

Topic 11 Unbalanced designs (ST&D Section 9.6, p.219 & Chapter 18)

11. 1. The problem of missing data

Accidents often result in loss of data. Crops are destroyed on some plots, plants and animals die, etc. In the standard methods for handling missing data, it is assumed that missing items are due to mistakes and not to a failure of a treatment. To put it another way, any missing observation Y_{ij} is assumed to follow the same mathematical model as the observations that are present.

In a one-way design, the imbalance resulting from a missing data is not a problem. The only effect is the reduction of the sample sizes in the affected classes. However, missing values pose a problem for two-way classifications. Missing items destroy the symmetry and simplicity of the analysis, which becomes more complex if several Y_{ij} are missing.

11. 2. RCBD Example with one missing data

As an example, Table 1 (Snedecor & Cochran 1980 p 275) shows the yield in an experiment of four breeding lines of wheat in which we have supposed that the yield Y_{41} for line D in block 1 is missing.

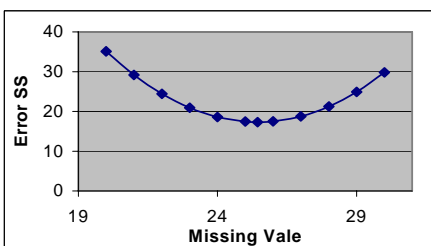
Line	Block					Total
	1	2	3	4	5	
A	32.3	34.0	34.3	35.0	36.5	172.1
B	33.3	33.0	36.3	36.8	34.5	173.9
C	30.8	34.3	35.3	32.3	35.8	168.5
D	...	26.0	29.8	28.0	28.8	112.6
Total	96.4	127.3	135.7	132.1	135.6	627.1

With a single missing value in a two-way classification, we obtain most of what we want by inserting the least-squares estimate of the missing value in the vacant cell and analyzing the complete data. This method gives least-squares estimates of every treatment mean and correct residual sum of squares.

If the missing value is in row i column j , and “ I ” is the number of treatments and “ J ” the number of blocks, the value to be inserted is calculated by the following formula:

$$\text{Estimated } Y_{ij} = (IY_{i.} + JY_{.j} - Y_{..}) / [(I-1)(J-1)]$$

The value to be inserted is: $[4 * 112.6 + 5 * 96.4 - 627.1] / 3 * 4 = 25.44$



If different ANOVAs are calculated using different values to replace the missing value and the SS_{error} is plotted against these values, a minimum are found at **25.44**.

This value is entered in the table as the missing plot. ANOVA is computed as usual.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	206.74650	29.53521	20.45	0.0001
Error	12	17.33100	1.44425		
Corrected Total	19	224.07750			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMNT	3	171.36150	57.12050	39.55	0.0001
BLOCK	4	35.38500	8.84625	6.13	0.0063

Two additional corrections are required:

- **The degrees of freedom in the total and error sums of squares are both reduced by 1.** In our example there are only 18 df for the total and 11 df for the error sums of squares,
- **Row SS and Column SS are both adjusted by a special correction before their MS is computed.**

Correction to be subtracted from the Treatment SS:

$$\text{Correction SS treatment} = [Y_j - (I-1) \cdot \text{estimated } Y_{ij}]^2 / I \cdot (I-1)$$

$$\text{Correction SS blocks} = [Y_i - (J-1) \cdot \text{estimated } Y_{ij}]^2 / J \cdot (J-1)$$

In the example from Table 1:

$$\text{Correction SS}_{\text{trtmnt}} = [96.4 - 3 \cdot 25.4]^2 / 4 \cdot 3 = 34.0033 \Rightarrow \text{SS}_{\text{trtmnt}} = 171.361 - 34.003 = 137.36$$

$$\text{Correction SS}_{\text{block}} = [112.6 - 4 \cdot 25.4]^2 / 5 \cdot 4 = 6.05 \Rightarrow \text{SS}_{\text{block}} = 35.38 - 6.05 = 29.33$$

The corrected ANOVA is:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	206.74650	29.53521	20.45	0.0001
Error	<u>11</u>	17.33100	<u>1.57554</u>		
Corrected Total	<u>18</u>	224.07750			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMNT	3	<u>137.36</u>	45.78	<u>29.06</u>	<u>0.0001</u>
BLOCK	4	<u>29.33</u>	7.33	<u>4.66</u>	<u>0.0192</u>

