Chapter 1 A Setting for Mixed Models Applications: Randomized Blocks Designs

1.1 Introduction

Blocking is a research technique that is 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. Then levels of the factor being investigated, called **treatments**, are **randomly assigned to units within the blocks**. An experiment conducted in this manner is called a **randomized blocks design**. The primary objectives usually are to estimate and compare treatment means. In most cases, the **treatment effects** are considered **fixed** because the treatments in the experiment are the only ones to which inference is to be made. That is, no conclusions will be drawn about treatments which were not employed in the experiment. **Block effects** are usually considered **random** because the blocks in the experiment are only a small subset of the larger set of blocks over which inference about treatment means is 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 hypothesis) that are valid in reference to the entire population of blocks, not just those 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 blocks design usually should contain fixed effects for treatment contributions and random effects for block contributions, making it a **mixed** model.

Section 1.2 presents the randomized blocks model as it is usually found in a basic statistical methods textbook. The standard analysis of variance methods are given, followed by an example to illustrate the standard methods. Section 1.3 illustrates using the MIXED procedure to obtain the results for the example, followed by results using the GLM procedure and the VARCOMP procedures for comparison. Then, basic mixed model theory for the randomized blocks design is given in section 1.4, including a presentation of the model in matrix notation. Section 1.5 presents an analysis of data from an incomplete blocks design to illustrate PROC MIXED and PROC GLM with unbalanced data.

1.2 Mixed Model for a Randomized Complete Blocks Design

A randomized blocks design that has each treatment applied in each block is called a **randomized complete blocks design** (RCBD). In the most common situation each treatment appears once in each block. Assume there are r blocks and t treatments and there will be one observation per experimental unit. Because each of the t treatments is applied in each of the r blocks, there are tr experimental units altogether. Letting y_{ij} denote the response from the experimental unit that received treatment i in block j, the equation for the model is

$$y_{ij} = \mu + \tau_i + b_j + e_{ij} \tag{1.1}$$

where

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

Assumptions for random effects are

- block effects are distributed **normally and independently** with mean 0 and variance σ_b^2; that is, the b_j are distributed iid N(0,σ_b^2)
- errors e_{ij} are distributed **normally and independently** with mean 0 and variance σ^2; that is, the e_{ij} are distributed iid N(0,σ^2).

These are the conventional assumptions for a randomized blocks model.

1.2.1 Means and Variances from Randomized Blocks Design

Recall that the usual objectives of a randomized blocks 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. It follows from model equation 1.1 that a treatment mean, for example, $\bar{y}_{1\bullet}$, can be written

$$\bar{y}_{1\bullet} = \mu_1 + \bar{b}_\bullet + e_{1\bullet} \tag{1.2}$$

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

$$\bar{y}_{1\bullet} - \bar{y}_{2\bullet} = \mu_1 - \mu_2 + \bar{e}_{1\bullet} - \bar{e}_{2\bullet} \tag{1.3}$$

From these expressions, you see that the variances of $\bar{y}_1$ and $\bar{y}_{1\bullet} - \bar{y}_{2\bullet}$ are

$$\text{Var}(\bar{y}_{1\bullet}) = (\sigma^2 + \sigma_b^2)/r \tag{1.4}$$

and

$$\text{Var}(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) = 2(\sigma^2)/r \quad (1.5)$$

Notice that the variance of a treatment mean $\text{Var}(\bar{y}_{1\bullet})$ contains the block variance component σ_b^2, but the variance of the difference between two means $\text{Var}(\bar{y}_{1\bullet} - \bar{y}_{2\bullet})$ does *not* contain σ_b^2. This is the manifestation of the RCBD controlling block variation; differences between treatments are estimated free of block variation.

1.2.2 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 blocks design. Our assumption is 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 means squares (Exp MS). Readers needing more information concerning analysis of variance may consult Littell, Freund, and Spector (1991), Milliken and Johnson (1984), or Winer (1991). Table 1.1 is a standard ANOVA table for the RCB, showing sources of variation, degrees of freedom, mean squares, and expected mean squares.

Table 1.1 *ANOVA Table for Randomized Complete Blocks Design*

Source of Variation	**df**	**MS**	**Exp MS**
Blocks	$r-1$	MS(Blks)	$\sigma^2 + t\sigma_b^2$
Treatments	$t-1$	MS(Trts)	$\sigma^2 + r\varphi^2$
Error	$(r-1)(t-1)$	MS(Error)	σ^2

1.2.3 Using Expected Mean Squares

Expected means squares are the quantities that are estimated by mean squares in an analysis of variance. They can be used to motivate test statistics. The basic idea is to examine the expected mean square for a factor and see how it differs under the null and alternative hypotheses. For example, the expected mean square for treatments, $E(\text{MS(Trts)}) = \sigma^2 + r\varphi^2$, can be used to determine how to set up a test statistic for treatment differences. The null hypothesis is $H_o{:}\mu_1=...=\mu_t$. The expression φ^2 in Exp MS(Trts) is $\varphi^2=\Sigma(\mu_i-\bar{\mu}_\bullet)^2/(t-1)$, where $\bar{\mu}_\bullet$ is the mean of $\mu_1,...,\mu_t$. Thus $\varphi^2=0$ is equivalent to $\mu_1=...=\mu_t$ So, if the null hypothesis $H_o{:}\mu_1=...=\mu_t$ is true, MS(Trts) simply estimates σ^2. On the other hand, if $H_o{:}\mu_1=...=\mu_t$ is false, then Exp MS(Trts) estimates a quantity larger than σ^2. Now, MS(Error) estimates σ^2 regardless of whether H_0 is true or false. Therefore, MS(Trts) and MS(Error) tend to be approximately the same magnitude if H_o is true, and MS(Trts) tends to be larger than MS(Error) if $H_o{:}\mu_1=...=\mu_t$ is false. So a comparison of MS(Trts) with MS(Error) is an indicator of whether $H_o{:}\mu_1=...=\mu_t$ 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 also can be used to estimate variance components, variances of treatment means, and differences between treatment means. They reveal that estimates of the variance components are

$$\hat{\sigma}^2 = \text{MS(Error)} \tag{1.6}$$

and

$$\hat{\sigma}_b^2 = [\text{MS(Blks)} - \text{MS(Error)}]/t \tag{1.7}$$

These are called **analysis of variance** estimates of the variance components. It follows that estimates of Var($\bar{y}_{1\bullet}$) and Var($\bar{y}_{1\bullet} - \bar{y}_{2\bullet}$) are

$$\hat{V}\text{ar}(\bar{y}_{1\bullet}) = (\hat{\sigma} + \hat{\sigma}_b^2)/r$$

$$= \text{MS(Blks)}/rt + (t-1)\text{MS(Error)}/rt \tag{1.8}$$

and

$$\hat{V}\text{ar}(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) = 2\text{MS(Error)}/r \tag{1.9}$$

The expression for Var($\bar{y}_{1\bullet}$) points out a common misconception that the estimate of the variance of a treatment mean from a randomized blocks design is simply MS(Error)/*r*. This misconception prevails in some text books and results in incorrect calculation of standard errors by computer program packages. The good news is that it gives a starting point for illustrating the MIXED procedure in the SAS System.

1.2.4 Example: A Randomized Complete Blocks Design

An example from Mendenhall, Wackerly, and Scheaffer (1990) is used to illustrate analysis of data from a randomized blocks design.

Data for an RCB designed experiment are presented in Data Set 1.1, "BOND," in Appendix 4, "SAS Data Sets." Blocks are ingots of a composition material and treatments are metals (nickel, iron, or copper). The response is the amount of pressure required to break a bond of two pieces of material from an ingot that used one of the metals as the bonding agent.

