LINEAR MODELS FOR SOME-CELLS-EMPTY DATA: THE CELL MEANS FORMULATION, A CONSULTANT'S BEST FRIEND

Shayle R. Searle[1)]

Biometrics Unit, Cornell University, Ithaca, N.Y.

BU-867-M February, 1985

ABSTRACT

Linear model analyses are well known for balanced data, for balanced data having a few missing observations, and for data exhibiting planned unbalancedness, such as those from latin squares and balanced incomplete blocks. For data of a more generally unbalanced nature, those that have all cells filled can be usefully analyzed using the weighted-squares-of-means analysis. For some-cells-empty data, analysis based on main-effects-only models are useful whenever interactions are to be ignored. But analyzing some-cells-empty data on the basis of models with interactions is best undertaken using cell means models. Whereas the essential concepts and arithmetic are then easy, the data gatherer and the consulting statistician must work together to decide on, to estimate, and to test hypotheses about, linear combinations of cell means that are of interest. Extensions of cell means models to excluding some (or even all) interactions, and to mixed models, are also available.

1. INTRODUCTION

Consider the analysis of variance of data classified by two different factors that shall be called rows and columns. Suppose there are a rows and b columns. Data in row i and column j (for i taking values i = 1,2,...,a and for j = 1,2,...,b) shall be described as being in cell i,j . Let y_{ijk} be the k'th observation in cell i,j, where there are n_{ij} observations in that cell, so that k = 1,2,...,n_{ij} .

[1)] This article was prepared while the author was at the University of Augsburg, Federal Republic of Germany, on a U.S. Senior Scientist Award from the Alexander von Humboldt-Stiftung.

Data wherein n_{ij} has the same value for every cell, i.e., $n_{ij} = n > 0$ for every cell i,j, are called balanced data. Means of the data for cell i,j, row i, column j and for all the data are then

$$\bar{y}_{ij.} = \sum_{k=1}^{n} y_{ijk}/n = y_{ij.}/n, \qquad \bar{y}_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}/bn = y_{i..}/bn$$

$$\bar{y}_{.j.} = \sum_{i=1}^{a}\sum_{b=1}^{n} y_{ijk}/an = y_{.j.}/an, \qquad \bar{y}... = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}/abn = y.../abn \tag{1}$$

In contrast, data where the n_{ij} are not all the same value are called unbalanced and the observed means are then

$$\bar{y}_{ij.} = \sum_{k=1}^{n_{ij}} y_{ijk}/n_{ij}$$

$$\bar{y}_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}/n_{i.} = y_{i..}/n_{i.} \quad \text{for} \quad n_{i.} = \sum_{j=1}^{b} n_{ij}$$

$$\bar{y}_{.j.} = \sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk}/n_{.j} = y_{.j.}/n_{.j} \quad \text{for} \quad n_{.j} = \sum_{i=1}^{a} n_{ij} \tag{2}$$

and

$$\bar{y}... = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}/n.. = y.../n.. \text{ for } n.. = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\ .$$

Clearly, when $n_{ij} = n$ for all i and j, the means in (2) reduce to those in (1).

Several subclasses of unbalanced data deserve distinction. First is when $n_{ij} = n$ for basically all i and j except that in a very few cells (one, two or three, say) the number of observations is one or two less than n. This is usually the case of a very few intended observations being lost or missing from the data due to experimental misadventure; maybe one or two laboratory animals died, or a petri dish got broken, or an experimental plot got eaten by a cattle beast that broke a fence. Under these circumstances, there are

well-known techniques for estimating such missing observations (e.g. Steel and Torrie, 1980, pp. 209,227 and 388); we refer to such cases of unbalanced data as data with missing observations.

Second, is what can be called planned unbalancedness. This is when there are no observations on certain, carefully planned combinations of levels of the factors involved in an experiment. Latin squares are examples of this: a Latin square of order n is a (1/n)'th replicate of a three-factor factorial experiment with each factor having n levels, an n^3 experiment. Balanced incomplete blocks are also examples of planned unbalancedness.

The third and final class of unbalanced data is where the numbers of observations in the sub-most cells (cells defined by one level of each factor) are not all equal, and may in fact be quite unequal. This can include some cells having no data but, in contrast to planned unbalancedness, with those cells occurring in an unplanned manner. Survey data are an example of this, where data are simply collected because they exist, and so the numbers of observations in the cells are just those that arise in the collection process. Records of most human activities are of this nature; e.g., yearly income for people classified by age, sex, education, education of each parent, and so on. This is the kind of data, that shall be called unbalanced data; and within this class of data we make two further divisions. One is for data in which all the sub-most cells contain data; none are empty. We call this all-cells-filled data. Complementary to this is some-cells-empty data, wherein several sub-most cells have no data.

2. BALANCED DATA

2.1. Analysis of variance

The analysis of variance was developed by R.A. Fisher as an analysis of differences among observed means. Its very basis, for the row-by-column example with means given in (1) is the simple algebraic identity

$$\begin{aligned}\sum\sum\sum(y_{ijk} - \bar{y}...)^2 \\ \equiv \sum\sum\sum(\bar{y}_{i..} - \bar{y}...)^2 + \sum\sum\sum(y_{.j.} - \bar{y}...)^2 \\ + \sum\sum\sum(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}...)^2 \\ + \sum\sum\sum(y_{ijk} - \bar{y}_{ij.})^2\end{aligned} \tag{3}$$

where, in each case, the triple summation is $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}$. Under the customary assumptions of homoscedasticity and normality, each sum of squares on the right-hand side of (3) is distributed proportional to a χ^2-variable, independently of the others; and from this come the familiar F-statistics. These and (3) are then summarized in tabular form as an analysis of variance table.

