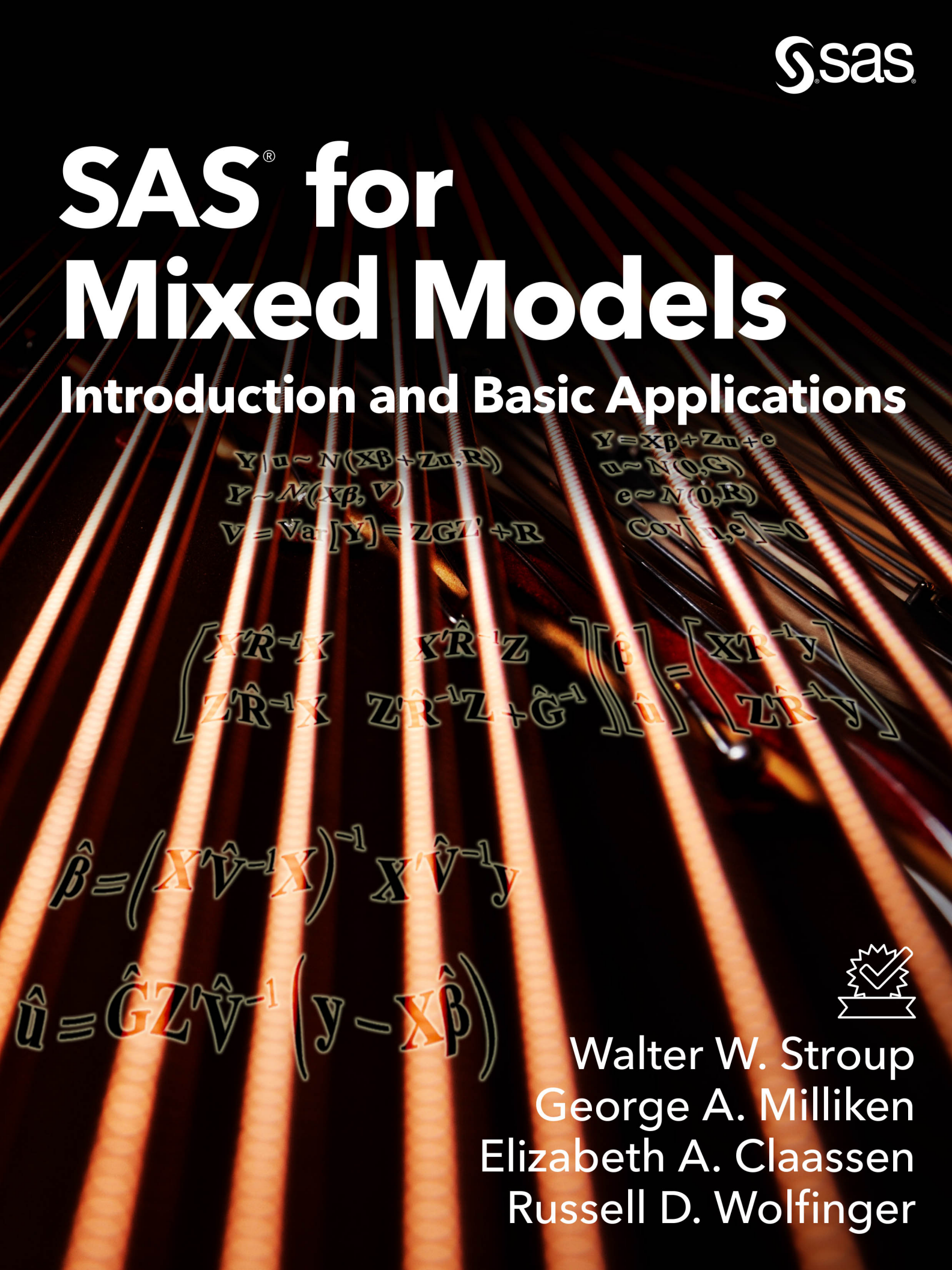
sas

SAS® for Mixed Models

Introduction and Basic Applications



Walter W. Stroup
George A. Milliken
Elizabeth A. Claassen
Russell D. Wolfinger

The correct bibliographic citation for this manual is as follows: Stroup, Walter W., George A. Milliken, Elizabeth A. Claassen, and Russell D. Wolfinger . 2018. *SAS® for Mixed Models: Introduction and Basic Applications*. Cary, NC: SAS Institute Inc.

SAS® for Mixed Models: Introduction and Basic Applications

978-1-63526-135-6 (Hardcopy)
978-1-63526-154-7 (Web PDF)
978-1-63526-152-3 (epub)
978-1-63526-153-0 (mobi)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

December 2018

Contents

About This Book

What Does This Book Cover?

During the past 25 years, mixed models have become an integral part of statistical methodology. Nearly all areas of application that use statistics use mixed models in some form. Mixed models are taught in graduate-level statistics courses, as well as disciplines outside traditional statistics. Mixed models are familiar to most statisticians. Nonetheless, many persons who are engaged in analyzing mixed model data have questions about the appropriate implementation of the methodology. In addition, given the rapid growth of degree programs in data science, as well as statistics, those who are new to the discipline and ready to extend their knowledge of statistical methods to mixed models need a place to start. Being an area of active research and methodological development, mixed models have ever-increasing new applications capabilities available. Those who studied the topic several years ago may not be aware of these developments and need a resource that makes these advances accessible. This edition is intended to address the needs of this diverse audience.

Like the first two editions of *SAS for Mixed Models*, this third publication presents mixed model methodology in a setting that is driven by applications. The scope is both broad and deep. Examples represent numerous areas of application and range from introductory examples to technically advanced case studies. The book is intended to be useful to as diverse an audience as possible, although persons with some knowledge of analysis of variance and regression analysis will benefit most.

The first chapter provides important definitions and categorizations and delineates mixed models from other classes of statistical models. Chapters 2 through 10 cover specific forms of mixed models and the situations in which they arise. Randomized block designs (Chapter 2) give rise to models with fixed treatment and random block effects—among the simplest mixed models. These enable us to introduce elementary mixed model concepts and operations, and to demonstrate the use of SAS mixed model procedures in this simple setting. An overview of mean comparison procedures for various treatment designs is presented in Chapter 3. The topic of "power and sample size" often means doing a power calculation for a designated design at the end of the planning process. However, power involves more than sample size—different designs with the same sample size can yield very different power characteristics. Mixed models provide a powerful methodology for comprehensive assessment of competing plausible designs. Mixed model power and precision analysis is introduced in Chapter 4. Studies with multiple levels, such as split-plot and hierarchical designs, are common in many areas of application. These give rise to models with multiple random effects. The analysis of the associated models is discussed in Chapter 5. Chapter 6 considers models in which all effects are random, and it covers variance component estimation and inference on random effects. Chapter 7 covers analysis of covariance in the mixed model setting. Repeated measures in time or space and longitudinal data give rise to mixed models in which the serial dependency among observations can be modeled directly; this is the topic of Chapter 8. Chapter 9 continues with inference on random effects, a topic begun in Chapter 6. Chapter 9 is devoted to statistical inference based on best linear unbiased prediction of random effects. This naturally leads us to random coefficient and multilevel linear models (Chapter 10).

The second edition of *SAS for Mixed Models* was published when the earliest version of the GLIMMIX procedure had just been released. Since then, new releases of PROC GLIMMIX have greatly expanded SAS capability to handle generalized linear mixed models (GLMMs), mixed models for non-Gaussian data. Although the first two editions of *SAS for Mixed Models* devoted a single chapter to GLMMs, this edition devotes three. The GLMM is introduced in Chapter 11 with binomial data. Chapter 12 introduces GLMMs for count data. Chapter 13 covers multilevel and repeated measures designs in a GLMM context.

In Chapter 14 we revisit power and precision. Chapter 4 concerns simple designs and Gaussian data only, whereas Chapter 14 considers more complex designs and non-Gaussian data. Chapter 14 showcases the full potential of GLMM-based precision and power analysis. Chapter 15 covers mixed model diagnostics based on residuals and influence analysis, as well as some trouble-shooting strategies.

Good statistical applications require a certain amount of theoretical knowledge. The more advanced the application, the more an understanding of mixed models' theoretical foundations will help. Although this book focuses on applications, theoretical developments are presented as well. Appendix A covers linear mixed model theory. Appendix B covers generalized linear mixed model theory. These appendices describe how mixed model methodology works, provide essential detail about the algorithms used by SAS mixed model software, and cover the assumptions underlying mixed model analysis. In addition to

describing how mixed models work, these appendices should help readers understand why things are not working in cases (hopefully few) where problems arise.

Topics included in *SAS for Mixed Models, Second Edition*, but not appearing in this volume are as follows:

- Bayesian analysis
- spatial variability
- heterogeneous variance models
- the NLMIXED procedure
- additional case studies

The authors have reserved these topics for a planned subsequent publication.

What's New in This Edition?

SAS for Mixed Models, Second Edition, has been *the* go-to book for practitioners, students, researchers and instructors on mixed model methodology for more than a decade. PROC GLIMMIX is *the* most comprehensive and sophisticated mixed model software on the market. The current version of PROC GLIMMIX was released in 2008, two years after the publication of the second edition. This publication will be a worthy update incorporating developments over the past decade, building on the *SAS for Mixed Models* go-to status and fully taking advantage of PROC GLIMMIX capabilities.

Some topics have been rearranged to provide a more logical flow, and new examples are introduced to broaden the scope of application areas. Nearly all examples have been updated to use PROC GLIMMIX as the "one-stop shop" for linear modeling, whether fixed effect only, linear mixed models, or generalized linear mixed models. The chapters on GLMMs greatly expand on *SAS for Mixed Models, Second Edition,* as knowledge and software capability have both improved over the past decade. Expanded power and precision chapters enhance the researcher's ability to plan experiments for optimal outcomes. Statistical graphics now utilize the modern SGPLOT procedures.

Is This Book for You?

SAS for Mixed Models: Introduction and Basic Applications is useful to anyone wanting to use SAS for analysis of mixed model data. It is meant to be a comprehensive reference book for data analysts working with mixed models. It is a good supplementary text for a statistics course in mixed models, or a course in hierarchical modeling or applied Bayesian statistics. Mixed model applications have their roots in agricultural research, the behavioral sciences, and medical research—aspects of mixed model methodology arose somewhat independently in these three areas. But the same or similar methodology has proven to be useful in other subject areas, such as the pharmaceutical, natural resource, engineering, educational, and social science disciplines. We assert that almost all data sets have features of mixed models.

Not everyone will want to read the book from cover to cover. Readers who have little or no exposure to mixed models will be interested in the early chapters and can progress through later chapters as their needs require. Readers with good basic skills may want to jump into the chapters on topics of specific interest and refer to earlier material to clarify basic concepts.

To gain the most benefit from this book, ideally readers will have intermediate knowledge of SAS. More importantly, knowledge of some statistical ideas, such as multiple regression, analysis of variance, and experimental design, will ensure that the reader gains the most value from the book.

What Should You Know about the Examples?

This book includes examples for you to follow to gain hands-on experience with SAS.

Software Used to Develop the Book's Content

The software products used to develop the content for this book are as follows:

Base SAS 9.4
SAS/STAT 14.3
SAS/GRAPH 9.4

Example Code and Data

You can access the example code and data for this book by linking to its author pages at https://support.sas.com/authors.

Output and Figures

The tabular and graphical output in this book was generated with a SAS Output Delivery System style customized for optimal book print quality; therefore, your output will differ in appearance. Color versions of Figures 3.11 and 3.13 are included with the example code and data: https://support.sas.com/authors.

SAS University Edition


This book is compatible with SAS University Edition. If you are using SAS University Edition, then begin here: https://support.sas.com/ue-data .

SAS Press Wants to Hear from You


Do you have questions about a SAS Press book that you are reading? Contact us at saspress@sas.com.


SAS Press books are written *by* SAS Users *for* SAS Users. Please visit sas.com/books to sign up to request information on how to become a SAS Press author.


We welcome your participation in the development of new books and your feedback on SAS Press books that you are using. Please visit sas.com/books to sign up to review a book


Learn about new books and exclusive discounts. Sign up for our new books mailing list today at https://support.sas.com/en/books/subscribe-books.html.


Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more:

https://support.sas.com/en/books/authors/walter-stroup.html

https://support.sas.com/en/books/authors/george-milliken.html

https://support.sas.com/en/books/authors/elizabeth-claassen.html

https://support.sas.com/en/books/authors/russell-wolfinger.html

Chapter 2: Design Structure I: Single Random Effect

2.1 Introduction

The simplest design structures that raise mixed model issues are those with blocking. *Blocking* is a research technique used to diminish the effects of variation among experimental units. The units can be people, plants, animals, manufactured mechanical parts, or numerous other objects that are used in experimentation. *Blocks* are groups of units that are formed so that units within the blocks are as nearly homogeneous as possible. Examples of blocking criteria include batches of manufactured items, plots or benches containing plants, matched pairs of people, day on which an assay is performed, etc. In a designed experiment, the levels of the factor being investigated, called *treatments*, are *randomly assigned to units within the blocks*. However, as noted in Chapter 1, blocking can more generally be understood as a grouping method used in survey sampling (e.g. strata or clustering), observational studies (e.g. matched pairs), and the like.

An experiment conducted using blocking is called a *randomized block design*. While the methods discussed in this chapter are presented in the context of randomized block designs, you can easily adapt these methods to survey or observational study contexts. Usually, the primary objectives are to estimate and compare the means of treatments (i.e. treatments as broadly defined). In most cases, the *treatment effects* are considered *fixed* because the treatments in the study are the only ones to which inference is to be made. That is, no conclusions will be drawn about treatments that were not used in the experiment. *Block effects* are usually considered *random* because the blocks in the study constitute only a small subset of a larger set of blocks over which inferences about treatment means are to be made. In other words, the investigator wants to estimate and compare treatment means with statements of precision (confidence intervals) and levels of statistical significance (from tests of hypotheses) that are valid in reference to the entire population of blocks, not just those blocks of experimental units in the experiment. To do so requires proper specification of random effects in model equations. In turn, computations for statistical methods must properly accommodate the random effects. The model for data from a randomized block design usually contains fixed effects for treatment contributions or factors and random effects for blocking factors contributions, making it a *mixed* model.

The issue of whether blocks effects are considered fixed or random becomes especially important in blocked designs with missing data, or incomplete block designs. Analysis with random block effects enables recovery of inter-block information, and the resulting analysis is called combined inter- and intra-block analysis. For estimation and testing treatment differences, analysis with or without recovery of inter-block information is identical only in the case of a

complete block design with no missing data. Otherwise, except where noted in this chapter, inter-block information adds efficiency and accuracy to the analysis.

Section 2.2 presents the randomized block model as it is usually found in basic statistical methods textbooks. The traditional analysis of variance (ANOVA) methods are given, followed by an example to illustrate the ANOVA methods. Section 2.3 illustrates mixed model analysis using the GLIMMIX and MIXED procedures to obtain the results for the example. Section 2.4 presents an analysis of data from an incomplete block design to illustrate similarities and differences between analyses with and without recovery of inter-block information with unbalanced data. Finally, Section 2.5 presents an example of an analysis with a negative block variance estimate. This presents a common dilemma for data analysts: does one allow the variance estimate to remain negative or does one set it to zero. This section presents the pros and cons of each alternative, as well as a general recommendation. Then, basic mixed model theory for the randomized block design is given in Section 2.6, including a presentation of the model in matrix notation.

2.2 Mixed Model for a Randomized Block Design

A design that groups experimental units into blocks is called a *randomized block design*. These have two forms: complete block and incomplete block. Complete block designs are generally referred to as *randomized complete block designs*, or by the acronym RCBD. In an RCBD, each treatment is applied to an experimental unit in each block. In incomplete block designs, only a subset of the treatments is assigned to experimental units in any given block. The balanced incomplete block and partially balanced incomplete block (acronym BIB and PBIB, respectively) are two common examples of this type of design. Blocked designs with missing data share modeling issues and approaches with incomplete block designs. In most—but not all—cases, each treatment is assigned to at most one experimental unit in a given block. See Section 2.4, Milliken and Johnson (2009) and Mead, et al. (2012) for complete discussions of block design strategy and structure.

Whether complete or incomplete, all randomized block designs share a common model. Assume that there are t treatments and r blocks, and that there is one observation per experimental unit. Once the treatments are selected to appear in a given block, each selected treatment is randomly assigned to one experimental unit in that block. In general, there will be N total experimental units. For complete block designs, because each of the t treatments is assigned to one experimental unit in each of the r blocks, there are $N = tr$ experimental units altogether. For incomplete block designs with the same number of experimental units in each block, there are $N = rk$ experimental units, where k denotes the number of experimental units per block.

The conventional assumptions for a randomized block model are as follows:

Letting y_{ij} denote the response from the experimental unit that received treatment i in block j, the equation for the model is as follows:

$$y_{ij} = \mu + \tau_i + b_j + e_{ij} \tag{2.1}$$

where the terms are defined as follows:

- $i = 1, 2, \ldots, t$
- $j = 1, 2, \ldots, r$
- μ and τ_i are fixed parameters such that the mean for the i^{th} treatment is $\mu_i = \mu + \tau_i$
- b_j is the random effect associated with the j^{th} block
- e_{ij} is the random error associated with the experimental unit in block j that received treatment i

Assumptions for random effects are as follows:

Block effects are distributed *normally and independently* with mean 0 and variance σ_b^2; that is, the b_j ($j = 1,2,\ldots,r$) are distributed *iid* $\mathrm{N}\left(0, \sigma_b^2\right)$.

Errors e_{ij} are distributed *normally and independently* with mean 0 and variance σ^2; that is, the e_{ij} ($i = 1,2,\ldots,t$; $j = 1,2,\ldots,r$) are distributed *iid* $\mathrm{N}\left(0, \sigma^2\right)$. The e_{ij} are also distributed independently of the b_j.

2.2.1 Means and Variances from Randomized Block Design

The usual objectives of a randomized block design are to estimate and compare treatment means using statistical inference. Mathematical expressions are needed for the variances of means and differences between means in order to construct confidence intervals and conduct tests of hypotheses. The following results apply to complete block designs. Once these results are in place, you can adapt them for incomplete blocks, as shown below.

For the RCBD, it follows from Equation 2.1 that a treatment mean, such as $\bar{y}_{1\cdot}$, can be written as follows:

$$\bar{y}_{1\cdot} = \mu_1 + \bar{b}_{\cdot} + \bar{e}_{1\cdot}$$

Likewise, the difference between two means, such as $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$, can be written as follows:

$$\bar{y}_{1\cdot} - \bar{y}_{2\cdot} = \mu_1 - \mu_2 + \bar{e}_{1\cdot} - \bar{e}_{2\cdot}$$

From these expressions, the variances of $\bar{y}_{1\cdot}$ and $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ are

$$\mathrm{Var}\left[\bar{y}_{1\cdot}\right] = \left(\sigma^2 + \sigma_b^2\right)/r$$

and

$$\mathrm{Var}\left[\bar{y}_{1\cdot} - \bar{y}_{2\cdot}\right] = 2\sigma^2/r$$

Notice that the variance of a treatment mean $\mathrm{Var}\left[\bar{y}_{1\cdot}\right]$ contains the block variance component σ_b^2, but the variance of the difference between two means $\mathrm{Var}\left[\bar{y}_{1\cdot} - \bar{y}_{2\cdot}\right]$ does *not* involve σ_b^2. This is the manifestation of the RCBD controlling block variation; the variances of differences between treatments are estimated free of block variation.

2.2.2 The Traditional Method: Analysis of Variance

Almost all statistical methods textbooks present analysis of variance (ANOVA) as a key component in analysis of data from a randomized block design. We assume that readers are familiar with fundamental concepts for analysis of variance, such as degrees of freedom, sums of squares (SS), mean squares (MS), and expected mean squares (E[MS]). Readers needing more information about analysis of variance may consult Littell, Stroup, and Freund (2002) or Milliken and Johnson (2009). Table 2.1 shows a standard ANOVA table for the RCBD, containing sources of variation, degrees of freedom, mean squares, and expected mean squares.

Table 2.1: ANOVA for Randomized Complete Blocks Design

Source of Variation	*df*	MS	E[MS]
Blocks	r – 1	MS(Blk)	$\sigma^2 + t\sigma_b^2$
Treatments	t – 1	MS(Trt)	$\sigma^2 + r\phi^2$
Error	(r – 1)(t – 1)	MS(Error)	σ^2

2.2.3 Expected Mean Squares

As the term implies, *expected mean squares* are the expectations of means squares. They are the quantities estimated by mean squares in an analysis of variance. The expected mean squares can be used to motivate test statistics, and to provide a way to estimate the variance components. For test statistics, the basic idea is to examine the expected mean square for a factor and see how it differs under null and alternative hypotheses. For example, the expected mean square for treatments, $\mathrm{E[MS(Trt)]} = \sigma^2 + r\phi^2$, can be used to determine how to set up a test statistic for treatment differences. The null hypothesis is H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$. The expression ϕ^2 in E[MS(Trt)] is

$$\phi^2 = (t-1)^{-1} \sum_{i=1}^{t} \left(\mu_i - \bar{\mu}_{\cdot}\right)^2$$

where $\bar{\mu}_{\cdot}$ is the mean of the μ_i. Thus, $\phi^2 = 0$ is equivalent to $\mu_1 = \mu_2 = \ldots = \mu_t$. So, if the null hypothesis is true, MS(Trt) simply estimates σ^2. On the other hand, if H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$. is false, then E[MS(Trt)] estimates a quantity larger than σ^2. Now, MS(Error) estimates σ^2 regardless of whether H_0 is true or false. Therefore, MS(Trt) and MS(Error) tend to be

approximately equal if H_0 is true, and MS(Trt) tends to be larger than MS(Error) if H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ is false. So a comparison of MS(Trt) with MS(Error) is an indicator of whether H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ is true or false. In this way the expected mean squares show that a valid test statistic is the ratio F = MS(Trt)/MS(Error).

Expected mean squares can also be used to estimate variance components, variances of treatment means, and variances of differences between treatment means. Equating the observed mean squares to the expected mean squares provides the following system of equations:

$$\text{MS(Blk)} = \hat{\sigma}^2 + t\hat{\sigma}_b^2$$
$$\text{MS(Error)} = \hat{\sigma}^2$$

The solution for the variance components is

$$\hat{\sigma}^2 = \text{MS}\left(\text{Error}\right)$$

and

$$\hat{\sigma}_b^2 = \frac{1}{t}\left[\text{MS(Blk)} - \text{MS(Error)}\right]$$

These are called *analysis of variance* estimates of the variance components. Using these estimates of the variance components, it follows that estimates of $\text{Var}\left[\bar{y}_{1\cdot}\right]$ and $\text{Var}\left[\bar{y}_{1\cdot} - \bar{y}_{2\cdot}\right]$ are

$$\begin{aligned}\hat{\text{Var}}\left[\bar{y}_{1\cdot}\right] &= \left(\hat{\sigma}^2 + \hat{\sigma}_b^2\right)/r \\ &= \frac{1}{rt}\text{MS(Blk)} + \frac{t-1}{rt}\text{MS(Error)}\end{aligned}$$

and

$$\hat{\text{Var}}\left[\bar{y}_{1\cdot} - \bar{y}_{2\cdot}\right] = \frac{2}{r}\text{MS(Error)}$$

The expression for

$$\text{Var}\left[\bar{y}_{1\cdot}\right]$$

illustrates a common misconception that the estimate of the variance of a treatment mean from a randomized block design is simply MS(Error)/r. This misconception prevails in some textbooks and results in incorrect calculation of standard errors by some computer software packages, as well as incorrect reporting in refereed journal articles

2.2.4 Example: A Randomized Complete Block Design

An example from Mendenhall, Wackerly, and Scheaffer (1996, p. 601) is used to illustrate analysis of data from a randomized block design.

Data for this example are presented as Data Set "Bond". There are seven blocks and three treatments. Each block is an ingot of a composition material. The treatments are metals (nickel, iron, or copper). Pieces of material from the same ingot are bonded using one of the metals as a bonding agent. The response is the amount of pressure required to break a bond of two pieces of material that used one of the metals as the bonding agent. Table 2.2 contains the analysis of variance table for the BOND data where the ingots form the blocks.