Table 1.2 *ANOVA Table for BOND Data*

Source of Variation	df	SS	MS	F	P
Ingots	6	268.29	44.72		
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 gives a test of the null hypothesis H_o: $\mu_c = \mu_i = \mu_n$. The significance probability for the F-test is $p = 0.0131$, indicating strong evidence of differences between metal means. Estimates of the variance components are $\hat{\sigma}^2 = 10.37$ and $\hat{\sigma}_b^2 = (44.72 - 10.37)/3 = 11.45$. Thus, an estimate of the variance of a metal mean is $(\hat{\sigma} + \hat{\sigma}_b^2)/7 = 3.11$, and the estimated standard error is $3.11^{1/2}=1.77$. An estimate of the variance of a difference between two metal means is $2\hat{\sigma}^2/7 = 2*10.37/7=2.96$, and the standard error is $(2.96)^{1/2}$.

A Note for PROC GLM Users

It might seem natural at this point to illustrate the use of PROC GLM to obtain the analysis described in the previous subsection. Indeed, this was strongly considered. However, the main focus of this book is on PROC MIXED, and we decided that this procedure should be the first introduced rather than forever linking PROC MIXED through PROC GLM. It is important to show relationships between the two procedures, and this is done throughout the book. Readers who want to see PROC GLM results for the RCBD first may turn to subsection 1.3.3.

1.3 Using PROC MIXED to Analyze RCBD Data

PROC MIXED is a procedure based on likelihood. That is, many of the estimation and inferential methods are implemented on the basis of the likelihood function and associated principles and theory. Most readers are probably more familiar with the analysis of variance approach described in the previous section. In fact, section 1.2 is presented as a frame of reference rather than as an indication of how PROC MIXED works. In section 1.3.1 results from PROC MIXED are shown that duplicate many of the results of the previous section.

1.3.1 Basic PROC MIXED Statements and Output

Program

Here are the basic PROC MIXED statements for the RCBD data analysis:

```
proc mixed data=rcb;
   class ingot metal;
   model pres=metal;
   random ingot;
run;
```

The PROC MIXED statement calls the procedure.

The CLASS statement specifies that INGOT and METAL are classification variables as opposed to 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.)

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

The MODEL and RANDOM statements are the core essential statements for many mixed model applications. Results from these statements appear in Output 1.1.

Results

Output 1.1 *Results of the RCBD Data Analysis*

```
                        The MIXED Procedure

                      Class Level Information❶

                  Class     Levels  Values

                  INGOT          7  1 2 3 4 5 6 7
                  METAL          3  c i n


                 REML Estimation Iteration History❷

         Iteration  Evaluations     Objective     Criterion

                 0            1   79.32809232
                 1            1   74.70841482    0.00000000

                     Convergence criteria met.


               Covariance Parameter Estimates (REML)❸

    Cov Parm          Ratio      Estimate     Std Error       Z

    INGOT        1.10376333   11.44777778    8.72036577    1.31
    Residual     1.00000000   10.37158730    4.23418279    2.45


               Covariance Parameter Estimates (REML)

               Pr > |Z|

                 0.1893
                 0.0143


                 Model Fitting Information for PRES❹
              Description                          Value

              Observations                       21.0000
              Variance Estimate                  10.3716
              Standard Deviation Estimate         3.2205
              REML Log Likelihood               -53.8951
              Akaike's Information Criterion    -55.8951
              Schwarz's Bayesian Criterion      -56.7855
              -2 REML Log Likelihood            107.7902

                     Tests of Fixed Effects❺

             Source      NDF   DDF  Type III F  Pr > F

             METAL         2    12        6.36  0.0131
```

Interpretation

The first feature you notice in Output 1.1 is that the PROC MIXED output reflects its likelihood orientation, as opposed to the analysis of variance orientation of the ANOVA and GLM procedures. Here are annotations of key portions of the output.

❶ **Class Level Information** lists the variables in the CLASS statement and their levels.

❷ **REML Estimation Iteration History** 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.

❸ **Covariance Parameter Estimates (REML)** show estimates of the variance component parameters. The estimate of σ_b^2, the block variance component, is 11.45 (labeled INGOT), and the estimate of σ^2, the error variance component, is 10.37 (labeled Residual). For this example of a balanced data set, these variance component estimates are exactly the same as the estimates obtained from the analysis of variance method. That is, the estimate of σ^2 is MS(Error) from the ANOVA table, and the estimate of σ_b^2 is (MS(Blocks)−MS(Error))/7. In more complicated unbalanced data sets, the REML estimates are not necessarily equal to the ANOVA estimates.

❹ **Model Fitting Information for PRES** shows the "Number of Observations" equal to 21. Next is the "Variance Estimate," 10.3716, which is the same as the Residual variance component. This is followed by the Standard Deviation Estimate, 3.2205, which is simply the square root of the Variance Estimate, equal to $10.3716^{1/2}$. Remaining computations are more complicated to explain and are not essential to this analysis.

❺ **Tests of Fixed Effects** is like an abbreviated ANOVA table showing a line of computations for each term in the MODEL statement, in this example, METAL. Included is an F-test for testing the null hypothesis H_0: $\mu_c = \mu_i = \mu_n$. With 2 numerator and 12 denominator degrees of freedom, the F-value of 6.36 is significant at the p=0.0131 level. 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 Table 1.2.

In summary, the basic PROC MIXED computations are based on likelihood principles, but many of the statistical computions are the same as those obtained from analysis of variance methods for a balanced data set.

1.3.2 Estimating and Comparing Means: LSMEANS, ESTIMATE, and CONTRAST Statements

You can obtain treatment means from the LSMEANS (Least Squares MEANS) statement

```
lsmeans metal / pdiff;
```

Results appear in Output 1.2.

Results Using the LSMEANS Statement

Output 1.2 *Results of Using the LSMEANS Statement*

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
METAL c	70.18571429	1.76551753	12	39.75	0.0001
METAL i	75.90000000	1.76551753	12	42.99	0.0001
METAL n	71.10000000	1.76551753	12	40.27	0.0001

Differences of Least Squares Means

Level 1	Level 2	Difference	Std Error	DDF	T	Pr > |T|
METAL c	METAL i	-5.71428571	1.72142692	12	-3.32	0.0061
METAL c	METAL n	-0.91428571	1.72142692	12	-0.53	0.6050
METAL i	METAL n	4.80000000	1.72142692	12	2.79	0.0164

Interpretation

For the case of balanced data, the LS means are simply the averages for treatments. Also printed are standard errors for the means. The value is 1.766 for each of the means. This estimate is equal to $((\hat{\sigma} + \hat{\sigma}_b^2)/7)^{1/2}$, which is a valid estimate of the true standard error $((\sigma^2+\sigma_b^2)/7)^{1/2}$ because it uses the correct linear combinations of variance components.

Pairwise comparisons of means are obtained by the PDIFF option on the LSMEANS statement. In Output 1.3 you see pairwise differences and standard errors of the differences. Each standard error has the value 1.721, equal to $(2\,\hat{\sigma}^2/7)^{1/2}$. You also see results of *t*-tests for the statistical significance of the difference between the means in each pair. The results declare both copper and nickel different from iron but copper and nickel not different from each other.

Linear combinations of means can be estimated with the ESTIMATE statement. For illustration, consider the linear combination equal to the nickel mean, μ_n. First of all, express the nickel mean as a linear combination of the model parameters, $\mu_n=\mu+\tau_n$. More explicitly, $\mu_n=1\mu+0\tau_c+0\tau_i+1\tau_n$. Then insert these coefficients of the model parameters into the ESTIMATE statement:

```
estimate 'nickel mean' intercept 1 metal 0 0 1;
```

In a similar fashion, the difference between the means for copper and iron is $\mu_c - \mu_i = \tau_c - \tau_i$, so the ESTIMATE statement is

```
estimate 'copper vs iron' metal 0 1 -1;
```

Results of these ESTIMATE statements appear in Output 1.3.

