Gains in scores on a standardized test were recorded for 1,515 fourth-grade students in all schools in a district. *Gain* is defined as the score at the end of the year minus the score at the beginning of the year. The students' genders and ethnicities, as well as the identification numbers of the students' teachers, were also recorded. The primary objective was to evaluate and compare the schools in the gain scores. A secondary objective was to assess the effects of gender and ethnicity. This is a data set from an observational study.

The data consist of the following effects:

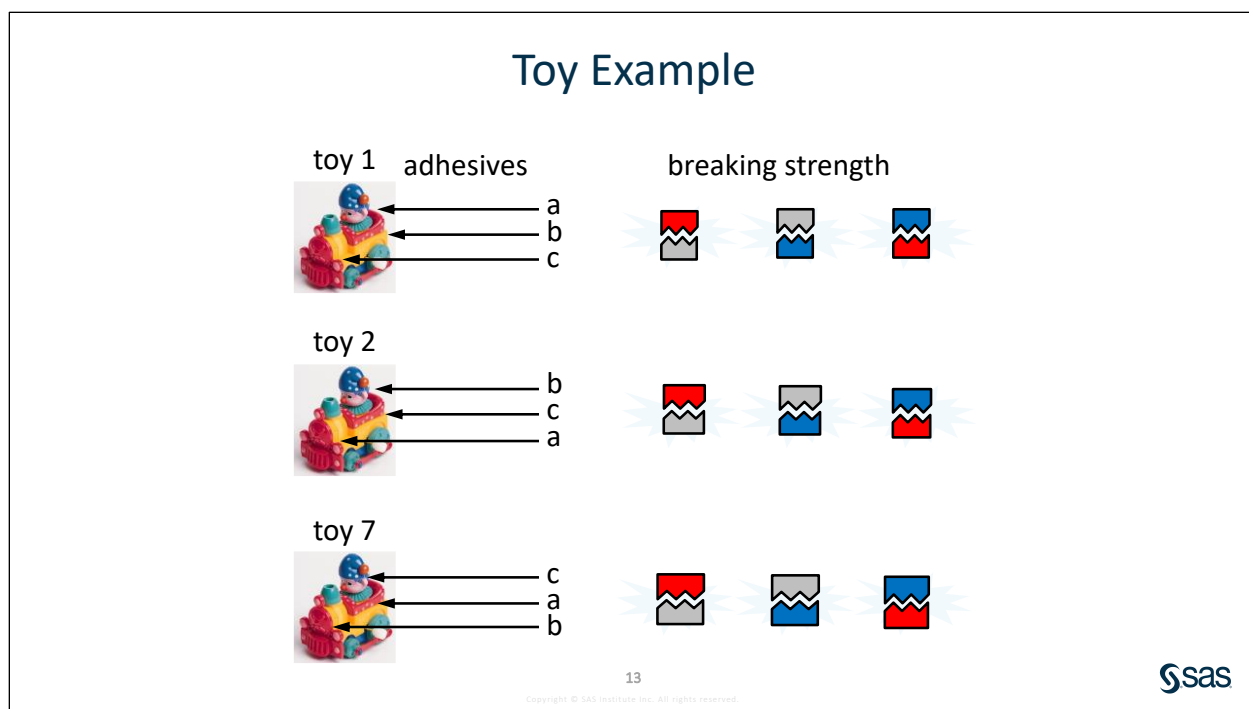
- school** is a **fixed** effect because only the schools included in the study are of interest.
- ethnicity** is a **fixed** effect.
- gender** is a **fixed** effect.
- teacher** is a **random** effect because the teachers represent a sample of the population of teachers who could teach at the schools.

# 1.2 Introduction to Mixed Models and Theory

---

## Objectives

- Compare MIXED and GLM.
- Describe the basic theory about general linear models and general linear mixed models.
- Compare the ML and REML methods.
- Describe the GLS estimation method for the fixed effects.



An engineer wants to test the strength of three adhesives that are used as bonding agents. Seven toys are randomly selected from a population of toys and are used for this strength test. Three brands of adhesives, *a*, *b*, and *c*, are used to glue parts from each toy. The amount of pressure that is required to break the bond is recorded. Data are stored in the SAS data set **aglm.toy**.

This *randomized complete block (RCB)* design consists of the following effects:

**adhesive** a treatment effect. This is a **fixed** effect because only three adhesives (*a*, *b*, and *c*) are used in the study. The engineer is interested only in making an inference about these three adhesives.

**toys** a blocking effect. This is a **random** effect because the seven toys are randomly selected from a population of toys. The inference about the treatment means is made over the entire population of toys.

**Note:** The treatments are assumed not to interact with the blocking variable.

The purpose of such an experiment is to accomplish the following:

- estimate and compare the treatment means over the entire population of blocks
- account for the variability in the response variable due to the blocks

## The Data

Obs	toy	adhesive	pressure
1	1	a	72.2
2	2	a	66.4
3	3	a	74.5
4	4	a	67.3
5	5	a	73.2
6	6	a	68.7
7	7	a	69.0
8	1	b	71.9
9	2	b	68.8
10	3	b	82.6
11	4	b	78.1
12	5	b	74.2
13	6	b	70.8
14	7	b	84.9
15	1	c	67.0
16	2	c	67.5
17	3	c	76.0
18	4	c	72.7
19	5	c	73.1
20	6	c	65.8
21	7	c	75.6

14

Copyright © SAS Institute Inc. All rights reserved.





## Fitting a Mixed Model Using the Mixed Models Task

Before performing an analysis of variance, you should conduct an initial data exploration. This can be accomplished, in part, by looking at a series plot of the data.

The data set **aglm.toy** contains the following variables:

**adhesive**      the adhesives used as bonding agents  
**toy**              the toy number  
**pressure**      the pressure to break the sample material

Expand **Tasks and Utilities** in the Navigation pane. Then, expand **Tasks** ⇨ **Graph** by clicking the arrows to the left. Double-click the **Series Plot** task to open this task interface.

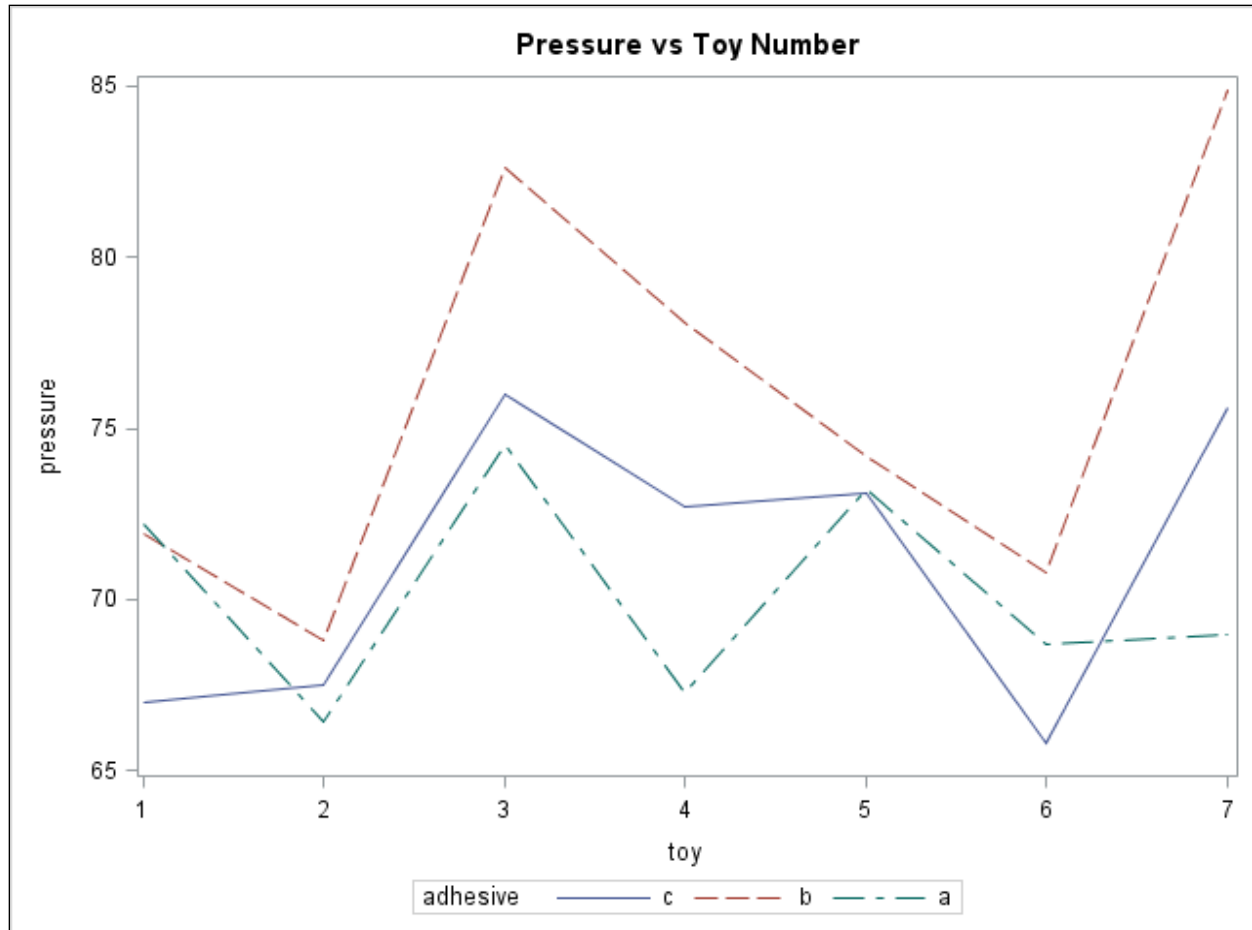
1. On the DATA tab, select the **toy** data set in the **AGLM** library.
2. Select **toy** as the X variable, **pressure** as the Y variable, and **adhesive** as the group variable.
3. Run the program. It is generated by clicking the **Running Person** icon.

The generated code is similar to that found in the QMM01d01.sas file. The Series Plot task generates code from the SGPLOT procedure.

```
proc sgplot data=aglm.toy;  
    series x=toy y=pressure / group=adhesive;  
    title 'Pressure vs Toy Number';  
run;  
title;
```

The SERIES statement requests a plot in which the data values are connected. X=**toy** specifies **toy** number to be the x-axis variable, and Y=**pressure** specifies **pressure** to be the y-axis variable. The GROUP=**adhesive** option specifies a variable that is used to group the data. A separate plot is created for each **adhesive**. The plot elements for each adhesive are automatically distinguished by different visual attributes.





There is a substantial amount of variability among the seven blocks. Also, **b** seems to be the strongest adhesive brand in six of the seven toys.

You can also examine the distribution of the data using a side-by-side box plot. This can be accomplished by using the Box Plot task.

1. In the Tasks area of Tasks and Utilities, double-click the **Box Plot** task within the Graph area to launch that task.
2. On the DATA tab, select the **toy** data set in the **AGLM** library.
3. Select **pressure** as the analysis variable and **adhesive** as the category variable.
4. Run the task.

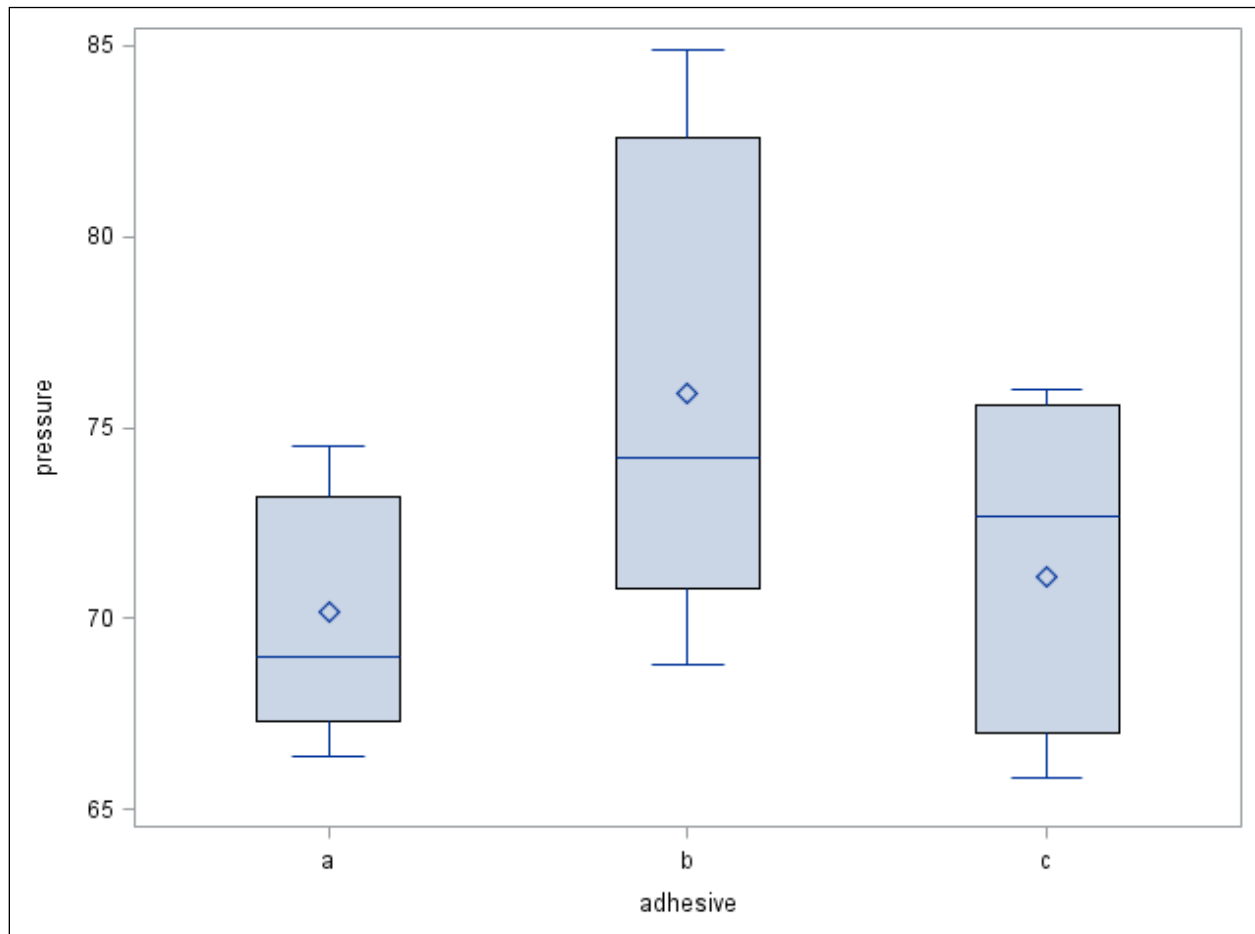
The generated code is similar to that found in the QMM01d01.sas file. The Box Plot task generates code from the SGPLOT procedure.

```
proc sgplot data=aglm.toy;
  vbox pressure / category=adhesive;
run;
```

Selected VBOX statement option:

**CATEGORY** specifies the category variable for the plot. A box plot is created for each distinct value of the category variable.

## PROC SGPLOT Output



There might be some differences among the three treatment means. The mean for adhesive **b** seems to be bigger than that for the other two adhesives. The variation for **b** seems to be larger as well. The variability in **pressure** for each adhesive is due primarily to the block (**toy**) variability.

Suppose you want to determine whether there are significant differences in the mean breaking pressures of bonds that are made using the three adhesives. You can use the Mixed Models task to analyze the **aglm.toy** data, and treat the block effect **toy** as a random effect.

1. In the Tasks and Utilities area, expand **Statistics**. Double-click the **Mixed Models** task.
2. On the DATA tab, select the **toy** data set in the **AGLM** library.
3. Select **pressure** as the dependent variable, and select **toy** and **adhesive** as classification variables.
4. On the MODEL tab, click the **Edit** button under Fixed Effects. Select **adhesive** and select **Add** to include it as a fixed effect in the model. Click **OK** to return to the task.
5. Click **Add Random**. Click the **Edit** button under Random Effects. Select **toy** and select **Add** to include it as a random effect in the model. Click **OK** to return to the task.
6. Run the task.

The Mixed Models task generates code from the MIXED procedure.

```
proc mixed data=aglm.toy;
  class adhesive toy;
  model pressure=adhesive / e3;
  random toy;
run;
```

All classification variables (fixed and random effects) are listed in the CLASS statement, but only the fixed effect (for example, **adhesive**) is listed in the MODEL statement. The random effect (for example, **toy**) is specified in the RANDOM statement.

**Note:** To add the e3 option, click the **Edit** button above the code area. This opens the task-generated code in a new window, where the code can be edited.

Selected MODEL statement option:

**E3** requests that Type III L matrix coefficients be displayed for all specified effects. This helps you interpret the Type 3 test for the fixed effects.