11. 2. 1 Same RCBD Example using SAS

Missing data are indicated in SAS by a “.”. The SAS program for the previous example is:

```

Data lec_11SC;
  do trtmnt= 1 to 4;
    do block= 1 to 5;
      input yield @@;
      output;
    end;
  end;
cards;
32.3      34.0      34.3      35.0      36.5
33.3      33.0      36.3      36.8      34.5
30.8      34.3      35.3      32.3      35.8
.         26.0      29.8      28.0      28.8
;
proc glm;
class trtmnt block;
model yield=trtmnt block;
run; quit;

```

The output for this program is:

```

Class      Levels      Values
TRTMNT      4      1 2 3 4
BLOCK      5      1 2 3 4 5
Number of observations in data set = 20

```

NOTE: Due to missing values, only 19 observations can be used in this analysis.

Dependent Variable: YIELD

Source	DF	Squares	Sum of Square	Mean Square	F Value	Pr > F
Model	7	151.79952	21.68565	13.76		0.0001
Error	11	17.32996	1.57545			
Corrected Total	18	169.12947				

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TRTMNT	3	122.46347	40.82116	25.91	0.0001
BLOCK	4	29.33604	7.33401	4.66	0.0192

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRTMNT	3	137.35921	45.78640	29.06	0.0001
BLOCK	4	29.33604	7.33401	4.66	0.0192

Note that the Type III SS produce exactly the same result as the one we obtained by replacing the missing value with its least-squares estimate based on block and column totals.

11. 2. 2 Effect of the order of the factors in the model statement: differences between Type I and Type III sum of squares.

In the previous SAS program if we replace

```
model yield=trtmnt block;
```

by

```
model yield= block trtmnt;
```

We obtain the following output:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BLOCK	4	14.44031	3.61008	2.29	0.1248
TRTMNT	3	137.35921	45.78640	29.06	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BLOCK	4	29.33604	7.33401	4.66	0.0192
TRTMNT	3	137.35921	45.78640	29.06	0.0001

Note that the Type III SS produce exactly the same result as before, but the TYPE I SS is different. In the first case, where the block effect is the last factor in the model, the TYPE I SS_{blocks} is equal to the TYPE III SS_{blocks} . In the second example, where the treatment effect is the last factor in the model, the TYPE I SS_{trtmnt} is equal to the TYPE III SS_{trtmnt} .

The difference between TYPE I and TYPE III sum of squares is that TYPE I lists the SS for each variable as if it were entered one at a time into the model, in the order they are specified in the model statement. Hence they can be thought of as incremental SS. This may or may not be desirable. The TYPE III SS gives the sum of squares that would be obtained for each variable if it were entered last into the model. That is, the effect of each variable is evaluated after all other factors have been accounted for.

11.3. Effects of unbalanced data on the estimation of differences between means

The computational formulas for PROC GLM that use the various treatment means provide correct statistics for *balanced* or *orthogonal* data, e.g., data with an equal number of observations ($n_{ij} = n$ for all u, j) for each treatment combination. When data are not balanced, sums of squares computed from these means can contain functions of the other parameters of the model.

To illustrate the effects of unbalanced data on the estimation of differences between means and computation of sums of squares, consider the data in this two-way table:

Data		B		Mean
		1	2	
A	1	7, 9	5	7
	2	8	4, 6	6
		8	5	

Means		B		Mean
		1	2	
A	1	8	5	6.5
	2	8	5	6.5
		8	5	

Within level 1 of B, the cell mean for each level of A is 8, hence there is no evidence of a difference between the levels of A within level 1 of B. Likewise, there is no evidence of a difference between levels of A within level 2 of B because both means are 5. Hence we may conclude that there is no evidence in the table of a difference between the levels of A. However, the marginal means for A are: 7 and 6 and the difference of $7 - 6 = 1$ between these marginal

means may be erroneously interpreted as measuring an overall effect of the factor A. The problem is that because the design is unbalanced and the effect of B gets mixed up in the calculation of the effect of A

The observed difference between the marginal means for the two levels of A measures the effect of factor B in addition to the effect of factor A.

This can be verified by expressing the observations in terms of the analysis-of-variance model $y_{ij} = \mu + \alpha_i + \beta_j$. (For simplicity, the interaction and error terms have been left out of the model.)