Table 2.2: ANOVA Table for BOND Data

Source of Variation	*df*	SS	MS	*F*	*p*-value
Ingots	6	268.29	44.72	4.31	0.0151
Metal	2	131.90	65.95	6.36	0.0131
Error	12	124.46	10.37		

The ANOVA table and the metal means provide the essential computations for statistical inference about the population means.

The ANOVA $F = 6.36$ for metal provides a statistic to test the null hypothesis H_0: $\mu_c = \mu_i = \mu_n$. The significance probability for the F test is $p = 0.0131$, indicating strong evidence that the metal means are different. Estimates of the variance components are $\hat{\sigma}^2 = 10.37$ and $\hat{\sigma}_b^2 = \left(44.72 - 10.37\right)/3 = 11.45$. Thus, an estimate of the variance of a metal mean is $\left(\hat{\sigma}^2 + \hat{\sigma}_b^2\right)/7 = 3.11$, and the estimated standard error is $\sqrt{3.11} = 1.77$. An estimate of the variance of a difference between two metal means is $2\hat{\sigma}^2/7 = 2 \times 10.37/7 = 2.96$, and the standard error is $\sqrt{2.96} = 1.72$.

2.3 The MIXED and GLIMMIX Procedures to Analyze RCBD Data

PROC GLIMMIX and PROC MIXED are procedures with several capabilities for different methods of analyzing mixed models. PROC MIXED can be used for linear mixed models (LMMs), i.e., when you can assume that the response variable has a Gaussian distribution. PROC MIXED enables you to estimate the variance components using sums of squares and expected mean squares, as described in the previous section or by using likelihood methods. PROC GLIMMIX can be used for LMMs and generalized linear mixed models (GLMMs; i.e., for both Gaussian and non-Gaussian response variables). PROC GLIMMIX uses only likelihood-based methods.

For the randomized block examples presented in this chapter, and for more complex LMM applications presented in Chapters 5 through 10, analyses obtained using PROC MIXED or PROC GLIMMIX are essentially identical. For certain advanced LMMs, not presented in this volume, PROC MIXED offers specialized capabilities that are not available in PROC GLIMMIX. On the other hand, for GLMMs with non-Gaussian data, discussed in Chapters 11 through 13, and for inference on variance components, presented in Chapter 6, PROC GLIMMIX provides capabilities that are not available in PROC MIXED. For this reason, in this section, analyses of an RCBD are shown using both procedures, but all subsequent examples in this volume use PROC GLIMMIX.

In both PROC MIXED and PROC GLIMMIX, many of the estimation and inferential methods are implemented on the basis of the likelihood function and associated principles and theory (see Appendix A, "Linear Mixed Model Theory," for details). Readers may be more familiar with the analysis of variance approach described in the previous section; those results are obtained and presented in Section 2.3.1. The likelihood method results are presented in Section 2.3.2. Output from both PROC MIXED and PROC GLIMMIX are presented so readers can see that the results are the same, but the presentation format is slightly different. The results of the analysis of variance and likelihood methods are compared and are shown to duplicate many of the results of the previous section.

There are extensive post-processing options for mean comparison estimation, testing, and plotting available with both procedures. Presentation of these options, focusing on the more extensive options available with PROC GLIMMIX, are deferred to Chapter 3.

2.3.1 PROC MIXED Analysis Based on Sums of Squares

This section contains the code to provide the analysis of the RCBD with PROC MIXED using the sums of squares approach as described in Section 2.2.4. The METHOD=TYPE3 option is used to request that Type 3 sums of squares be computed along with their expected mean squares. Those mean squares and expected mean squares are used to provide estimates of the variance components and estimates of the standard errors associated with the means and comparisons of the means.

Program

Program 2.1 shows the basic PROC MIXED statements for the RCBD data analysis.

Program 2.1

```
proc mixed data=bond cl method=type3;
   class ingot metal;
   model pres = metal;
   random ingot;
   lsmeans metal;
run;
```

The PROC MIXED statement calls the procedure. The METHOD=TYPE3 option requests that the Type 3 sums of squares method be used in estimating the variance components. You can request Type 1, 2, or 3 sums of squares. See

Milliken and Johnson (2009) or Littell and Stroup (2002) for additional detail. The CLASS statement specifies that INGOT and METAL are classification variables, not continuous variables.

The MODEL statement is an equation whose left-hand side contains the name of the response variable to be analyzed, in this case PRES. The right-hand side of the MODEL statement contains a list of the fixed effect variables, in this case the variable METAL. In terms of the statistical model, this specifies the τ_i parameters. (The intercept parameter μ is implicitly contained in all models unless otherwise declared by using the NOINT option.)

The RANDOM statement contains a list of the random effects, in this case the blocking factor INGOT, and represents the b_j terms in the statistical model.

The MODEL and RANDOM statements are the core essential statements for many mixed model applications, and the terms in the MODEL statement do not appear in the RANDOM statement, and vice versa.

Results

Results from the MODEL and RANDOM statements about the methods used appear in Output 2.1.

Output 2.1: Results of RCBD Data Analysis from PROC MIXED Using Type 3 Sums of Squares

Model Information	
Data Set	WORK.BOND
Dependent Variable	pres
Covariance Structure	Variance Components
Estimation Method	Type 3
Residual Variance Method	Factor
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information		
Class	Levels	Values
ingot	7	1 2 3 4 5 6 7
metal	3	c i n

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

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

Interpretation

The "Model Information" table contains the model specifications for the data set being used, the response variable, the methods used to estimate the variance components, the approximate degrees of freedom, and the standard errors for the fixed effects.

The "Class Level Information" table lists the levels for each of the variables declared in the class statement. You should be sure that these levels are specified consistently with how the study was conducted.

The "Dimensions" table shows how many columns are in the fixed effects matrix (**X**) and in the random effects matrix (**Z**) parts of the model, where the linear predictor is $\mathbf{X\beta} + \mathbf{Zu}$ (see Section 1.7). For this study there are three levels of the treatment factor (metal) plus an intercept, which accounts for four columns in the **X** matrix. There are seven ingots (blocks), thus there are seven columns in the **Z** matrix. The inclusion of the RANDOM statement means that there is one

variance component for the ingot effects, plus the residual variance, providing two parameters in the covariance structure of the model. There is no SUBJECT= option used in this RANDOM statement, so PROC MIXED assumes that all observations are from the same subject, a quantity that can be ignored here.

The "Number of Observations" table indicates how many observations are in the data set and how many of those observations had valid data values for all variables used in the analysis. The difference between the number in the data set and the number used is the number of observations not used in the analysis. The information in these dimension specifications must match the information that is expected from the design being analyzed. Checking these values can help determine if there are data errors that need to be addressed, because they can cause the analysis to fail.

Results

Statistical results from the MODEL and RANDOM statements appear in Output 2.2.

Output 2.2: Results of the RCBD Data Analysis from PROC MIXED Using Type 3 Sums of Squares to Estimate the Variance Components

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
metal	2	131.900952	65.950476	Var(Residual) + Q(metal)	MS(Residual)	12	6.36	0.0131
ingot	6	268.289524	44.714921	Var(Residual) + 3 Var(ingot)	MS(Residual)	12	4.31	0.0151
Residual	12	124.459048	10.371587	Var(Residual)	.	.	.	.

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
ingot	11.4478	8.7204	1.31	0.1893	0.05	-5.6438	28.5394
Residual	10.3716	4.2342	2.45	0.0072	0.05	5.3332	28.2618

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

Interpretation

The "Type 3 Analysis of Variance" table is the usual analysis of variance table with degrees of freedom, sums of squares, mean squares, expected mean squares, error terms for effects other than the residual, *F* tests, and significance levels for these tests. The terms Var(Residual) and Var(ingot) denote the variance components σ^2 and σ_b^2, respectively. See the discussion of the "Tests of Fixed Effects" table for more detail.

The "Covariance Parameter Estimates" table gives estimates of the variance component parameters obtained by solving the set of equations from equating the observed mean squares to the expected mean squares. The estimate of σ_b^2, the block variance component, is 11.4478 (labeled "ingot"), and the estimate of σ^2, the error variance component, is 10.3716 (labeled "Residual"). The confidence intervals for the variance components are Wald confidence intervals.

The "Tests of Fixed Effects" table is like an abbreviated ANOVA table, showing a line of computations for each term in the MODEL statement. In this example, only METAL is included in the MODEL statement. The *F* statistic is used to test the null hypothesis H_0: $\mu_c = \mu_i = \mu_n$ vs. H_a (not H_0). With 2 numerator and 12 denominator degrees of freedom, the *F* value of 6.36 is significant at the 5% level (*p*-value is 0.0131). If the true METAL means are equal, then an *F*-value as large as 6.36 would occur less than 131 times in 10,000 by chance. This is the same *F* test that was obtained from the analysis of variance.

In summary, these basic PROC MIXED computations are based on sums of squares and provide the same statistical computations obtained from analysis of variance methods for a balanced data set.

2.3.2 Basic PROC MIXED Analysis Based on Likelihood

Both PROC MIXED and PROC GLIMMIX, by default, provide maximum likelihood estimates (acronym MLE) of model effects, and REML estimates of variance components. REML stands for *REsidual (or REstricted) Maximum Likelihood* (Patterson and Thompson 1971). A fundamental strength of likelihood-based methodology is its adaptability. For randomized block models, analysis of variance and likelihood-based methods produce identical results, but analysis of variance methods cannot be applied to most cases that are even slightly more complex than the randomized block, whereas likelihood-based methods can be applied to arbitrarily complex models.

When comprehensive mixed model software first became widely available—in the early 1990s—some questioned the use of REML as the default variance estimation method. Specifically, why not maximum likelihood (ML) estimates of the variance? While not shown here, you can obtain ML variance estimates by using METHOD=ML in PROC MIXED or METHOD=MSPL in PROC GLIMMIX. The resulting variance estimates will be less that the corresponding REML estimates, and the resulting confidence intervals will be narrower and test statistics will be greater. This reflects the well-known fact that ML estimates of variance components are biased downward. For example, in the one-sample case, when $y_1, y_2, ..., y_n$ is a random sample from $N(\mu, \sigma^2)$, the ML estimate of the variance is as follows:

$$\sum_i (y_i - \bar{y})^2 \Big/ n$$

whereas the sample variance—which is the simplest REML variance estimate—is as follows:

$$\sum_i (y_i - \bar{y})^2 \Big/ (n-1)$$

We know that the latter is unbiased and universally regarded as the preferred variance estimate. One can easily show that the use of ML variance estimates results in upwardly biased type I error rates (rejection rates as high as 25% for a nominal $\alpha = 0.05$), and inadequate confidence interval coverage.

Program

Program 2.2 uses the default REML method for estimating the variance components. One could exclude METHOD=REML in the PROC MIXED statement, and achieve the same results. The assumptions of normality of the various terms in the model Equation 2.1 are required in order to construct the appropriate likelihood function that is maximized. The code to provide the likelihood-based analysis is identical to that of the sums of squares method, except for the method specification.

Program 2.2

```
proc mixed data=bond method=reml;
   class ingot metal;
   model pres=metal;
   random ingot;
run;
```

The PROC MIXED statement invokes the procedure for the default method of estimation, REML. The CLASS, MODEL, and RANDOM statements are identical to those in Section 2.3.1.

Results

The results of Program 2.2 appear in Output 2.3.

Output 2.3: Results of RCBD Data Analysis from PROC MIXED METHOD=REML

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

Class Level Information		
Class	**Levels**	**Values**
ingot	7	1 2 3 4 5 6 7
metal	3	c i n

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

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

Iteration History			
Iteration	**Evaluations**	**-2 Res Log Like**	**Criterion**
0	1	112.40987952	
1	1	107.79020201	0.00000000

Convergence criteria met.

Covariance Parameter Estimates	
Cov Parm	**Estimate**
ingot	11.4478
Residual	10.3716

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

Differences between results in Output 2.3 and Output 2.2 include the following:

- The "Model Information" table shows that REML is the specified method of estimating the variance components.
- The "Iteration History" table shows the sequence of evaluations to obtain (restricted) maximum likelihood estimates of the variance components. This portion of the output is not critical to most applications, such as the present RCBD analysis.
- The "Covariance Parameter Estimates" table gives estimates of the variance component parameters. The REML estimate of σ_b^2, the block variance component, is 11.4478 (labeled "ingot"), and the estimate of σ^2, the error variance component, is 10.3716 (labeled "Residual"). For this example of a balanced data set, these variance component estimates are identical to the estimates obtained from the analysis of variance method.