PROC MIXED Output

Model Information	
Data Set	AGLM.TOY
Dependent Variable	pressure
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

The Model Information table describes the model, some of the variables that it involves, and the method that is used to fit it. The default covariance structure fitted by PROC MIXED is variance components (VC) (constant variances and zero covariance). It is also referred to as *independent covariance structure*. Restricted or residual maximum likelihood (REML) is the default variance component estimation method that is used by PROC MIXED.

There are four methods for handling the residual variance in the model.

- The *profile* method concentrates the residual variance out of the optimization problem. This means solving analytically for the optimal residual variance and putting this expression back into the likelihood formula. This reduces the number of optimization parameters by one and can improve convergence properties.
- The *fit* method retains the variance as a parameter in the optimization.
- The *factor* method keeps the residual fixed.
- *None* is displayed when a residual variance is not a part of the model.

The row labeled *Fixed Effects SE Method* in this table describes the method that is used to compute the approximate standard errors for the fixed-effects parameter estimates and related functions of them. The default method is *Model-Based*, which can be changed to *Empirical* when the EMPIRICAL option in the PROC MIXED statement is used, or *Prasad-Rao-Jeske-Kackar-Harville* when the DDFM=KR2 option is specified in the MODEL statement.

**Note:** DDFM=KR2 applies the (prediction) standard error and degrees-of-freedom correction that are detailed by Kenward and Roger (2009). This correction reduces the bias of the precision estimator for fixed effects under nonlinear covariance structures.

The row labeled *Degrees of Freedom Method* in this table lists the method that is used for estimating the denominator degrees of freedom for the fixed effect. Seven possibilities for this are Containment, Between-Within, Residual, Satterthwaite, Kenward-Roger, Kenward-Roger (Firstorder), and Kenward-Roger2. Containment is the default method when the RANDOM statement is specified. You can use the DDFM= option in the MODEL statement to specify other methods.

Class Level Information			
Class	Levels	Values	
adhesive	3	a b c	
toy	7	1 2 3 4 5 6 7	

The Class Level Information table lists the levels of every variable that is specified in the CLASS statement. You should check this information to make sure that the data are correct. You can adjust the order of the CLASS variable levels with the ORDER= option in the PROC MIXED statement.

Dimensions	
Covariance Parameters	2
Columns in X	4
Columns in Z	7
Subjects	1
Max Obs per Subject	21

The Dimensions table lists the sizes of relevant matrices. This table can be useful for determining CPU time and memory requirements.

Number of Observations	
Number of Observations Read	21
Number of Observations Used	21
Number of Observations Not Used	0

The Number of Observations table shows the total number of observations in the data set and how many are used for fitting the model.

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	112.40987952	
1	1	107.79020201	0.00000000
Convergence criteria met.			

The Iteration History table describes the optimization of the residual log likelihood function. The optimization is performed using a ridge-stabilized Newton-Raphson algorithm, and the rows of this table describe the iterations that this algorithm takes to minimize the objective function.

Covariance Parameter Estimates	
Cov Parm	Estimate
toy	11.4478
Residual	10.3716

The variance component estimates are as follows:

$$\hat{\sigma}_b^2 = 11.4478 \text{ and } \hat{\sigma}^2 = 10.3716$$

Fit Statistics	
-2 Res Log Likelihood	107.8
AIC (Smaller is Better)	111.8
AICC (Smaller is Better)	112.6
BIC (Smaller is Better)	111.7

The Fit Statistics table provides information for goodness of model fit. All three information criteria consider both the fitness of the model and the model complexity. The model with more parameters receives a larger penalty. Bayesian information criterion (BIC) tends to produce a larger penalty than both the Akaike information criterion (AIC) and the finite-sample corrected Akaike information criterion (AICC). The detailed formula for the computations of these information criteria is in the SAS online documentation. For all three information criteria, smaller values indicate a better model.

Type 3 Coefficients for adhesive			
Effect	adhesive	Row1	Row2
Intercept			
adhesive	a	1	
adhesive	b		1
adhesive	c	-1	-1

The Type 3 Coefficients for **adhesive** prints the Type 3 **L** matrix that is used for the test for the **adhesive** effect. In this example,  $\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$ .

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
adhesive	2	12	6.36	0.0131

The Type 3 Tests of Fixed Effects table contains the *F* test for **adhesive** ( $F=6.36$ ), with a *p*-value of 0.0131. Therefore, you conclude that there is a difference between the mean breaking pressures for **adhesive** across all toys at a 5% significance level.

**Note:** PROC MIXED does not compute the sums of squares as PROC GLM does. Therefore, you do not see sums of squares in the Test of Fixed Effects table. The *F* statistic for the fixed

effect is computed as  $F = \frac{\hat{\beta}' \mathbf{L} [\mathbf{L}(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{L}' \mathbf{L} \hat{\beta}]}{\text{rank}(\mathbf{L})}$ , where

$\hat{\beta}$  is the vector of the fixed-effect estimates.

$\mathbf{L}$  is the coefficient matrix discussed earlier.

$\mathbf{X}$  is the design matrix for the fixed effects.

$\hat{\mathbf{V}}$  is the estimated covariance matrix for an observation using the REML method.

This statistic accounts for all variance components in the model as indicated by  $\hat{\mathbf{V}}$  in the equation.

**End of Demonstration**

## MIXED versus GLM

The syntax of these two procedures is very similar.

- *GLM* uses the ordinary least squares (OLS) method to make inferences about fixed effects. Therefore, the inferences are based on a fixed-effects-only model, even when you specify a *RANDOM* statement.
- *MIXED* uses the generalized least squares (GLS) method to make inferences about fixed effects. Therefore, the inferences directly incorporate the variance-covariance structure that you specify.
- *GLM* computes the analysis of variance to assess variations.
- *MIXED* uses the maximum likelihood approach to estimate the variance components. (The restricted maximum likelihood (REML) method is the default.)

16

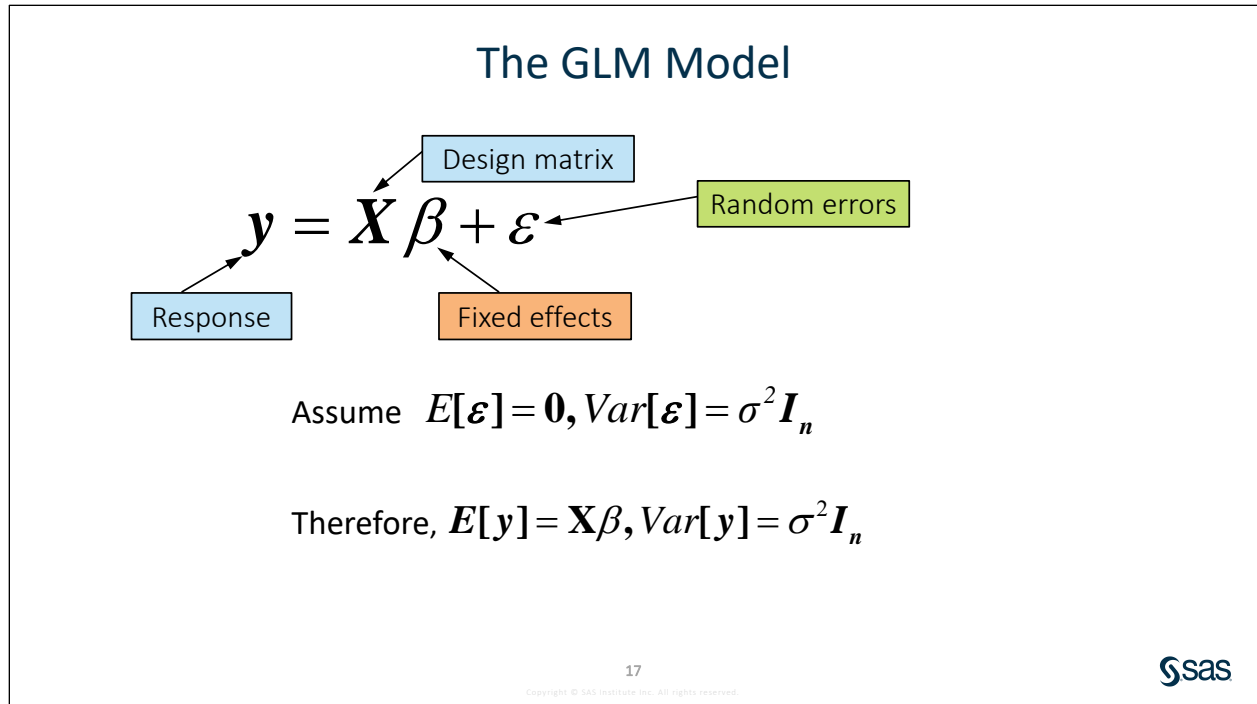
Copyright © SAS Institute Inc. All rights reserved.



The generalized least squares (GLS) method accounts for the variance-covariance structure that is modeled by the *MIXED* procedure. Theoretically, the GLS method is superior to the ordinary least squares (OLS) method for fixed effects inferences.

Although REML is the default method for estimating the variance parameters in *PROC MIXED*, you can use the method-of-moments techniques to estimate the variance components by specifying the *METHOD=* option in the *PROC MIXED* statement.

The *GLM* procedure is a fixed-effects procedure, so you should use the *MIXED* procedure to analyze mixed models.



$y$  is the vector of observed response data values.

$X$  is the known design matrix based on the model specification.

$\beta$  is the vector of unknown fixed-effect parameters.

$\varepsilon$  is the vector of random errors.

Assume that  $\varepsilon$  is a vector of independent random variables that are independently and normally distributed with a mean of 0 and a variance of  $\sigma^2$ . Therefore, for the general linear model,

$$E(\varepsilon) = \mathbf{0} \text{ and } \text{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n, \text{ where } \mathbf{I}_n \text{ is an } n \times n \text{ identity matrix } \begin{bmatrix} 1 & & 0 \\ & 1 & \\ & & \ddots \\ 0 & & & 1 \end{bmatrix}_{n \times n}.$$

Then it follows that  $E(y) = X\beta$  and  $\text{Var}(y) = \sigma^2 \mathbf{I}_n$ , where  $\mathbf{I}_n$  is an  $n \times n$  identity matrix.

**Note:** PROC GLM always models all effects as fixed, even when the RANDOM statement is used. The RANDOM statement is used to produce expected mean squares.

## The MIXED Model

$$y = X\beta + Z\gamma + \varepsilon$$

Assume  $\gamma \sim N(0, G)$  and  $\varepsilon \sim N(0, R)$

$\Rightarrow E(y) = X\beta, \quad Var(y) = ZGZ' + R = V$

18

sas

Copyright © SAS Institute Inc. All rights reserved.

$y$  is the vector of observed response data values.

$X$  is the known design matrix for the fixed effects that result from the MODEL statement in PROC MIXED.

$\beta$  is the vector of unknown fixed-effect parameters.

$Z$  is the known design matrix for the random effects that result from the RANDOM statement in PROC MIXED.

$\gamma$  is the vector of unknown random effects parameters.

$\varepsilon$  is the vector of random errors.

Assume that  $\gamma$  and  $\varepsilon$  are independently and normally distributed with the following:

$$E \begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} = \mathbf{0} \text{ and } \text{Var} \begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

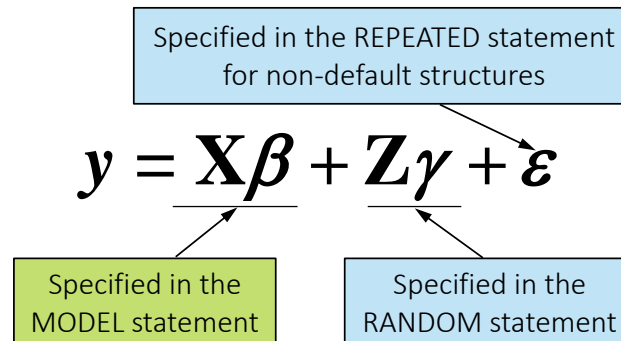
In the **aglm.toy** data set example,  $G = \sigma_b^2 \mathbf{I}_7$  and  $R = \sigma^2 \mathbf{I}_{21}$ . PROC MIXED enables you to specify various covariance structures for both the  $G$  and  $R$  matrices. The default is the variance component (or simple) structure as used in the **toy** example.

It follows that for the observed response variable  $y$ , you have  $E(y) = X\beta$  and  $Var(y) = ZGZ' + R = V$ .

**Note:** The mixed model extends the general linear model by enabling a more general specification of the covariance matrix of  $y$ . The mixed model allows random factors in the model and random elements of errors  $\varepsilon$  to be correlated. The general linear model is a special case of the mixed model with  $Z = \mathbf{0}$  (which means  $Z\gamma$  disappears from the model) and  $R = \sigma^2 \mathbf{I}_n$ .



## Statements in the MIXED Procedure



19

Copyright © SAS Institute Inc. All rights reserved.



In the MIXED procedure, you use the MODEL statement to specify the fixed effects, the RANDOM statement to specify the random effects, and the REPEATED statement to specify the variance-covariance structure of the errors, which is not the default  $\sigma^2 I_n$ . You should be careful when you specify both RANDOM and REPEATED statements in PROC MIXED. In some cases, one statement (for example, the REPEATED statement) with the specified variance-covariance structure captures all the variations in your data, and the other statement (for example, the RANDOM statement) is not needed.

## Linear Mixed Model Assumptions

- Random effects and residuals are normally distributed with mean zero and covariance matrices **G** and **R**, respectively.
- Random effects and residuals are independent of each other.
- The means (expected values) of the responses are linearly related to the predictor variables (linear in terms of fixed-effects parameters).

20

Copyright © SAS Institute Inc. All rights reserved.



Because normal data can be modeled entirely in terms of its means and variances or covariances, the two sets of parameters in a mixed linear model specify the complete probability distribution of the data. The parameters of the mean model are referred to as *fixed-effects parameters*. The parameters of the variance-covariance model are referred to as *covariance parameters*.

continued...

## Define the Mixed Model – Toy Example

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

$$\mathbf{y}_{21 \times 1} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{17} \\ y_{21} \\ \vdots \\ y_{37} \end{bmatrix} \quad \boldsymbol{\beta}_{4 \times 1} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \quad \boldsymbol{\gamma}_{7 \times 1} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \end{bmatrix} \quad \boldsymbol{\varepsilon}_{21 \times 1} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{17} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{37} \end{bmatrix}$$

21

Copyright © SAS Institute Inc. All rights reserved.



There is a total of 21 observations because  $3 \times 7 = 21$  observations are taken.

- four fixed effect parameters (overall mean plus three adhesives)
- seven random effect parameters (seven blocks)
- 21 random errors corresponding to each of the observations.

## Define the Mixed Model – Toy Example

$$\mathbf{X}_{21 \times 4} = \begin{bmatrix} 1_7 & 1_7 & 0_7 & 0_7 \\ 1_7 & 0_7 & 1_7 & 0_7 \\ 1_7 & 0_7 & 0_7 & 1_7 \end{bmatrix} \quad \mathbf{Z}_{21 \times 7} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_7 \\ \mathbf{I}_7 \end{bmatrix}$$

22

Copyright © SAS Institute Inc. All rights reserved.



The design matrices for the fixed effects ( $\mathbf{X}$ ) and the random effects ( $\mathbf{Z}$ ) are constructed based on the model specifications and your data. The subscript 7 represents seven identical rows.

## Covariance Matrices – Toy Example

$$\mathbf{G}_{7 \times 7} = \sigma_b^2 \mathbf{I}_7$$

$$\mathbf{R}_{21 \times 21} = \sigma^2 \mathbf{I}_{21}$$

$$\mathbf{V}_{21 \times 21} = \text{Var}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R} \rightarrow$$

$$\begin{cases} \text{Var}(y_{ij}) = \sigma_b^2 + \sigma^2 \\ \text{Cov}(y_{ij}, y_{kl}) = \begin{cases} \sigma_b^2 & i \neq k \text{ and } j = l \\ 0 & j \neq l \end{cases} \end{cases}$$

23

Copyright © SAS Institute Inc. All rights reserved.



The random effect vector  $\gamma$  has a multivariate normal distribution with mean vector 0 and variance-covariance matrix  $\mathbf{G} = \sigma_b^2 \mathbf{I}_7$ .

The random error vector  $\varepsilon$  has a multivariate normal distribution with mean 0 and variance-covariance matrix  $\mathbf{R} = \sigma^2 \mathbf{I}_{21}$ .

Therefore, the variance-covariance matrix of  $y$  is  $\text{Var}(y) = \mathbf{ZGZ}' + \mathbf{R} = \mathbf{V}$ .

**Note:**  $\mathbf{G}$  denotes the covariance matrix for the random effects.  $\mathbf{R}$  denotes the covariance matrix for random errors.  $\mathbf{V}$  denotes the covariance matrix for the observations.