		B	
		1	2
A	1	7 = $\mu + \alpha_1 + \beta_1$ 9 = $\mu + \alpha_1 + \beta_1$	5 = $\mu + \alpha_1 + \beta_2$
	2	6 = $\mu + \alpha_2 + \beta_1$	4 = $\mu + \alpha_2 + \beta_2$ 6 = $\mu + \alpha_2 + \beta_2$

The difference between marginal means for A_1 and A_2 is shown with a little algebra to be:

$$\begin{aligned} \text{Means } (A_1 - A_2) &= 1/3 [(\alpha_1 + \beta_1) + (\alpha_1 + \beta_1) + (\alpha_1 + \beta_2)] - 1/3[(\alpha_2 + \beta_1) + (\alpha_2 + \beta_2) + (\alpha_2 + \beta_2)] \\ &= (\alpha_1 - \alpha_2) + 1/3 (\beta_1 - \beta_2) \end{aligned}$$

Thus, instead of estimating the difference between means A_1 and A_2 , the difference between the marginal means of A estimates $(\alpha_1 - \alpha_2)$ plus a function of the factor B parameters: $1/3 (\beta_1 - \beta_2)$. In other words:

The difference between the A marginal means is biased by factor B effects.

The null hypothesis about A we would normally wish to test is:

$$H_0: \alpha_1 - \alpha_2 = 0.$$

However, the sum of squares for A computed by Type I SS in PROC GLM actually tests the hypothesis:

$$H_0: \alpha_1 - \alpha_2 + 1/3 (\beta_1 - \beta_2) = 0,$$

which involves the factor B difference in addition to the factor A difference. The β 's get mixed up there, too. In summary, the problem with unbalanced designs in multifactor analyses is that the **factors get mixed up with each other in the calculations.**

11. 3. 1. Effects of unbalanced data on the estimation of the marginal means

In terms of the μ model $y_{ij} = \mu_{ij} + \epsilon_{ijk}$, we usually want to estimate

$$(\mu_{11} + \mu_{12})/2 \quad \text{and} \quad (\mu_{21} + \mu_{22})/2.$$

However, the A marginal means estimate $(2\mu_{11} + \mu_{22})/3$ and $(\mu_{21} + 2\mu_{22})/3$, respectively. These estimates are functions of the usually irrelevant cell frequencies and may be useless.

For example the expected marginal mean for A=1 is:

$$[(\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_2)]/3 = [3\mu + 3\alpha_1 + 2\beta_1 + \beta_2]/3 = \mu + \alpha_1 + 2/3\beta_1 + 1/3\beta_2$$

The LSMEANS statement produces the least-squares estimates of class variable means; these are sometimes referred to as adjusted means. Least-squares means should not, in general, be confused with ordinary means, which are available with a MEANS statement. The MEANS statement produces simple, unadjusted means for all observations in each class or treatment. Except for one-way designs, and some nested and balanced factorial structures, these unadjusted means are generally not equal to the least-squares means.

In the previous example from Table 1, means and least-squares means can be obtained in SAS by:

```
proc GLM;
model yield=trtmnt block;
  means trtmnt block;
  lsmeans trtmnt block / pdiff;
```

The PDIFF option after the slash prints all possible probability values for the hypothesis $H_0: LSM_i = LSM_j$. These tests are analogous to the LSD in the balanced case. The experiment-wise error rate is not controlled. To compare LSMEANS using other multiple comparison techniques use / pdiff adjust= tukey. Tukey option can be replaced for other tests described on topic 5.

Table 2. Comparison of means and LS means using data from Table 1 with the missing data and with the missing data replaced by its mean squares estimate.

	Missing value as “25.44166”		Missing value as “.”	
	Means	LS Means	Means	LS Means
Treatment A	34.4200	34.4200	34.4200	34.4200
Treatment B	34.7800	34.7800	34.7800	34.7800
Treatment C	33.7000	33.7000	33.7000	33.7000
Treatment D	27.6083	27.6083	28.1500	27.6083
Block 1	30.4604	30.4604	32.1333	30.4604
Block 2	31.8250	31.8250	31.8250	31.8250
Block 3	33.9250	33.9250	33.9250	33.9250
Block 4	33.0250	33.0250	33.0250	33.0250
Block 5	33.9000	33.9000	33.9000	33.9000

When the missing value is replaced by the least-squares estimate (25.44166) the design is “balanced” again and the means and LS means are identical. When the missing value is not replaced (“.” is used) unadjusted means are not equal to the least-squares means for treatment D and block 1, where the missing data is located. The means of unbalanced data are a function of

sample sizes; the LS means are not. The LS means produce values that are identical to those obtained by replacing the missing data by its least-squares estimate.