Notice that the essential output you would report; that is, the variance component estimates and the test statistics for the null hypothesis of no treatment effect—in essence, the *F* value, 6.36, and *p*-value, 0.0131—are identical to the results using analysis of variance.

In summary, the default PROC MIXED computations are based on likelihood principles, but many of the results are the same as those obtained from analysis of variance methods for the RCBD.

2.3.3 Basic PROC GLIMMIX Analysis

You can use PROC GLIMMIX to compute the same analysis as PROC MIXED METHOD=REML. Because PROC GLIMMIX is the most general of the SAS mixed model procedures, most examples from this point forward use PROC GLIMMIX. The RCBD data set is shown using both procedures to enable you to see the similarities, as well as some minor differences in the format of the results.

Program

Program 2.3 shows the PROC GLIMMIX program corresponding to PROC MIXED Program 2.2 in Section 2.3.2.

Program 2.3

```
proc glimmix data=bond method=rspl;
   class ingot metal;
   model pres=metal;
   random ingot;
run;
```

The only difference is that RSPL replaces REML in the METHOD option. RSPL (Residual Subject-specific Pseudo Likelihood) is a generalized form of the REML algorithm that can be used for generalized linear mixed models (GLMMs), essentially mixed models with non-Gaussian response variable. The more general algorithm is required to enable PROC GLIMMIX to accommodate non-Gaussian data. Chapters 11, 12 and 13 cover GLMMs. Details of the RSPL algorithm are given in Appendix B. The distinction between RSPL and REML is only relevant in those chapters. With Gaussian response variables—in essence, when the data are assumed to have a normal distribution—the RSPL algorithm reduces to REML. For Gaussian data, RSPL and REML are one and the same.

Results

Output 2.4 shows selected results.

Output 2.4: Results of RCBD Data Analysis from PROC GLIMMIX

Model Information	
Data Set	WORK.BOND
Response Variable	pres
Response Distribution	Gaussian
Link Function	Identity
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Restricted Maximum Likelihood
Degrees of Freedom Method	Containment

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
ingot	11.4478	8.7204
Residual	10.3716	4.2342

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

Default output from PROC GLIMMIX is similar to default REML output from PROC MIXED. They differ in that the PROC GLIMMIX table of "Covariance Parameter Estimates," includes a column for the standard error whereas PROC MIXED does not. For small data sets, the standard error of the variance component estimate is not too useful, because it is based on too few degrees of freedom. Confidence intervals for variance components based on a Satterthwaite approximation or the profile likelihood are useful when Wald type confidence intervals are not. Satterthwaite and profile likelihood confidence intervals are discussed in the next section.

2.3.4 Confidence Intervals for Variance Components

Confidence intervals can be used when it is of interest to access the uncertainty about the variance components in the model. A $(1 - \alpha) \times 100\%$ confidence interval about σ^2 can be constructed by using the chi-square distribution, as

$$\frac{(b-1)(t-1)\hat{\sigma}^2}{\chi^2_{(1-\alpha/2),(b-1)(t-1)}} \leq \sigma^2 \leq \frac{(b-1)(t-1)\hat{\sigma}^2}{\chi^2_{(\alpha/2),(b-1)(t-1)}}$$

where

$$\chi^2_{(1-\alpha/2),(b-1)(t-1)}$$

and

$$\chi^2_{\alpha/2,(b-1)(t-1)}$$

are the lower and upper $\alpha/2$ percentage points of a central chi-square distribution with $(b-1) \times (t-1)$ degrees of freedom, respectively. When the estimate of σ_b^2 is positive, an approximate $(1 - \alpha) \times 100\%$ confidence interval about σ_b^2 can be constructed using a Satterthwaite (1946) approximation. The estimate of σ_b^2 is a linear combination of mean squares, which in general can be expressed as

$$\hat{\sigma}_b^2 = \sum_{i=1}^{s} q_i MS_i$$

where the i^{th} mean square is based on f_i degrees of freedom and q_i is the constant by which the i^{th} mean square is multiplied to obtain $\hat{\sigma}_b^2$. The approximate number of Satterthwaite degrees of freedom associated with $\hat{\sigma}_b^2$ is as follows:

$$v = \frac{\left(\hat{\sigma}_b^2\right)^2}{\sum_{i=1}^{s}\left[\left(q_i MS_i\right)^2\right]/f_i}$$

For the randomized complete block, the expression is the following:

$$\hat{\sigma}_b^2 = \frac{1}{t}\left(\text{MS(Blk)} - \text{MS(Error)}\right)$$

The approximate number of degrees of freedom is as follows:

$$v = \frac{\left(\hat{\sigma}_b^2\right)^2}{\frac{\left(t^{-1}\text{MS(Blk)}\right)^2}{b-1} + \frac{\left(t^{-1}\text{MS(Error)}\right)^2}{(b-1)(t-1)}}$$

A $(1 - \alpha) \times 100\%$ confidence interval about σ_b^2 can be constructed using the chi-square distribution, as the following, where $\chi^2_{(1-\alpha/2),v}$ and $\chi^2_{\alpha/2,v}$ are the lower and upper $\alpha/2$ percentage points with v degrees of freedom, respectively:

$$\frac{v\hat{\sigma}_b^2}{\chi^2_{(1-\alpha/2),v}} \leq \sigma_b^2 \leq \frac{v\hat{\sigma}_b^2}{\chi^2_{\alpha/2,v}}$$

Program to Obtain Satterthwaite Approximation Confidence Intervals

You can use either PROC GLIMMIX or PROC MIXED to obtain Satterthwaite approximation confidence intervals about σ_b^2 and σ^2. With PROC MIXED, use the COVTEST and CL options in the PROC statement. With PROC GLIMMIX, use the COVTEST *statement* with the CL option. Program 2.4 shows the PROC GLIMMIX statements.

Program 2.4

```
proc glimmix data=bond;
   class ingot metal;
   model pres=metal;
   random ingot;
   covtest / cl;
run;
```

Results

The results of computing the estimate of the variance components and using the Satterthwaite approximation to construct the confidence interval about σ_b^2 are given in Output 2.5.

Output 2.5: Wald Confidence Intervals for Block and Residual Variance from the PROC GLIMMIX COVTEST CL Option

Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Wald 95% Confidence Bounds	
ingot	11.4478	8.7204	3.8811	121.55
Residual	10.3716	4.2342	5.3332	28.2618

The 95% confidence interval for the block (INGOT) variance is (3.88, 121.55) and for the residual variance is (5.33, 28.26). The confidence intervals denoted as Wald confidence intervals are in fact Satterthwaite approximate confidence intervals. The Satterthwaite degrees of freedom are computed as $df = 2 * Z^2$, where Z = Estimate/(Standard Error). The confidence interval is as follows:

$$\frac{df * \text{Estimate}}{\chi^2_{1-\alpha/2,df}} \leq \sigma^2 \leq \frac{df * \text{Estimate}}{\chi^2_{\alpha/2,df}}$$

For the Ingot variance component, terms are as follows:

- $Z = 11.4478 / 8.7204 = 1.313$
- $df = 2 * (1.313^2) = 3.45$
- $\chi^2_{0.025,3.45} = 0.3246$
- $\chi^2_{0.975,3.45} = 10.17$

The 95% confidence for the Ingot variance is as follows:

$$\frac{3.45 * 11.4478}{10.17} \leq \sigma^2_{\text{Ingot}} \leq \frac{3.45 * 11.4478}{0.3246}$$

or

$$3.881 \leq \sigma^2_{\text{Ingot}} \leq 121.55$$

which is the same as shown in Output 2.5.

For all but very large data sets, the Satterthwaite confidence bounds are more accurate than Wald confidence bounds and therefore recommended. You can obtain Satterthwaite bounds using either PROC GLIMMIX or PROC MIXED. An alternative procedure, available only with PROC GLIMMIX, uses the likelihood ratio. Let $\boldsymbol{\sigma}$ denote the vector of covariance parameters, and $\log L(\hat{\boldsymbol{\sigma}})$ denote the restricted log likelihood given the REML estimates of the parameters of $\boldsymbol{\sigma}$. For the ingot example,

$$\boldsymbol{\sigma}' = \begin{bmatrix} \sigma_b^2 & \sigma^2 \end{bmatrix}$$

Let

$$\log L\left(\hat{\boldsymbol{\sigma}} \mid \tilde{\sigma}_b^2\right)$$

denote the restricted log likelihood for a given value—not necessarily the REML or ML estimate—of the block variance, denoted $\tilde{\sigma}_b^2$ and the estimate of the other variance components, in this case σ^2, given $\tilde{\sigma}_b^2$. We know that $-2\log(\Lambda)$ where Λ denotes the likelihood ratio, can be written as follows:

$$2\left\{\log L\left(\hat{\boldsymbol{\sigma}}\right) - \log L\left(\hat{\sigma}^2 \mid \tilde{\sigma}_b^2\right)\right\}$$

And we know that it has an approximate χ^2 distribution. Just as the Satterthwaite approximation confidence interval contains all variance component values such that the test statistic, $\nu\hat{\sigma}_b^2 / \sigma_b^2$, is between upper and lower quantiles of the χ_ν^2 distribution, you can form a 95% confidence interval for σ_b^2 from the set of all $\tilde{\sigma}_b^2$ such that the likelihood ratio test statistic,

$$2\left\{\log L\left(\hat{\boldsymbol{\sigma}}\right) - \log L\left(\hat{\sigma}^2 \mid \tilde{\sigma}_b^2\right)\right\} < \chi^2$$

You can obtain profile likelihood confidence intervals for a variance component in two ways. The *profile likelihood ratio* (PLR) re-estimates all the other covariance parameters for each new value of the parameter for which the confidence interval is being determined. The *empirical likelihood ratio* (ELR) uses the REML estimate of σ^2 to calculate the likelihood ratio for all values of σ_b^2 being evaluated. The latter is computationally simpler and is adequate for blocked designs. You can obtain empirical profile likelihood confidence intervals using the following modification to the COVTEST statement.

```
covtest / cl(type=elr);
```

Output 2.6 shows the result.

Output 2.6: Estimated Profile Likelihood Confidence Intervals for Block and Residual Variance

Covariance Parameter Estimates						
Cov Parm	Estimate	Standard Error	Estimated Likelihood 95% Confidence Bounds			
			Lower		Upper	
			Bound	Pr > Chisq	Bound	Pr > Chisq
ingot	11.4478	8.7204	2.2907	0.0500	56.4772	0.0500
Residual	10.3716	4.2342	5.1386	0.0500	25.2825	0.0500

Notice that the confidence bounds are noticeably more precise, especially for the block variance.

2.3.5 Comments on Using PROC GLM for the Analysis of Mixed Models

Prior to the advent of mixed model methods—PROC MIXED was introduced in the early 1990s—PROC GLM was the principal SAS procedure for analyzing mixed models, even though the basic computations of PROC GLM are for fixed effects models. Statistical methods textbooks continued to present the analysis of blocked designs using PROC GLM well into the 2000s. For the *complete* block designs with *no missing data,* the GLM procedure produces results similar to the PROC MIXED analysis of variance output shown in Section 2.3.1. However, this is not true for *incomplete* blocks designs or *any* blocked designs (complete or incomplete) with missing data. PROC GLM was not designed to solve mixed model estimating equations or to compute mixed model inferential statistics. Specifically, the RANDOM statement in PROC GLM does not modify estimation or inference as do RANDOM statements in PROC GLIMMIX and PROC MIXED. The RANDOM statement in PROC GLM merely assigns sums of squares to be used to construct *F* values and standard errors. The sums of squares, however, are computed as if all effects are fixed.

As a result, you *cannot* use PROC GLM to implement *any* of the mixed models analyses shown subsequently in this book. In many cases, PROC GLM *does* implement an analysis that would have been considered state of the art in the 1970s. However, these analyses are known to be less accurate than the corresponding mixed model analyses. In many cases, the standard errors and test statistics obtained by PROC GLM do not correctly account for random effects. PROC GLM is an excellent tool when used for what it was intended (fixed-effects-only models), but we *emphatically*

discourage its use for all applications beyond the RCBD, beginning with incomplete block designs or RCBDs with missing data. Refer to Littell, Stroup, and Freund (2002) for more detailed PROC GLM coverage.