### Details

$$\mathbf{G} = \sigma_b^2 \mathbf{I}_7 = \begin{bmatrix} \sigma_b^2 & 0 & \cdots & 0 \\ 0 & \sigma_b^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_b^2 \end{bmatrix}_{7 \times 7}, \quad \mathbf{R} = \sigma^2 \mathbf{I}_{21} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}_{21 \times 21}, \text{ and}$$

$$\mathbf{V} = \begin{bmatrix} \sigma_b^2 + \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_b^2 + \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & \ddots & & & & & & \ddots & & & & & & & & \ddots & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 + \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 \\ \vdots & & \vdots & & \vdots & & \ddots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\ \vdots & & \vdots & & \vdots & & & \ddots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 + \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 + \sigma^2 & 0 & 0 & 0 & 0 \\ \vdots & & \ddots & & & & & \ddots & & & & & & & & \ddots & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 + \sigma^2 \end{bmatrix}_{21 \times 21}$$

## Estimation Methods for the Covariance Parameters: ML versus REML

- Both methods are likelihood-based, and therefore, are consistent, asymptotically normal, and efficient.
- Both methods require numerical optimization.
- In general, REML estimators of the variance components are less biased than the ML estimators.
- In general, REML solutions are the ANOVA estimators for balanced data.
- REML is less sensitive to outliers in the data than ML.

The difference between the maximum likelihood (ML) and the restricted or residual maximum likelihood (REML) methods is the construction of the likelihood function. REML constructs the likelihood based on residuals and obtains maximum likelihood estimates of the variance components from this restricted or residual likelihood function.

By default, ML and REML constrain the parameter estimates to be nonnegative values. This creates an upward bias in both estimators. REML estimates are less biased than ML estimates and are generally preferred. Usually, the differences between the ML and REML estimations increase as the number of fixed effects in the model increases.

There seems to be a growing preference for REML rather than ML, for obtaining covariance parameter estimates (McCulloch and Searle 2001). It is sensible for balanced data for which REML solutions are the ANOVA estimators (PROC GLM), which are minimal variance unbiased estimators. Also, the REML estimators consider the degrees of freedom for the fixed effects in the model. Therefore, they correct the downward bias that is produced by the ML estimators. Finally, REML is less sensitive to outliers in the data than ML.

### Details

PROC MIXED constructs an objective function that is associated with ML or REML and maximizes it over all unknown parameters. The corresponding log likelihood functions are as follows:

$$ML: l(G, R) = -\frac{1}{2} \log |V| - \frac{1}{2} r' V^{-1} r - \frac{n}{2} \log 2\pi$$

$$REML: l_R(G, R) = -\frac{1}{2} \log |V| - \frac{1}{2} \log |X' V^{-1} X| - \frac{1}{2} r' V^{-1} r - \frac{n-p}{2} \log 2\pi$$

where  $r = y - X(X' V^{-1} X)^{-1} X' V^{-1} y$  and  $p = \text{rank}(X)$ .

By default, PROC MIXED uses a ridge-stabilized Newton-Raphson algorithm to find the parameter estimates that minimize  $-2$  times the log likelihood functions.

## Estimation Methods for the Covariance Parameters: ML versus REML

- The fit statistics that are based on REML can be used to compare different covariance models that are based on the same mean model
- The fit statistics that are based on ML can be used to compare different mean models that are based on the same covariance model.

Geometrically, REML projects the data onto the space orthogonally to the fixed-effects design matrix,  $\mathbf{X}$ . You must use the same MODEL statement to compare different covariance structures using the REML method. In other words, the fit statistics (for example, AICC values) from the REML method are not comparable if your fixed-effects models are different. You can use AICC values from the ML method to compare models with different fixed-effects if you specified the same covariance model.

## The Estimation Method for Fixed Effects

The generalized least squares (GLS) method has the following characteristics:

- takes into account the covariance matrices  $\mathbf{G}$  and  $\mathbf{R}$
- requires a reasonable estimate of  $\mathbf{G}$  and  $\mathbf{R}$
- produces the estimated GLS solutions when  $\mathbf{G}$  or  $\mathbf{R}$  is unknown

<div style="background-color: #0070C0; color: white; padding: 2px 10px; display: inline-block; margin-bottom: 5px;"><b>Estimated GLS</b></div> $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ $\hat{\boldsymbol{\gamma}} = (\hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$	<div style="background-color: #0070C0; color: white; padding: 2px 10px; display: inline-block; margin-bottom: 5px;"><b>OLS</b></div> $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
---	---

The generalized least squares (GLS) estimates consider the covariance matrices  $\mathbf{G}$  and  $\mathbf{R}$ , but they also depend on the reasonable estimates of  $\mathbf{G}$  and  $\mathbf{R}$  being obtained.

The MIXED procedure computes the GLS solutions for the fixed effects  $\boldsymbol{\beta}$ . Notice that the true value of  $\mathbf{V}$  is usually unknown. Replacing  $\mathbf{V}$  by an estimator  $\hat{\mathbf{V}}$  yields the estimated GLS solutions.

Recall the mixed models notation  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ . The variance matrix of the data is given by  $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$ . The GLS solutions for  $\boldsymbol{\beta}$  are obtained by minimizing the following:

$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$  for  $\boldsymbol{\beta}$ , yielding the estimated GLS solution for  $\boldsymbol{\beta}$  as  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$

The estimated GLS solutions for the random effects  $\boldsymbol{\gamma}$  can be obtained by  $\hat{\boldsymbol{\gamma}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ .

The ordinary least squares (OLS) solutions for a fixed effect model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  is given by  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ , which is a special case of the GLS solution with  $\mathbf{V} = \sigma^2\mathbf{I}_n$ .

See the SAS online documentation for more details.

**Note:** The maximum likelihood estimator for  $\boldsymbol{\beta}$  in the mixed model equals the generalized least squares estimator.

## GLS versus OLS

- *Ordinary least squares (OLS)* estimates and standard errors are based on the assumption that the errors are independently, normally distributed with a common variance, that is,  $R = \sigma^2 I_n = V$ .
- *Generalized least squares (GLS)* estimates and standard errors are based on the **G** and **R** matrices that can take a variety of forms.
- OLS is a special case of GLS.
- For balanced data, estimates from OLS and GLS generally agree, but the standard errors do not for mixed models.
- If the covariance model is incorrectly specified, sometimes OLS can perform better than GLS.

27

Copyright © SAS Institute Inc. All rights reserved.



In general linear models, you assume that all effects are fixed, and the random errors are independently, normally distributed with a mean of zero and a constant variance of  $\sigma^2$ , in other words,  $R = \sigma^2 I$ . The ordinary least squares (OLS) solutions to the normal equations contain nice statistical properties such as linear, unbiased, minimum variance, and so on.

In mixed models, you assume that random effects and random errors are independent of each other and follow a multivariate normal distribution with a mean of zero and covariance matrices **G** and **R**, respectively. **G** and **R** can take a variety of forms, including heterogeneous variances and interdependent covariances.

Likelihood-based methods are generally used for obtaining estimated **G** and **R** matrices. The solutions to the mixed model equations are generalized least squares (GLS) estimates. For balanced data, the estimates from OLS and GLS generally agree. However, the standard errors do not for mixed models. For unbalanced data, neither the estimates nor the standard errors are the same between OLS and GLS. In general, GLS estimates are superior to OLS estimates for mixed models.

**Note:** The GLM procedure produces OLS estimates and standard errors. The MIXED procedure produces GLS estimates and standard errors.

## Inferences about the Fixed Effects

- The variance-covariance matrix of the GLS fixed effect estimates  $\hat{\beta}_{GLS}$  is given by the following:

$$\text{Var}(\hat{\beta}_{GLS}) = (X'V^{-1}X)^{-1}$$

- The variance-covariance matrix of the estimated GLS fixed effect estimates  $\hat{\beta}_{EGLS}$  is given by the following:

$$\text{Var}(\hat{\beta}_{EGLS}) = (X'\hat{V}^{-1}X)^{-1}$$

- The variance-covariance matrix of the OLS estimates, by comparison, is as follows:

$$\text{Var}(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$$

28



Copyright © SAS Institute Inc. All rights reserved.

Because  $V$  is usually unknown, the estimator of the variance-covariance matrix of the fixed effects is  $(X'\hat{V}^{-1}X)^{-1}$ . This estimator is biased downward because the variability introduced by working with estimated variance components in  $V$ , rather than with the known variance components, is not taken into account in the approximation. PROC MIXED does not compute any inflation factors by default, but rather accounts for the downward bias by using the approximate  $t$  statistic and  $F$  statistic.

However, the DDFM=KR2 option in the MODEL statement does introduce an inflation factor for the estimated variance-covariance matrix of the fixed and random effects. This method was proposed by Prasad and Rao (1990) and Harville and Jeske (1992), with the Satterthwaite-based degrees of freedom that was suggested by Kenward and Roger (2009).



# 1.3 Lesson Summary

---

Almost any data are produced by designed experiments, sample surveys, or observational studies. When you build statistical models to analyze your data, you frequently must identify fixed effects and random effects.

*Fixed effects* are those factors whose levels are selected by a nonrandom process or whose levels consist of the entire population of possible levels. Factors are fixed if all levels of interest are included in the data set.

*Random effects* are those factors whose levels represent a random sample of levels from a population of possible levels. Inference can be made about a set of levels larger than those included in the data set. Variances associated with random effects are known as *variance components*.

Models in which some factors are fixed effects and other factors are random effects are called *mixed models*.

One of the simplest mixed model examples is the one that is fit to a randomized complete block design. In this, each treatment is applied to experimental units in each block, and the treatments are randomly assigned to the experimental units within each block. *Blocks* are groups of experimental units that are formed so that units within blocks are as homogeneous as possible. Blocks are almost always random effects.

The MIXED procedure should be used for analyzing mixed models. The GLM procedure was developed for fixed-effects models. Therefore, it is not suitable for mixed model analyses.

The MIXED procedure uses the residual or restricted maximum likelihood method to estimate the variance-covariance parameters. It uses the generalized least squares (GLS) method to estimate fixed effects parameters and standard errors. The MIXED procedure uses the MODEL statement to model the fixed effects (or the mean model), uses the RANDOM statement to model the random effects (the covariance model), and uses the REPEATED statement to model the non-default error structures (the covariance model).

The variance-covariance matrix for random effects is denoted as **G**, the variance-covariance matrix for the residuals is denoted as **R**, and the covariance-covariance matrix for the observed data is denoted as **V**. It can be shown that  $V = ZGZ' + R$ . The GLS estimation method takes into account the covariance matrices **G** and **R**, but requires a reasonable estimate of **G** and **R**. In general, GLS is superior to ordinary least squares (OLS) assuming an appropriate covariance structure.

SAS/GRAPH statistical graphics procedures (SG procedures) enable you to easily create complex statistical graphics that use the principles of effective graphics to accurately communicate the results of your analysis to your consumers. The SG procedures, including PROC SGSCATTER, PROC SGPLOT, PROC SGPANEL, and PROC SGRENDER require minimal coding, which enables you to focus on your statistical analysis instead of the visual appearance of your graphs.

You can use the GLIMMIX procedure to obtain the design matrix for a mixed model.

General form of the MIXED procedure:

```
PROC MIXED options;  
  CLASS variables;  
  MODEL dependent=fixed-effects / options;  
  RANDOM random-effects / options;  
  CONTRAST 'label' fixed-effect values |  
             random-effect values / options;  
  ESTIMATE 'label' fixed-effect values |  
            random-effect values / options;  
  LSMEANS fixed-effects / options;  
RUN;
```

# Lesson 2      Examples of Mixed Models

<b>2.1</b>	<b>Two-Way Mixed Models .....</b>	<b>2-3</b>
	Demonstration: Plotting the Data .....	2-6
	Demonstration: Fitting the Two-Way Mixed Model .....	2-8
<b>2.2</b>	<b>Analysis of Covariance with Random Effects .....</b>	<b>2-13</b>
	Demonstration: Fitting an ANCOVA Model .....	2-18
<b>2.3</b>	<b>Introduction to Repeated Measures Analysis.....</b>	<b>2-24</b>
	Demonstration: Producing Profile Plots for the Three Drugs.....	2-29



## 2.1 Two-Way Mixed Models

### Objectives

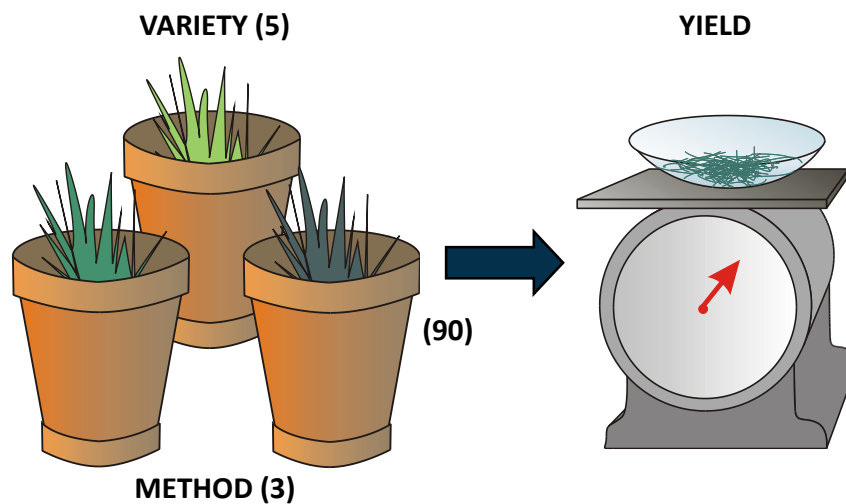
- Identify a two-way mixed model design.
- Using the MIXED procedure, analyze two-way mixed models.

3

Copyright © SAS Institute Inc. All rights reserved.



### Grass Example



4

Copyright © SAS Institute Inc. All rights reserved.



Three seed growth methods are applied to seeds from each of five varieties of turf grass. Six pots are planted with seeds from each method-by-variety combination. The 90 pots are randomly placed in a uniform growth chamber. Dry matter yields are measured from clippings at the end of four weeks.

This is an example of a completely randomized experiment with a factorial arrangement of treatments. The 15 treatments are the combinations of levels of the two factors, **variety** (five levels) and **method** (three levels). Assume that the five varieties were randomly chosen from a broader population of varieties. Interest is not in these particular five varieties but in the population from which they were chosen. The variable **variety** is considered a **random** effect. The variable **method** is considered a **fixed** effect, because the interest is only in these three methods. The **method\*variety** interaction is a random effect.

Because both random and fixed effects are involved, the model is defined as being mixed. The varieties are random. The inference about method means that differences should apply across all varieties.

### The Data

method	variety	yield
A	1	22.1
A	1	24.1
A	1	19.1
A	1	22.1
A	1	25.1
A	1	18.1
A	2	29.1
A	2	17.1
A	2	21.6
A	2	28.6
A	2	17.1
A	2	26.6
A	3	25.3
A	3	25.8
A	3	22.8
A	3	28.3
A	3	21.3
A	3	18.3
A	4	19.8
A	4	28.3
...	...	...



Data are recorded in a SAS data set **aglm.grass**. The following variables are in the data set:

**method** the growth method

**variety** the variety of turf grass

**yield** the amount of dried turf growth in each pot after four weeks

## Define the Mixed Model

$$y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \varepsilon_{ijk}$$

method effect,  
fixed

variety effect,  
random

method\*variety  
effect, random

$$b_j \sim N(0, \sigma_b^2), (\alpha b)_{ij} \sim N(0, \sigma_{\alpha b}^2), \varepsilon_{ijk} \sim N(0, \sigma^2)$$

Therefore,

$$E(y_{ijk}) = \mu + \alpha_i$$

$$Var(y_{ijk}) = \sigma_b^2 + \sigma_{\alpha b}^2 + \sigma^2$$

7

Copyright © SAS Institute Inc. All rights reserved.



- $y_{ijk}$      $k^{\text{th}}$  observation ( $k=1$  to 6) for the  $i^{\text{th}}$  method ( $i=1, 2, 3$ ) and the  $j^{\text{th}}$  variety ( $j=1$  to 5)
- $\mu$     overall mean and an unknown fixed parameter
- $\alpha_i$     effect for the  $i^{\text{th}}$  method and an unknown fixed parameter
- $b_j$     effect of the  $j^{\text{th}}$  variety, a random effect,  $b_j \sim \text{i.i.d. } N(0, \sigma_b^2)$
- $(\alpha b)_{ij}$     interaction between the  $i^{\text{th}}$  method and the  $j^{\text{th}}$  variety, a random effect  $(\alpha b)_{ij} \sim \text{i.i.d. } N(0, \sigma_{\alpha b}^2)$
- $\varepsilon_{ijk}$     experimental error,  $\varepsilon_{ijk} \sim \text{i.i.d. } N(0, \sigma^2)$

The effects  $b_j$ ,  $(\alpha b)_{ij}$ , and  $\varepsilon_{ijk}$  are assumed to be independent random variables.