In summary, a major problem in the analysis of unbalanced data is the contamination of means and differences between factor means by effects of other factors. The solution to this problem is to adjust the means to remove the contaminating effects using LSMEANS and the use of Type III SS.

CONTRASTS: In an unbalanced crossed design SAS uses automatically the TYPE III Sum of Squares (SAS SYSTEM for Linear Models page 164-167).

```
proc GLM;
model yield=trtmnt block;
lsmeans trtmnt block / pdiff;
contrast 'A vs BC' trtmnt -2 1 1 0;
```

11. 4. Sums of Squares Computed by PROC GLM

PROC GLM recognizes different theoretical approaches to analysis of variance by providing four types of sums of squares and associated statistics. The four types of sums of squares in PROC GLM are called Type I, Type II, Type III, and Type IV (Goodnight 1978).

Though we are going to use only Type I and Type III SS during this course a description of all four types is included.

11. 4. 1. Type I

Type I sums of squares correspond to adding each source (factor) sequentially to the model in the order listed. For example, the Type I sum of squares for the first factor listed is the same as PROC ANOVA would compute for that effect. It reflects differences between unadjusted means of that factor as if the data consist of a one-way structure. The Type I SS may not be particularly useful for analysis of unbalanced multi way structures but may be useful for nested models, polynomial models, and certain tests involving the homogeneity of regression coefficients. Also, comparing Type I and other types of sums of squares provides some information on the effect of the lack of balance.

11. 4. 2. Type II

Type II sums of squares are more difficult to understand. Generally, the Type II SS for an effect U, which may be a main effect or interaction, is adjusted for an effect V if and only if V does not contain U. Specifically, for a two-factor structure with interaction, the main effects, A and B, are not adjusted for the A*B interactions because the symbol A*B contains both A and B. Factor A is adjusted for B because the symbol B does not contain A. Similarly, B is adjusted for A, and the A*B interaction is adjusted for the two main effects. Type II SS is adjusted for all factors that do not contain the **complete** set of letters in the effect

Type II analysis relates to the following general guidelines often given in applied statistical texts. First, test for the significance of the A*B interaction. If A*B is insignificant, delete it from the

model and analyze main effects, each adjusted for the other. If A*B is significant, then abandon main-effects analysis and focus your attention on simple effects.

11. 4. 3. Type III

In this model every effect is adjusted for all other effects. This is the closest thing to a "standard" for ANOVA. The Type III sums of squares will produce the same SS as a Type I SS for a data set in which the missing data are replaced by the least-squares estimates of the values (See topic Topic 11.2.2). The Type III SS correspond to Yates' weighted squares of means analysis. One use of this SS is in situations that require a comparison of main effects even in the presence of interaction.

Type III sums of squares are **partial sums of squares**: each effect is adjusted for all other effects.

In particular, main effects A and B are adjusted for the interaction A*B if all these terms are in the model. If the model contains only main effects, then Type II and Type III analyses are the same.

11. 4. 4. Type IV

The Type IV functions were designed primarily for situations where there are empty cells. The principles underlying the Type IV sums of squares are quite involved and can be discussed only in a framework using the general construction of estimable functions. It should be noted that the Type IV functions are not necessarily unique when there are empty cells but are **identical to those provided by Type III when there are no empty cells**.

PROC GLM produces Type I and Type III SS as default. The four sums of squares can be requested in PROC GLM as options in the MODEL statement. For example, the following SAS statement specifies the printing of all four sums of squares.

```
model . . . / ss1 ss2 ss3 ss4;
```

11. 5. Unbalanced nested designs

The unbalance in subsample number in a nested design generates additional problems with the Expected Mean Squares (Topic 10). A complete example is available in ST&D page 168.

Example: Specific gravity of boards from several trees in three locations

Location	Location1						Location 3						Location 4			
Tree	1023	1096				1153	3008	3015	3020	4053	4067					
100xSG	55	53	50	51	54	58	45	48	52	48	52	62	59	55	60	

In this example both Trees and locations are RANDOM and trees are nested in location.

When you use the **RANDOM** statement, by default the GLM procedure produces the **Type III** expected mean squares (EMS) for model effects and for contrasts specified before the RANDOM statement in the program code.

You then use PROC VARCOMP to obtain appropriate estimates of the different components of variances. There are other different methods of estimation of variance components, which will produce similar results in a balanced design but different ones in unbalanced design. The Sequential or Type1 method is easier to understand and is presented in the example below (is generated by adding `method= Type1` after `proc varcomp`).