2.4 Unbalanced Two-Way Mixed Model: Examples with Incomplete Block Design

In some applications of blocking there are not enough experimental units in each block to accommodate all treatments. *Incomplete block designs* are designs in which only a subset of the treatments is applied in each block. The treatments that go into each block should be selected in order to provide the most information relative to the objectives of the experiment.

Three types of incomplete block designs are *balanced* incomplete block designs (BIBD), *partially balanced* incomplete block design (PBIBD), and *unbalanced* incomplete block design. The word "balanced" has a specific meaning for incomplete block designs. In design theory, the meaning of "balanced" for BIB and PBIB designs results in all treatment mean estimates having the same variances (and hence the same standard error). Also, the variances of estimated treatment mean differences are the same for all pairs of treatments with BIBDs and for sets of treatments with PBIBDs. As you may suspect, it is not possible to construct BIB or PBIB designs for all possible numbers of treatments and blocks. Discovery of numbers of blocks and treatments for which BIBDs and PBIBDs can be constructed was once an active area of statistical research. With the advent of fast computers and good statistical software, the existence of BIBDs and PBIBDs for given numbers of blocks and treatments has become a less important problem. For example, you can use PROC OPTEX or the optimal design software in JMP to construct approximately balanced incomplete block designs. These designs are commonly used in many fields of research. Mead et al. (2011) have an excellent discussion of this issue.

This section presents the two most commonly used analyses for incomplete block designs. In one, called *intra-block* analysis, block effects are assumed to be fixed. In the pre-mixed-model era of statistical software, intra-block analysis was the only available method. In the other method, called *combined inter- and intra-block* analysis, block effects are assumed to be random. In most cases, using information provided by the block variance, called *recovery of inter-block information*, improves the accuracy and precision of the resulting analysis. However, the intra-block analysis is useful for introducing the distinction between Least Squares treatment means, also called *adjusted means*, and unadjusted arithmetic means, and the associated distinction between Type I tests of hypotheses and Type III tests.

You can use PROC GLM, PROC GLIMMIX or PROC MIXED to implement intra-block (fixed block effect) analysis. To do combined inter- and intra-block (random block effect) analysis, you must use either PROC GLIMMIX or PROC MIXED. PROC GLM was not designed to perform the required computations for recovery of inter-block information.

For consistency, both types of analyses are demonstrated using PROC GLIMMIX. Data from a PBIBD is used to illustrate the similarities and differences between intra-block and combined inter- and intra-block analyses. Note that the intra-block analysis shown in Section 2.5.1 is identical to the analysis that you would get if you use PROC GLM or PROC MIXED (assuming block effects fixed). The combined inter- and intra-block analysis in Section 2.5.2 is identical to the results using PROC MIXED (assuming random block effects). Finally, although the example is a PBIBD, data analysis methods in this section apply to incomplete block designs in general.

As noted above, models for an incomplete block design are the same as for an RCBD. That is, the model equation is

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}$$

where $\mu_i = \mu + \tau_i$ denotes the treatment mean, b_j denotes the block effects b_j and the residual, or experimental error effects e_{ij} are assumed *iid* N(0, σ^2). An analysis of variance table for an incomplete block design is shown in Table 2.3.

Table 2.3: Type III Analysis of Variance Table for Incomplete Blocks Design

Source of Variation	*df*	*F*
Blocks (adjusted for treatments)	r – 1	
Treatments (adjusted for blocks)	t – 1	MS(Trts adj.) / MS(Residual)
Residual	N – r – t + 1	

In the table, *r* is the number of blocks, *t* is the number of treatments, and *N* is the total number of observations. Notice that the treatments source of variation is adjusted for blocks (Littell and Stroup 2002). The treatments cannot be

compared simply on the basis of the usual sum of squared differences between treatment means, because this would contain effects of blocks as well as treatment differences. Instead, a sum of squared differences must be computed between treatment means that have been adjusted to remove the block effects. The difference between the adjusted and unadjusted analyses is illustrated in Section 2.4.1.

Most statistics textbooks that cover BIBD and PBIBD present intra-block analyses. A few also present combined intra- and inter-block analysis. In older textbooks, combined inter- and intra-block analysis appears needlessly daunting. This is especially true of textbooks written before mixed model software was available, i.e., before PROC MIXED was introduced in the early 1990s, and it was not recognized that recovery of inter-block information is simply random block effect mixed model analysis. Textbooks that do cover both types of analysis are all over the map regarding advice about when to use which method of analysis. In Section 2.4.3, we address this question.

2.4.1 Intra-block Analysis of PBIB Data

Data Set PBIB contains data from Cochran and Cox (1957, p. 456). The design is a PBIBD with fifteen blocks, fifteen treatments, and four treatments per block. Data are pounds of seed cotton per plot. The *block size* is the number of treatments per block. This PBIBD has a block size of four. Each treatment appears in four blocks. Some pairs of treatments appear together within a block (e.g., treatments 1 and 2), and other treatments do not appear together in the same blocks (e.g., treatments 1 and 6).

The data appear in multivariate form; that is, with one data line per block, and the four treatment identifiers and responses given as separate variables. To arrange the data in the univariate form in which each observation has a single data line, as required for SAS mixed model procedures, use the following DATA step:

```
data pbib;
  input blk @@;
   do eu=1 to 4;
    input treat y @@;
    output;
   end;
datalines;
```

Program for Intra-block Analysis

An intra-block analysis of the PBIBD data is obtained from Program 2.5.

Program 2.5

```
proc glimmix data=pbib;
  class treat blk;
  model y=treat blk;
  lsmeans treat;
run;
```

Results

Selected results from this PROC GLIMMIX run appear in Output 2.7.

Output 2.7: Incomplete Block Design: PROC GLIMMIX Output for Intra-block Analysis

Fit Statistics	
-2 Res Log Likelihood	46.33
AIC (smaller is better)	106.33
AICC (smaller is better)	1966.33
BIC (smaller is better)	149.35
CAIC (smaller is better)	179.35
HQIC (smaller is better)	120.36
Pearson Chi-Square	2.67
Pearson Chi-Square / DF	0.09

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treat	14	31	1.23	0.3012
blk	14	31	2.76	0.0090

treat Least Squares Means					
treat	Estimate	Standard Error	DF	t Value	Pr > |t|
1	2.8456	0.1634	31	17.41	<.0001
2	2.4128	0.1634	31	14.76	<.0001
3	2.4517	0.1634	31	15.00	<.0001
4	2.6833	0.1634	31	16.42	<.0001
5	2.8067	0.1634	31	17.17	<.0001
6	2.9039	0.1634	31	17.77	<.0001
7	2.7711	0.1634	31	16.96	<.0001
8	2.8100	0.1634	31	17.19	<.0001
9	2.9333	0.1634	31	17.95	<.0001
10	2.5150	0.1634	31	15.39	<.0001
11	2.8539	0.1634	31	17.46	<.0001
12	3.0128	0.1634	31	18.44	<.0001
13	2.6683	0.1634	31	16.33	<.0001
14	2.5333	0.1634	31	15.50	<.0001
15	2.8483	0.1634	31	17.43	<.0001

Interpretation

As with the Fit Statistics output for the RCBD, only the last two lines are relevant to interpreting these results. The Pearson Chi-Square is equivalent to the residual sum of squares in an ANOVA table, and hence the Pearson Chi-square/DF gives the MS(residual) and is thus the estimated residual variance, $\hat{\sigma}^2 = 8.62$. The F value for differences between (adjusted) treatment differences is given in the Type III Tests of Fixed Effects: F= 1.23 and its associated p-value is 0.3012.

The least-squares means, obtained from the LSMEANS statement, are usually called adjusted means in standard textbooks. In complete block designs, the LSMEANS and the usual arithmetic means that you would calculate by hand are the same. This is not true for incomplete blocks designs, or for complete block designs with missing data. Both are examples of "unbalanced" designs in the standard design sense as defined above. Data for a given treatment in a block design with unbalance come from only a subset of the blocks. Each treatment is observed on a potentially unique subset of blocks. For example, in the PBIB example treatment 1 is observed in blocks 1, 2, 3 and 6, whereas treatment 2 is observed in blocks 3, 4, 9, and 12. If you compared unadjusted sample means of these two treatments, they would be confounded with blocks. In other words, if the sample means differ, you could not say whether it was a treatment 1 versus 2 difference, or a blocks 1, 2 and 6 versus blocks 4, 9, and 12 difference. A least squares mean adjusts for the fact that each treatment is observed on a different subset of blocks by taking the estimates of the intercept, treatment effect, and the average of all block effects. In other words, it is an estimate of what the treatment mean would have been if it had been observed in all blocks. For the PBIB, the LSMEAN for the i^{th} treatment is defined as the estimate of $\mu + \tau_i + (1/15)\sum_j b_j$.

Program to Compare Unadjusted and Adjusted Sample Means

Program 2.6 enables you to see the difference between unadjusted sample means and adjusted, or least squares, means and the inference associated with them.

Program 2.6

```
proc glimmix data=pbib;
  class treat blk;
  model y=treat blk/htype=1,3;
```

```
  lsmeans treat / e;
  lsmeans treat / bylevel e;
run;
```

The first LSMEANS statement causes PROC GLIMMIX to compute adjusted means. The E option enables you to see which linear combination of model parameters is being used to calculate these means. The BYLEVEL option in the second LSMEANS statement causes PROC GLIMMIX to compute unadjusted sample means, and the associated E option enables you to see how these means are calculated. The HTYPE=1,3 statement obtains TYPE I and TYPE III tests of treatment effects. If you put TREAT first in the MODEL statement, the Type I tests for treatment are not adjusted for blocks, whereas the TYPE III tests are.

Results

Selected results appear in Output 2.8.

Output 2.8: Adjusted versus Unadjusted Means with Intra-block Analysis

Type I Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treat	14	31	2.48	0.0172
blk	14	31	2.76	0.0090

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treat	14	31	1.23	0.3012
blk	14	31	2.76	0.0090

Obs	Effect	treat	adj_mean	adj_stderr	unadj_mean	unadj_stderr
1	treat	1	2.84556	0.16343	2.775	0.14676
2	treat	2	2.41278	0.16343	2.400	0.14676
3	treat	3	2.45167	0.16343	2.450	0.14676
4	treat	4	2.68333	0.16343	2.950	0.14676
5	treat	5	2.80667	0.16343	2.800	0.14676
6	treat	6	2.90389	0.16343	2.925	0.14676
7	treat	7	2.77111	0.16343	2.825	0.14676
8	treat	8	2.81000	0.16343	2.725	0.14676
9	treat	9	2.93333	0.16343	2.825	0.14676
10	treat	10	2.51500	0.16343	2.450	0.14676
11	treat	11	2.85389	0.16343	2.975	0.14676
12	treat	12	3.01278	0.16343	3.125	0.14676
13	treat	13	2.66833	0.16343	2.525	0.14676
14	treat	14	2.53333	0.16343	2.425	0.14676
15	treat	15	2.84833	0.16343	2.875	0.14676

Obs	Effect	treat	blk	adj_coef4	adj_coef5	unadj_coef4	unadj_coef5
1	Intercept	_	_	1.00000	1.00000	1.00	1.00
2	treat	1	_	0.00000	0.00000	0.00	0.00
3	treat	2	_	0.00000	0.00000	0.00	0.00
4	treat	3	_	0.00000	0.00000	0.00	0.00
5	treat	4	_	1.00000	0.00000	1.00	0.00
6	treat	5	_	0.00000	1.00000	0.00	1.00

Obs	Effect	treat	blk	adj_coef4	adj_coef5	unadj_coef4	unadj_coef5
7	treat	6	_	0.00000	0.00000	0.00	0.00
8	treat	7	_	0.00000	0.00000	0.00	0.00
9	treat	8	_	0.00000	0.00000	0.00	0.00
10	treat	9	_	0.00000	0.00000	0.00	0.00
11	treat	10	_	0.00000	0.00000	0.00	0.00
12	treat	11	_	0.00000	0.00000	0.00	0.00
13	treat	12	_	0.00000	0.00000	0.00	0.00
14	treat	13	_	0.00000	0.00000	0.00	0.00
15	treat	14	_	0.00000	0.00000	0.00	0.00
16	treat	15	_	0.00000	0.00000	0.00	0.00
17	blk	_	1	0.06667	0.06667	0.00	0.00
18	blk	_	2	0.06667	0.06667	0.00	0.25
19	blk	_	3	0.06667	0.06667	0.00	0.00
20	blk	_	4	0.06667	0.06667	0.00	0.00
21	blk	_	5	0.06667	0.06667	0.25	0.00
22	blk	_	6	0.06667	0.06667	0.25	0.00
23	blk	_	7	0.06667	0.06667	0.00	0.00
24	blk	_	8	0.06667	0.06667	0.00	0.25
25	blk	_	9	0.06667	0.06667	0.25	0.25
26	blk	_	10	0.06667	0.06667	0.00	0.00
27	blk	_	11	0.06667	0.06667	0.00	0.00
28	blk	_	12	0.06667	0.06667	0.00	0.00
29	blk	_	13	0.06667	0.06667	0.00	0.25
30	blk	_	14	0.06667	0.06667	0.00	0.00
31	blk	_	15	0.06667	0.06667	0.25	0.00

The first two tables show the unadjusted (Type I) and adjusted (Type III) test of overall treatment effect. Notice that the *p*-values are noticeably different. Based on the adjusted test, with $p = 0.3012$, you would conclude that the treatment effect is not statistically significant; based on the Type I $p = 0.0172$, you find a statistically significant difference among the treatments. However, a careful examination of the next two tables reveals that the Type I test is confounded with block effects.