Therefore,

- $E(y_{ijk}) = \mu + \alpha_i$  is the mean for method  $i$  averaged across all varieties in the population.
- $Var(y_{ijk}) = \sigma_b^2 + \sigma_{\alpha b}^2 + \sigma^2$  is the variance of an observation. The variance components are  $\sigma_b^2$  (**variety** variance),  $\sigma_{\alpha b}^2$  (**method\*variety** variance), and  $\sigma^2$  (random errors).



## Plotting the Data

---

Before fitting the mixed model to the **aglm.grass** data set, you might want to perform initial data explorations. You can use the Line Chart task to examine the yield across the five varieties for the three methods.

1. Expand the **Graph** area under Tasks within Tasks and Utilities.
2. Double click the **Line Chart** task to launch it.
3. On the DATA tab, select the **grass** data set in the **AGLM** library.
4. Select **variety** as the category variable, **yield** as the response variable, and **method** as the group variable. Do not change the statistic. Retain mean as the statistic.
5. Run the task.

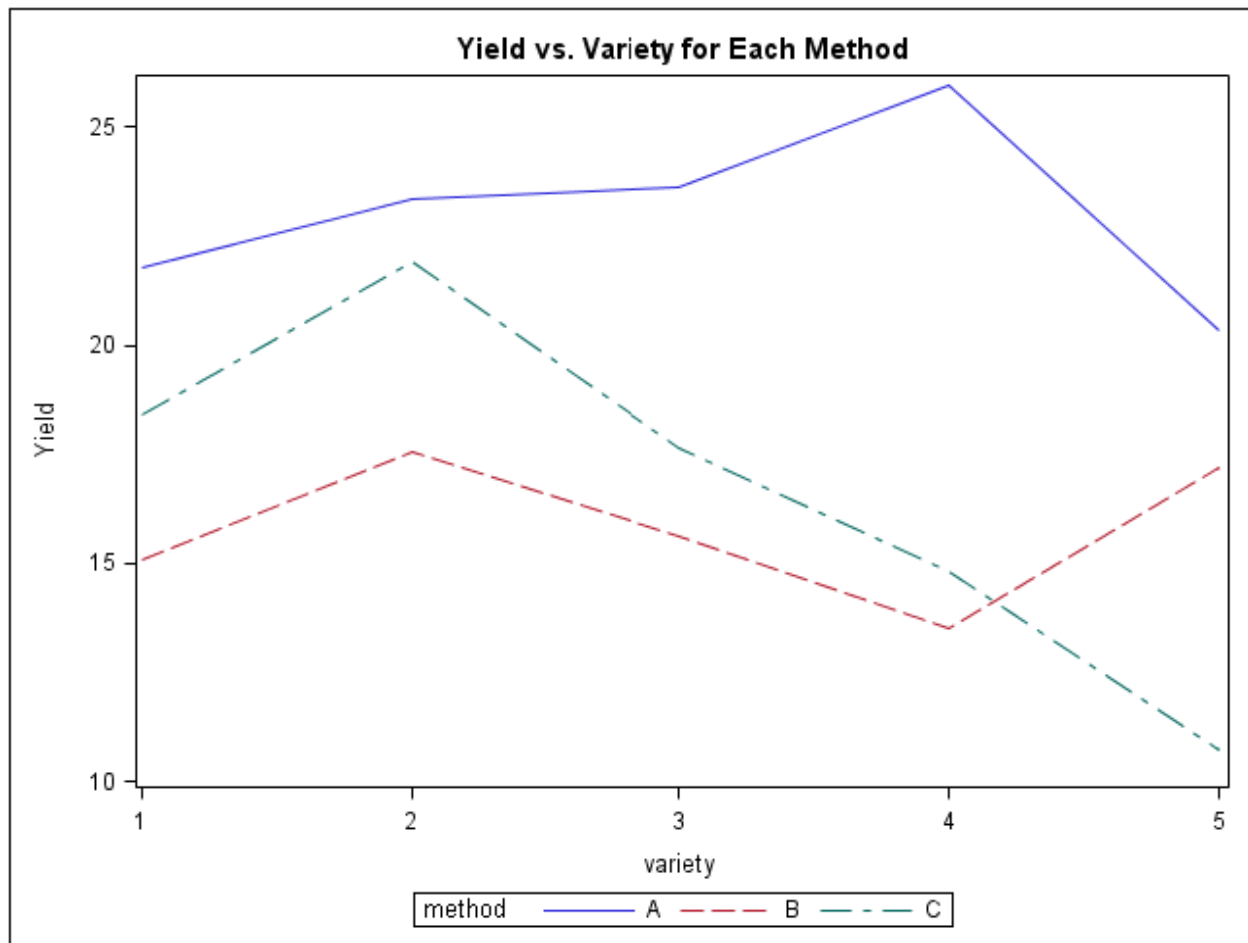
The Line Chart task generates code from the SGPLOT procedure.

```
proc sgplot data=aglm.grass;  
    vline variety / group=method stat=mean response=yield;  
    title 'Yield vs. Variety for Each Method';  
run;  
title;
```

The MEANS procedure was used to compute the mean yield values for each method and variety combination. The NWAY option in PROC MEANS specifies that the output data set contains only the statistics that correspond to the combination of all class variables, that is, every **method** and **variety** combination.



## PROC SGPLOT Output



The three lines, one for each level of **method**, are not parallel. This indicates a **method\*variety** interaction. Method A seems to produce the largest yield for all five varieties.

**End of Demonstration**



## Fitting the Two-Way Mixed Model

Now you can use the Mixed Model task to fit this two-way mixed model. Notice that the interaction term is a random effect if at least one factor involved in the interaction is random.

1. In the Tasks and Utilities area, expand **Statistics**. Double-click the **Mixed Models** task.
2. On the DATA tab, select the **grass** data set in the **AGLM** library.
3. Select **yield** as the dependent variable. Select **method** and **variety** as classification variables.
4. On the MODEL tab, click the **Edit** button under Fixed Effects. Select **method** and click **Add** to include it as a fixed effect in the model. Click **OK** to return to the task.
5. Click **Add Random**. Click the **Edit** button under Random Effects. Select **variety** and click **Add** to include it as a random effect in the model. Select both **variety** and **method**. Click **Cross** to add the interaction as a random effect in the model. Click **OK** to return to the task.
6. On the OPTIONS tab, expand the **Details** section. Select **Kenward and Roger's method 2** as the method to compute denominator degrees of freedom.
7. Run the task.

The included CONTRAST and ESTIMATE statements can be added to the generated code by clicking the **Edit** button above the generated code.

```
proc mixed data=aglm.grass;
  class method variety;
  model yield=method / ddfm=kr2;
  random variety method*variety;
  contrast 'A vs B and C' method 2 -1 -1;
  estimate 'A vs B and C' method 2 -1 -1 / divisor=2 cl alpha=0.02;
  estimate 'Method A mean' intercept 1 method 1 0 0;
run;
```

Selected MODEL statement option:

**DDFM=** specifies the method for computing the denominator degrees of freedom for the tests of fixed effects that result from the MODEL, CONTRAST, ESTIMATE, and LSMEANS statements. Possible values are CONTAIN, BETWITHIN, SATTERTH, RESIDUAL, KENWARDROGER/KR, KENWARDROGER(FIRSTORDER), and KENWARDROGER2/KR2.

The DDFM=KENWARDROGER2 option performs the degrees-of-freedom calculations detailed by Kenward and Roger (2009). This approximation involves inflating the estimated variance-covariance matrix of the fixed and random effects by the method proposed by Prasad and Rao (1990) and Harville and Jeske (1992). (Refer also to Kackar and Harville (1984).) Satterthwaite-type degrees of freedom are then computed based on this adjustment. KR2 is an improvement to the original method that was outlined by Kenward and Roger (1997). The correction under DDFM=KR2 reduces the bias of the precision estimator for the fixed effects under nonlinear covariance structures when using DDFM=KR.

**Note:** Studies show that the KENWARDROGER2 method of estimating the denominator degrees of freedom for the fixed effects is the most appropriate method in almost all cases. Therefore, it is a highly recommended DDFM method for your mixed model analysis.

Selected ESTIMATE statement option:

DIVISOR= specifies a value by which to divide all coefficients so that fractional coefficients can be entered as integers.

CL requests that *t*-type confidence limits be constructed. The confidence level is 0.95 by default. This can be changed with the ALPHA= option.

The CONTRAST statement was used to compare method A versus methods B and C. The null hypothesis is  $H_0: \mu_A = \frac{1}{2}(\mu_B + \mu_C)$ . The first ESTIMATE statement was used to estimate the difference in mean yield between method A and the average of methods B and C. The second ESTIMATE statement was used to compute the average yield for method A.

#### Partial PROC MIXED Output

The Mixed Procedure	
Model Information	
Data Set	AGLM.GRASS
Dependent Variable	yield
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Kenward-Roger2
Degrees of Freedom Method	Kenward-Roger2
Class Level Information	
Class	Levels      Values
method	3      A B C
variety	5      1 2 3 4 5
Dimensions	
Covariance Parameters	3
Columns in X	4
Columns in Z	20
Subjects	1
Max Obs Per Subject	90
Number of Observations	
Number of Observations Read	90
Number of Observations Used	90
Number of Observations Not Used	0

The Model Information, Class Level Information, Dimensions, and Number of Observations tables provide basic descriptive information about the data and model fitting.

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	528.89057283	
1	1	522.49142693	0.00000000
Convergence criteria met.			

The Iteration History table summarizes the iterative process of finding the REML estimates of the parameters.

Covariance Parameter Estimates	
Cov Parm	Estimate
variety	0.4285
method*variety	4.7715
Residual	18.4347

The variance component estimates obtained by REML are as follows:

$$\hat{\sigma}_b^2 = 0.4285$$

$$\hat{\sigma}_{ab}^2 = 4.7715$$

$$\hat{\sigma}^2 = 18.4347$$

**Note:** It can be shown that the **G** matrix is a diagonal matrix of dimension 20 by 20, with five diagonal elements of 0.4285 and 15 diagonal elements of 4.7715. The **R** matrix is a diagonal matrix of dimension 90 by 90, with the diagonal element of 18.4347. The **V** matrix follows the pattern described below.

$$\begin{bmatrix} \mathbf{V}_1 & \mathbf{0} & \dots & \mathbf{C}_1 & \mathbf{0} & \dots & \mathbf{C}_1 & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{V}_2 & \mathbf{0} & \dots & \mathbf{C}_2 & \mathbf{0} & \dots & \mathbf{C}_2 & \dots \\ \vdots & & \ddots & & & \ddots & & & \mathbf{0} \\ \mathbf{C}_1 & \mathbf{0} & \dots & \mathbf{V}_6 & \mathbf{0} & \dots & \mathbf{C}_6 & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{C}_2 & & & \mathbf{V}_7 & & & \mathbf{C}_7 & \\ \vdots & & \ddots & & & \ddots & & & \ddots \\ \mathbf{C}_1 & & & \mathbf{C}_6 & & & \mathbf{V}_{11} & & \\ \mathbf{0} & \mathbf{C}_2 & & & \ddots & & & \ddots & \\ \vdots & & \ddots & & & \mathbf{C}_{10} & & & \mathbf{V}_{15} \end{bmatrix}_{90 \times 90}$$

where  $\mathbf{V}_1 = \mathbf{V}_2 = \dots = \mathbf{V}_{15}$

$$= \begin{bmatrix} \sigma^2 + \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 \\ \sigma_b^2 + \sigma_{ab}^2 & \sigma^2 + \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 \\ \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma^2 + \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 \\ \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma^2 + \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 \\ \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma^2 + \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 \\ \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma_b^2 + \sigma_{ab}^2 & \sigma^2 + \sigma_b^2 + \sigma_{ab}^2 \end{bmatrix}_{6 \times 6}$$

$$\text{and } \mathbf{C}_1 = \mathbf{C}_2 = \dots = \mathbf{C}_{10} = \begin{bmatrix} \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \end{bmatrix}_{6 \times 6}$$

You can use the G and V options in the RANDOM statement to request the **G** and **V** matrices from the MIXED procedure.

Fit Statistics	
-2 Res Log Likelihood	522.5
AIC (smaller is better)	528.5
AICC (smaller is better)	528.8
BIC (smaller is better)	527.3

The Fit Statistics table provides information for goodness of model fit. All three criteria consider both the fitness of the model and the model complexity. These statistics are useful when you compare different models. Smaller values typically indicate a better model fit.

Type 3 Tests of Fixed Effects								
Effect			Num DF	Den DF	F Value	Pr > F		
method			2	8	9.84	0.0070		
Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr >  t	Alpha	Lower	Upper
A vs B and C	6.7583	1.5340	8	4.41	0.0023	0.02	2.3151	11.2015
Method A mean	23.0100	1.2863	11.9	17.89	<.0001	0.05	20.2058	25.8142
Contrasts								
Label			Num DF	Den DF	F Value	Pr > F		
A vs B and C			1	8	19.41	0.0023		

The Type 3 test for the fixed effect shows that the  $F$  value for **method** produced by PROC MIXED is 9.84, with a  $p$ -value of 0.0070. The **method** means, averaged across all varieties, are statistically, significantly different for at least one of the methods.

The estimate for **A** versus **B** and **C** is 6.7583, and the standard error of the estimate is 1.5340. The average yield produced by **method A** is 6.7583 higher than the average **yield** produced by **B** and **C**. This difference is significant ( $p$ -value=0.0023). The 98% confidence limits for this difference are [2.3151, 11.2015]. The estimate for the **method A** mean is 23.01, the standard error is 1.2863, and the 95% confidence limits for the **method A** mean is [20.2058, 25.8142]. Notice that the confidence limits were generated for both ESTIMATE statements, although the CL option was specified in only one of the ESTIMATE statements. However, you must specify the ALPHA= option in each CONTRAST or ESTIMATE statement if you want to change the significance level.

The contrast for **A** versus **B** and **C** agrees with the result from the ESTIMATE statement.

**End of Demonstration**

## 2.2 Analysis of Covariance with Random Effects

### Objectives

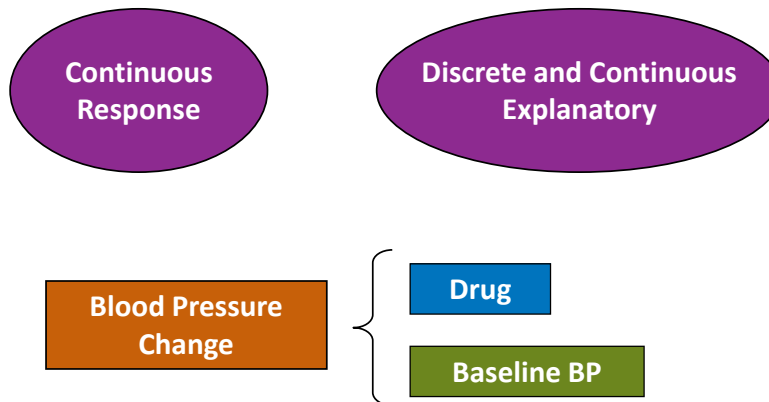
- Use the Mixed Models Task to perform an analysis of covariance.
- Interpret the parameter estimates from an analysis of covariance.

12

Copyright © SAS Institute Inc. All rights reserved.



### Analysis of Covariance

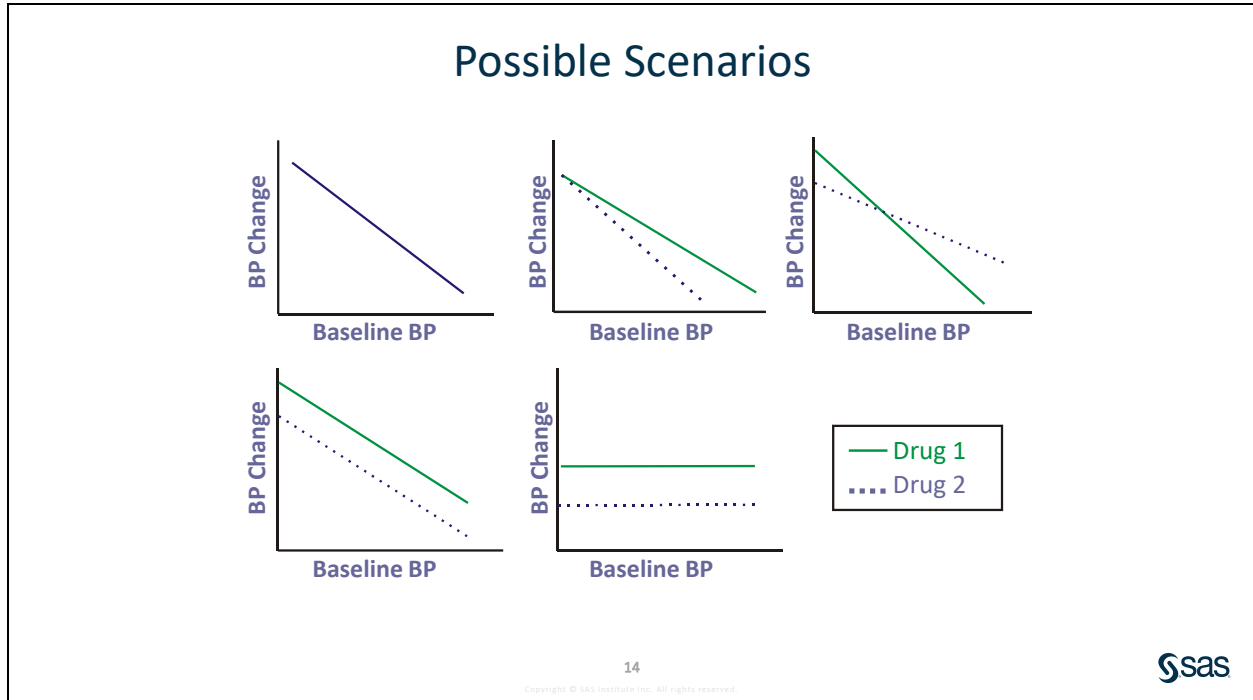


13

Copyright © SAS Institute Inc. All rights reserved.



