

5

Designing experiments—keeping it simple

5.1 Three fundamental principles of experimental design

The concept of blocking in experimental design is introduced. To analyse a blocked experiment, the categorical variable BLOCK is included in the model formula. In this example, we compare two analyses: one in which BLOCK has been statistically eliminated, and one in which it has not. The second of these examples illustrates how to include two categorical variables in a model. All the elements of the command and menu language required have been introduced in previous chapters.

SAS COMMANDS FOR BOX 5.1

```
Commands  proc glm data=gandh.Chapter5;
           class BLOCK BEAN;
           model YIELD = BEAN;
           run;
```

```
Menu route  Statistics > Anova > Linear Models...
           YIELD → Dependent
           BEAN → Class
```

SAS OUTPUT FOR BOX 5.1 Analysis of bean yields assuming a fully randomised design						
The GLM Procedure						
Dependent Variable: YIELD						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	5	444.4350000	88.8870000	14.59	<.0001	
Error	18	109.6900000	6.0938889			
Corrected Total	23	554.1250000				
	R-Square	Coeff Var	Root MSE	YIELD Mean		
	0.802048	14.80408	2.468580	16.67500		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
BEAN	5	444.4350000	88.8870000	14.59	<.0001	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
BEAN	5	444.4350000	88.8870000	14.59	<.0001	

SAS COMMANDS FOR BOX 5.2 Analysis of a blocked experiment	
Commands	<pre>proc glm data=gandh.Chapter5; class BLOCK BEAN; model YIELD = BLOCK BEAN; run;</pre>
Menu route	Statistics > Anova > Linear Models... YIELD → Dependent BLOCK BEAN → Class <div style="border: 1px solid black; display: inline-block; padding: 2px;">Statistics</div> <input checked="" type="checkbox"/> Type I

SAS OUTPUT FOR BOX 5.2 **Analysis of bean yields assuming a randomised block design**

The GLM Procedure

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	497.3300000	62.1662500	16.42	<.0001
Error	15	56.7950000	3.7863333		
Corrected Total	23	554.1250000			

R-Square	Coeff Var	Root MSE	YIELD Mean
0.897505	11.66927	1.945850	16.67500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	3	52.8950000	17.6316667	4.66	0.0171
BEAN	5	444.4350000	88.8870000	23.48	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	3	52.8950000	17.6316667	4.66	0.0171
BEAN	5	444.4350000	88.8870000	23.48	<.0001

A Latin Square design is an example of using two blocking factors in an experimental design. Both categorical blocking variables are added to the model formula. The following analysis uses the *oilseed rape* dataset.

SAS COMMANDS FOR BOX 5.3 **Analysing a latin square design**

```
Commands  proc glm data=gandh.Chapter5;
           class COLUMN ROW TREATMT;
           model SEEDS = COLUMN ROW TREATMT;
           run;
```

Menu route Statistics > Anova > Linear Models...

SEEDS → Dependent

COLUMN ROW TREATMT → Class

Statistics

Type I

SAS OUTPUT FOR BOX 5.3 **Analysis of a latin square design**

The GLM Procedure

Dependent Variable: SEEDS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	5072.750000	563.638889	11.16	0.0041
Error	6	303.000000	50.500000		
Corrected Total	15	5375.750000			

R-Square	Coeff Var	Root MSE	SEEDS Mean
0.943636	6.527059	7.106335	108.8750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
COLUMN	3	1332.250000	444.083333	8.79	0.0129
ROW	3	1090.250000	363.416667	7.20	0.0206
TREATMT	3	2650.250000	883.416667	17.49	0.0023

Source	DF	Type I SS	Mean Square	F Value	Pr > F
COLUMN	3	1332.250000	444.083333	8.79	0.0129
ROW	3	1090.250000	363.416667	7.20	0.0206
TREATMT	3	2650.250000	883.416667	17.49	0.0023

5.2 The geometrical analogy for blocking

Calculating the fitted model for two categorical variables

Revisiting the *Beans* dataset, the full coefficient table can be obtained by altering the level of output requested (see SAS supplement chapter 3 and below).

SAS COMMANDS FOR BOX 5.4 **Coefficients table for two categorical variables**

```
Commands  proc glm data=gandh.Chapter5;
           class BLOCK BEAN;
           model YIELD = BLOCK BEAN / solution;
           run;
```

Menu route Statistics > Anova > Linear Models...

