ANALYSIS OF VARIANCE WITH STRUCTURAL ZEROES

Thomas Hans Chelius, Raymond G. Hoffmann, Medical College of Wisconsin
Thomas Hans Chelius, Medical College of WI, 8701 Watertown Plank Rd, Milwaukee, WI, 53226

Key Words: Missing Data, Imputation, Type IV Sum of Squares, Estimability

Abstract

Structural zeroes, or unobservable data, create a problem when performing Analysis of Variance. Replacing the missing cells with estimated values is inappropriate in this case because these cells cannot have data in them. The software packages BMDP and SPSS, up until recently, had no methods for dealing with Structural Zeroes. SAS uses the Type IV Sum of Squares to handle the problem of missing cells, but is very sensitive to how the data is organized. Two alternatives are a reconstrained Least Squares approach or a decomposition of the problem into smaller complete blocks. These two methods are compared with the Type IV Sum of Squares approach.

Section 1: Introduction

Section 1.1: The Problem

One of the problems one confronts when doing an Analysis of Variance (ANOVA) is missing values. An example of such a problem is when an investigator cannot observe data because of some equipment malfunction. If such a malfunction occurred for an entire treatment combination, then a **missing cell** has occurred. The standard normal equations for ANOVA do not allow for missing cells. There are various methods for dealing with missing cells, such as filling in missing cells (Kirk, 1968) or breaking up the data into complete subsets (Searle, 1987).

Let us take the problem of missing cells one step further. The previous problem assumes that the missing cells **could** be observed under better conditions. If the particular treatment combination cannot be observed under any ideal conditions, then the missing cell is called a **structural zero**. An example of such data is a SPECT scan of the brain. Here, instead of treatment levels, we have the horizontal and vertical locations of the observations. Clearly, the brain is an irregular region. If one uses ANOVA to analyze this particular type of data, one would have many cells that could not be observed, because there is no tissue that could be scanned. Structural zeroes provide a different problem from that of missing cells, since structural zeroes can never be observed, and methods for filling in the missing cells are not appropriate.

Section 1.2: Analysis of Variance

The model for a **two-way Analysis of Variance** can be written as:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk},$$

where the ε_{ijk}'s are independent and follow a Normal Distribution with mean 0 and variance σ^2. This model is known as the **effects model**. The random variable, y, is the outcome variable. The parameter μ is the overall mean effect. The parameter α_i corresponds to the effect of treatment A_i on the outcome. The parameter β_j corresponds to the effect of treatment B_j on the outcome. The parameter $\alpha\beta_{ij}$ represents the **interaction** effects of treatments A and B. The values of *k* represent replicate observations. The following figure illustrates an example of the rectangular nature of the data if the data is complete. The $y_{ij\bullet}$ notation represents more than one observation in each cell.

Figure 1.1: Data Table for ANOVA

		Treatment A			
		1	2	3	4
Treat-	1	$y_{11\bullet}$	$y_{21\bullet}$	$y_{31\bullet}$	$y_{41\bullet}$
ment	2	$y_{12\bullet}$	$y_{22\bullet}$	$y_{32\bullet}$	$y_{42\bullet}$
	3	$y_{13\bullet}$	$y_{23\bullet}$	$y_{33\bullet}$	$y_{43\bullet}$
B	4	$y_{14\bullet}$	$y_{24\bullet}$	$y_{34\bullet}$	$y_{44\bullet}$

When structurally missing data occurs, the figure could look like the following table.

Figure 1.2: Data Table with Structurally Zeroes

		Treatment A			
		1	2	3	4
Treat-	1	$y_{11\bullet}$	$y_{21\bullet}$	$y_{31\bullet}$	$y_{41\bullet}$
ment	2	$y_{12\bullet}$	$y_{22\bullet}$	$y_{32\bullet}$	$y_{42\bullet}$
	3	$y_{13\bullet}$	$y_{23\bullet}$	$y_{33\bullet}$	$y_{43\bullet}$
B	4	$y_{14\bullet}$	$y_{24\bullet}$	$y_{34\bullet}$	$y_{44\bullet}$

The shaded area represents the cells that do not exist. The values of in these cells are effectively zero.

Section 2: Methods for Dealing with Missing Data

Section 2.1: Decomposition as a Solution

Using subsets is appealing for the structured zero problem, because it does not assume that the missing cells exist. Unfortunately, a decomposition as proposed by Searle, destroys the structure of the problem. However, one can modify the decomposition to include overlapping subsets. Let us assume we have the data matrix in Figure 1.2. This figure can be decomposed into the following two overlapping subsets.