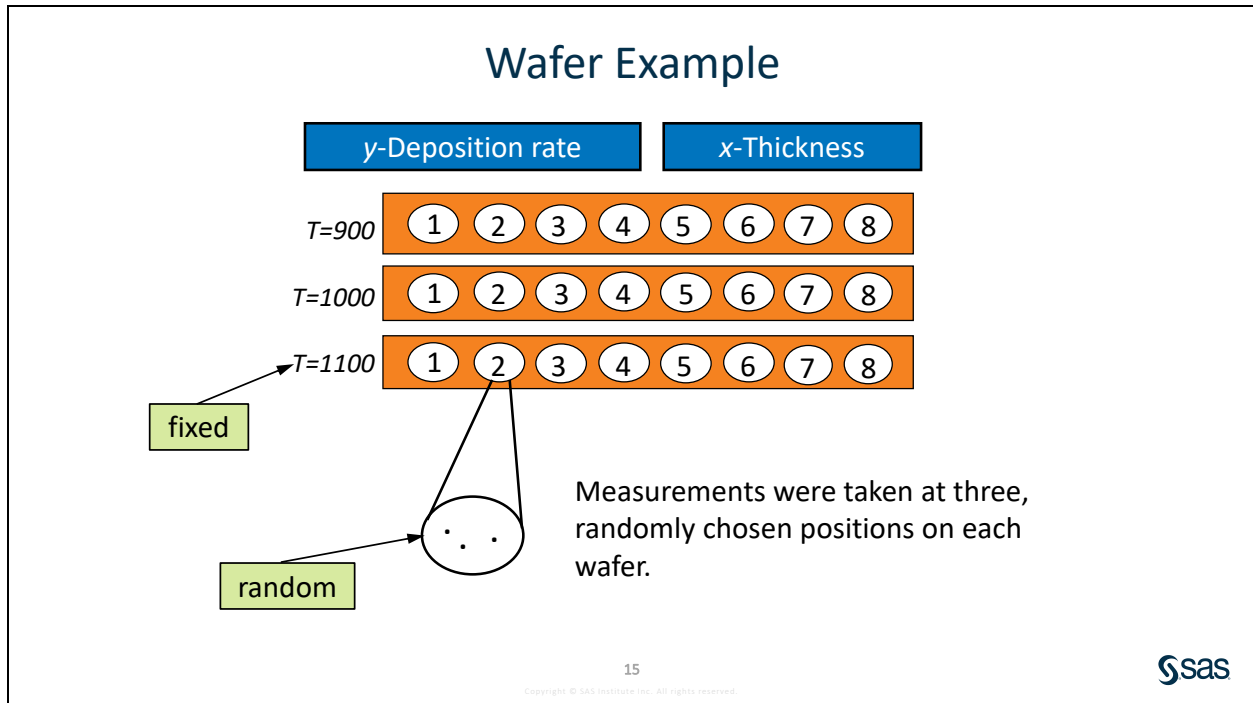
*Analysis of covariance* refers to an analysis where the response variable is continuous, and the independent variables include one or more continuous variables in addition to classification variables. The goal of the analysis is often to examine the treatment effects considering the variability that is associated with the continuous variables, called *covariates*. A more global view of analysis of covariance describes it as a methodology to compare a series of regression models.



Suppose two drugs were evaluated for the effect of reducing blood pressure. A baseline blood pressure was taken. Then changes in blood pressure, after administering one of the two drugs, were measured for each subject. The relationship between the response variable, the treatment, and the covariate can be one of the following possible scenarios:

- The slopes and intercepts for the treatments are the same.
- The slopes are different, but the intercepts are the same.
- The slopes and intercepts are different.
- The intercepts are different, but the slopes are the same.
- The intercepts are different, but all slopes are zero (a special case of the previous scenario).





The **aglm.wafer4** data set was obtained from the semiconductor industry. The experiment was designed to study the effect of temperature on the deposition rate of a layer of polysilicon in the fabrication of wafers. It was thought that the wafer thickness before the deposition process was applied might influence the deposition rate. Therefore, the average thickness of each wafer (**thick**) was determined and used as a possible covariate.

A random sample of 24 wafers was collected and used in the experiment. Wafers were randomly assigned to one of the three levels of temperature (900°F, 1000°F, and 1100°F). Thus, eight wafers were assigned to each level of temperature. The amounts of deposited material from each wafer at three randomly chosen sites were measured.

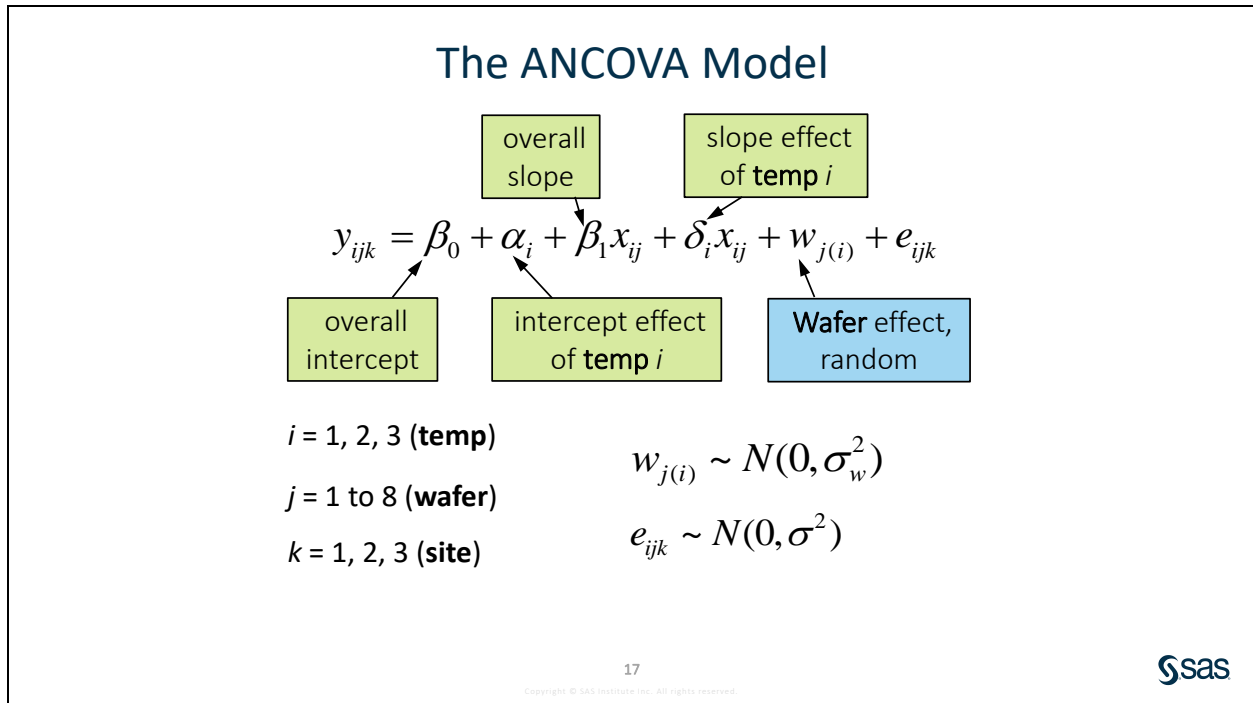
## The Data

temp	wafer	site	deposit	thick
900	1	1	291	1919
900	1	2	295	1919
900	1	3	294	1919
900	2	1	318	2113
900	2	2	315	2113
900	2	3	315	2113
900	3	1	306	1841
900	3	2	302	1841
900	3	3	305	1841
900	4	1	342	2170
900	4	2	341	2170
900	4	3	336	2170
900	5	1	318	2019
900	5	2	323	2019
900	5	3	323	2019
900	6	1	307	1872
900	6	2	308	1872
900	6	3	308	1872
900	7	1	295	1862
900	7	2	297	1862
...	...	...	...	...



The following variables are in the **aglm.wafer4** data set:

- temp**      temperature (900°F, 1000°F, and 1100°F)
- wafer**     wafers randomly selected and assigned to one of the three temperatures
- site**       sites on each wafer where the response measurements were taken (1, 2, and 3)
- deposit**   the amount of deposited material at each site
- thick**      the average thickness of each wafer before the deposition process



- $y_{ijk}$  the deposition rate for the  $k^{\text{th}}$  site from the  $j^{\text{th}}$  wafer assigned to the  $i^{\text{th}}$  temperature.
- $\beta_0$  the overall intercept.
- $\alpha_i$  the coefficient for the  $i^{\text{th}}$  temperature effect of the intercept.
- $\beta_1$  the overall slope.
- $\delta_i$  the coefficient for the  $i^{\text{th}}$  temperature effect of the slope.
- $x_{ij}$  the covariate **thick** measured on the  $j^{\text{th}}$  wafer assigned to the  $i^{\text{th}}$  temperature.
- $w_{j(i)}$  the wafer effect,  $w_{j(i)} \sim \text{i.i.d. } N(0, \sigma_w^2)$ . This random effect is identified by **wafer** nested within **temp**.
- $e_{ijk}$  the site effect,  $e_{ijk} \sim \text{i.i.d. } N(0, \sigma^2)$ . This term corresponds to the random error.



## Fitting an ANCOVA Model

---

First, use the Scatter Plot task to plot the data.

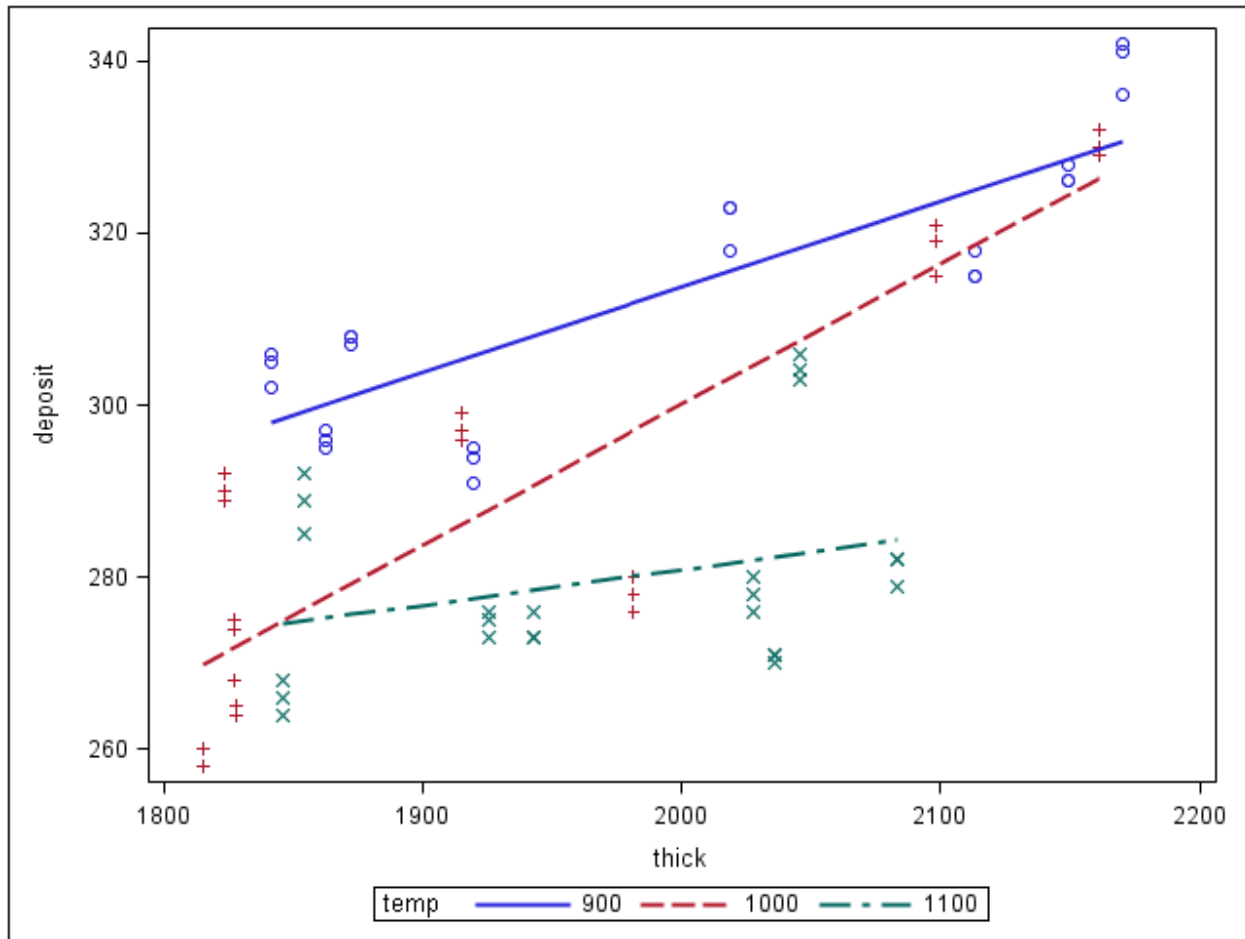
1. Expand the **Graph** area under Tasks within Tasks and Utilities.
2. Double-click the **Scatter Plot** task to launch it.
3. On the DATA tab, select the **wafer4** data set in the **AGLM** library.
4. Select **thick** as the X variable, **deposit** as the Y variable, and **temp** as the group variable.
5. Expand **FIT PLOTS**. Select the check box next to **Regression**.
6. Run the task.

The Scatter Plot task generates code from the SGPLOT procedure.

```
proc sgplot data=aglm.wafer4;  
    reg y=deposit x=thick / group=temp;  
run;
```

The REG statement in PROC SGPLOT creates a fitted regression line or curve. The GROUP= option specifies a variable that is used to group the data. A separate plot is created for each unique value of the grouping variable. The plot elements for each group value are automatically distinguished by different visual attributes.

## PROC SGPLOT Output



There seems to be a positive relationship between the deposition rate and the thickness of the wafers across all three temperatures. On average, the deposition rate at 900°F seems to be higher than the deposition rate at other temperatures within the range of the data. The slopes do not seem to be the same across the three temperatures.

To implement the model discussed above in the Mixed Model task, you specify **temp** for the term  $\alpha_i$ , **thick** for the term  $\beta_1$ , and **thick\*temp** for the term  $\delta_i$ .

1. In the Tasks and Utilities area, expand **Statistics**. Double-click the **Mixed Models** task.
2. On the DATA tab, select the **wafer4** data set in the **AGLM** library.
3. Select **deposit** as the dependent variable, and select **temp** and **wafer** as classification variables. Select **thick** as a continuous variable.
4. On the MODEL tab, click the **Edit** button under Fixed Effects. Select **temp** and **thick**. Click **Two-way Factorial** to include each main and the interaction effect as a fixed effect in the model. Click **OK** to return to the task.
5. Click **Add Random**. Click the **Edit** button under Random Effects. Select **temp** and **wafer** and click **Nest**. Place **wafer** as the outer term and **temp** to be nested within outer. Click **Add**. Then click the **x** to return to the Random Effects Builder window. Click **OK** to return to the task.

6. On the OPTIONS tab, expand the **Details** section. Select **Kenward and Roger's method 2** as the method to compute denominator degrees of freedom.
7. Select the default and additional statistics under Select statistics to display. Expand the **Parameter Estimates** section. Under Fixed Effects, select the check box next to **Show parameter estimates**.
8. Run the task.

```
proc mixed data=aglm.wafer4;
  class temp wafer;
  model deposit=temp thick thick*temp / ddfm=kr2 solution;
  random wafer(temp);
run;
```

Selected MODEL statement option:

SOLUTION requests that a solution for the fixed-effects parameters be produced.

Partial PROC MIXED Output

Class Level Information						
Class	Levels	Values				
temp	3	900	1000	1100		
wafer	8	1	2	3	4	5 6 7 8

The variable **thick** is a continuous variable, so it is not listed in this table.

Covariance Parameter Estimates		
Cov Parm	Estimate	
wafer(temp)	132.54	
Residual	4.1944	

The estimated wafer-to-wafer variance is 132.54. The residual variance (the site-to-site variance) is 4.1944.