YIELD → Dependent

BLOCK BEAN → Class

Statistics

Type I

Parameter Estimates

This would give the following ANOVA and coefficients table:

SAS OUTPUT FOR BOX 5.4 ANOVA and Coefficient tables for a randomised blocked design

The GLM Procedure

Dependent Variable: YIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	497.3300000	62.1662500	16.42	<.0001
Error	15	56.7950000	3.7863333		
Corrected Total	23	554.1250000			

	R-Square	Coeff Var	Root MSE	YIELD Mean
	0.897505	11.66927	1.945850	16.67500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	3	52.8950000	17.6316667	4.66	0.0171
BEAN	5	444.4350000	88.8870000	23.48	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	3	52.8950000	17.6316667	4.66	0.0171
BEAN	5	444.4350000	88.8870000	23.48	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	9.49166667 B	1.19158508	7.97	<.0001
BLOCK 1	1.00000000 B	1.12343719	0.89	0.3875
BLOCK 2	3.35000000 B	1.12343719	2.98	0.0093
BLOCK 3	-0.51666667 B	1.12343719	-0.46	0.6522
BLOCK 4	0.00000000 B	.	.	.
BEAN 1	11.30000000 B	1.37592393	8.21	<.0001
BEAN 2	11.92500000 B	1.37592393	8.67	<.0001
BEAN 3	5.62500000 B	1.37592393	4.09	0.0010
BEAN 4	5.97500000 B	1.37592393	4.34	0.0006
BEAN 5	2.52500000 B	1.37592393	1.84	0.0864
BEAN 6	0.00000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

There are four blocks and six varieties of bean, and a coefficient is given for each level, but that for the final level of each variable equals zero exactly and is assigned no standard error. This is because the final level of each categorical variable is aliased (as first discussed in chapter 3). The aliasing convention followed in the main text is the not the same as that followed by SAS. To calculate the full set of fitted values, we therefore need a small amendment to the method of the text. The amendment is the way in which the aliased value is calculated, and used in subsequent fitted value calculations.

The coefficients for each level are given with reference to the last aliased value. In other words, in this example, block two has a mean 3.35 units higher than block four, and variety three has a mean 5.625 units higher than variety six. This allows a rapid visual inspection of the means of all other blocks with reference to the last block, and all other varieties of bean with reference to variety six. With only one categorical variable, the reference level represented by the final level of a variable is simply the constant, as we saw earlier. But with two categorical variables, the reference level for each variable is a bit more complicated. We need to go into this a bit further in order to show how the coefficients are linked to the fitted values.

Table 5.1 illustrates the derivation of the fitted values for this particular example. The fourth fitted value for BLOCK involves the mean of the deviations of the variety coefficients from variety six; and the sixth fitted value for BEAN involves the mean of the deviations of the block coefficients from block four. This seems rather counterintuitive. However, a little algebra illustrates how it works.

Let B_1 = the mean of block 1; B_2 for block 2; V_1 = the mean for variety 1 etc. Then the grand mean may be expressed as:

$$\frac{V_1 + V_2 + V_3 + V_4 + V_5 + V_6}{6}$$

or alternatively as

$$\frac{B_1 + B_2 + B_3 + B_4}{4}.$$

However, we choose to make B_4 and V_6 as reference points. Every other block mean can be expressed as a function of B_4 and parameters e.g. $B_2 = B_4 + \alpha_2$, and similarly $V_3 = V_6 + \beta_3$. If we substitute these functions into our two expressions for the grand mean, we find

$$B_4 + \frac{\alpha_1 + \alpha_2 + \alpha_3}{4} = V_6 + \frac{\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5}{6}.$$

Re-arranging for B_4 , we get

$$B_4 = \left\{ V_6 - \frac{\alpha_1 + \alpha_2 + \alpha_3}{4} \right\} + \frac{\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5}{6}$$

or re-arranging for V_6 , we get

$$V_6 = \left\{ B_4 - \frac{\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5}{6} \right\} + \frac{\alpha_1 + \alpha_2 + \alpha_3}{4}.$$

However, the expressions inside the curly brackets are equal—so these are set to μ . This gives us the same expression for the fitted values of B_4 and V_6 as in SAS Table 5.1. The intercept, μ , can be thought of as the mean of variety six adjusted for differences between the blocks, or the mean for block four, adjusted for differences between the varieties. Setting the coefficient for the final level of a variable to zero is therefore not always as simple as it seems!

To summarise Table 5.1 as a fitted value equation equivalent to the version given on page 87 of the main text, we introduce the symbols α_{ave} and β_{ave} for the average values of α and β . This gives us the following:

$$y = \mu + \alpha_{ave} + \beta_{ave} + \begin{pmatrix} \text{Block} \\ 1 & \alpha_1 \\ 2 & \alpha_2 \\ 3 & \alpha_3 \\ 4 & 0 \end{pmatrix} + \begin{pmatrix} \text{Variety} \\ 1 & \beta_1 \\ 2 & \beta_2 \\ 3 & \beta_3 \\ 4 & \beta_4 \\ 5 & \beta_5 \\ 6 & 0 \end{pmatrix} + \varepsilon$$

The following table shows how this formula works numerically for our example.