Output 1.3 *Inference about Linear Combinations of Means*

```
                                ESTIMATE Statement Results

Parameter      Estimate        Std Error     DDF     T       Pr > |T|
nickel mean    71.10000000    1.76551753     12      40.27   0.0001
copper vs iron -5.71428571    1.72142692     12      -3.32   0.0061

                                CONTRAST Statement Results

Source           NDF              DDF      F       Pr > F
copper vs iron   1               12       11.02     0.0061
```

You see that the estimates of the nickel mean and differences between the copper and iron means are the same as those obtained from the LSMEANS statement in Output 1.2.

The CONTRAST statement is a companion to the ESTIMATE statement. It is used to test hypotheses about linear combinations of model parameters. To test the null hypothesis $H_o{:}\mu_c - \mu_i = 0$, submit the statements

```
contrast 'copper vs iron' metal 1 -1 0;
```

This CONTRAST statement produces the F-test shown in Output 1.3.

The F-value of 11.02 is equal to the square of the *t*-value from the corresponding ESTIMATE statement.

The general use of the ESTIMATE and CONTRAST statements is to estimate and test linear combinations of *all* terms in the mixed model, including random effects.

1.3.3 Comparison of PROC MIXED with PROC GLM for the RCBD Data

PROC GLM was the principal SAS procedure for analyzing mixed models data prior to the advent of PROC MIXED, even though the basic computations of PROC GLM are for fixed effect models. The GLM procedure uses statements similar to those used by PROC MIXED. In this section you will see differences and similarities in the statements and output. However, you will not see complete coverage of PROC GLM capabilities. Refer to Littell, Freund, and Spector (1991) for more detailed PROC GLM coverage.

Program

Statements for PROC GLM to obtain the ANOVA table, mean estimates, and comparisons analogous to those discussed in Section 1.3.1 and 1.3.2 are

```
proc glm data=rcb;
   class ingot metal;
   model pres=ingot metal;
   lsmeans metal/stderr pdiff;
   estimate 'nickel mean' intercep 1 metal 0 0 1;
   estimate 'copper vs iron' metal 1 -1 0;
   contrast 'copper vs iron' metal 1 -1 0;
   random ingot;
run;
```

Results of these statements appear in Output 1.4.

Results

Output 1.4 *Randomized Blocks Analysis with PROC GLM*

```
                  General Linear Models Procedure
                      Class Level Information

                  Class    Levels    Values

                  INGOT         7    1 2 3 4 5 6 7

                  METAL         3    c i n


              Number of observations in data set = 21


                  General Linear Models Procedure

Dependent Variable: PRES

Source                  DF    Sum of Squares    F Value     Pr > F

Model                    8      400.19047619       4.82     0.0076

Error                   12      124.45904762

Corrected Total         20      524.64952381

                  R-Square              C.V.             PRES Mean

                  0.762777          4.448490            72.3952381


Source                  DF         Type I SS    F Value     Pr > F

INGOT                    6      268.28952381       4.31     0.0151
METAL                    2      131.90095238       6.36     0.0131

Source                  DF       Type III SS    F Value     Pr > F
```

```
INGOT                    6      268.28952381       4.31     0.0151
METAL                    2      131.90095238       6.36     0.0131


                  General Linear Models Procedure
                        Least Squares Means

      METAL          PRES       Std Err     Pr > |T|   LSMEAN
                   LSMEAN        LSMEAN   H0:LSMEAN=0   Number

      c        70.1857143     1.2172327        0.0001     1
      i        75.9000000     1.2172327        0.0001     2
      n        71.1000000     1.2172327        0.0001     3

                  Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

                       i/j     1       2       3
                       1   .      0.0061  0.6050
                       2  0.0061   .      0.0164
                       3  0.6050  0.0164   .

NOTE: To ensure overall protection level, only probabilities
      associated with pre-planned comparisons should be used.

                  General Linear Models Procedure

Source      Type III Expected Mean Square

INGOT       Var(Error) + 3 Var(INGOT)

METAL       Var(Error) + Q(METAL)



Contrast               Contrast Expected Mean Square

copper vs iron         Var(Error) + Q(METAL)

                  General Linear Models Procedure

Dependent Variable: PRES

Contrast                DF       Contrast SS    F Value     Pr > F

copper vs iron           1      114.28571429      11.02     0.0061

                                T for H0:    Pr > |T|   Std Error of
Parameter         Estimate    Parameter=0                 Estimate

copper vs iron  -5.7142857          -3.32     0.0061     1.72142692
nickel mean     71.1000000          58.41     0.0001     1.21723265
```

Interpretation

Following is a comparison of syntax and output for PROC GLM and PROC MIXED statements:

- First, both procedures use the same CLASS statements; if a variable is a classification variable in PROC GLM, then so it is with PROC MIXED.
- You see that the MODEL statements of PROC MIXED and PROC GLM are *not* exactly the same. This is a very important distinction between the procedures: In PROC MIXED, you list only the *fixed* effects in the right-hand side of the MODEL

statement. But in PROC GLM you list all effects, both fixed and random, although PROC GLM does not really treat the random effects as random. The options in PROC GLM for inference accommodating random effects are adaptations of the fixed effect computations for this procedure. PROC MIXED, on the other hand, was conceived from the outset for mixed models. The distinction between MODEL statements in PROC MIXED and PROC GLM carries over to the output from the two procedures. In Output 1.6, from PROC GLM, you see an ANOVA table listing all the terms in the MODEL statement, with no distinction between fixed and random. But in Output 1.2, from PROC MIXED, the list of Tests of Fixed Effects contains only terms in the MODEL statement.

- The LSMEANS statements for the two procedures are essentially the same, except that you need the STDERR option in the LSMEANS statement for PROC GLM. Here is another important distinction between the procedures: In comparing Output 1.3 with Output 1.4, you see that the LSMEANS estimates are the same for the two procedures, but their standard errors are *not* the same. This is due to the inherent fixed-effect nature of PROC GLM. The standard errors for the LSMEANS computed by PROC GLM are $(\hat{\sigma}^2/7)^{1/2}$, which would be appropriate if INGOT (the blocks) were fixed. Recall from section 1.3.1 that PROC MIXED performed correct computations for the LSMEAN standard errors with INGOT random.

- Syntax for ESTIMATE statements is the same in PROC GLM as in PROC MIXED for estimating linear combinations of fixed effects. The remarks comparing standard errors of LSMEANS estimates from PROC MIXED and PROC GLM also hold pertaining to ESTIMATE statements. You see in Output 1.5 from PROC GLM that the standard error of the estimate of the nickel mean is the same as for the nickel LSMEAN, which previously was stated to be incorrect.

  Estimates, their standard errors and tests of the difference between the COPPER and IRON means are the same for PROC GLM (Output 1.4) and PROC MIXED (Outputs 1.2 and 1.3). This is true for the present example, but not for all mixed model data sets, as you will see in the case of an incomplete block design in subsection 1.5.2.

- The RANDOM statements for PROC MIXED and PROC GLM represent another major distinction between the two procedures although they have the same appearance for the present example. In PROC MIXED, listing INGOT in the RANDOM statement causes all standard errors and test statistics to incorporate the information that the effect is random. This is not true in PROC GLM. The RANDOM statement in PROC GLM (as used here) merely computes expected mean squares for terms in the MODEL statement and for linear combinations in the CONTRAST statement. You must then digest the information in the expected means squares table and formulate appropriate tests. (The TEST option in the PROC GLM RANDOM statement will do this automatically for terms in the MODEL statement, but not for CONTRAST statements.) In the RCBD example, the default tests computed by PROC GLM are correct, so no modification is needed for the test of differences from the MODEL and CONTRAST statements.

These comparisons of PROC MIXED and PROC GLM are summarized in Table 1.3 in the next section.

