Experimental Design Models

By J. Leroy Folks*

INTRODUCTION

While the discussion in this paper is limited mainly to the two-way classification for ease and compactness of presentation, it applies to most of the models encountered in experimental design.

If we were to assume a linear relationship between x and y described by the model $y = \alpha + \beta x + e$ it is unlikely that we would consider writing the model as $y = a + bx + cx + e$. It is even more unlikely that we would apply the least squares principle by minimizing Σe^2 with respect to a, b, and c. Yet a similar thing happens in experimental design. In fact, it is common practice to use less than full-rank models where the parameters are not defined and, in cases where they are defined, to minimize Σe^2 with respect to the full set of parameters which are not functionally independent.

Certainly, it is valuable to have more than one way of looking at a given situation and it seems that one's understanding is deepened by having many perspectives. However, it is the author's conviction that in many cases the use of the less than full-rank models has not aided understanding. From a computational viewpoint, their use has led to the situation where people build unreasonably large sets of equations and then proceed to impose conditions in order to get a solution. This is often done with a complete lack of understanding of the consequences. Alternatively, we build a large set of equations and then proceed to go through some process of reducing the number of equations (Harvey, 1960). This can be avoided. The model can be written in the first place as a full-rank model — the principle of least squares yielding immediately a smaller set of linearly independent equations.

POSSIBLE MODELS

Model I:

$$y_{ijk} = \mu_{ij} + e_{ijk}$$
$$i = 1, 2, \ldots, r$$
$$j = 1, 2, \ldots, s$$
$$k = 0, 1, 2, \ldots\ n_{ij}$$

This simply means that in cell i-j, we have n_{ij} observations from that population. This is a perfectly workable model and would suffice for testing any hypothesis about cell means for the two-way classification.

*Statistical Laboratory, Oklahoma State University.

Model II:
$$y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$
$$\alpha_i = \mu_i$$
$$\beta_j = \mu_{.j} - \mu_{..}$$
$$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$$

The dots indicate averages over the respective subscripts. It should be emphasized that the additive property of the model, as far as parameters are concerned, is not merely assumed but follows directly from the identity:

$$\mu_{ij} = \mu_{i.} + (\mu_{.j} - \mu_{..}) + (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}).$$

It is a consequence of the definition of parameters that

$$\sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

If we assume that $(\alpha\beta)_{ij} = 0$ for all i and j, the non-interaction model follows directly since $(\alpha\beta)_{ij} = 0$ for all i and j implies

$$\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..} = 0$$
$$\mu_{ij} = \mu_{i.} + (\mu_{.j} - \mu_{..})$$
$$= \alpha_i + \beta_j$$

Alternative definitions for α_i, βj, and $(\alpha\beta)_{ij}$ in terms of μ_{ij} are sometimes given either explicitly or in terms of conditions the parameters must satisfy. These alternative definitions will not be discussed in this paper.

Model III:
$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$
$$\mu = \mu_{..}$$
$$\alpha_i = \mu_{i.} - \mu_{..}$$
$$\beta_j = \mu_{.j} - \mu_{..}$$
$$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$$

Again, it should be emphasized that the additive property of the parameters is not assumed but follows directly from their definition in terms of the μ's. Also $\sum \alpha_i = \sum_i \beta_j = \sum (\alpha\beta)_{ij} = \sum_i (\alpha\beta)_{ij} = 0$ is a consequence of definition. The assumption that $(\alpha\beta)_{ij} = 0$ for all i and j leads immediately to the model

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

As with Model II, we shall not consider alternative definitions in this paper.

Model IV: $y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$

Definitions in terms of μ's and the conditions satisfied by parameters are not stated.

Since the set of equations

$$\mu_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij},\ i = 1, 2, \ldots, r$$
$$j = 1, 2, \ldots, s$$

has no unique solution, a model of this type simply means that the parameters have not been defined in terms of the μ's. This seems to be an unnatural way of conceiving of a model. Yet such models are in common usage.

Model V: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$

Definitions in terms of μ's and conditions satisfied by parameters are not stated.

As with model IV, the parameters are simply not defined.

All of the above models describe the so-called fixed effect situations. That is, the μ_{ij}'s involved in the experiment represent the entire population of μ_{ij}'s to which inference is to be made.

THE MATTER OF ESTIMABILITY

If an observation is obtained in cell i-j, μ_{ij} can be estimated. Any parameter which is a function of estimable μ_{ij}'s is estimable. Thus, with no missing cells, all parameters in models I, II, and III are estimable. With missing cells,

(1) The parameters must be redefined in terms of cell means represented in the data, or

(2) The assumption of no interaction permits estimates of the parameters in models I, II, and III if the data set is a connected set.

The question of estimability occupies a large part of the literature in the design of experiments, particularly for models IV and V. Since the parameters in these models are not explicitly defined, they are not estimable. While this fact is obvious, it seems somehow to be obscured by the array of theorems and methods for handling such models (Graybill 1960). Any function of these parameters which is uniquely defined in terms of estimable μ_{ij}'s is also estimable. This fact also seems less profound than when it is obscured by all sorts of matrix manipulations.

It is common to speak of reparametrization of models IV and V. This amounts to defining parameters in terms of estimable μ_{ij}'s. Any such parameter is then estimable. Actually, it would be better to refer to this process as parametrization instead of reparametrization since the α's, β's, etc., of models IV and V are not even defined.