Figure 2.1a: Overlapping (Upper) Block 1

		Treatment A			
		1	2	3	4
Treat.	1	$y_{11\bullet}$	$y_{21\bullet}$	$y_{31\bullet}$	$y_{41\bullet}$
B	2	$y_{12\bullet}$	$y_{22\bullet}$	$y_{32\bullet}$	$y_{42\bullet}$

Figure 2.1b: Overlapping (Right) Block 2

		Treatment A	
		3	4
Treat-	1	$y_{31\bullet}$	$y_{41\bullet}$
ment	2	$y_{32\bullet}$	$y_{42\bullet}$
	3	$y_{33\bullet}$	$y_{43\bullet}$
B	4	$y_{34\bullet}$	$y_{44\bullet}$

With this decomposition into the maximum sized overlapping rectangles,

- we still keep the structure of the problem and avoid filling in the missing cells,
- we have the **maximum size blocks** to test for inconsistencies due to interactions, and
- we can estimate a **complete set of interactions** for the combined problem and for each subset.

This method has its advantages and disadvantages. It is certainly easy to implement in such software packages like SAS. One is also working with complete blocks. On the down side, it clouds the idea of interaction in the overlap. In the example above, does one use the Upper Block, the Right Block or some combination of both? Also, when using this method, one has effectively turned one analysis into multiple analyses. Finally, the data needs to be connected. Otherwise, the decomposition cannot be made into complete blocks.

Section 2.2: SAS and Type IV Sum of Squares

The statistical software package SAS suggests using **Type IV Sum of Squares** to handle missing cell data. Type IV Sum of Squares can be calculated by using the R() notation (Speed, Hocking, Hackney;1988). We will use the two-way ANOVA with interaction as our model. The Type IV Sum of Squares for the main effect A is $R(\alpha \mid \mu, \beta, \alpha\beta)$, for main effect B, $R(\beta \mid \mu, \alpha, \alpha\beta)$, and for the interaction $R(\alpha\beta \mid \mu, \alpha, \beta)$ (Little, Freund, Spector; 1993). These are identical to Type III Sum of Squares in all cases except when missing cells occur. Unfortunately, Type IV Sum of Squares can be different depending on how your data is organized. Using the identical data set, one can change a main effect sum of squares by reorganizing the data set and moving the missing cells. The following tables give the results of using Type IV analysis in SAS and switching rows and columns.

Figure 2.2: Different Scenarios for Testing Type IV Analysis

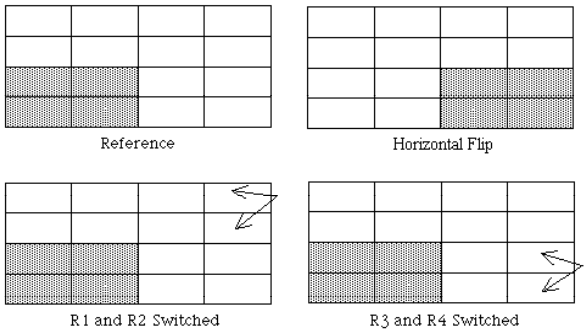

Table 2.1: Type IV Sum of Squares

	Reference	Horizontal Flip	R1,R2 switch	R3,R4 switch
	Sum of Squares			
SSA	36.02	40.89	36.02	36.02
SSB	40.16	40.16	40.16	40.16
SSAB	18.57	18.57	18.57	18.57

Table 2.2: SAS's Parameter Estimates

	True Value	Refer.	Horiz. Flip	R1,R2 switch	R3,R4 switch
	Parameter Estimates				
μ	5	9.96	4.54	9.96	6.81
$\alpha 1$	-2	-2.55	0.00	-5.56	-2.55
$\alpha 2$	-1	-3.50	-0.94	-2.60	-3.50
$\alpha 3$	1	-2.87	2.56	-2.87	1.29
$\alpha 4$	2	0.00	5.42	0.00	0.00
$\beta 1$	-2	-4.38	-4.52	-4.38	-1.22
$\beta 2$	-1	-4.52	-1.65	-4.52	-1.37
$\beta 3$	1	-3.16	1.00	-3.16	0.00
$\beta 4$	2	0.00	0.00	0.00	3.16
$\alpha\beta 11$	-1	-3.01	0.00	0.00	-3.01
$\alpha\beta 12$	1	0.00	0.00	3.01	0.00
$\alpha\beta 21$	1	0.89	3.90	0.00	0.89
$\alpha\beta 22$	-1	0.00	0.00	-0.89	0.00
$\alpha\beta 31$	-.5	0.76	0.90	0.76	-3.40
$\alpha\beta 32$	.5	2.87	0.00	2.87	-1.29
$\alpha\beta 33$	1	4.16	0.00	4.16	0.00
$\alpha\beta 34$	-1	0.00	0.00	0.00	-4.16
$\alpha\beta 41$	.5	0.00	0.14	0.00	0.00
$\alpha\beta 42$	-.5	0.00	-2.87	0.00	0.00
$\alpha\beta 43$	-1	0.00	-4.16	0.00	0.00
$\alpha\beta 44$	1	0.00	0.00	0.00	0.00