1.3.4 Comparison of PROC VARCOMP with PROC MIXED and PROC GLM for the RCBD Data Program

The VARCOMP procedure is used to obtain estimates of the variance components in a mixed model. For the RCBD data, submit the statements

```
proc varcomp method=reml data=rcb;
   class ingot metal;
   model pres=metal ingot / fixed=1;
run;
```

PROC VARCOMP uses the same CLASS statement as PROC GLM and PROC MIXED. The MODEL statement in PROC VARCOMP contains a list of all fixed-effect variables followed by all random-effect variables. The designation FIXED=1 tells VARCOMP how many of the variables are fixed. Only classification variables are allowed in the MODEL statement of PROC VARCOMP. Results appear in Output 1.5.

Results

Output 1.5 *Variance Component Estimation Procedure*

```
                Variance Components Estimation Procedure
                        Class Level Information

                    Class    Levels    Values

                    INGOT         7    1 2 3 4 5 6 7

                    METAL         3    c i n


                Number of observations in data set = 21


             REML Variance Components Estimation Procedure

Dependent Variable: PRES


    Iteration        Objective       Var(INGOT)       Var(Error)

        0          50.87068437      11.44777778      10.37158730
        1          50.87068437      11.44777778      10.37158730

                       Convergence criteria met.

      Asymptotic Covariance Matrix of Estimates

                              Var(INGOT)        Var(Error)

             Var(INGOT)      76.04477922       -5.97610129
             Var(Error)      -5.97610129       17.92830386
```

Interpretation

You see iterative computations toward estimation of the variance components σ_b^2 and σ^2. The estimates are 11.45 and 10.37. These are the same estimates produced by PROC MIXED. The “Asymptotic Covariance Matrix of Estimates” contains estimates of the large sample variances and covariances of the variance component estimates. The variance estimate of the INGOT variance component estimate is 76.04; its square root of 8.72 is the standard error of the variance component estimate printed by PROC MIXED in Output 1.1. Likewise, the variance estimate of the ERROR variance component estimate is 17.93, and its square root of 4.23 is the standard error of the variance component estimate labeled Residual in Output 1.1. The estimate of the covariance between the variance component estimates is −5.98.

Comparisons of PROC VARCOMP, PROC GLM, and PROC MIXED for the RCBD data are summarized in Table 1.3.

Table 1.3 *Summary Comparison of Syntax and Output for PROC GLM, PROC MIXED, and PROC VARCOMP*

Statement	PROC MIXED	PROC GLM	PROC VARCOMP
CLASS	list classification variables	same	same
MODEL	specify dependent variable and list fixed effect	specify dependent variable and list all terms in model	specify dependent variable and list all terms in model, fixed effects preceding random
RANDOM	specify random effects	obtain table of expected mean squares	not applicable
LSMEANS	estimate means for fixed effect factors	estimate means for fixed effect factors	not applicable
ESTIMATE	estimate linear combination of model terms	estimate linear combination of model terms	not applicable
CONTRAST	test set of linear combinations of model terms	test set of linear combination of model terms	not applicable

1.4 Introduction to the Theory of Linear Models

You have seen an application of PROC MIXED to the RCBD data set. Before introducing other example data sets, a brief introduction to mixed model theory is presented to help you understand the basis for applications. Also, this introduction to theory helps you understand the differences between results produced by PROC MIXED and PROC GLM. This is a very brief and incomplete presentation of mixed model theory. Complete results are presented in Appendix 1. The use of matrix notation makes the results more concise and easily comprehensible than the use of summation notation.

1.4.1 Some Basic Theory Results

You are probably already familiar with the standard linear regression model in matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{1.10}$$

where

$\mathbf{Y}$ is the vector of observations
$\mathbf{X}$ is the matrix of values of independent variables
$\boldsymbol{\beta}$ is the vector of regression parameters
$\mathbf{e}$ is a vector of errors.

Ordinary least squares (OLS) estimates of the parameters in $\boldsymbol{\beta}$ are given by solving the normal equations

$$\mathbf{X'X\boldsymbol{\beta} = X'Y} \tag{1.11}$$

A solution is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-}\mathbf{X'Y}) \tag{1.12}$$

where $(\mathbf{X'X})^{-}$ is a generalized inverse of $\mathbf{X'X}$.

Notice that we have not made any assumptions about the probability distribution of the vector **e** of errors. The estimator $\hat{\boldsymbol{\beta}}$ is an OLS estimator of $\boldsymbol{\beta}$ regardless of the distribution of **e**. Now assume that the errors are independently and normally distributed with common variance σ^2. Then the covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\mathbf{X'X})^{-}$. Also, $\hat{\boldsymbol{\beta}}$ is the best estimate of $\boldsymbol{\beta}$ in the following sense: if **K** is a vector of coefficients such that the linear combination $\mathbf{K'}\boldsymbol{\beta}$ is estimable, then the best linear unbiased estimate (BLUE) of $\mathbf{K'}\boldsymbol{\beta}$ is $\mathbf{K'}\hat{\boldsymbol{\beta}}$. These results provide the foundation for statistical inference about linear combinations of the parameter vector $\boldsymbol{\beta}$.

The regression model is a special case of the general linear mixed model (GLMM), which has the equation

$$\mathbf{Y = X\boldsymbol{\beta} + Zu + e} \tag{1.13}$$

In this equation **u** is distributed multivariate normal with mean vector **0** and covariance matrix **G** (which we denote MVN(**0**,**G**)), and **e** is distributed MVN(**0**,**R**). The only requirement of the matrices **G** and **R** is that they be positive definite (that is, they are covariance matrices). Then the covariance matrix of **Y** is

$$\mathbf{V(Y) = ZGZ' + R} \tag{1.14}$$

The **RANDOM** statement in PROC MIXED **defines G**, and the **REPEATED** statement, which will be introduced in Chapter 4, **defines R**.

It follows that **Y** is distributed MVN($\mathbf{X\boldsymbol{\beta}}$,$\mathbf{ZGZ' + R}$). Thus the generalized least squares (GLS) estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'V^{-1}X})^{-}\mathbf{X'V^{-1}Y} \tag{1.15}$$

where $\mathbf{V = V(Y) = ZGZ' + R}$. Also,

$$\mathrm{V}(\hat{\boldsymbol{\beta}}) = (\mathbf{X'V^{-1}X})^{-} \tag{1.16}$$

If $\mathbf{K'}\boldsymbol{\beta}$ is estimable, then the BLUE of $\mathbf{K'}\boldsymbol{\beta}$ is $\mathbf{K'}\hat{\boldsymbol{\beta}}$, and the variance of $\mathbf{K'}\hat{\boldsymbol{\beta}}$ is $\mathbf{K'(X'V^{-1}X)^{-}K}$.

Analogous to the regression model, the preceding results provide the foundation for statistical inference about linear combinations of the parameter vector $\boldsymbol{\beta}$. That is, they can be used to test hypotheses and to construct confidence intervals about linear combinations of parameters. These are some of the basic theoretical results used by PROC MIXED.

In reality, the covariance matrices **G and R are usually functions of unknown parameters**, which must be estimated. Typically, the parameters are variance components or correlation parameters. PROC MIXED uses either ML or REML to estimate the parameters of **G** and **R** (ML and REML are described in Appendix 1). Then these estimates are substituted in place of the true parameters values in **G** and **R** to compute estimates of $\boldsymbol{\beta}$ and $\mathbf{V}(\boldsymbol{\beta})$. Also, the matrix $\mathbf{X'V^{-1}X}$ typically is singular so that a generalized inverse is used.

1.4.2 The RCBD Model in Matrix Notation

The RCBD model in equation (1.1) can be written in matrix notation. In explicit detail, the model equation is

$$\begin{bmatrix} Y_{11} \\ \cdot \\ \cdot \\ \cdot \\ Y_{t1} \\ \cdot \\ \cdot \\ \cdot \\ Y_{1r} \\ \cdot \\ \cdot \\ \cdot \\ Y_{tr} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 1 & 0 & & & & 1 \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & & \cdot & \cdot \\ 1 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 1 & 0 & & & & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \cdot \\ \cdot \\ \cdot \\ \tau_t \end{bmatrix} + \begin{bmatrix} 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_r \end{bmatrix} + \begin{bmatrix} e_{11} \\ \cdot \\ \cdot \\ \cdot \\ e_{t1} \\ \cdot \\ \cdot \\ \cdot \\ e_{1r} \\ \cdot \\ \cdot \\ \cdot \\ e_{tr} \end{bmatrix} \qquad (1.17)$$

with terms defined following equation (1.1). In more compact matrix notation the equation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad (1.18)$$

where

- **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 experimental errors.