Solution for Fixed Effects						
Effect	temp	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept		199.07	94.8319	18	2.10	0.0502
temp	900	-84.6703	114.31	18	-0.74	0.4684
temp	1000	-225.80	113.35	18	-1.99	0.0618
temp	1100	0	.	.	.	.
thick		0.04091	0.04809	18	0.85	0.4061
thick*temp	900	0.05879	0.05774	18	1.02	0.3221
thick*temp	1000	0.1225	0.05781	18	2.12	0.0483
thick*temp	1100	0	.	.	.	.

The SOLUTION option produces the estimates for the fixed effect parameters. Because this is an over-parameterized model, which means more parameters need to be estimated (8) than there are independent pieces of information (6), PROC MIXED (and many other SAS procedures) set the estimate of the last level of factors, **temp** and **thick\*temp**, at zero. Therefore, the following conditions occur:

- The intercept term corresponds to the intercept for the last level of the group variable (in this case, **temp 1100°F**). The estimated intercept coefficient for **temp 1100°F** is 199.07.
- The estimate for **temp 900°F** corresponds to  $\alpha_1 - \alpha_3$ . It is the difference between the intercept coefficients (effects) for **temp 900°F** and **temp 1100°F**. The nonsignificant *p*-value indicates that these two intercept coefficients are **not** different from each other. The coefficient for the **temp 900°F** intercept is  $199.07 - 84.6703 = 114.40$ .
- The estimate for **temp 1000°F** corresponds to  $\alpha_2 - \alpha_3$ . It is the difference between the intercept coefficients (effects) for **temp 1000°F** and **temp 1100°F**. The marginally significant *p*-value indicates that these two intercept coefficients might be different from each other. The coefficient for the **temp 1000°F** intercept is  $199.07 - 225.80 = -26.73$ .
- The estimate for **thick** corresponds to the slope for the last level of the group variable (in this case, **temp 1100°F**). The estimated slope for **temp 1100°F** is 0.04091.
- The estimate for **thick\*temp 900°F** corresponds to the difference between the slope coefficients for **temp 900°F** and **temp 1100°F**. The nonsignificant *p*-value indicates that these two slope coefficients are **not** different from each other. The slope for **temp 900°F** is  $0.04091 + 0.05879 = 0.0997$ .
- The estimate for **thick\*temp 1000°F** corresponds to the difference between the slope coefficients for **temp 1000°F** and **temp 1100°F**. The significant *p*-value indicates that these two slope coefficients are different from each other. The slope for **temp 1000°F** is  $0.04091 + 0.1225 = 0.1634$ .

The three regression lines are as follows:

- For **temp 900°F**, **deposit** =  $(199.07 - 84.6703) + (0.04091 + 0.05879) * \text{thick}$   
=  $114.40 + 0.0997 * \text{thick}$ .
- For **temp 1000°F**, **deposit** =  $(199.07 - 225.80) + (0.04091 + 0.1225) * \text{thick}$   
=  $-26.73 + 0.1634 * \text{thick}$ .
- For **temp 1100°F**, **deposit** =  $199.07 + 0.0409 * \text{thick}$ .

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
temp	2	18	2.38	0.1209
thick	1	18	21.18	0.0002
thick*temp	2	18	2.44	0.1155

The test for **thick\*temp** corresponds to the differential effects of **thick** at different levels of **temp**. In other words, it tests whether the slopes are equal across all three levels of temperatures. The nonsignificant *p*-value (0.1155) indicates that the slopes are not significantly different from each other at  $\alpha=0.05$ . There might not be enough evidence to warrant an unequal slope model.

However, in the Solutions for Fixed Effects table, you found a significant difference in the slopes between temperatures 1000°F and 1100°F. Some statisticians feel that if the overall  $F$  test is nonsignificant, then you might not examine the individual tests. Others feel that the individual test might be more powerful than the overall test in some situations and that you also need to examine the graph.

You might decide to keep the unequal slope model, given the graph and the solutions results. Alternatively, if you decide that the evidence to support an unequal slope model is not enough ( $p$ -value=0.1155), you might want to remove the **thick\*temp** interaction term and fit a common slope model.

1. In the task area, click the **MODEL** tab.
2. Under Fixed Effects, click the **Edit** button to enter the effect builder.
3. Under Model Effects, on the right, select the interaction term and click the **Trash Can** icon. Confirm the deletion on the menu that appears. This removes the interaction from the model. Click **OK** to return to the task.
4. Run the task.

```
proc mixed data=aglm.wafer4;
  class temp wafer;
  model deposit=temp thick / solution ddfm=kr2;
  random wafer(temp);
run;
```

#### Partial PROC MIXED Output

		Covariance Parameter Estimates	
	Cov Parm	Estimate	
	wafer(temp)	151.82	
	Residual	4.1944	

The covariance parameter estimate for **wafer(temp)** is 151.82. This is different from the estimates in the previous model (132.54). You specified a different mean model and that likely changes the parameter estimates for the fixed effects as well.

Solution for Fixed Effects						
Effect	temp	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept		53.0989	43.3872	20	1.22	0.2352
temp	900	30.7988	6.2097	20	4.96	<.0001
temp	1000	13.6520	6.2477	20	2.19	0.0410
temp	1100	0	.	.	.	.
thick		0.1150	0.02191	20	5.25	<.0001



The three regression lines are as follows:

- For **temp 900°F**, **deposit** = (30.7988 + 53.0989) + 0.1150 \* **thick**  
= 83.8977 + 0.1150 \* **thick**.
- For **temp 1000°F**, **deposit** = (13.6520 + 53.0989) + 0.1150 \* **thick**  
= 66.7509 + 0.1150 \* **thick**.
- For **temp 1100°F**, **deposit** = 53.0989 + 0.1150 \* **thick**.

Type 3 Tests of Fixed Effects					
Effect	Num DF	Den DF	F Value	Pr > F	
temp	2	20	12.33	0.0003	
thick	1	20	27.55	<.0001	

The tests for the fixed effects suggest that the common slope (indicated by **thick**) significantly differs from zero ( $p$ -value < 0.0001).

**Note:** The intercept estimate for each temperature can also be obtained by specifying ESTIMATE statements in PROC MIXED. This can be added to the generated code by clicking the **Edit** button above the code area.

For example, the following code produces the intercept coefficient for each temperature:

```
ods select estimates;
proc mixed data=aglm.wafer4;
  class temp wafer;
  model deposit=temp thick / ddfm=kr2;
  random wafer(temp);
  estimate 'Intercept for temp 900' intercept 1 temp 1 0 0;
  estimate 'Intercept for temp 1000' intercept 1 temp 0 1 0;
  estimate 'Intercept for temp 1100' intercept 1 temp 0 0 1;
run;
```

Related Output

Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept for temp 900	83.8977	43.8914	20	1.91	0.0704
Intercept for temp 1000	66.7510	42.5371	20	1.57	0.1323
Intercept for temp 1100	53.0989	43.3872	20	1.22	0.2352

**End of Demonstration**

## 2.3 Introduction to Repeated Measures Analysis

---

### Objectives

- Describe basic issues that occur in repeated measures analysis.
- List the advantages of the mixed model approach.
- Produce a group profile plot for repeated measures data.

21

Copyright © SAS Institute Inc. All rights reserved.



### Basic Issues in Repeated Measures Data

- *Repeated measures* refer to multiple measurements on the same experimental unit (or subject).
- Measures taken on the same subject tend to be more similar than measures taken on different subjects.
- Measures made close in time on the same subject tend to be more highly correlated than measures made far apart in time.
- The analysis of repeated measures data accounts for the presence of correlation between the observations that are obtained on the same subject and for possible nonconstant variances.

22

Copyright © SAS Institute Inc. All rights reserved.



Usually, repeated measures are made over time, but they can be over space as well. A common example is to apply the treatments to experimental units in a random fashion. Measurements are taken at each of several times. The basic objectives for repeated measures data are essentially those of any non-repeated measures data, that is, to examine interaction between factors and to assess either the simple or main effects of the factors. The distinguishing feature of a repeated measures model is in the variance and covariance structure of the data. That is, the independent and homogeneous error variances are no longer valid assumptions. This is because the time points are not randomly assigned to subjects. Therefore, it is not reasonable to assume that the random errors  $\varepsilon_{ijk}$  are independent.

There are often two aspects of the covariance structure in the errors.

- First, two measures on the same subject are more likely to be closer to each other than two measures on different subjects. They are positively correlated simply because they contain common effects from that same subject. This is basically the phenomenon of measures on the same whole-plot unit in a split-plot type of experiment.
- Second, two measures made close in time on the same subject are likely to be more highly correlated than two measures made far apart in time. This distinguishes a repeated measures covariance from a split-plot covariance structure. In a split-plot experiment, levels of subplot treatment are randomly assigned to subplot units within whole-plot units. This results in equal correlation between all pairs of measurements in the same whole-plot unit.

The analysis of repeated measures data accounts for correlations between the observations that are obtained on the same subject, and also, some possible heterogeneous variances among observations that are obtained on the same subject.

## Traditional Approaches

- Univariate ANOVA using GLM
  - treats repeated measures as split-plot in time
  - assumes equal correlations regardless of distance in time
  - risks a higher Type I error rate for fixed effect tests
- Multivariate ANOVA using GLM
  - requires that data points are the same across subjects
  - uses the complete case analysis
  - avoids modeling the covariance structures, which potentially causes the tests to be less powerful

Two traditional methods are the *univariate ANOVA* and the *multivariate ANOVA* using the GLM procedure. The univariate ANOVA approach basically treats repeated measures data as split-plot data. That is, the experimental units to which treatments are assigned are considered as the whole-plot units, and the experimental units at a given time are considered as the subplot units. Thus, this approach accommodates the first aspect of repeated measures covariance (between-subject variation), but not the second. (Measures close in time are more highly correlated than measures far apart in time.) This approach is sometimes referred to as a *split-plot in time ANOVA*.

The univariate ANOVA approach assumes that differences between each pair of measures have equal variances. This condition, also called the *Huynh-Feldt (H-F) condition*, is a necessary condition for univariate ANOVA to be valid. The more commonly used covariance structure, the *compound symmetry (CS)* structure, assumes that the covariances of the repeated measures of the same subject obey the following conditions:

- Measures at each time have equal variances.
- The correlations between any two measures are the same.

This CS covariance structure is a special case of the H-F condition. Although the H-F condition is mathematically more general than CS, it is often not more general from a practical point of view.

However, the equal correlation assumption is frequently unrealistic for repeated measures data. Measurements close in time are often more highly correlated than measurements far apart in time. Therefore, this approach risks underestimating the standard errors of mean comparisons at different times. This results in an excessive Type I error rate.

The multivariate ANOVA, or analysis of contrasts, computes one or more linear combinations of data on each of the subjects, and then analyzes the linear combination as data. One example of the linear combination is the contrast between levels of the factor (often **time**) and a reference level.

Multivariate ANOVA requires balanced data with the same time points for all subjects. The number of repeated measures on each subject is the same, and measurements occur at the same time points for all subjects. It uses the complete case analysis. That is, if one repeated measure is missing for a subject, all the data for this subject are eliminated from the analysis. For unbalanced data, this might result in data that are not feasible for a meaningful analysis. In addition, if repeated measures are not obtained at the same time points across all subjects (which occurs in most observational studies), the complete case requirement probably results in less data or even no data that are available for the analysis.

In addition, multivariate ANOVA is a device for avoiding the covariance problem in repeated measures analysis. The method does not directly accommodate the covariance structure. It is based on an unstructured within-subject variance-covariance matrix, whose parameter estimates are obtained by the method of moments. A simpler covariance structure is probably sufficient to characterize the repeated measures data. In this case, you sacrifice power and efficiency for the fixed effects tests.

## Advantages of Mixed Model Analyses

- Mixed model analyses enable a flexible approach to modeling covariance structures.
- Mixed model analyses handle unbalanced data with unequally spaced time points within and across subjects.
- The generalized least squares (GLS) method is generally superior to the ordinary least squares (OLS) method, if you assume that an appropriate covariance structure exists.
- The presence of missing data poses less of a problem for the MIXED approach than for the multivariate ANOVA approach.

24

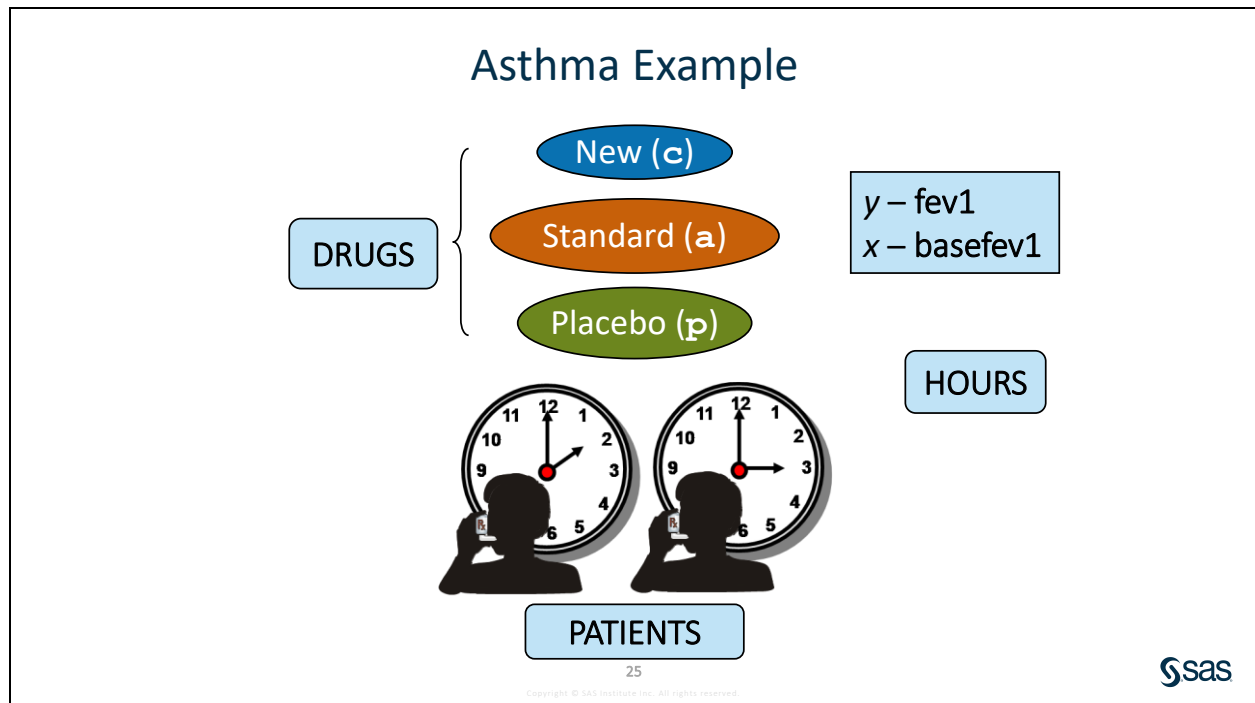
Copyright © SAS Institute Inc. All rights reserved.



The mixed model procedure enables a flexible approach to modeling covariance structure. It results in a parsimonious covariance model that adequately accounts for within-subject correlations. Therefore, it is superior to other approaches. In addition, the MIXED procedure uses generalized least squares (GLS) to estimate and test fixed effects. If reasonable covariance structure is specified, this method accounts for all the covariance parameters that you modeled for your data. Therefore, it is superior to the method used by PROC GLM, that is, ordinary least squares (OLS).

PROC MIXED provides a variety of modeling techniques to handle unequally spaced time points within each subject or subjects that are observed at different sets of time points.

In the presence of missing repeated measures for a subject, the MIXED procedure does not exclude this subject from the analysis. Instead, it uses all the available data. This method (likelihood-based ignorable analysis) leads to a valid analysis when the missing data can be assumed to be missing at random (MAR). For more information about missing data, MAR and MCAR (missing completely at random), refer to Little and Rubin (1987).



A pharmaceutical company wants to examine the effects of three drugs on the respiratory ability of asthma patients. The three drugs are labeled **a**, **c**, and **p**. Drug **a** is a standard drug that is used to treat asthma. Drug **c** is a potential competitor that was developed by the pharmaceutical company. Drug **p** is a placebo.

Each of the three drugs is randomly assigned to 24 patients. A total of 72 patients are included in the study. The assigned drug is administered to each patient. Then a standard measure of respiratory ability called **fev1** (forced exhaled volume in one second) is measured hourly for eight hours after treatment. Also, a baseline **fev1** is measured immediately before administering a drug.

The data are stored in two SAS data sets, **aglm.fev1mult** and **aglm.fev1uni**.

- The **aglm.fev1mult** data set is in a multivariate, or wide, arrangement. It is used to illustrate the multivariate analysis of contrasts using the REPEATED statement in the GLM procedure.
- The **aglm.fev1uni** data set contains the **fev1** data in a univariate, or long, arrangement. It is used to illustrate the univariate ANOVA and mixed model approaches to the analysis of repeated measures data.