SAS VERSION OF TABLE 5.1 **Calculating fitted values from coefficients**

Term	Level	Coefficient from SAS output		Fitted Value
Intercept		μ	9.492	
Block	1	α_1	1.000	$\mu + \beta_{ave} + \alpha_1 =$ 9.492 + 6.225 + 1
	2	α_2	3.350	$\mu + \beta_{ave} + \alpha_2 =$ 9.492 + 6.225 + 3.35
	3	α_3	-0.517	$\mu + \beta_{ave} + \alpha_3 =$ 9.492 + 6.225 - 0.517
	4	α_4 (aliased)	0	$\mu + \beta_{ave} + \alpha_4 =$ 9.492 + 6.225 + 0
		α_{ave}	$= \frac{1 + 3.35 - 0.517 + 0}{4} = 0.958$	
Bean	1	β_1	11.300	$\mu + \alpha_{ave} + \beta_1 =$ 9.492 + 0.958 + 11.3
	2	β_2	11.925	$\mu + \alpha_{ave} + \beta_2 =$ 9.492 + 0.958 + 11.925
	3	β_3	5.625	$\mu + \alpha_{ave} + \beta_3 =$ 9.492 + 0.958 + 5.625
	4	β_4	5.975	$\mu + \alpha_{ave} + \beta_4 =$ 9.492 + 0.958 + 5.975
	5	β_5	2.525	$\mu + \alpha_{ave} + \beta_5 =$ 9.492 + 0.958 + 2.525
	6	β_6 (aliased)	0	$\mu + \alpha_{ave} + \beta_6 =$ 9.492 + 0.958 + 0
		β_{ave}	$= \frac{11.3 + 11.925 + 5.625 + 5.975 + 2.525 + 0}{6}$ $= 6.225$	

5.3 The concept of orthogonality

Loss of orthogonality leads to differences in adjusted and sequential SS for a variable. This may be illustrated by the *Beans* data set in which the variables MYIELD, MBLOCK and MBEAN are shortened versions of the original variables, due to loss of orthogonality.

SAS COMMANDS FOR BOX 5.5 **Loss of orthogonality for two categorical variables**

```
Commands  proc glm data=gandh.Chapter5;
           class MBLOCK MBEAN;
           model MYIELD = MBLOCK MBEAN / solution;
           run;
```

Menu route Statistics > Anova > Linear Models...

MYIELD → Dependent

MBLOCK MBEAN → Class

Statistics

Type I

Parameter Estimates

SAS OUTPUT FOR BOX 5.5 **Slight loss of orthogonality in a randomised block design**

The GLM Procedure

Dependent Variable: MYIELD

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	498.7038194	62.3379774	24.34	<.0001
Error	13	33.2961806	2.5612447		
Corrected Total	21	532.0000000			

	R-Square	Coeff Var	Root MSE	MYIELD Mean
	0.937413	9.526124	1.600389	16.80000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
MBLOCK	3	49.2908333	16.4302778	6.41	0.0067
MBEAN	5	449.4129861	89.8825972	35.09	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MBLOCK	3	49.6896528	16.5632176	6.47	0.0065
MBEAN	5	449.4129861	89.8825972	35.09	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	8.65833333 B	1.30671207	6.63	<.0001
MBLOCK 1	0.89722222 B	1.06692594	0.84	0.4156
MBLOCK 2	3.24722222 B	1.06692594	3.04	0.0094
MBLOCK 3	-0.61944444 B	1.06692594	-0.58	0.5714
MBLOCK 4	0.00000000 B	.	.	.
MBEAN 1	12.21041667 B	1.24395283	9.82	<.0001
MBEAN 2	13.90000000 B	1.30671207	10.64	<.0001
MBEAN 3	6.53541667 B	1.24395283	5.25	0.0002
MBEAN 4	6.88541667 B	1.24395283	5.54	<.0001
MBEAN 5	3.43541667 B	1.24395283	2.76	0.0162
MBEAN 6	0.00000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

5.5 Exercises

These exercises do not involve any new commands.

Growing carnations

SAS COMMANDS FOR BOX 5.6 Analysis of the number of tulip blooms with bed, water and shade

```

Commands  proc glm data=gandh.Chapter5;
           class BED WATER SHADE;
           model SQBLOOMS = BED WATER SHADE / solution;
           run;

```

Menu route Statistics > Anova > Linear Models...