The equation 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 experimental errors **e**.

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

The variance of the observation vector **Y** is

$$\mathbf{V} = \mathbf{V}(\mathbf{Y}) = \mathbf{ZGZ'} + \mathbf{R}$$

$$= \begin{bmatrix} \sigma_b^2\mathbf{J}_r & \Phi_r & \cdots & \Phi_r \\ \Phi_r & \sigma_b^2\mathbf{J}_r & \cdots & \Phi_r \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \Phi_r & \Phi_r & \cdots & \sigma_b^2\mathbf{J}_r \end{bmatrix} + \sigma^2\mathbf{I}_{tr}$$

where $\mathbf{V}_b = \sigma_b^2\mathbf{J}_r + \sigma^2\mathbf{I}_r$ is the covariance matrix of all the observations in a particular block, $\mathbf{\Phi}_r$ is an $r \times r$ matrix of zeros, and $\mathbf{J}_r$ is an $r \times r$ matrix of 1's.

The matrix $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is singular, so a generalized inverse must be used to obtain a GLS estimate **β** of the fixed effect parameter vector. But the treatment means and differences between treatment means are estimable parameters. Thus, no matter what generalized inverse is used, there will be a **K** vector for which **K'β** is equal to a mean or a difference between means. For example, choosing $\mathbf{K}'=(1,1,0,...,0)$ gives $\mathbf{K}'\boldsymbol{\beta} = \mu + \tau_1 = \mu_1$. Then the general theory gives $\mathbf{V}(\mathbf{K}'\boldsymbol{\beta}) = (\sigma_b^2 + \sigma^2)/r$. Likewise, $\mathbf{K}'=(0,1,-1,0...,0)$ gives $\mathbf{K}'\boldsymbol{\beta} = \mu_1 - \mu_2$, and $\mathbf{V}(\mathbf{K}'\boldsymbol{\beta}) = 2(\sigma^2)/r$. These are the expressions presented in section 1.2.1.

In the case of a relatively simple, balanced design such as an RCBD, the variance

$$= \begin{bmatrix} \mathbf{V}_b & \Phi_r & \cdots & \Phi_r \\ \Phi_r & \mathbf{V}_b & \cdots & \Phi_r \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \Phi_r & \Phi_r & \cdots & \mathbf{V}_b \end{bmatrix}$$

expressions can be derived directly from the model as was done in subsection 1.2.1. But in more complicated unbalanced situations, the general theoretical results must be invoked. In this subsection, we have illustrated the general results in the RCBD setting to confirm their validity and to assist you in becoming more comfortable in using the general linear mixed model.

1.5 Example of an Unbalanced Two-way Mixed Model: 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 are 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.

Two types of incomplete block designs are the so-called **balanced** incomplete block design (BIBD) and the **partially balanced** incomplete block design (PBIBD). This does not mean "balanced" in the more common sense of the word, in which each treatment appears the same number of times in each block. In fact, any incomplete block design is "unbalanced" by the common definition.

The BIB and PBIB designs result in all treatments having the same variance (and hence the same standard error). Also, the variances of differences between two treatment means 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. Mead (1988) has an excellent discussion of this issue.

This section presents analyses for a PBIBD using PROC GLM and PROC MIXED to further illustrate some of the similarities and differences between the two procedures. You can see some distinctions between PROC MIXED and PROC GLM that did not occur in the analyses of the RCB design. Although the example is a PBIBD, data analysis methods of this section apply to incomplete block designs in general.

Model

The equation for the model of an incomplete blocks design is the same as for an RCBD. That is, the response Y_{ij} that results from applying treatment i in block j is assumed to be equal to a treatment mean $\mu_i = \mu + \tau_i$ plus a block effect b_j, plus experimental error e_{ij}. Thus the equation

$$Y_{ij} = \mu + \tau_i + b_j + e_{ij} \qquad (1.19)$$

where the block effects b_j are iid $N(0,\sigma_b^2)$, the experimental errors e_{ij} are iid $N(0,\sigma^2)$, and the b_j are independent of the e_{ij}. An analysis of variance table for an incomplete blocks design is shown in Table 1.4.

Table 1.4 *Analysis of Variance Table for Incomplete Blocks Design*

Source of Variation	df	F
Blocks	r-1	
Treatments (adjusted for blocks)	t-1	MS(Trts adj.)/MS(Error)
Error	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. 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.

Analyses of BIBD and PBIBD data that are presented in most statistics textbooks are called **intra-block** analyses because treatments are compared on the basis of differences computed within blocks. You can perform this type of analysis with PROC GLM. It is discussed first in subsection 1.5.1 because this is the analysis with which you are most likely familiar. In subsection 1.5.2, an analysis is presented using PROC MIXED that utilizes information about treatment means contained in differences between blocks. This type of analysis combines intra- and inter-block information.

1.5.1 The Usual Intra-block Analysis of PBIB Data Using PROC GLM

Data Set 1.5.1, PBIB, in Appendix 4, "SAS Data Sets," 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 in one block (e.g., treatments 1 and 2) and others do not appear together in the same blocks (e.g., treatments 1 and 6).

Program

An intra-block analysis of the PBIBD data is obtained from submitting the statements

```
proc glm data=pbib;
   class blk treat;
   model response=blk treat;
   means treat;
   lsmeans treat / stderr pdiff;
   estimate 'treat 1 mean' intercep 1 treat 1;
   estimate 'trt 1 mean' intercep 15 treat 15
             blk 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 / divisor=15;
   estimate 'trt 1 blk 1' intercep 1 treat 1 blk 1;
   estimate 'trt 1 vs trt 2' treat 1 -1;
   contrast 'trt 1 vs trt 2' treat 1 -1;
   random block;
run;
```

Results appear in Output 1.6.

Results

Output 1.6 *Incomplete Blocks Design: PROC GLM Analysis*

```
                  General Linear Models Procedure
                      Class Level Information

       Class    Levels    Values

       BLK          15    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

       TREAT        15    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


              Number of observations in data set = 60


                  General Linear Models Procedure

Dependent Variable: RESPONSE

Source                 DF    Sum of Squares    F Value     Pr > F

Model                  28        6.32855556       2.62     0.0050

Error                  31        2.67077778

Corrected Total        59        8.99933333

                  R-Square              C.V.         RESPONSE Mean

                  0.703225          10.72546            2.73666667


Source                 DF       Type III SS    F Value     Pr > F

BLK                    14        3.33422222       2.76     0.0090
TREAT                  14        1.48922222       1.23     0.3012



                  General Linear Models Procedure

            Level of       -----------RESPONSE----------
            TREAT      N       Mean               SD

            1          4     2.77500000       0.17078251
            2          4     2.40000000       0.21602469
            3          4     2.45000000       0.23804761
            4          4     2.95000000       0.36968455
            5          4     2.80000000       0.14142136
            6          4     2.92500000       0.80156098
            7          4     2.82500000       0.17078251
            8          4     2.72500000       0.34034296
            9          4     2.82500000       0.51881275
            10         4     2.45000000       0.12909944
            11         4     2.97500000       0.32015621
            12         4     3.12500000       0.29860788
            13         4     2.52500000       0.37749172
            14         4     2.42500000       0.05000000
            15         4     2.87500000       0.55000000
                  General Linear Models Procedure
                        Least Squares Means

      TREAT      RESPONSE       Std Err      Pr > |T|   LSMEAN
                   LSMEAN        LSMEAN   H0:LSMEAN=0   Number
```