Looking at Table 2.1, one realizes that the horizontal flip changes the Sum of Squares for Treatment A. The data has not changed, only the position, yet Type IV analysis gives inconsistent results. This is due primarily to the fact that the main effects hypotheses change when data is reorganized, although SAS does not tell you what those hypotheses are. When examining Table 2.2, one notices an odd behavior. Take, for instance, the difference α_1-α_2. In all but one case, the difference is approximately 0.95. The case where rows 1 and 2 are switched leads to a difference of 2.96. We can see that this is dramatically different from the other three cases. This is particularly surprising since we did not change the positioning of the missing cells.

The advantages and disadvantages for SAS Type IV Sum of Squares should be clear. Ease of implementation and the variety of options SAS offers make SAS an appealing option. Unfortunately, with different arrangements come different Type IV hypotheses. SAS does not inform the user what hypotheses it is testing. It is quite possible that the hypotheses being tested might be of no interest to the investigator. Fortunately one can use the CONTRAST statement to test appropriate hypotheses. The SOLUTION statement is misleading. It yields biased, non-unique results that are also dependent on data arrangement. One needs to consider what is estimable and use the appropriate ESTIMATE statement. Freund (1980) comments that using Type IV analysis may cause sufficient confusion that one might give up altogether. Clearly, Type IV analysis should be used carefully for the problem of structural zeroes.

Section 2.3: The Structural Zero Least Squares Approach

When one is trying to estimate parameters in an ANOVA model, one needs to use the least squares equation,

$$X'X\underline{\beta} = X'\underline{Y},$$

where X is the design matrix, $\underline{\beta}$ is the vector of parameters and $\underline{Y}$ is the vector of the observed data. Solving for $\underline{\beta}$ yields,

$$\hat{\underline{\beta}} = (X'X)^{-} X'\underline{Y}.$$

Without constraints, (X' X) is a singular matrix, and one needs to find a generalized inverse to solve this equation. Another way of handling this problem is to put constraints on $\underline{\beta}$. The general form for constraints on $\underline{\beta}$ is,

$$K\underline{\beta} = \underline{0}.$$

In the example of a balanced two-way ANOVA with interaction, one usually uses the following constraints,

$$\sum_i \tau_i = 0,$$
$$\sum_j \beta_j = 0,$$
$$\sum_i \alpha\beta_{ij} = \sum_j \alpha\beta_{ij} = 0.$$

One can then write some of the parameters in terms of the others, and reduce the design matrix X to a nonsingular form.

When dealing with structural zeroes, one needs to modify these constraints to take into account that there are no parameters in the holes Take for instance the example in Figure 1.2. Here, $\alpha\beta_{13} \equiv \alpha\beta_{23} \equiv \alpha\beta_{14} \equiv \alpha\beta_{24} \equiv 0$. Since the degrees of freedom left for interaction after estimating the main effects are

now different as a result of the missing cells, we will need to impose additional constraints to keep the design matrix X nonsingular. Using Figure 1.2, we see that we do not need to adjust the constraints for the main effects. We do, however, need to adjust the constraints for interaction: $\alpha\beta_{11}$= -$\alpha\beta_{12}$, $\alpha\beta_{21}$= -$\alpha\beta_{22}$, $\alpha\beta_{33}$= -$\alpha\beta_{43}$, and $\alpha\beta_{34}$= -$\alpha\beta_{44}$. So, we need to work around the hole by setting the interaction terms of the hole to 0 and adjusting the constraints on the remaining terms.