The third table shows the adjusted and unadjusted means and their respective standard errors. Notice that these means are not the same. In particular, consider treatments 4 and 5. The unadjusted means are 2.95 and 2.80, respectively, whereas the adjusted means are 2.68 and 2.81, respectively. With the unadjusted analysis, you would conclude that there is a treatment effect and that the mean of treatment 4 is greater than the mean of treatment 5. With the adjusted means, you would conclude that the mean of treatment 5 is greater, but there is insufficient evidence to conclude that a treatment effect exists. The fourth table clarifies the problem with the unadjusted means. This table shows the results of the E option for both sets of means; in the interest of space, only the coefficients of treatments 4 and 5 are given. In the usual PROC GLIMMIX output, these variables are named ROW4 and ROW5—here they are re-named ADJ_COEF4, UNADJ_COEF4, etc. The values in each column give the coefficients of the model effects used to compute the respective mean. For example, the adjusted mean for treatment 4 is computed as follows:

$$\mu + \tau_4 + 0.06667\left(\sum_{j=1}^{15} b_j\right)$$

The unadjusted mean is computed as $\mu + \tau_4 + 0.25(b_5 + b_6 + b_9 + b_{16})$. These tell you what each mean estimates. You can see that if you take the difference between the adjusted means, you estimate $\tau_4 - \tau_5$, whereas if you take the difference between the unadjusted means you estimate $\tau_4 - \tau_5 + 0.25\left(b_5 + b_6 + b_{15} - b_2 - b_8 - b_{13}\right)$. With the latter, you have no way of knowing if treatments 4 and 5 are different, or if blocks 5, 6, and 15 differ from blocks 2, 8, and 13. This is why you use treatments results adjusted for blocks—in essence, Type III tests of fixed effects for treatment and default LSMEANS—and not Type I tests or hand-calculated sample means.

The means and their standard errors in intra-block analysis stem from the ordinary least squares (OLS) estimation. Thus, they do not take into account the fact that blocks are random. The adjustment of treatment means to remove block effects is a computation that treats blocks simply as another fixed effect. The intra-block analysis does not use all available information about the treatment effects, and thus it is suboptimal compared to the combined intra- and inter-block estimators provided by PROC GLIMMIX and PROC MIXED.

2.4.2 Combined Intra- and Inter-block PBIB Data Analysis with PROC GLIMMIX

When blocks are really treated as random, the result is the combined intra- and inter-block analysis. You can obtain this analysis with either the GLIMMIX or MIXED procedure.

Program

The PROC GLIMMIX statements are given in Program 2.7.

Program 2.7

```
proc glimmix data=pbib;
   class blk treat;
   model response=treat;
   random blk;
   lsmeans treat/diff;
run;
```

The primary difference between these statements and those for intra-block analysis is that BLK appears in the RANDOM statement instead of the MODEL statement. You could add the HTYPE=1,3 option to the MODEL statement, and a second LSMEANS statement using the BYLEVEL option, as shown in the previous section. You will find that the Type I and III tests, and the default and BYLEVEL means are identical with block effects assumed to be random. This is because with random block effects, the estimable function to LSMEANS is $\mu + \tau_i$ and does not require coefficients for the b_j terms. The resulting LSMEANS are adjusted, but the adjustment occurs differently than it does in intra-block analysis. This is explained in more detail below. The DIFF option causes all possible pairwise differences—there are $(15 \times 14)/2 = 105$ of them—to be computed. These are computed in this section to illustrate the role of the standard error of the difference in defining what "partially balanced" means in a PBIBD. Other mean comparison options are presented in detail in Chapter 3.

Results

Selected PROC GLIMMIX results appear in Output 2.9 for the combined intra- and inter-block analysis.

Output 2.9: Incomplete Block Design: PROC GLIMMIX Analysis

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
blk	0.04652	0.02795
Residual	0.08556	0.02158

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treat	14	31	1.53	0.1576

treat Least Squares Means					
treat	Estimate	Standard Error	DF	t Value	Pr > |t|
1	2.8175	0.1664	31	16.93	<.0001
2	2.4053	0.1664	31	14.45	<.0001
3	2.4549	0.1664	31	14.75	<.0001
4	2.7838	0.1664	31	16.73	<.0001
5	2.8049	0.1664	31	16.86	<.0001

treat Least Squares Means					
treat	Estimate	Standard Error	DF	t Value	Pr > |t|
6	2.9107	0.1664	31	17.49	<.0001
7	2.7890	0.1664	31	16.76	<.0001
8	2.7816	0.1664	31	16.72	<.0001
9	2.8913	0.1664	31	17.37	<.0001
10	2.4911	0.1664	31	14.97	<.0001
11	2.8987	0.1664	31	17.42	<.0001
12	3.0528	0.1664	31	18.34	<.0001
13	2.6178	0.1664	31	15.73	<.0001
14	2.4913	0.1664	31	14.97	<.0001
15	2.8592	0.1664	31	17.18	<.0001

Interpretation

Information about the effect of blocks moves from the test of fixed effects output to the Covariance Parameter Estimates. The estimated block variance is $\sigma_b^2 = 0.04562$. The REML estimate of the residual variance component is 0.08556, compared to 0.086154 from the intra-block analysis (Output 2.9). Although PROC GLIMMIX output gives standard errors of variance component estimates, these are asymptotic standard errors.

The *F* statistic in the "Type 3 Tests of Fixed Effects" table is 1.53 with a *p*-value of 0.1576. Compare this to the results from the intra-block analysis (Output 2.8, $F = 1.23$, $p = 0.3012$). This smaller *p*-value in the mixed model analysis is the result of increased power associated with the combined intra- and inter-block estimates of the treatment effects.

The Least Squares Mean estimates of the treatment means are similar, but not identical, to the adjusted means in the intra-block analysis in Section 2.4.1. For example, the estimate of the treatment 1 mean is 2.817, compared with the intra-block estimate of 2.846. The latter is an ordinary least squares (OLS) estimate, whereas the former is a mixed model estimate, equivalent to (estimated) generalized least squares (GLS). Theoretically, the GLS estimate is superior, because it accounts for BLK being random and computes the estimate of the best linear unbiased estimate (EBLUE) accordingly, substituting estimates of the variance components for block and residual. Likewise, the standard errors in the combined inter- and intra-block analysis are different from those in Section 2.4.1. The standard error of the OLS estimate is 0.163 whereas the GLS estimate is 0.166. The former is not a valid estimate of the true standard error, for the same reason that the fixed-block-effect analysis did not compute a valid standard error estimate for a treatment mean for the RCBD data in Section 2.2.1: the random effects of blocks were ignored.

Program

In the combined inter- and intra-block PROC GLIMMIX run (Program 2.7), the differences of the least-squares means were saved to a data set with the ODS OUTPUT statement. We now want to carry out additional processing on these differences. Program 2.8 shows how. First, a data set (PAIRS) is created that contains the pairs of observations that occur together in a block in this partially balanced incomplete block design.

Program 2.8

```
data pairs;
   set pbib_mv;
   array tx{4} trt1-trt4;
   array yy{4} y1-y4;
   do i=1 to 3; do j=(i+1) to 4;
      treat  = min(tx{i},tx{j});
      _treat = max(tx{i},tx{j});
      output;
   end; end;
   keep blk treat _treat;
run;
proc sort data=pairs nodupkey; by treat _treat; run;
proc print data=pairs(obs=23); run;
```

The PAIRS data set is created from the original data in multivariate format. The variables TREAT and _TREAT are set up to match the variables by the same name in the DIFMIX data set that was created in the PROC GLIMMIX call.

Results

Output 2.10 shows the first 23 observations of the PAIRS data set. These observations correspond to the pairings of treatments within a block that involve the first two treatments.

Output 2.10: Pairs within a Block Involving Treatments 1 and 2

Obs	blk	treat	_treat
1	3	1	2
2	6	1	3
3	6	1	4
4	2	1	5
5	2	1	7
6	2	1	8
7	1	1	9
8	3	1	10
9	6	1	12
10	1	1	13
11	3	1	14
12	1	1	15
13	4	2	3
14	9	2	4
15	9	2	5
16	12	2	6
17	12	2	8
18	12	2	9
19	3	2	10
20	4	2	11
21	9	2	13
22	3	2	14
23	4	2	15

Interpretation

Treatment 1 occurs with all other treatments somewhere in a block, except for treatments 6 and 11. Similarly, treatment 2 appears with all but treatments 7 and 12. Pairs of treatments that never appear in the same block are called "disconnected pairs."

Next, the output data set of treatment mean differences was sorted by StdErr. This reveals that there are two values of standard errors. Output 2.11 shows all pairs with the greater standard error value. Output 2.12 shows a subset of the standard errors of differences with the lower standard error, specifically differences between treatment 1 or 2 and all other treatments for which the standard error is at the lower level.

Output 2.11: Least-Squares Means Differences for Disconnected Pairs

treat	_treat	Estimate	StdErr	Probt
1	6	-0.09317	0.2272	0.6846
1	11	-0.08118	0.2272	0.7233
2	7	-0.3837	0.2272	0.1013
2	12	-0.6475	0.2272	0.0077
3	8	-0.3267	0.2272	0.1605
3	13	-0.1628	0.2272	0.4789
4	9	-0.1075	0.2272	0.6395
4	14	0.2925	0.2272	0.2075
5	10	0.3138	0.2272	0.1771

treat	_treat	Estimate	StdErr	Probt
5	15	-0.05434	0.2272	0.8126
6	11	0.01199	0.2272	0.9582
7	12	-0.2638	0.2272	0.2544
8	13	0.1638	0.2272	0.4762
9	14	0.4000	0.2272	0.0882
10	15	-0.3682	0.2272	0.1153

Output 2.12: Least-Squares Means Differences for Connected Pairs Involving Treatments 1 and 2

treat	_treat	Estimate	StdErr	Probt
1	2	0.4122	0.2221	0.0729
1	3	0.3626	0.2221	0.1126
1	4	0.03369	0.2221	0.8804
1	5	0.01262	0.2221	0.9550
1	7	0.02854	0.2221	0.8986
1	8	0.03592	0.2221	0.8726
1	9	-0.07379	0.2221	0.7419
1	10	0.3265	0.2221	0.1516
1	12	-0.2353	0.2221	0.2975
1	13	0.1998	0.2221	0.3753
1	14	0.3262	0.2221	0.1519
1	15	-0.04171	0.2221	0.8522
2	3	-0.04963	0.2221	0.8246
2	4	-0.3785	0.2221	0.0983
2	5	-0.3996	0.2221	0.0817
2	6	-0.5054	0.2221	0.0299
2	8	-0.3763	0.2221	0.1002
2	9	-0.4860	0.2221	0.0363
2	10	-0.08575	0.2221	0.7020
2	11	-0.4934	0.2221	0.0337
2	13	-0.2125	0.2221	0.3461
2	14	-0.08600	0.2221	0.7012
2	15	-0.4539	0.2221	0.0495

In Output 2.11, all of the standard errors of differences between disconnected pairs are 0.2272, whereas in Output 2.12, all standard errors for differences between connected pairs are 0.2221. Although only the results for the connected pairs of treatments 1 and 2 are shown in Output 2.12, similar results are obtained for the other treatments. These standard errors differ because the treatment pairs in Output 2.11 were observed together in the same block a different number of times—in this case zero—than the pairs in Output 2.12. This is a defining characteristic of a PBIBD—there are exactly two levels of standard error of the difference. In a BIBD, there is only one level—if there is more than one level, the design is not a BIBD. The more times that treatment pairs appear together in the same block in a given design, the lower the standard error will be. Although in this example the difference is small, it is an important difference, because it reflects the decreased precision that is the result of disconnected treatment pairs. Contrasts involving treatments that do not appear in the same block are not estimated with the same precision as contrasts involving treatments that do appear in the same block. Chapter 4 covers procedures that include these principles in the design of experiments.