However, this Type1 method is valid only in pure nested designs (as in the example below). If there are crossed factors, the unbalance requires Type III SS and the analysis should NOT `INCLUDE method= Type1`. In the absence of this specification SAS uses the default method, **MIVQUE(0) Estimates**, which uses Type III SS.

The VARCOMP procedure handles general linear models that have random effects. Random effects are classification effects with levels that are assumed to be randomly selected from an infinite population of possible levels. **PROC VARCOMP** estimates the contribution of each of the random effects to the variance of the dependent variable. You can specify certain effects as fixed (nonrandom) by putting them first in the MODEL statement and indicating the number of fixed effects with the `FIXED=` option. Except for the effects specified as fixed, all other effects are assumed to be random.

SAS program for the ST&D example in page 168:

```
data STD170;
input location tree data;
cards;
1 1023 55
1 1096 53
1 1096 50
1 1096 51
1 1153 54
1 1153 58
3 3008 45
3 3008 48
3 3015 52
3 3015 48
3 3020 52
4 4053 62
4 4067 59
4 4067 55
4 4067 60
;
proc glm;
class location tree;
model data = location tree (location);
random tree(location) location /test;
proc varcomp method= Type1;
class location tree;
model data = location tree (location);
run; quit;
```

Dependent Variable: data

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
--------	----	----------------	-------------	---------	--------

Model	7	286.5666667	40.9380952	7.32	0.0088
Error	7	39.1666667	5.5952381		
Corrected Total	14	325.7333333			

R-Square	Coeff Var	Root MSE	data Mean
0.879758	4.424113	2.365426	53.46667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
location	2	222.2333333	111.1166667	19.86	0.0013
tree(location)	5	64.3333333	12.8666667	2.30	0.1539

Source	DF	Type III SS	Mean Square	F Value	Pr > F
location	2	198.9514652	99.4757326	17.78	0.0018
tree(location)	5	64.3333333	12.8666667	2.30	0.1539

Source	Type III	Expected Mean Square
location	Var(Error) + 1.5412	Var(tree(location)) + 4.0549
tree(location)	Var(Error) + 1.6733	Var(tree(location))

Tests of Hypotheses for Random Model Analysis of Variance

Dependent Variable: data

Source	DF	Type III SS	Mean Square	F Value	Pr > F
location	2	198.951465	99.475733	8.09	0.0239
Error	5.3744	66.065089	12.292523		

Error: $0.921 * MS(\text{tree}(\text{location})) + 0.079 * MS(\text{Error})$

Note that MSE $12.29 = 5.6 + 1.54 * 4.35$

This is exactly the error requested by the Expected Mean Squares to test **loc!**

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tree(location)	5	64.333333	12.866667	2.30	0.1539
Error: MS(Error)	7	39.166667	5.595238		

Variance Components Estimation Procedure Same as above

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
location	2	222.23	111.12	Var(Error) + 2.22
tree(location)	5	64.33	12.87	Var(Error) + 1.67
Error	7	39.17	5.59	Var(Error)
Corrected Total	14	325.73		

Variance Component	Estimate
Var(loc)	19.44
Var(tree(loc))	4.35
Var(Error)	5.60

This last result represents one of the main objectives of the nested design: ESTIMATE THE MAGNITUDE OF THE DIFFERENT VARIANCE COMPONENTS

Final comment:

For **unbalanced mixed models** with crossed factors it is necessary to use a different SAS procedure called **PROC MIXED** (ST&D page 411) that will not be covered in this class. The syntax is similar to PROC GLM but the output is more complex. Information about PROC MIXED is available at: <http://software.ucdavis.edu/sas/sashtml/stat/chap41/index.htm>

PROC MIXED is also useful when you need to test contrasts for a factor that has a complex synthetic denominator. In PROC GLM contrasts are not corrected by the RANDOM statement, and there is no way provided for the user to specify a synthetic denominator for a contrast. PROC MIXED will automatically give appropriate tests for all model effects and, unlike GLM, will give appropriate tests for contrasts.

In PROC MIXED the fixed factor effects and random factor variance components are estimated by a method known as **Restricted Maximum Likelihood (REML)**.

PROC MIXED has similar CLASS, MODEL, CONTRAST, and LSMEANS statements as GLM; but their RANDOM and REPEATED statements differ.