(The programs are presented in an appendix.)



## Producing Profile Plots for the Three Drugs

In this demo, we will look directly at code. These variables are in **aglm.fev1mult**:

<b>patient</b>	patient identification number
<b>basefev1</b>	baseline <b>fev1</b> measurement taken before administering the treatment
<b>fev11h</b>	<b>fev1</b> measurement one hour after administering the drug
<b>fev12h</b>	<b>fev1</b> measurement two hours after administering the drug
<b>fev13h</b>	<b>fev1</b> measurement three hours after administering the drug
<b>fev14h</b>	<b>fev1</b> measurement four hours after administering the drug
<b>fev15h</b>	<b>fev1</b> measurement five hours after administering the drug
<b>fev16h</b>	<b>fev1</b> measurement six hours after administering the drug
<b>fev17h</b>	<b>fev1</b> measurement seven hours after administering the drug
<b>fev18h</b>	<b>fev1</b> measurement eight hours after administering the drug
<b>drug</b>	drug administered ( <b>a</b> , <b>c</b> , or <b>p</b> )

These variables are in **aglm.fev1uni**:

<b>patient</b>	patient identification number
<b>basefev1</b>	baseline <b>fev1</b> measurement taken before administering the treatment
<b>drug</b>	the drug administered ( <b>a</b> , <b>c</b> , or <b>p</b> ) to a patient
<b>hour</b>	number of hours that the measurement was taken after administering the drug
<b>fev1</b>	<b>fev1</b> measurement

```
proc format;
  value $drug
    'a'='a-Standard'
    'c'='c-New'
    'p'='p-Placebo'
;
proc print data=aglm.fev1mult;
  format drug $drug.;
proc print data=aglm.fev1uni;
  format drug $drug.;
run;
```

The PROC FORMAT statement formats each of the drugs. The PROC PRINT statements create listing outputs for the two data sets, **aglm.fev1mult** and **aglm.fev1uni**.

Partial PROC PRINT Output for **aglm.fev1mult** (a multivariate data structure)

Obs	patient	basefev1	fev11h	fev12h	fev13h	fev14h	fev15h	fev16h	fev17h	fev18h	drug
1	201	2.46	2.68	2.76	2.50	2.30	2.14	2.40	2.33	2.20	a-Standard
2	202	3.50	3.95	3.65	2.93	2.53	3.04	3.37	3.14	2.62	a-Standard
3	203	1.96	2.28	2.34	2.29	2.43	2.06	2.18	2.28	2.29	a-Standard
4	204	3.44	4.08	3.87	3.79	3.30	3.80	3.24	2.98	2.91	a-Standard
5	205	2.80	4.09	3.90	3.54	3.35	3.15	3.23	3.46	3.27	a-Standard
6	206	2.36	3.79	3.97	3.78	3.69	3.31	2.83	2.72	3.00	a-Standard
7	207	1.77	3.82	3.44	3.46	3.02	2.98	3.10	2.79	2.88	a-Standard
8	208	2.64	3.67	3.47	3.19	2.19	2.85	2.68	2.60	2.73	a-Standard
9	209	2.30	4.12	3.71	3.57	3.49	3.64	3.38	2.28	3.72	a-Standard
10	210	2.27	2.77	2.77	2.75	2.75	2.71	2.75	2.52	2.60	a-Standard
11	211	2.44	3.77	3.73	3.67	3.56	3.59	3.35	3.32	3.18	a-Standard
12	212	2.04	2.00	1.91	1.88	2.09	2.08	1.98	1.70	1.40	a-Standard
13	214	2.77	3.36	3.42	3.28	3.30	3.31	2.99	3.01	3.08	a-Standard
14	215	2.96	4.31	4.02	3.38	3.31	3.46	3.49	3.38	3.35	a-Standard
15	216	3.11	3.88	3.92	3.71	3.59	3.57	3.48	3.42	3.63	a-Standard
16	217	1.47	1.97	1.90	1.45	1.45	1.24	1.24	1.17	1.27	a-Standard
17	218	2.73	2.91	2.99	2.87	2.88	2.84	2.67	2.69	2.77	a-Standard
18	219	3.25	3.59	3.54	3.17	2.92	3.48	3.05	3.27	2.96	a-Standard
19	220	2.73	2.88	3.06	2.75	2.71	2.83	2.58	2.68	2.42	a-Standard
20	221	3.30	4.04	3.94	3.84	3.99	3.90	3.89	3.89	2.98	a-Standard

Partial PROC PRINT Output for **aglm.fev1uni** (a univariate data structure)

Obs	patient	basefev1	drug	hour	fev1
1	201	2.46	a-Standard	1	2.68
2	201	2.46	a-Standard	2	2.76
3	201	2.46	a-Standard	3	2.50
4	201	2.46	a-Standard	4	2.30
5	201	2.46	a-Standard	5	2.14
6	201	2.46	a-Standard	6	2.40
7	201	2.46	a-Standard	7	2.33
8	201	2.46	a-Standard	8	2.20
9	202	3.50	a-Standard	1	3.95
10	202	3.50	a-Standard	2	3.65
11	202	3.50	a-Standard	3	2.93
12	202	3.50	a-Standard	4	2.53
13	202	3.50	a-Standard	5	3.04
14	202	3.50	a-Standard	6	3.37
15	202	3.50	a-Standard	7	3.14
16	202	3.50	a-Standard	8	2.62
17	203	1.96	a-Standard	1	2.28
18	203	1.96	a-Standard	2	2.34
19	203	1.96	a-Standard	3	2.29
20	203	1.96	a-Standard	4	2.43
21	203	1.96	a-Standard	5	2.06
22	203	1.96	a-Standard	6	2.18
23	203	1.96	a-Standard	7	2.28
24	203	1.96	a-Standard	8	2.29
25	204	3.44	a-Standard	1	4.08
26	204	3.44	a-Standard	2	3.87
27	204	3.44	a-Standard	3	3.79
28	204	3.44	a-Standard	4	3.30



You can submit the program below to produce the group profile plot for the three drugs. A *group profile plot* is a plot of the mean responses for different treatments versus the time. This type of plot can assist you in visualizing whether the treatments, the time, or both have a significant effect on the mean response variable.

```
proc sgplot data=aglm.fevluni;
  vline hour / group=drug stat=mean response=fev1;
  title 'Average FEV1 vs. Hour by Drug';
run;
title;
```

Selected PROC SGPLOT statement:

VLINE *category-variable* </option(s)> creates a vertically stacked line graph.

Selected VLINE arguments:

Required Argument:

*category-variable* specifies the variable whose variable determines the categories of data represented by the lines.

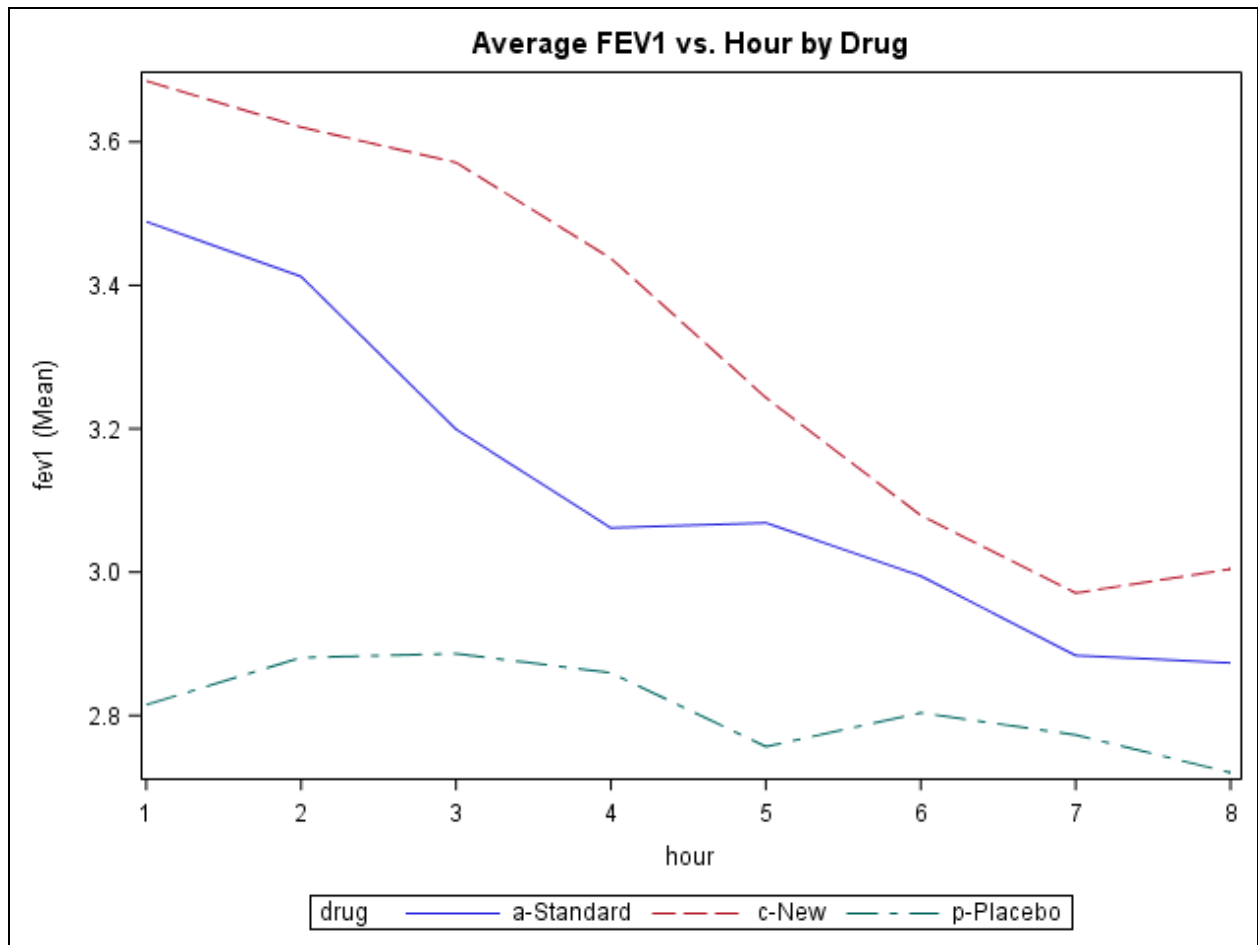
Optional Arguments:

GROUP=*variable* specifies a variable that is used to group the data.

RESPONSE=*response=variable*  
specifies a numeric response variable for the plot. The summarized values of the response variable are displayed on the vertical axis.

STAT=  
specifies the statistic for the vertical axis. Choices include *freq*, *mean*, *median*, *percent*, and *sum*.

## PROC SGPLOT Output



The plot shows that the average **fev1** measurements of the patients differ across the three drugs. There appears to be a time effect for at least drugs **a** and **c**.

**End of Demonstration**

## Four-Step Procedure for Mixed Model Analyses

1. Model the mean structure, usually by specification of the fixed effects.
2. Specify the covariance structures for between-subject and within-subject effects.
3. Fit the mean model. Using GLS, allow for the covariance structure. This step might include making the mean model more parsimonious.
4. Make statistical inferences based on the results of Step 3. This step might also include making the mean model more parsimonious.

27

Copyright © SAS Institute Inc. All rights reserved.



As recommended by Littell, Pendergast, and Natarajan (2000), you generally follow the four steps stated above when you use PROC MIXED to analyze repeated measures data. Another excellent reference is Verbeke and Molenberghs (2000).



# Lesson 3      LMIXED Procedure in SAS Viya

<b>3.1</b>	<b>LMIXED Procedure in SAS Viya.....</b>	<b>3-3</b>
	Demonstration: Using PROC LMIXED within CAS.....	3-11



## 3.1 LMIXED Procedure in SAS Viya

### Objectives

- Compare and contrast mixed modeling in SAS®9 and CAS.

2

Copyright © SAS Institute Inc. All rights reserved.



```
PROC MIXED options;  
  CLASS variables;  
  MODEL dependent=fixed-effects / options;  
  RANDOM random-effects / options;  
  CONTRAST 'label' fixed-effect values |  
             random-effect values / options;  
  ESTIMATE 'label' fixed-effect values |  
            random-effect values / options;  
  LSMEANS fixed-effects / options;  
  ...and more  
RUN;
```

3

Copyright © SAS Institute Inc. All rights reserved.



```
PROC LMIXED options;  
  CLASS variables;  
  MODEL dependent=fixed-effects / options;  
  RANDOM random-effects / options;  
  CONTRAST 'label' fixed-effect values |  
             random-effect values / options;  
  ESTIMATE 'label' fixed-effect values |  
            random-effect values / options;  
  LSMEANS fixed-effects / options;  
  ...  
RUN;
```

4

Copyright © SAS Institute Inc. All rights reserved.



```
PROC LMIXED options;  
  CLASS variables;  
  MODEL dependent=fixed-effects / options;  
  RANDOM random-effects / options;  
  ...  
  OPTIMIZATION <options>;  
  PARMS (value-list) / options;  
  BLUP OUT=CAS-libref.data-table <options>;  
  ... and more  
RUN;
```

5

Copyright © SAS Institute Inc. All rights reserved.





## Features in the LMIXED Procedure

The LMIXED procedure provides easy accessibility to numerous linear mixed models that are useful in many common statistical analyses.

Here are the main features of PROC LMIXED:

- The RANDOM statement supports many covariance structures, including variance components, compound symmetry, unstructured, AR(1), Toeplitz, factor analytic, and so on.
- Both the MODEL statement and the RANDOM statement are supported for model specification, as in the MIXED procedure.
- Inference features include standard errors and  $t$  tests for fixed and random effects.
- A subject effect for blocking is supported.
- Both REML and ML estimation methods are supported; they are implemented with a variety of optimization algorithms.
- It handles unbalanced data.
- Specialized dense and sparse matrix algorithms are provided.
- The OUTPUT statement produces output data tables that contain predicted values, residuals, studentized residuals, confidence limits, and influence statistics.
- The PARMS statement enables you to fit a linear mixed model that has known covariance values, or to set boundary values for the parameters.

Because PROC LMIXED runs on SAS Cloud Analytic Services, it also does the following:

- enables you to run on a cluster of machines that distribute the data and the computations
- enables you to run in single-machine mode on CAS
- exploits all the available cores and concurrent threads. For information about how the LMIXED procedure uses threads

### Massive Mixed Models

Two Types of “Massive”

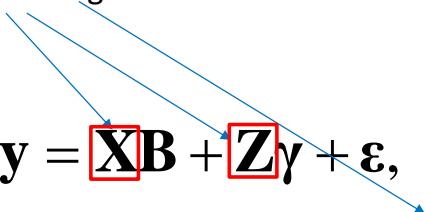
$$\mathbf{y} = \mathbf{XB} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$$

## Massive Mixed Models

### Two Types of “Massive”

#### 1. Large and sparse design matrices



$$\mathbf{y} = \mathbf{X}\mathbf{B} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

9

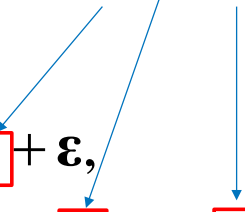
Copyright © SAS Institute Inc. All rights reserved.



## Massive Mixed Models

### Two Types of “Massive”

#### 2. Large covariance matrix



$$\mathbf{y} = \mathbf{X}\mathbf{B} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

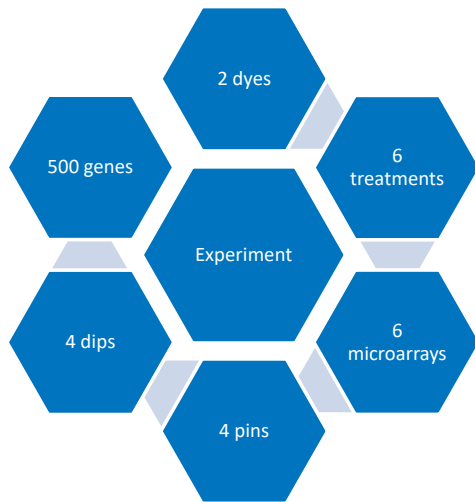
10

Copyright © SAS Institute Inc. All rights reserved.



## Massive Mixed Models

### 1. Large and Sparse Design Matrices



$X = 4,507$  columns

$Z = 3,054$  columns

```
proc mixed data=microarray;
  class marray dye trt gene pin dip;
  model log2i = dye trt gene dye*gene
    trt*gene pin;
  random int gene dip pin
    /subject=marray s;
run;
```

Takes about two hours  
to run

12

Copyright © SAS Institute Inc. All rights reserved.