1	2.84555556	0.16342514	0.0001	1
2	2.41277778	0.16342514	0.0001	2
3	2.45166667	0.16342514	0.0001	3
4	2.68333333	0.16342514	0.0001	4
5	2.80666667	0.16342514	0.0001	5
6	2.90388889	0.16342514	0.0001	6
7	2.77111111	0.16342514	0.0001	7
8	2.81000000	0.16342514	0.0001	8
9	2.93333333	0.16342514	0.0001	9
10	2.51500000	0.16342514	0.0001	10
11	2.85388889	0.16342514	0.0001	11
12	3.01277778	0.16342514	0.0001	12
13	2.66833333	0.16342514	0.0001	13
14	2.53333333	0.16342514	0.0001	14
15	2.84833333	0.16342514	0.0001	15

Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

i/j	1	2	3	4	5	6	7	8
1	.	0.0711	0.0989	0.4887	0.8677	0.8093	0.7500	0.8789
2	0.0711	.	0.8677	0.2515	0.0989	0.0420	0.1450	0.0962
3	0.0989	0.8677	.	0.3248	0.1354	0.0599	0.1776	0.1450
4	0.4887	0.2515	0.3248	.	0.5981	0.3482	0.7072	0.5882
5	0.8677	0.0989	0.1354	0.5981	.	0.6774	0.8789	0.9886
6	0.8093	0.0420	0.0599	0.3482	0.6774	.	0.5705	0.6879
7	0.7500	0.1450	0.1776	0.7072	0.8789	0.5705	.	0.8677
8	0.8789	0.0962	0.1450	0.5882	0.9886	0.6879	0.8677	.
9	0.7072	0.0318	0.0458	0.3049	0.5882	0.8996	0.4887	0.5981
10	0.1634	0.6619	0.7863	0.4727	0.2328	0.1031	0.2772	0.2121
11	0.9725	0.0661	0.0923	0.4669	0.8397	0.8361	0.7231	0.8509
12	0.4756	0.0178	0.0214	0.1648	0.3802	0.6414	0.3211	0.3879
13	0.4498	0.2782	0.3729	0.9488	0.5545	0.3168	0.6602	0.5587
14	0.1873	0.6063	0.7267	0.5360	0.2468	0.1196	0.3124	0.2412
15	0.9905	0.0694	0.0967	0.4814	0.8631	0.8120	0.7410	0.8696

Pr > |T| H0: LSMEAN(i)=LSMEAN(j)

i/j	9	10	11	12	13	14	15
1	0.7072	0.1634	0.9725	0.4756	0.4498	0.1873	0.9905
2	0.0318	0.6619	0.0661	0.0178	0.2782	0.6063	0.0694
3	0.0458	0.7863	0.0923	0.0214	0.3729	0.7267	0.0967
4	0.3049	0.4727	0.4669	0.1648	0.9488	0.5360	0.4814
5	0.5882	0.2328	0.8397	0.3802	0.5545	0.2468	0.8631
6	0.8996	0.1031	0.8361	0.6414	0.3168	0.1196	0.8120
7	0.4887	0.2772	0.7231	0.3211	0.6602	0.3124	0.7410
8	0.5981	0.2121	0.8509	0.3879	0.5587	0.2412	0.8696
9	.	0.0805	0.7338	0.7338	0.2612	0.1052	0.7160
10	0.0805	.	0.1533	0.0395	0.5127	0.9374	0.1742
11	0.7338	0.1533	.	0.4977	0.4290	0.1761	0.9810
12	0.7338	0.0395	0.4977	.	0.1469	0.0468	0.4829
13	0.2612	0.5127	0.4290	0.1469	.	0.5641	0.4428
14	0.1052	0.9374	0.1761	0.0468	0.5641	.	0.1835
15	0.7160	0.1742	0.9810	0.4829	0.4428	0.1835	.

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

General Linear Models Procedure

Source	Type III Expected Mean Square
BLK	Var(Error) + 3.2143 Var(BLK)

```
TREAT       Var(Error) + Q(TREAT)



Contrast                Contrast Expected Mean Square

trt1 vs trt2            Var(Error) + Q(TREAT)


                   General Linear Models Procedure

Dependent Variable: RESPONSE

Contrast                 DF       Contrast SS    F Value     Pr > F

trt1 vs trt2              1        0.30101240       3.49     0.0711

                                T for H0:    Pr > |T|   Std Error of
Parameter         Estimate    Parameter=0                  Estimate

treat 1 mean    2.84555556          17.41     0.0001     0.16342514
trt 1 mean      2.84555556          17.41     0.0001     0.16342514
trt 1 blk 1     2.39666667          11.74     0.0001     0.20406165
trt 1 vs trt 2  0.43277778           1.87     0.0711     0.23153188
```

From Output 1.6, you can construct the analysis of variance table, as shown in Table 1.5.

Table 1.5 *Analysis of Variance Table for Incomplete Blocks Design from Cochran and Cox (1957)*

Source of Variation	df	SS	MS	F	*p*
Blocks	14	3.334	0.238		
Treatments (adjusted for blocks)	14	1.489	0.106	2.62	0.3012
Error	31	2.671	0.086		

Interpretation

The F-test for differences between (adjusted) treatment differences has a significance probability of *p*=0.3012, which presents no evidence of differences between treatments. In agreement with this conclusion, the table of significance probabilities for the lsmeans shows only six (out of 120) *p*-values less that 0.05. You would expect that many by chance.

Least squares means, obtained from the LSMEANS statement, are usually called **adjusted means** in standard textbooks. These means and their standard errors come from the OLS estimation of the treatment means. 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. This analysis of variance, along with adjusted treatment means and differences between them and their standard errors, comprise the so-called **intra-block** analysis of PBIBD data.

In the PROC GLM statements following the LSMEANS statement, you see three ESTIMATE statements and a CONTRAST statement. The first ESTIMATE statement (with the label “treat1 mean”) specifies coefficients for the INTERCEPT and the TREAT 1 parameter. By default, PROC GLM averages across the BLK parameters to form the linear combination of model terms. The second ESTIMATE statement (with the label “trt 1 mean”) explicitly specifies the same linear combination of model terms; that is, it specifies coefficients 1/15 for each of the BLK terms. The results of these two ESTIMATE statements are identical, including their standard errors. Moreover, they duplicate the LSMEAN for TREAT 1 and its standard error. These standard errors do not take into account the random block effects. Subsection 1.5.2 shows analogous ESTIMATE statements in PROC MIXED that produce different results. This is the purpose for showing these first two ESTIMATE statements. The fourth ESTIMATE statement (with label “trt 1 vs trt 2”) simply computes the difference between the lsmeans for TREAT 1 and TREAT 2. The CONTRAST statement with the label “trt 1 vs trt 2” computes an F-test that is equivalent to the *t*-test from the ESTIMATE statement with the same label.

The RANDOM statement in GLM, as already mentioned, causes only expected mean squares to be computed. The expected mean squares table in Output 1.7 shows that the correct denominator for the F-test for TREAT is MS(Error).

1.5.2 The Combined Intra- and Inter-block Analysis of PBIB Data Using PROC MIXED

Program

When blocks are treated as random, the result is then combined intra- and inter-block analysis. You can obtain this using PROC MIXED by submitting the statements:

```
proc mixed data=pbib;
   class blk treat;
   model response=treat;
   random block;
   lsmeans treat / pdiff;
   estimate 'treat 1 mean' intercept 1 treat 1;
   estimate 'trt 1 mean' intercept 15 treat 15 |
             blk 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 / divisor=15;
   estimate 'trt 1 blk 1' intercept 1 treat 1 | blk 1;
   estimate 'trt 1 vs trt 2' treat 1 -1;
   contrast 'trt 1 vs trt 2' treat 1 -1;
run;
```

Results appear in Output 1.7 for the combined intra- and inter-block analysis.