As noted above, textbooks that do include sections on combined inter- and intra-block analysis, i.e., assuming random block effects, often include cautionary warnings about using this analysis. Textbooks vary in the nature and extent of these warnings. Some appear to dismiss mixed model analysis altogether, while some warn against mixed model analysis when the number of blocks is "small." The definition of "small" varies. Stroup (2015) reported a simulation study on the behavior of mixed model analysis of incomplete block designs. In most cases, even when the assumptions of the randomized block mixed model are violated (including cases where the block effect distribution is bimodal or beta with most of the probability density at zero or one, the performance of the combined inter- and intra-block analysis was still

equal to or superior to intra-block analysis, even with incomplete block designs with 4 blocks. The only case for which any caution seems valid is if the number of blocks is small (≤ 6) *and* the block variance is small relative to the residual variance ($\sigma_b^2/\sigma^2 < 0.5$). Otherwise, the warnings appear to be more of a holdover from an era before PROC MIXED appeared when recovery of inter-block information was difficult, and research comparing intra-block analysis to random block effect analysis focused more on "is it worth the trouble?" rather than "because it is easy with mixed-model software, is there any harm in using it?" Taking the latter perspective—the more relevant perspective given easy access to mixed model software—our answer is as follows: with the one exception noted above, no, there is no harm: you are never worse off, and you are usually better off, if you use the mixed model, random block effect approach.

2.5 Analysis with a Negative Block Variance Estimate: An Example

This section focuses on cases when the default algorithms in PROC GLIMMIX and PROC MIXED set the block variance estimate to zero. This is accompanied by a warning in the SAS LOG, "Estimated G matrix is not positive definite." Users ask if this is a problem and the answer is, "Yes, it is." This section covers an example, discusses why this occurs, why it is a problem, and what to do about it.

2.5.1 Illustration of the Problem

With this example we demonstrate a case in which the estimate of the block variance has been set to zero.

Program

The data set titled "RCBD with Negative Variance Estimate" contains data analyzed using the statements in Program 2.9.

Program 2.9

```
proc glimmix data=zero_v_ex;
 class blk trt;
 model y=trt;
 random blk;
run;
```

Results

This produces the result shown in Output 2.13.

Output 2.13: Variance Components Estimates Illustrating Set-to-Zero Default

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
block	0	.
Residual	8.4792	2.4477

Notice that the block variance estimate has been set to zero. Rerunning the analysis using PROC MIXED with option METHOD=TYPE3 produces an insight into why this happens.

Output 2.14: ANOVA Table Generating Negative Block Variance Estimate

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
trt	5	59.184969	11.836994	Var(Residual) + Q(trt)	MS(Residual)	20	1.18	0.3542
block	4	2.663237	0.665809	Var(Residual) + 6 Var(block)	MS(Residual)	20	0.07	0.9913
Residual	20	200.838496	10.041925	Var(Residual)	.	.	.	.

The MS(blk) is less than MS(residual). Recalling the ANOVA estimate of the block variance from Section 2.3.1, you can see that this results in a negative estimate. Because variance cannot be negative, the traditional approach is to set the variance estimate to zero, the lowest number within the block variance's parameter space.

Unfortunately, while setting the estimate to zero solves the problem of trying to explain how a variance can be negative (it can't), doing so has undesirable consequences. Specifically, for tests of hypotheses about treatment effects, it raises the Type I error rate relative to the nominal α-level, and it reduces the accuracy of confidence interval coverage. Using a simulation study, Littell and Stroup (2003) document the consequences of the set-to-zero default.

The reason these problems occur can be seen from the discrepancy between the residual variance estimate in the default output and the MS(residual). We know that the MS(residual) is an unbiased estimate of the residual variance. When the set-to-zero default is invoked, MS(blk) is pooled with MS(residual), as are the degrees of freedom for block and residual. The result is a downward bias in the residual variance estimate and an upward bias in the *t* and *F* statistics used to test hypotheses. The default *F* statistic for TRT is 1.40 versus 1.18 using the ANOVA MS(residual).

There are two ways to avoid this problem without abandoning the benefits of mixed model analysis. These are discussed in the next two sections.

First, some comments about negative variance estimates that have more to do with design than with analysis. A zero estimate may suggest a number of things. It may indicate that variation associated with the criterion used to block is relatively small compared to background noise. If so, the likelihood of data producing a MS(blk) less than MS(residual) is rather high. In such cases, the options given in the next two sections are strongly recommended. On the other hand, a zero variance estimate may suggest a flawed design. Often it means that blocking was not done in a manner consistent with the blocking criterion. See the discussion in Chapter 4 on effective versus ineffective blocking strategies. If flawed blocking is the case, all bets are off. Before proceeding, you should always do a retrospective and be willing to ask hard questions about how the design was implemented and how the data were collected.

To conclude, one common practice that we strongly discourage is pooling block and error sources of variation. This is equivalent to the set-to-zero approach, and, as noted above, is a recipe for inflated Type I error rates and poor confidence interval coverage. If the design used blocking, the data analyst must respect the design.

2.5.2 Use of NOBOUND to Avoid Loss of Type I Error Control

In this approach, you override the set-to-zero default using the option NOBOUND in the PROC statement. You can use NOBOUND in either PROC GLIMMIX or PROC MIXED.

Program

The PROC GLIMMIX statements are shown in Program 2.10.

Program 2.10

```
proc glimmix data=zero_v_ex nobound;
 class block trt;
 model y=trt;
 random block;
run;
```

Results

Output 2.15: Variance Estimates Obtained Using the PROC GLIMMIX NOBOUND Option

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
block	-1.5627	0.5350
Residual	10.0419	3.1755

Notice that the variance estimate for BLK corresponds to $(1/t)\left[\text{MS}(blk) - \text{MS}(residual)\right] = (1/6)\left[0.66 - 10.04\right]$ from the ANOVA table above. The residual variance estimate is now equal to the MS(*residual*). NOBOUND results in the following *F* statistic for treatment:

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
trt	5	20	1.18	0.3542

As with the variance estimates, this result agrees with the ANOVA table. The main problem with the NOBOUND option is that it makes interpreting the block variance awkward.

2.5.3 Re-parameterization of the Model as Compound Symmetry

As an alternative to NOBOUND, you can re-parameterize the randomized block model as follows to avoid the need to report a negative variance. The re-parameterized model is called a *compound symmetry* model, which is an important tool for mixed model analysis.

Re-write the model from Equation 2.1 as $y_{ij} = \mu + \tau_i + w_{ij}$, instead of $y_{ij} = \mu + \tau_i + b_j + e_{ij}$. That is, let $w_{ij} = b_j + e_{ij}$. You can easily show that $\text{Var}(w_{ij}) = \sigma_b^2 + \sigma^2$ and $\text{Cov}(w_{ij}, w_{i'j}) = \sigma_b^2$. It follows that the correlation between any two observations on different treatments in the same block is $\rho = \sigma_b^2/(\sigma_b^2 + \sigma^2)$. This is called the *intra-class correlation*. The model equation $y_{ij} = \mu + \tau_i + w_{ij}$, with $\text{Cov}(w_{ij}, w_{i'j})$ redefined as the intra-class covariance and denoted σ_w, is the simplest version of the compound symmetry covariance model. If $\sigma_w \geq 0$, the compound symmetry and randomized block model, with b_j defined as a random effect, are equivalent. However, unlike σ_b^2 in the randomized block model, σ_w, being a covariance, is not required to be nonnegative. The compound symmetry model enables you to interpret an apparently negative variance as a covariance. In fact, in many experiments, there are competition effects among experimental units within blocks, making negative covariance an unsurprising result.

Program

You can implement the compound symmetry model for randomized block designs using Program 2.11.

Program 2.11

```
proc glimmix;
 class blk trt;
 model y=trt;
 random trt / subject=blk type=cs residual;
run;
```

Read the RANDOM statement beginning with SUBJECT=BLK. This signifies that the residuals are assumed to be correlated within each block. TYPE=CS signifies that the correlation structure is compound symmetry. RANDOM TRT does *not* mean that TRT effects are random—it merely signifies that TRT identifies each observation within a block, and that the number of treatment levels determines the dimension of the covariance structure within each subject level, i.e., block in this case. The word RESIDUAL signifies that this covariance is part of the residual variance structure, not a random model effect. An equivalent way to write the RANDOM statement is as follows:

```
 random _residual_ / subject=blk type=cs;
```

The equivalent PROC MIXED statements are:

```
proc mixed;
 class blk trt;
 model y=trt;
 repeated / subject=blk type=cs;
run;
```

Results

The results from PROC GLIMMIX appear in Output 2.15.

Output 2.16: Compound Symmetry Covariance Estimates Obtained Using PROC GLIMMIX

Covariance Parameter Estimates			
Cov Parm	**Subject**	**Estimate**	**Standard Error**
CS	block	-1.5627	0.5350
Residual		10.0419	3.1755

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
trt	5	20	1.18	0.3542

Notice that the variance estimates are identical to those obtained using the NOBOUND option, as is the test statistic for TRT. The only difference is that the block variance has been relabeled as the CS covariance associated with block.

The compound symmetry model is the simplest form of a *marginal* model. Mixed models in which all sources of variation over and above residual appear as random effects in the linear predictor are called *conditional* mixed models. The name "conditional" is somewhat misleading: the best way to understand the distinction is that conditional mixed models enable you to compute predictions (called "Best Linear Unbiased Predictors") from linear combinations of fixed and random effects, whereas marginal models have only fixed effects in the linear predictor and account for all random variation through the residual covariance structure.

Note that the above comments are primarily applicable in cases involving designed experiments or observational studies with a clearly defined design structure (e.g. matched-pairs). In such cases, it is important to respect the design in the sense that sources of variation that are part of the design structure must be accounted for by the model in some form. On the other hand, in observational studies in which a term is included in the model because it is *suspected* to be a source of variation, but not *known* to be a source of variation, dropping the effect if it produces a set-to-zero variance estimate is preferable to the NOBOUND or compound symmetry approach.

2.6 Introduction to Mixed Model Theory

The randomized complete block design presents one of the simplest applications of mixed models. It has one fixed effect (treatments) and one random effect (blocks). In this section, we use the RCBD to introduce essential theory that underlies the mixed model. Refer to Appendix A, "Linear Mixed Model Theory," for the general setting and for additional details.

2.6.1 Review of Regression Model in Matrix Notation

The standard equation for the linear regression model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$$

In an application there would be n observed values of y and corresponding values of the predictor variables $x_1,\ldots, x_k$. Often the values of y are considered to be independent realizations with equal variance. These can be represented in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.2}$$

where the terms are defined as follows:

- $\mathbf{Y}$ is the n vector of observations.
- $\mathbf{X}$ is an $n \times (k + 1)$ matrix comprising a column of 1s and columns of observed values of $x_1,\ldots, x_k$.
- $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \beta_2 & \ldots & \beta_k\end{bmatrix}'$ is the n vector of regression coefficients.
- $\mathbf{e}$ is a vector of realizations of the errors e.

At this point we assume only that the error vector $\mathbf{e}$ has mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$, denoted as $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. This covariance matrix reflects the assumption of uncorrelated errors. In linear mixed models, we add the assumptions that the errors are normally distributed, denoted as $\mathbf{e} \sim \mathrm{N}(0, \sigma^2\mathbf{I})$. Note that when the normality assumption is added, lack of correlation among the errors is tantamount to independence of the errors. Refer to the expression $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ as the *model equation* form of the fixed effects linear model.