THE HANDLING OF NORMAL EQUATIONS

In the opinion of the author, the following two practices seem to be widespread and will be referred to as the usual approach:

1. Using models IV and V and imposing conditions upon the estimates, and

2. Using models II and III, ignoring the functional dependence of the parameters, and imposing conditions upon the estimates. The conditions imposed may or may not be the same as the condition imposed by the parameters in the model.

It seems to be ignored that by using models II and III and recognizing the functional dependence, we are led at once to a smaller set of normal equations than by the usual approach — a set which, in addition, is generally simpler to solve and is of full rank.

Example 1. Two-Way Classification without Interaction
Suppose $r = 3,\ s = 2,\ n_{ij} = 1.$

The usual approach with models III and V leads to the normal equations

$$\begin{bmatrix} 6 & 2 & 2 & 2 & 3 & 3 \\ 2 & 0 & 2 & 0 & 1 & 1 \\ 2 & 0 & 0 & 2 & 1 & 1 \\ 3 & 1 & 1 & 1 & 3 & 0 \\ 3 & 1 & 1 & 1 & 0 & 3 \\ 2 & 2 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} Y_{..} \\ Y_{1.} \\ Y_{2.} \\ Y_{3.} \\ Y_{.1} \\ Y_{.2} \end{bmatrix}$$

Here we have six equations, four of which are linearly independent.

If alternatively we work with model III and recognize the functional relationships, the observation equations are:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

The normal equations are:

$$\begin{bmatrix} 6 & 0 & 0 & 0 \\ 0 & 4 & 2 & 0 \\ 0 & 2 & 4 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} Y_{..} \\ Y_{1.} - Y_{3.} \\ Y_{2.} - Y_{3.} \\ Y_{.1} - Y_{.2} \end{bmatrix}$$

We again have four linearly independent equations which are simpler to solve than those given by the usual approach.

Example 2. Two-way classification with interaction, one observation per cell.

Of course, in this example, we have no estimate of error, but the example suffices to illustrate the problems of handling the normal equations.

Suppose $r = 3$, $s = 2$. The usual approach with models III or V leads to the normal equations:

$$\begin{bmatrix} 6 & 2 & 2 & 2 & 3 & 3 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 3 & 1 & 1 & 1 & 3 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ (\hat{\alpha}\beta)_{11} \\ (\hat{\alpha}\beta)_{12} \\ (\hat{\alpha}\beta)_{21} \\ (\hat{\alpha}\beta)_{22} \\ (\hat{\alpha}\beta)_{31} \\ (\hat{\alpha}\beta)_{32} \end{bmatrix} = \begin{bmatrix} Y_{..} \\ Y_{1.} \\ Y_{2.} \\ Y_{3.} \\ Y_{.1} \\ Y_{.2} \\ y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}$$

We then have twelve equations, six of which are linearly independent. Alternatively if we recognize the functional relationships in model III and express our model in terms of the functionally independent parameters, we have the observation equations:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{21} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

This leads us immediately to the normal equations:

$$\begin{bmatrix} 6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 2 & 0 & 0 & 0 \\ 0 & 2 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 2 \\ 0 & 0 & 0 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_1 \\ (\hat{\alpha\beta})_{11} \\ (\hat{\alpha\beta})_{21} \end{bmatrix} = \begin{bmatrix} Y_{..} \\ Y_{1.} - Y_{3.} \\ Y_{2.} - Y_{3.} \\ Y_{.1} - Y_{.2} \\ y_{11} + y_{32} - y_{12} - y_{31} \\ y_{21} + y_{32} - y_{22} - y_{31} \end{bmatrix}$$

We then have six linearly independent equations which are simple to solve.

Example 3. N-Way Classification with Interaction. The following problem is a problem with which the author came in contact.

An experiment involving seven factors, each with three levels, had been run and had resulted in unequal subclass numbers with no missing cells. Three factor interactions and higher were assumed negligible. A set of normal equations had been constructed by the usual approach, resulting in 85 equations. Of course, only 57 of these were linearly independent. By recognizing the functional dependence of the parameters in the first place and expressing the observation equations in terms of the functionally independent parameters, we were led at once to 57 linearly independent equations.

Example 4. Two-way Classification with Empty Cells.

Suppose we wish to test whether the levels of factor B have the same effect in the following classification:

		Factor B Level	
	1	2	3
1	3	4	x
Factor A Level 2	4	6	x
3	x	x	7

Proceeding in the usual manner, we would set up the normal equations, impose conditions, and obtain a sum of squares for testing the levels of factor B. A little reflection, however, will show that in reality the non-centrality factor of our test is $\lambda = (\mu_{.1} - \mu_{.2})^2/2$. That is, it does not involve the third level of B at all. In a simple example of this type, we recognize this readily, but in an n-way classification with missing data the procedure of imposing conditions may mislead us in that the non-centrality factor may not be at all what we think it is.

REFERENCES

[1] Harvey, Walter R., *Least-square analysis of data with unequal subclass numbers*. U.S. Agricultural Research Service, United States Department of Agriculture.

[2] Graybill, Franklin A., 1961, *An Introduction to linear statistical models*, vol. I, McGraw-Hill, New York.