## How Does PROC LMIXED Help?

### PROC LMIXED

- fits linear mixed models by using sparse matrix storage and sparse matrix computations
- is particularly suitable for problems in which the  $X$  and  $Z$  matrices have many columns and the crossproducts matrix contains many zeros.

```
proc lmixed data=cas.microarray dmmethod=sparse;
  class marray dye trt gene pin dip;
  model log2i = dye trt gene dye*gene trt*gene pin;
  random int gene dip pin/subject=marray s;
run;
```

Takes about 50 seconds  
to run

14




Copyright © SAS Institute Inc. All rights reserved.



## Massive Mixed Models

### 2. Large Covariance Matrix

#### Experiment

- 100 
- 3 
- 5000 

$X = 3$  columns

$Z = 100$  columns

$G = 1 \times 1$

$R = 15,000 \times 15,000$

```
proc mixed data=WeekSim;
  class Gender Clinic Patient Time;
  model Measurement = Gender;
  random Clinic / s;
  repeated Time / sub=Patient type=un;
run;
```

15

Copyright © SAS Institute Inc. All rights reserved.



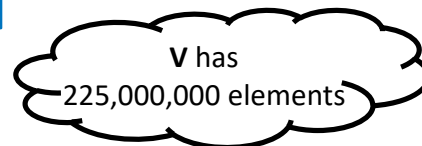
## How Does PROC LMIXED Help?

PROC LMIXED takes about 40 seconds on a single compute node.

- fits linear mixed models by using specialized efficient algorithms
- is particularly suitable for models that have thousands of levels in a common subject effect

```
proc lmixed data=cas.WeekSim;
  class Gender Clinic Patient Time;
  model Measurement = Gender;
  random Clinic / s;
  repeated Time / sub=Patient type=un;
run;
```

• • •



16

Copyright © SAS Institute Inc. All rights reserved.

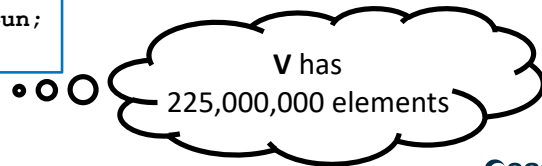


## Most Recent Release of SAS Viya

PROC MIXED take about 40 seconds, too!

- updated algorithms in the most recent release of PROC MIXED
- fits linear mixed models by using specialized efficient algorithms
- is particularly suitable for models that have many levels in a common subject effect

```
proc mixed data=cas.WeekSim noclprint;
  class Gender Clinic Patient Time;
  model Measurement = Gender/ddfm=kr2;
  random Clinic / s;
  repeated Time / sub=Patient type=un;
run;
```



17

Copyright © SAS Institute Inc. All rights reserved.



## A Note about Linear Mixed Models

You want to choose a covariance structure that is not too simple and not too complex.

Penalized fit statistics are useful for comparing models.\*

If the data do not have enough variability to support the structure, the model might not be useful.

Consider a simpler covariance structure.

18

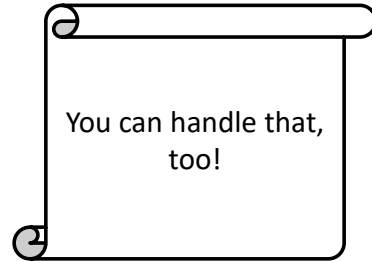
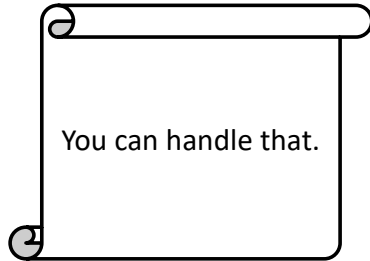
Copyright © SAS Institute Inc. All rights reserved.



## Massive Mixed Models

### Two Types of “Massive”

1. Large and sparse design matrices
2. Large covariance matrix





## Using PROC LMIXED within CAS

```

libname cat '.';

%let narray = 6;
%let ndye = 2;
%let nrow = 4;
%let ngene = 500;
%let ntrt = 6;
%let npin = 4;
%let ndip = 4;
%let no = %eval(&ndye*&nrow*&ngene);
%let tno = %eval(&narray*&no);

data cat.microarray;
    keep Gene MArray Dye Trt Pin Dip log2i;
    array PinDist{&tno};
    array DipDist{&tno};
    array GeneDist{&tno};

    array ArrayEffect{&narray};
    array ArrayGeneEffect{%eval(&narray*&ngene)};
    array ArrayDipEffect{%eval(&narray*&ndip)};
    array ArrayPinEffect{%eval(&narray*&npin)};

    do i = 1 to &tno;
        PinDist{i} = 1 + int(&npin*ranuni(12345));
        DipDist{i} = 1 + int(&ndip*ranuni(12345));
        GeneDist{i} = 1 + int(&ngene*ranuni(12345));
    end;

    igrand = 0;
    idip = 0;
    ipin = 0;
    do i = 1 to &narray;
        ArrayEffect{i} = sqrt(0.014)*rannor(12345);
        do j = 1 to &ngene;
            igrand = igrand+1;
            ArrayGeneEffect{igrand} = sqrt(0.0017)*rannor(12345);
        end;
        do j = 1 to &ndip;
            idip = idip + 1;
            ArrayDipEffect{idip} = sqrt(0.0033)*rannor(12345);
        end;
        do j = 1 to &npin;
            ipin = ipin + 1;
            ArrayPinEffect{ipin} = sqrt(0.037)*rannor(12345);
        end;
    end;

```

```

end;

i = 0;
do MArray = 1 to &narray;
  do Dye = 1 to &ndye;
    do Row = 1 to &nrow;
      do k = 1 to &ngene;
        if MArray=1 and Dye = 1 then do;
          Trt = 0;
          trtc = 0;
        end;
        else do;
          if trtc >= &no then trtc = 0;
          if trtc = 0 then do;
            Trt = Trt + 1;
            if Trt >= &ntrt then do;
              Trt = 0;
              trtc = 0;
            end;
          end;
          end;
          trtc = trtc + 1;
        end;
        i = i + 1;
        Pin = PinDist{i};
        Dip = DipDist{i};
        Gene = GeneDist{i};
        a    = ArrayEffect{MArray};
        ag   = ArrayGeneEffect{ (MArray-1)*&ngene+Gene};
        ad   = ArrayDipEffect{ (MArray-1)*&ndip+Dip};
        ap   = ArrayPinEffect{ (MArray-1)*&npin+Pin};
        log2i = 1 +
              + Dye
              + Trt
              + Gene/1000.0
              + Dye*Gene/1000.0
              + Trt*Gene/1000.0
              + Pin
              + a
              + ag
              + ad
              + ap
              + sqrt(0.02)*rannor(12345);

        output;
      end;
    end;
  end;
end;
run;

```



```

cas mySession ;
libname Caslib cas;
proc casutil;
    load data=cat.microarray outcaslib="casuser"
        casout="microArray" replace;
run;

* Sparse data: takes 1 hour 30 minutes;
proc mixed data=microarray;
    class marray dye trt gene pin dip;
    model log2i = dye trt gene dye*gene trt*gene pin;
    random int gene dip pin/subject=marray s;
    * ods output solutionr=BLUPs;
run;

* Sparse data: takes 53 seconds;
proc lmixed data=caslib.microarray dmmethod=sparse;
    class marray dye trt gene pin dip;
    model log2i = dye trt gene dye*gene trt*gene pin;
    random int gene dip pin/subject=marray;
    * ods output solutionr=BLUPs;
run;

```

```

%let NClinic = 100;
  %let NPatient = %eval(&NClinic*50);
  %let NTime = 3;
  %let SigmaC = 2.0;
  %let SigmaP = 4.0;
  %let SigmaE = 8.0;
  %let Seed = 12345;

libname cat '.';

data WeekSim;
  keep Gender Clinic Patient Time Measurement;
  array PGender{&NPatient};
  array PClinic{&NPatient};
  array PEffect{&NPatient};
  array CEffect{&NClinic};
  array GEeffect{2};

  do Clinic = 1 to &NClinic;
    CEffect{Clinic} = sqrt(&SigmaC)*rannor(&Seed);
  end;

  GEeffect{1} = 10*ranuni(&Seed);
  GEeffect{2} = 10*ranuni(&Seed);

  do Patient = 1 to &NPatient;
    PGender{Patient} = 1 + int(2 * ranuni(&Seed));
    PClinic{Patient} = 1 + int(&NClinic*ranuni(&Seed));
    PEffect{Patient} = sqrt(&SigmaP)*rannor(&Seed);
  end;

  do Patient = 1 to &NPatient;
    Gender = PGender{Patient};
    Clinic = PClinic{Patient};
    Mean = 1 + GEeffect{Gender} + CEffect{Clinic} +
PEffect{Patient};
    do Time = 1 to &NTime;
      Measurement = Mean + sqrt(&SigmaE)*rannor(&Seed);
      output;
    end;
  end;
run;

cas mySession ;
libname Caslib cas;
proc casutil;
  load data=cat.weekSim outcaslib="casuser"
  casout="weekSim" replace;
run;

```

```
* runs for 20 seconds ;
proc mixed data=cat.WeekSim;
  class Gender Clinic Patient Time;
  model Measurement = Gender;
  random Clinic / s;
  repeated Time / sub=Patient type=un;
run;

*runs for 41 seconds;
proc lmixed data=caslib.WeekSim;
  class Gender Clinic Patient Time;
  model Measurement = Gender;
  random Clinic / s;
  repeated Time / sub=Patient type=un;
run;
```

**End of Demonstration**



# Appendix A References

A.1	References.....	A-3
-----	-----------------	-----



# A.1 References

---

- Agresti, A. 1990. *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.
- Akaike, H. 1974. "A new look at the statistical model identification." *IEEE Transaction on Automatic Control*. 19(6):716–723.
- Beckman, R. J., Nachtsheim, C. J., and Cook, R. D. 1987. "Diagnostics for Mixed-Model Analysis of Variance." *Technometrics*, 29(4):413–426.
- Box, G. E. P. and Tiao, G. C. 1973. *Bayesian Inference in Statistical Analysis*. Wiley Classics Library Edition Published 1992. New York: John Wiley & Sons, Inc.
- Breslow, N. E. and Clayton, D. G. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of American Statistical Association*. 88:9–25.
- Breslow, N. E. and Lin, X. 1995. "Bias Correction in Generalized Linear Mixed Models with a Single Component of Dispersion." *Biometrika*. 82(1):81–91.
- Brown, H. and Prescott, R. 1999. *Applied Mixed Models in Medicine*. Chichester, England: John Wiley & Sons, Inc.
- Bryk, A. S. and Raudenbush, S. W. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: SAGE Publications, Inc.
- Burnham, K. P. and Anderson, D. R. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.
- Cochran, W. G. and Cox, G. M. 1957. *Experimental Designs*. New York: John Wiley & Sons, Inc.
- Cressie, N. A. C. 1993. *Statistics for Spatial Data*. New York: John Wiley & Sons, Inc.
- Diggle, P. J., Liang, K. Y., and Zeger, S. L. 1994. *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- Goldstein, H. and Rasbash, J. 1996. "Improved Approximations for Multilevel Models with Binary Responses." *Journal of Royal Statistical Society. A*. 159:505–513.
- Gregoire, T. G., Schabenberger, O., and Barrett, J. P. 1995. "Linear Modelling of Irregularly Spaced, Unbalanced, Longitudinal Data from Permanent Plot Measurements." *Canadian Journal of Forest Research*. 25:137–156.
- Guerin, L. and Stroup, W. W. 2000. "A Simulation Study to Evaluate PROC MIXED Analysis of Repeated Measures Data." *Proceedings of the 2000 Conference on Applied Statistics in Agriculture*. Manhattan, KS: Kansas State University.
- Harville, D. A. and Jeske, D. R. 1992. "Mean Squared Error of Estimation of Prediction Under a General Linear Model." *Journal of the American Statistical Association*. 87:724–731.
- Healy M. J. R. 1995. *Matrices for Statistics*. Oxford: Clarendon Press.
- Hocking, R. R. 1984. *The Analysis of Linear Models*. Monterey, CA: Brooks-Cole Publishing Co.
- John, P. W. M. 1971. *Statistical Design and Analysis of Experiments*. New York: The Macmillan Company.

- Kackar, R. N. and Harville, D. A. 1984. "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models." *Journal of the American Statistical Association*. 79:853–862.
- Kenward, M. G. and Roger, J. H. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics*. 53:983–997.
- Kenward, M. G. and Roger, J. H. 2009. "An Improved Approximation to the Precision of Fixed Effects from Restricted Maximum Likelihood." *Computational Statistics and Data Analysis*. 53:2583–2595.
- Liang, K. Y. and Zeger, S. L. 1986. "Longitudinal Data Analysis using Generalized Linear Models." *Biometrika*. 73:13–22.
- Lin, X. and Breslow, N. E. 1996. "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion." *Journal of American Statistical Association*. 91(435):1007–1016.
- Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. 1996. *SAS® System for Mixed Models*. Cary, NC: SAS Institute Inc.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. 2006. *SAS® for Mixed Models, Second Edition*. Cary, NC: SAS Institute Inc.
- Littell, R. C., Pendergast, J., and Natajan, R. 2000. "Modeling Covariance Structure in the Analysis of Repeated Measures Data." *Statistics in Medicine*. 19: 1793–1819.
- Littell, R. C., Stroup, W. W., and Freund, R. J. 2002. *SAS® System for Linear Models*. Cary, NC: SAS Institute Inc.
- Little, R. J. A. and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- McCullagh, P. 1983. "Quasi-Likelihood Functions." *Annals of Statistics*. 11:59–67.
- McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models, Second Edition*. London: Chapman and Hall.
- McCulloch, C. E. and Searle, S. R. 2001. *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons, Inc.
- Milliken, G. A. and Johnson, D. E. 1992. *Analysis of Messy Data, Volume 1: Designed Experiments*. New York: Chapman & Hall.
- Montgomery, D. C. 1997. *Design and Analysis of Experiments, Fourth Edition*. New York: John Wiley & Sons, Inc.
- Nelder, J. A. and Wedderburn, R. W. M. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society A*. 135:370–384.
- Pinheiro, J. C. and Bates, D. M. 1995. "Approximation to the Log-likelihood Function in the Nonlinear Mixed-effects Model." *Journal of Computational and Graphical Statistics*. 4:12–35.
- Prasad, N. G. N. and Rao, J. N. K. 1990. "The Estimation of Mean Squared Error of Small-Area Estimators." *Journal of the American Statistical Association*. 85:163–171.
- Rodriguez, G. and Goldman, N. 1995. "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses." *Journal of Royal Statistical Society, A*. 158:73–90.



- SAS Institute Inc. 1996. *SAS/STAT® Technical Report: Spatial Prediction Using the SAS® System*. Cary, NC: SAS Institute Inc.
- SAS Institute Inc. 1999. *SAS/GRAPH® Software: Reference, Version 8, Volume 1*. Cary, NC: SAS Institute Inc.
- SAS Institute Inc. 1999. *SAS/STAT® User's Guide, Version 8*. Cary, NC: SAS Institute Inc.
- Schabenberger, O. and Pierce, F. J. 2002. *Contemporary Statistical Models for the Plant and Soil Sciences*. Boca Raton, FL: CRC Press LLC.
- Schaalje, G. B., McBride, J. B., and Fellingham, G. W. 2001. "Approximation to Distributions of Test Statistics in Complex Mixed Linear Models Using SAS PROC MIXED." *SAS Users Group International (SUGI26)*. Cary, NC: SAS Institute Inc.
- Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics*. 6:461–464.
- Searle, S. R. 1982. *Matrix Algebra Useful for Statistics*. New York: John Wiley & Sons, Inc.
- Searle, S. R. 1987. *Linear Models for Unbalanced Data*. New York: John Wiley & Sons, Inc.
- Searle, S. R., Casella, G., and McCulloch, C. E. 1992. *Variance Components*. New York: John Wiley & Sons, Inc.
- Singer, J. D. 1998. "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics*. 24(4):323–355.
- Swallow, W. H. and Monahan, J. F. 1984. "Monte Carlo Comparison of ANOVA, MIVQUE, REML and ML Estimators of Variance Components." *Technometrics*. 28:47–57.
- Verbeke, G. and Molenberghs, G. 1997. *Linear Mixed Models in Practice: A SAS-Oriented Approach*. New York: Springer-Verlag.
- Verbeke, G. and Molenberghs, G. 2000. *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.
- Wedderburn, R. W. M. 1974. "Quasilikelihood Methods, Generalized Linear Models, and the Gauss-Newton Method." *Biometrika*. 61:439–447.
- Wolfinger, R. D. 1998. "Towards Practical Application of Generalized Linear Mixed Models." *Proceedings of the 13<sup>th</sup> International Workshop on Statistical Modeling*. 388–395.
- Wolfinger, R. D. and Kass, R. E. 2000. "Nonconjugate Bayesian Analysis of Variance Component Models." *Biometrics*. 56:768–774.
- Wolfinger, R. D. and O'Connell, M. 1993. "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach." *Journal of Statistical Computation and Simulation*. 48:233–243.