Alternatively, you can rewrite Equation 2.2 as follows:

$$\mathbf{Y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right). \tag{2.3}$$

Refer to this as the *probability distribution* form of the fixed effects linear model. There are two advantages to the probability distribution form. First, Equation 2.3 makes it clear that $\mathbf{X}\boldsymbol{\beta}$ models $\mathrm{E}\left(\mathbf{Y}\right)$. Second, the probability distribution form can be generalized to describe linear models for non-Gaussian data, whereas the model equation form cannot. The model equation form is useful for theoretical development of the mixed model with Gaussian (normally

distributed) data, so we will continue to use it when appropriate in this book. However, it general, Equation 2.3 is the preferred form.

2.6.2 The RCBD Model in Matrix Notation

The RCBD model in Equation 2.1 can be written in matrix notation. In explicit detail, the model equation is as follows:

$$\begin{bmatrix} y_{11} \\ \cdot \\ \cdot \\ \cdot \\ y_{t1} \\ \cdot \\ \cdot \\ \cdot \\ y_{1r} \\ \cdot \\ \cdot \\ \cdot \\ y_{tr} \end{bmatrix} = \begin{bmatrix} 1\,1 \ldots 0 \\ \cdots \\ \cdots \\ \cdots \\ 1\,0 \quad 1 \\ \cdots \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \\ 1\,1 \ldots 0 \\ \cdots \\ \cdots \\ \cdots \\ 1\,0 \quad 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \cdot \\ \cdot \\ \cdot \\ \tau_t \end{bmatrix} + \begin{bmatrix} 1 \ldots 0 \\ \cdots \\ \cdots \\ \cdots \\ 1 \ldots 0 \\ \cdots \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \\ 0 \ldots 1 \\ \cdots \\ \cdots \\ \cdots \\ 0 \ldots 1 \end{bmatrix} \begin{bmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_r \end{bmatrix} + \begin{bmatrix} e_{11} \\ \cdot \\ \cdot \\ \cdot \\ e_{t1} \\ \cdot \\ \cdot \\ \cdot \\ e_{1r} \\ \cdot \\ \cdot \\ \cdot \\ e_{tr} \end{bmatrix}$$

The terms are defined as in Equation 2.1. In more compact matrix notation the equation is as follows:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} \tag{2.4}$$

The definitions are as follows:

- **Y** is the vector of observations
- **X** is the treatment design matrix
- **β** is the vector of treatment fixed effect parameters
- **Z** is the block design matrix
- **u** is the vector of random block effects
- **e** is the vector of residuals

The model Equation 2.4 states that the vector **Y** of observations can be expressed as a sum of fixed treatment effects **Xβ**, random block effects **Zu**, and random experimental errors **e**. The **Xβ** portion is defined by the MODEL statement, and the **Zu** portion is defined by the RANDOM statement. It is not necessary in this example to define the residuals **e**.

For the RCBD model in matrix notation, the random vector **u** has a multivariate normal distribution with mean vector **0** and covariance matrix $\sigma_b^2\mathbf{I}_t$, $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \sigma_b^2\mathbf{I}_t)$, and the random vector **e** is distributed $\mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_{tr})$.

As with the fixed effects linear model, you can express Equation 2.4 in probability distribution form as

$$\mathbf{Y} \mid \mathbf{u} \sim \mathrm{N}\left(\mathbf{X\beta} + \mathbf{Zu}, \mathbf{R}\right);\ \mathbf{u} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{G}\right) \tag{2.5}$$

For the randomized block design as presented in this chapter, $\mathbf{R} = \sigma^2\mathbf{I}$, and $\mathbf{G} = \sigma_b^2\mathbf{I}_t$. As with Equation 2.3, Equation 2.5 can be adapted for mixed models with non-Gaussian data. Notice that with a mixed model, the distribution of the observation vector, **Y**, as conditional on the random model effects, and the mixed model linear predictor is used to estimate the conditional expectation, $\mathrm{E}\left(\mathbf{Y} \mid \mathbf{u}\right)$.

You can also write the marginal distribution of **Y** in model form as follows:

$$\mathbf{Y} \sim \mathrm{N}\left(\mathbf{X\beta}, \mathbf{V}\right)$$

where $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$. The specific form of $\mathbf{V}$ for the randomized block design is

$$\mathbf{V} = \begin{bmatrix} \sigma_b^2\mathbf{J}_t & \mathbf{0}_{t\times t} & \cdots & \mathbf{0}_{t\times t} \\ \mathbf{0}_{t\times t} & \sigma_b^2\mathbf{J}_t & \cdots & \mathbf{0}_{t\times t} \\ \mathbf{0}_{t\times t} & \mathbf{0}_{t\times t} & \ddots & \vdots \\ \mathbf{0}_{t\times t} & \mathbf{0}_{t\times t} & \cdots & \sigma_b^2\mathbf{J}_t \end{bmatrix} + \sigma^2\mathbf{I}_{tr}$$

$$= \begin{bmatrix} \sigma_b^2\mathbf{J}_t + \sigma^2\mathbf{I}_t & \mathbf{0}_{t\times t} & \cdots & \mathbf{0}_{t\times t} \\ \mathbf{0}_{t\times t} & \sigma_b^2\mathbf{J}_t + \sigma^2\mathbf{I}_t & \cdots & \mathbf{0}_{t\times t} \\ \mathbf{0}_{t\times t} & \mathbf{0}_{t\times t} & \ddots & \vdots \\ \mathbf{0}_{t\times t} & \mathbf{0}_{t\times t} & \cdots & \sigma_b^2\mathbf{J}_t + \sigma^2\mathbf{I}_t \end{bmatrix}$$

where $\sigma_b^2\mathbf{J}_t + \sigma^2\mathbf{I}_t$ is the covariance matrix of the observations in a particular block, $\mathbf{0}_{t\times t}$ is a $t \times t$ matrix of zeros, and $\mathbf{J}_t$ is a $t \times t$ matrix of ones.

Alternatively, you can redefine σ_b^2 as the compound symmetry covariance, denoted as σ_{cs} or you can denote $\text{Var}(y_{ij}) = \sigma_y^2$, define the intra-class correlation as $\rho = \sigma_{cs}/\sigma_y^2$ and write Var(**Y**) as the compound symmetry covariance matrix:

$$\mathbf{V} = \sigma_y^2 \begin{bmatrix} \rho\mathbf{J}_t + (1-\rho)\mathbf{I}_t & \mathbf{0}_{t\times t} & \ldots & \mathbf{0}_{t\times t} \\ \mathbf{0}_{t\times t} & \rho\mathbf{J}_t + (1-\rho)\mathbf{I}_t & \ldots & \mathbf{0}_{t\times t} \\ \mathbf{0}_{t\times t} & \mathbf{0}_{t\times t} & \ldots & \ldots \\ \mathbf{0}_{t\times t} & \mathbf{0}_{t\times t} & \ldots & \rho\mathbf{J}_t + (1-\rho)\mathbf{I}_t \end{bmatrix}$$

Note that in theory, $\sigma_b^2 = \sigma_{cs}$, $\sigma_y^2 = \sigma_b^2 + \sigma^2$ and $\rho = \sigma_b^2/(\sigma_b^2 + \sigma^2)$. You see the main advantage of the compound symmetry form in Section 2.5: it provides a useful way to deal with the problem of negative block variance estimates.

2.6.3 Inference Basics for the Randomized Block Mixed Model

If you want to estimate and perform inference on the fixed effects only—the treatment effects—you can use the fact that the estimate of $\boldsymbol{\beta}$ from the mixed model equations is equivalent to the solution from the generalized least squares (GLS) estimating equation $\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{V}^{-1}\mathbf{y}$. The matrix $\mathbf{X}$ is not of full column rank and so $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is singular. You must use a generalized inverse to obtain a GLS solution of the fixed effect parameter vector $\boldsymbol{\beta}$. But the treatment means, differences between treatment means, and contrasts are estimable. Thus, no matter what generalized inverse is used, there will be a vector $\mathbf{K}$ for which $\mathbf{K}'\boldsymbol{\beta}$ is equal to the mean, difference or contrast of interest. For example, choosing $\mathbf{K}'$ = [1,1,0,...,0] gives $\mathbf{K}'\boldsymbol{\beta} = \mu + \tau_1 = \mu_1$. Then the general theory gives

$$\text{Var}\left[\mathbf{K}'\hat{\boldsymbol{\beta}}\right] = \left(\sigma_b^2 + \sigma^2\right)/r$$

where $\hat{\boldsymbol{\beta}}$ is the generalized least-squares estimate. Likewise, $\mathbf{K}'$ = [0,1,–1,0...,0] gives $\mathbf{K}'\boldsymbol{\beta} = \mu_1 - \mu_2$, and $\text{Var}[\mathbf{K}'\hat{\boldsymbol{\beta}}] = 2\sigma^2/r$. These are the expressions presented in Section 2.2.1.

In the case of a relatively simple, balanced design such as an RCBD, the variance expressions can be derived directly from the model. This was the approach in Section 2.2.1. But more complicated, unbalanced situations require you to use the general theoretical result, $\text{Var}[\mathbf{K}'\hat{\boldsymbol{\beta}}] = \mathbf{K}'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{K}$. Given that the variance components are generally unknown and must be estimated, in practice you use the estimated variance, denoted as $\mathbf{K}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{K}$, for the following inferential statistics.

If $\mathbf{k}$ is a vector (e.g. for estimating a treatment difference or a contrast), then you can use a t statistic,

$$t = \mathbf{k}'\hat{\boldsymbol{\beta}} \Big/ \sqrt{\mathbf{k}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{k}}$$

Also, if $\mathbf{k}$ is a vector, you can obtain a confidence interval for $\mathbf{K}'\boldsymbol{\beta}$ with $\mathbf{k}'\hat{\boldsymbol{\beta}} \pm t_{\nu,\alpha/2}\sqrt{\mathbf{k}'\left(\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{k}}$, where ν denotes the degrees of freedom that are associated with the estimate of $\mathbf{k}'\left(\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{k}$.

If $\mathbf{K}$ is a matrix, which it is for testing any hypothesis with more than 1 numerator degree of freedom, square the *t* ratio to get the estimated Wald statistic, $\left(\mathbf{K}'\hat{\boldsymbol{\beta}}\right)\left[\mathbf{K}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{K}\right]^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}$. For complete block designs, when $\mathbf{K}'\boldsymbol{\beta}$ defines treatment difference or differences, the estimated Wald statistic reduces to $\left(\mathbf{K}'\hat{\boldsymbol{\beta}}\right)\left[\mathbf{K}'\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{K}\right]^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}\big/\hat{\sigma}^2$, which is equal to $\mathrm{SS}(\mathbf{K}'\boldsymbol{\beta})/\mathrm{MSE}$. Dividing by $\mathrm{rank}\left(\mathbf{K}\right)$ gives you $\mathrm{MS}(\mathbf{K}'\boldsymbol{\beta})/\mathrm{MSE}$, an *F* statistic.

These are the general theoretical results in the RCBD setting. They are provided to assist readers with a matrix and mathematical statistics background to better understand the methodology used in this chapter.

2.7 Summary

Chapter 2 begins with an example of a randomized block design with fixed treatments and random blocks. The importance of accounting for random effects in such a basic situation, to correctly compute the variance of a treatment mean, is demonstrated. The use of PROC GLIMMIX and PROC MIXED is introduced with explanations of how to set up the MODEL and RANDOM statements. Then, PROC GLM is briefly discussed, with emphasis on the fact that GLM is intended for fixed effect only models. We emphasize the basic applications that are handled appropriately by PROC MIXED and PROC GLIMMIX but not by PROC GLM. Then, an incomplete block design is used in Section 2.4 to illustrate some of the issues associated with unbalanced mixed model data. These include recovery of inter-block information, the difference between intra-block (fixed block) and combined inter-/intra-block (random block) analysis, and the difference between the arithmetic, or sample, mean and the least squares mean. The issue of negative "estimates" of variance components, why they matter and what to do about them, is discussed in Section 2.5. The chapter concludes with a section intended for readers with a matrix and mathematical statistics background, introducing mixed model theory relevant to estimation and inference for blocked designs.