SQBLOOMS → Dependent

BED WATER SHADE → Class

Statistics

Type I

Parameter Estimates

SAS OUTPUT FOR BOX 5.6 Analysis of the number of tulip blooms with bed, water and shade

The GLM Procedure

Dependent Variable: SQBLOOMS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	9.49409153	1.35629879	6.21	0.0002
Error	28	6.11733944	0.21847641		
Corrected Total	35	15.61143097			

R-Square	Coeff Var	Root MSE	SQBLOOMS Mean
0.608150	11.60118	0.467415	4.029028

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BED	2	4.13228106	2.06614053	9.46	0.0007
WATER	2	3.71534439	1.85767219	8.50	0.0013
SHADE	3	1.64646608	0.54882203	2.51	0.0789

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BED	2	4.13228106	2.06614053	9.46	0.0007
WATER	2	3.71534439	1.85767219	8.50	0.0013
SHADE	3	1.64646608	0.54882203	2.51	0.0789

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.353666667 B	0.22034136	15.22	<.0001
BED 1	0.504416667 B	0.19082121	2.64	0.0133
BED 2	0.822916667 B	0.19082121	4.31	0.0002
BED 3	0.000000000 B	.	.	.
WATER 1	-0.448916667 B	0.19082121	-2.35	0.0259
WATER 2	0.335250000 B	0.19082121	1.76	0.0899
WATER 3	0.000000000 B	.	.	.
SHADE 1	0.367333333 B	0.22034136	1.67	0.1066
SHADE 2	0.564222222 B	0.22034136	2.56	0.0161
SHADE 3	0.151666667 B	0.22034136	0.69	0.4969
SHADE 4	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

58 Designing experiments—keeping it simple

The second analysis does not include BED as a block.

SAS COMMANDS FOR BOX 5.7 The carnation bloom analysis without bed used as a block

```

Commands  proc glm data=gandh.Chapter5;
           class WATER SHADE;
           model SQBLOOMS = WATER SHADE / solution;
           run;
    
```

Menu route Statistics > Anova > Linear Models...

SQBLOOMS → Dependent

WATER SHADE → Class

Statistics

- Type I
- Parameter Estimates

SAS OUTPUT FOR BOX 5.7 The carnation bloom analysis without bed used as a block

The GLM Procedure

Dependent Variable: SQBLOOMS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5.36181047	1.07236209	3.14	0.0214
Error	30	10.24962050	0.34165402		
Corrected Total	35	15.61143097			

R-Square	Coeff Var	Root MSE	SQBLOOMS Mean
0.343454	14.50751	0.584512	4.029028

Source	DF	Type I SS	Mean Square	F Value	Pr > F
WATER	2	3.71534439	1.85767219	5.44	0.0097
SHADE	3	1.64646608	0.54882203	1.61	0.2086

Source	DF	Type III SS	Mean Square	F Value	Pr > F
WATER	2	3.71534439	1.85767219	5.44	0.0097
SHADE	3	1.64646608	0.54882203	1.61	0.2086

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.796111111 B	0.23862593	15.91	<.0001
WATER 1	-0.448916667 B	0.23862593	-1.88	0.0697
WATER 2	0.335250000 B	0.23862593	1.40	0.1703
WATER 3	0.000000000 B	.	.	.
SHADE 1	0.367333333 B	0.27554149	1.33	0.1925
SHADE 2	0.564222222 B	0.27554149	2.05	0.0494
SHADE 3	0.151666667 B	0.27554149	0.55	0.5861
SHADE 4	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

In the third analysis, three plots have been removed.

SAS COMMANDS FOR BOX 5.8 Analysis of the carnation blooms with three plot values removed

```
Commands  proc glm data=gandh.Chapter5;
           class B2 W2 S2;
           model SQ2 = B2 W2 S2 / SS3;
           run;
```

Menu route Statistics > Anova > Linear Models...
 SQ2 → Dependent
 B2 W2 S2 → Class

SAS OUTPUT FOR BOX 5.8 Analysis of the carnation blooms with three plot values removed

The GLM Procedure

Dependent Variable: SQ2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	8.64908650	1.23558379	7.50	<.0001
Error	25	4.12129768	0.16485191		
Corrected Total	32	12.77038418			

	R-Square	Coeff Var	Root MSE	SQ2 Mean
	0.677277	9.916329	0.406020	4.094455

Source	DF	Type I SS	Mean Square	F Value	Pr > F
B2	2	2.76258115	1.38129057	8.38	0.0016
W2	2	5.07932135	2.53966068	15.41	<.0001
S2	3	0.80718400	0.26906133	1.63	0.2072

Source	DF	Type III SS	Mean Square	F Value	Pr > F
B2	2	2.64902965	1.32451482	8.03	0.0020
W2	2	4.67637706	2.33818853	14.18	<.0001
S2	3	0.80718400	0.26906133	1.63	0.2072

The dorsal crest of the male smooth newt

See SAS output for this exercise in the answers for exercises.