To have a nonsingular design matrix, one needs to reduce the number of parameters down to the number of degrees of freedom of each effect. In our example, we clearly have 3 degrees of freedom for Treatment A and three degrees of freedom for Treatment B. We then need only estimate α_2, α_3, and α_4 for Treatment A and β_2, β_3, and β_4 for Treatment B. By the constraints, we have

$$\alpha_1 = -(\alpha_2 + \alpha_3 + \alpha_4),$$
$$\beta_1 = -(\beta_2 + \beta_3 + \beta_4).$$

For the interaction effect, we would normally have 9 degrees of freedom. By reconstraining the problem, we now have only 5 degrees of freedom for interaction as a result of removing the 4 parameters (2x2 hole) and setting them to 0. In other words we only have 5 linearly independent parameters, given our constraints. The method is similar for 1x1 holes and 3x3 holes. The following figure is one possible way to choose the 5 degrees of freedom (marked with a †).

Figure 2.4: Degrees of Freedom for the Interaction Effect

		Treatment A			
		1	2	3	4
Treat-	1	$\alpha\beta_{11}^{\dagger}$	$\alpha\beta_{21}^{\dagger}$	$\alpha\beta_{31}$	$\alpha\beta_{41}^{\dagger}$
ment	2	$\alpha\beta_{12}$	$\alpha\beta_{22}$	$\alpha\beta_{32}$	$\alpha\beta_{42}$
	3	$\alpha\beta_{13}$	$\alpha\beta_{23}$	$\alpha\beta_{33}$	$\alpha\beta_{43}^{\dagger}$
B	4	$\alpha\beta_{14}$	$\alpha\beta_{24}$	$\alpha\beta_{34}$	$\alpha\beta_{44}^{\dagger}$

For the given pattern, the remainder of the parameters can be written from the constraints as,

$$\alpha\beta_{12} = -\alpha\beta_{11} \qquad \alpha\beta_{22} = -\alpha\beta_{21}$$
$$\alpha\beta_{33} = -\alpha\beta_{43} \qquad \alpha\beta_{34} = -\alpha\beta_{44}$$
$$\alpha\beta_{31} = -(\alpha\beta_{11} + \alpha\beta_{21} + \alpha\beta_{41})$$
$$\alpha\beta_{42} = -(\alpha\beta_{41} + \alpha\beta_{43} + \alpha\beta_{44})$$
$$\alpha\beta_{32} = \alpha\beta_{11} + \alpha\beta_{21} + \alpha\beta_{41} + \alpha\beta_{43} + \alpha\beta_{44} \ .$$

Clearly, this is not the only pattern one can choose. Choosing different patterns will result in different equations.

The advantages to this method are clear. The full rank design allows for unique, unbiased parameter estimates. Also, results are not dependent on arrangement since all that changes is column order in the design matrix (and rows in the parameter matrix). Finally, depending on how this method is implemented, parameter estimates and sum of squares are readily extractable. SAS does not extract parameter estimates easily. One the other hand, since this method is not currently automated, implementation is difficult and time consuming.

Section 3: Discussion

Upon reviewing the various methods, there is no clear winner when dealing with Analysis of Variance with Structural Zeroes. As with all statistical analyses, the investigator needs to be aware of what hypotheses are being tested and what hypotheses need to be tested. Although the easiest to use, Type IV Sum of Squares analysis does not tell the investigator what is being tested. Instead, it is up to the investigator to supply the appropriate hypotheses relative to what is being studied. The Structural Zeroes Least Squares method gives the investigator the advantage of working with a full rank model, but implementation can be difficult depending on the level of the investigator and diagnostics even more cumbersome. The Overlapping Block Decomposition method is probably the least favorable. While easy to implement, it does not work for all cases. Interaction becomes difficult to understand in the overlap and neither of the treatments are fully expressed in either block. The deciding factors on which method to use are how much time and what resources are available to perform the analyses and what analyses are required.

References

1. Dodge, Y. (1985), *Analysis of Experiments with Missing Data*. New York: John Wiley and Sons, Inc.

2. Freund, R.J. (1980), “The Case of the Missing Cell.” *The American Statistician*, **34**, 94-98.

3. Kirk, R. (1968), *Experimental Design: Procedures for the Behavioral Sciences*. Belmont: Brooks/Cole Publishing Company.

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

5. Schluchter, M.D., *BMDP 5V -- Unbalanced Repeated Measures Models with Structured Covariance Matrices*. Tech Report 86. Los Angeles: BMDP Statistical Software, Inc.

6. Searle, S.R. (1987), *Linear Models for Unbalanced Data*. New York: John Wiley and Sons, Inc.

7. Speed, F.M., Hocking, R.R., and Hackney, O.P. (1978), “Methods of Analysis of Linear Models with Unbalanced Data.” *Journal of the American Statistical Association*, **73**, 105-112.

8. Yates, F., “The Analysis of Replicated Experiments when Field Results are Incomplete.” *Empire Journal of Experimental Agriculture*, 1933, **1**, 129-142.