Results

Output 1.7 *Incomplete Block Design: PROC MIXED Analysis*

```
                       The MIXED Procedure

                     Class Level Information

          Class     Levels  Values

          BLK           15  1 2 3 4 5 6 7 8 9 10 11 12 13
                            14 15
          TREAT         15  1 2 3 4 5 6 7 8 9 10 11 12 13
                            14 15


                REML Estimation Iteration History

         Iteration  Evaluations     Objective     Criterion

                 0            1  -24.83873612
                 1            3  -30.71608590    0.00022046
                 2            1  -30.71956742    0.00000043
                 3            1  -30.71957397    0.00000000

                    Convergence criteria met.

               Covariance Parameter Estimates (REML)

    Cov Parm          Ratio      Estimate     Std Error       Z  Pr > |Z|

    BLK          0.54373914    0.04652189    0.02795193    1.66    0.0960
    Residual     1.00000000    0.08555921    0.02157637    3.97    0.0001

              Model Fitting Information for RESPONSE

             Description                          Value

             Observations                       60.0000
             Variance Estimate                   0.0856
             Standard Deviation Estimate         0.2925
             REML Log Likelihood               -25.9924
             Akaike's Information Criterion    -27.9924
             Schwarz's Bayesian Criterion      -29.7991
             -2 REML Log Likelihood             51.9849
```

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
TREAT	14	31	1.53	0.1576

ESTIMATE Statement Results

Parameter	Estimate	Std Error	DDF	T	Pr > |T|
treat 1 mean	2.81752245	0.16641271	31	16.93	0.0001
trt 1 mean	2.81752245	0.15681751	31	17.97	0.0001
trt 1 blk 1	2.52880832	0.18454599	31	13.70	0.0001
trt 1 vs trt 2	0.41220751	0.22206087	31	1.86	0.0729

CONTRAST Statement Results

Source	NDF	DDF	F	Pr > F
trt 1 vs trt 2	1	31	3.45	0.0729

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
TREAT 1	2.81752245	0.16641271	31	16.93	0.0001
TREAT 2	2.40531494	0.16641271	31	14.45	0.0001
TREAT 3	2.45494317	0.16641271	31	14.75	0.0001
TREAT 4	2.78383061	0.16641271	31	16.73	0.0001
TREAT 5	2.80489797	0.16641271	31	16.86	0.0001
TREAT 6	2.91069390	0.16641271	31	17.49	0.0001
TREAT 7	2.78898509	0.16641271	31	16.76	0.0001
TREAT 8	2.78160548	0.16641271	31	16.72	0.0001
TREAT 9	2.89131103	0.16641271	31	17.37	0.0001
TREAT 10	2.49106159	0.16641271	31	14.97	0.0001
TREAT 11	2.89869913	0.16641271	31	17.42	0.0001
TREAT 12	3.05282147	0.16641271	31	18.34	0.0001
TREAT 13	2.61776910	0.16641271	31	15.73	0.0001
TREAT 14	2.49131103	0.16641271	31	14.97	0.0001
TREAT 15	2.85923304	0.16641271	31	17.18	0.0001

Differences of Least Squares Means

Level 1	Level 2	Difference	Std Error	DDF	T
TREAT 1	TREAT 2	0.41220751	0.22206087	31	1.86
TREAT 1	TREAT 3	0.36257929	0.22206087	31	1.63
TREAT 1	TREAT 4	0.03369184	0.22206087	31	0.15
TREAT 1	TREAT 5	0.01262448	0.22206087	31	0.06
TREAT 1	TREAT 6	-0.09317145	0.22720031	31	-0.41
TREAT 1	TREAT 7	0.02853736	0.22206087	31	0.13
TREAT 1	TREAT 8	0.03591697	0.22206087	31	0.16
TREAT 1	TREAT 9	-0.07378858	0.22206087	31	-0.33
TREAT 1	TREAT 10	0.32646086	0.22206087	31	1.47
TREAT 1	TREAT 11	-0.08117667	0.22720031	31	-0.36
TREAT 1	TREAT 12	-0.23529902	0.22206087	31	-1.06
TREAT 1	TREAT 13	0.19975335	0.22206087	31	0.90
TREAT 1	TREAT 14	0.32621142	0.22206087	31	1.47
TREAT 1	TREAT 15	-0.04171059	0.22206087	31	-0.19
TREAT 2	TREAT 3	-0.04962822	0.22206087	31	-0.22
TREAT 2	TREAT 4	-0.37851567	0.22206087	31	-1.70

TREAT 2	TREAT 5	-0.39958303	0.22206087	31	-1.80
TREAT 2	TREAT 6	-0.50537896	0.22206087	31	-2.28
TREAT 2	TREAT 7	-0.38367015	0.22720031	31	-1.69
TREAT 2	TREAT 8	-0.37629054	0.22206087	31	-1.69
TREAT 2	TREAT 9	-0.48599609	0.22206087	31	-2.19
TREAT 2	TREAT 10	-0.08574665	0.22206087	31	-0.39
TREAT 2	TREAT 11	-0.49338418	0.22206087	31	-2.22
TREAT 2	TREAT 12	-0.64750653	0.22720031	31	-2.85
TREAT 2	TREAT 13	-0.21245416	0.22206087	31	-0.96
TREAT 2	TREAT 14	-0.08599609	0.22206087	31	-0.39
TREAT 2	TREAT 15	-0.45391810	0.22206087	31	-2.04
TREAT 3	TREAT 4	-0.32888744	0.22206087	31	-1.48
TREAT 3	TREAT 5	-0.34995480	0.22206087	31	-1.58
TREAT 3	TREAT 6	-0.45575074	0.22206087	31	-2.05
TREAT 3	TREAT 7	-0.33404192	0.22206087	31	-1.50
TREAT 3	TREAT 8	-0.32666232	0.22720031	31	-1.44
TREAT 3	TREAT 9	-0.43636786	0.22206087	31	-1.97
TREAT 3	TREAT 10	-0.03611842	0.22206087	31	-0.16
TREAT 3	TREAT 11	-0.44375596	0.22206087	31	-2.00
TREAT 3	TREAT 12	-0.59787830	0.22206087	31	-2.69
TREAT 3	TREAT 13	-0.16282594	0.22720031	31	-0.72
TREAT 3	TREAT 14	-0.03636786	0.22206087	31	-0.16
TREAT 3	TREAT 15	-0.40428988	0.22206087	31	-1.82
TREAT 4	TREAT 5	-0.02106736	0.22206087	31	-0.09
TREAT 4	TREAT 6	-0.12686330	0.22206087	31	-0.57
TREAT 4	TREAT 7	-0.00515448	0.22206087	31	-0.02
TREAT 4	TREAT 8	0.00222513	0.22206087	31	0.01
TREAT 4	TREAT 9	-0.10748042	0.22720031	31	-0.47
TREAT 4	TREAT 10	0.29276902	0.22206087	31	1.32
TREAT 4	TREAT 11	-0.11486852	0.22206087	31	-0.52
TREAT 4	TREAT 12	-0.26899086	0.22206087	31	-1.21
TREAT 4	TREAT 13	0.16606150	0.22206087	31	0.75
TREAT 4	TREAT 14	0.29251958	0.22720031	31	1.29
TREAT 4	TREAT 15	-0.07540244	0.22206087	31	-0.34
TREAT 5	TREAT 6	-0.10579594	0.22206087	31	-0.48
TREAT 5	TREAT 7	0.01591288	0.22206087	31	0.07
TREAT 5	TREAT 8	0.02329249	0.22206087	31	0.10
TREAT 5	TREAT 9	-0.08641306	0.22206087	31	-0.39
TREAT 5	TREAT 10	0.31383638	0.22720031	31	1.38
TREAT 5	TREAT 11	-0.09380116	0.22206087	31	-0.42
TREAT 5	TREAT 12	-0.24792350	0.22206087	31	-1.12
TREAT 5	TREAT 13	0.18712887	0.22206087	31	0.84
TREAT 5	TREAT 14	0.31358694	0.22206087	31	1.41
TREAT 5	TREAT 15	-0.05433507	0.22720031	31	-0.24
TREAT 6	TREAT 7	0.12170882	0.22206087	31	0.55
TREAT 6	TREAT 8	0.12908842	0.22206087	31	0.58
TREAT 6	TREAT 9	0.01938288	0.22206087	31	0.09
TREAT 6	TREAT 10	0.41963231	0.22206087	31	1.89
TREAT 6	TREAT 11	0.01199478	0.22720031	31	0.05
TREAT 6	TREAT 12	-0.14212756	0.22206087	31	-0.64
TREAT 6	TREAT 13	0.29292480	0.22206087	31	1.32
TREAT 6	TREAT 14	0.41938288	0.22206087	31	1.89
TREAT 6	TREAT 15	0.05146086	0.22206087	31	0.23
TREAT 7	TREAT 8	0.00737961	0.22206087	31	0.03
TREAT 7	TREAT 9	-0.10232594	0.22206087	31	-0.46
TREAT 7	TREAT 10	0.29792350	0.22206087	31	1.34
TREAT 7	TREAT 11	-0.10971404	0.22206087	31	-0.49
TREAT 7	TREAT 12	-0.26383638	0.22720031	31	-1.16
TREAT 7	TREAT 13	0.17121599	0.22206087	31	0.77
TREAT 7	TREAT 14	0.29767406	0.22206087	31	1.34
TREAT 7	TREAT 15	-0.07024795	0.22206087	31	-0.32
TREAT 8	TREAT 9	-0.10970555	0.22206087	31	-0.49
TREAT 8	TREAT 10	0.29054389	0.22206087	31	1.31
TREAT 8	TREAT 11	-0.11709364	0.22206087	31	-0.53
TREAT 8	TREAT 12	-0.27121599	0.22206087	31	-1.22

TREAT 8	TREAT 13	0.16383638	0.22720031	31	0.72
TREAT 8	TREAT 14	0.29029445	0.22206087	31	1.31
TREAT 8	TREAT 15	-0.07762756	0.22206087	31	-0.35
TREAT 9	TREAT 10	0.40024944	0.22206087	31	1.80
TREAT 9	TREAT 11	-0.00738810	0.22206087	31	-0.03
TREAT 9	TREAT 12	-0.16151044	0.22206087	31	-0.73

Differences of Least Squares Means

Level 1	Level 2	Difference	Std Error	DDF	T
TREAT 9	TREAT 13	0.27354192	0.22206087	31	1.23
TREAT 9	TREAT 14	0.40000000	0.22720031	31	1.76
TREAT 9	TREAT 15	0.03207798	0.22206087	31	0.14
TREAT 10	TREAT 11	-0.40763754	0.22206087	31	-1.84
TREAT 10	TREAT 12	-0.56175988	0.22206087	31	-2.53
TREAT 10	TREAT 13	-0.12670751	0.22206087	31	-0.57
TREAT 10	TREAT 14	-0.00024944	0.22206087	31	-0.00
TREAT 10	TREAT 15	-0.36817145	0.22720031	31	-1.62
TREAT 11	TREAT 12	-0.15412234	0.22206087	31	-0.69
TREAT 11	TREAT 13	0.28093002	0.22206087	31	1.27
TREAT 11	TREAT 14	0.40738810	0.22206087	31	1.83
TREAT 11	TREAT 15	0.03946608	0.22206087	31	0.18
TREAT 12	TREAT 13	0.43505236	0.22206087	31	1.96
TREAT 12	TREAT 14	0.56151044	0.22206087	31	2.53
TREAT 12	TREAT 15	0.19358842	0.22206087	31	0.87
TREAT 13	TREAT 14	0.12645808	0.22206087	31	0.57
TREAT 13	TREAT 15	-0.24146394	0.22206087	31	-1.09
TREAT 14	TREAT 15	-0.36792202	0.22206087	31	-1.66

Interpretation

You will note several differences from the intra-block analysis given by PROC GLM.

First of all, referring to the results of the first two ESTIMATE statements (with labels “treat 1 mean” and “trt 1 mean”), the estimates of the treatment means are different. Granted, the differences are not major, but they are certainly numerically different. In other applications the distinction can be dramatic. The PROC MIXED estimate of the treatment 1 mean is 2.817, compared with the PROC GLM estimate of 2.846. The distinction is that the PROC GLM estimate is OLS, whereas the MIXED estimate is (estimated) GLS. Theoretically, the GLS estimate is superior. PROC MIXED accounts for BLK being random and computes the BLUE estimates accordingly. Estimates of the variance components are used to compute **V** in equation (1.15) because the true variance components are unknown. The standard errors in PROC MIXED likewise are different from those in PROC GLM. The standard error of the OLS estimate is 0.163 from GLM. This is not a valid estimate of the true standard error of the OLS estimate for the same reason that PROC GLM did not compute a valid standard error estimate for a treatment mean for the RCBD data in subsection 1.1.1; the random effects of blocks were ignored. You see different standard errors for the “treat 1 mean” and “trt 1 mean” estimates from PROC MIXED. The ESTIMATE statement with label “treat 1 mean” did not specify coefficients for the block terms, whereas the ESTIMATE statement with label “trt 1 mean” did specify coefficients for blocks. This made no difference with PROC GLM, but it does with PROC MIXED.

The standard error from the ESTIMATE statement labeled "treat 1 mean" correctly estimates the standard error of the GLS estimate considering blocks to be random. Thus it can be used to produce a confidence interval for the mean that would be valid for inference across the population of blocks from which those in the experiment were randomly drawn. The standard error from the ESTIMATE statement labelled "trt 1 mean," however, does not involve the block variance component. Thus a confidence interval based on this standard error is valid only for the blocks in the experiment. Standard errors of the LSMEANS are the same as for the "treat 1 mean" estimate. The "trt 1 mean" is an example of a *best linear unbiased predictor* (BLUP), and linear combination of fixed and random effects. BLUPs are unique to mixed model theory and are discussed in Chapter 6.

1.6 Summary

Chapter 1 begins with an example of a randomized blocks design with fixed treatments and random blocks. The importance of accounting for random effects in such a basic situation as computing a variance for a treatment mean is demonstrated. The use of PROC MIXED is introduced with explanations of how to set up the MODEL and RANDOM statements. The chapter continues with illustrations of CONTRAST, ESTIMATE, and LSMEANS statements. Then, PROC GLM is applied to the same example to illustrate similarities and differences of PROC GLM and PROC MIXED and to emphasize what basic applications are handled correctly by PROC MIXED and not by PROC GLM. A brief explanation of mixed model theory is presented in relation to the randomized blocks design, including explicit descriptions of the matrices in the general linear mixed model. Then, an incomplete block design is used to illustrate some of the issues confronted with unbalanced mixed model data.

1.7 References

Cochran, W.G. and Cox, G.M. (1957), *Experimental Designs, Second Edition*, New York: John Wiley & Sons.

Littell, R.C., Freund R.J., and Spector, P.C. (1991), *SAS System for Linear Models, Third Edition*, Cary, NC: SAS Institute Inc.

Mead, R. (1988), *The Design of Experiments*, New York: Cambridge University Press, 620 pp.

Mendenhall, W., Wackerly, D.D., and Scheaffer, R.L. (1990), *Mathematical Statistics with Applications*, 4th Edition, Belmont, CA: Duxbury Press.

Milliken, G.A. and Johnson, D.E. (1994), *Analysis of Messy Data, Volume 1: Designed Experiments*, London: Chapman Press.

Winer, B.J. (1971), *Statistical Principles in Experimental Design, Second Edition*, New York: McGraw-Hill, Inc.