Fisher has an interesting comment on this table. In a letter (on display at the 50'th Anniversary Conference of the Statistics Department at Iowa State University, June, 1983) dated "6/Jan/'34" Fisher writes to Snedecor that "the analysis of variance is (not a mathematical theorem but) a simple method of arranging arithmetical facts so as to isolate and display the essential features of a body of data with the utmost simplicity". That the analysis of variance table is indeed, as Fisher says, no more than "a simple method of arranging arithmetical facts" is worth emphasising in these days of computer-generated tables which too many computer package users are in-

clined to erroneously treat as sacrosanct.

2.2. Models

Fisher's starting point was (3). He seldom, if ever, began with a model, as we so often do to-day. Thus a current trend is to describe analysis of variance in terms of a model involving parameters in an additive (linear) manner, parameters which, as an aside to analysis of variance, we often seek to estimate. Thus for the row-by-column example we model y_{ijk} using the equation

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{4}$$

where μ is a general mean, α_i is an effect due to the i'th row, β_j is an effect due to the j'th column, γ_{ij} is an effect due to the interaction of the i'th row and j'th column, and E represents expectation over repeated sampling. Then, on defining $y_{ijk} - E(y_{ijk})$ as a random residual error term, e_{ijk}, the model equation is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \, . \tag{5}$$

The variance-covariance structure usually assumed for the e_{ijk} terms is that each of them has the same variance, σ^2 say, and that covariances between every two of them are zero. Under these conditions, the analysis of variance table based on the sums of squares in (3), for balanced data, can be described in terms of the model (4) and various sub-models thereof. For example, $\sum\sum\sum(\bar{y}_{i..} - \bar{y}...)^2$ is the difference between the sums of squares due to fitting $E(y_{ijk}) = \mu + \alpha_i + \beta_i$ and $E(y_{ijk}) = \mu + \beta_j$.

2.3 Estimation

One advantage of having (4) is that the parameters in that equation make it easy for us to be specific about what we might like to estimate; for example, concerning row effects, we might be interested in estimating such functions as α_1, $\alpha_1 - \alpha_2$ and $\frac{1}{3}(\alpha_1 + \alpha_2 + \alpha_3) - \alpha_4$. However, a counteracting difficulty is that (4) involves more parameters than there are observed means to estimate them from. There are $1 + a + b + ab$ parameters but only ab cell means, $\bar{y}_{ij\cdot}$ for i=1,...,a and j=1,...,b. Hence there are too many parameters for us to be able to estimate them all as linear functions of the means, the $\bar{y}_{ij\cdot}$. This is the feature of models such as (4) that is well known as overparameterization.

A consequence of overparameterization is that not all the parameters of a model can be estimated. To circumvent this situation we usually invoke one of two procedures: either we use estimable functions, which has us confine attention to only certain functions of the parameters that can be estimated satisfactorily from the data; or we use re-parameterization, wherein we define relationships among parameters of an overparameterized model which implicitly rewrites the model in terms of as many new parameters as can be estimated from the data. Each of these procedures (confinement to estimable functions, or reparameterization) is easily applied and easily interpreted with balanced data. Thus in the example, confinement to estimable functions limits us to the functions $\alpha_i - \alpha_n + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{n\cdot}$, $\beta_j - \beta_k + \bar{\gamma}_{\cdot j} - \bar{\gamma}_{\cdot k}$, and $\gamma_{ij} - \gamma_{ik} - \gamma_{nj} + \gamma_{hk}$ (and linear combinations thereof). Reparameterization in that same example is commonly achieved by using what are coming to be called the $\sum$-restrictions:

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_i = 0, \quad \sum_{j=1}^{b} \gamma_{ij} = 0 \ \forall\, i \quad \text{and} \quad \sum_{i=1}^{a} \gamma_{ij} = 0 \ \forall\, j. \qquad (6)$$

Since

$$\bar{\gamma}_{i\cdot} = \sum_{j=1}^{b} \gamma_{ij}/b \quad \text{and} \quad \bar{\gamma}_{\cdot j} = \sum_{i=1}^{a} \gamma_{ij}/a$$

the $\sum$-restrictions imply

$$\bar{\gamma}_{i\cdot} = 0 \quad \forall\, i \quad \text{and} \quad \bar{\gamma}_{\cdot j} = 0 \quad \forall\, j \,. \tag{7}$$

This can lead to results for the reparameterized model having a different appearance from those of the unreparameterized model, even when they stem from the same origin. For example, in the unreparameterized model, the BLUE (best linear unbiased estimator) of the estimable function $\alpha_i - \alpha_h + \bar{\gamma}_i - \bar{\gamma}_h$ is

$$\text{BLUE}\,(\alpha_i - \alpha_h + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{h\cdot}) = \bar{y}_{i\cdot\cdot} - \bar{y}_{h\cdot\cdot} \,. \tag{8}$$

But in the reparameterized model with the $\sum$-restrictions (6), and hence (7), this becomes

$$\text{BLUE}\,(\alpha_i - \alpha_h) = \bar{y}_{i\cdot\cdot} - \bar{y}_{h\cdot\cdot} \,. \tag{9}$$

Sometimes it is found helpful to have different notation for (9) in order to emphasize its distinction from (8), namely that in (9) the model includes $\sum$-restrictions whereas (8) does not. One way of achieving this is to write the model as

$$E(y_{ijk}) = \dot{\mu} + \dot{\alpha}_i + \dot{\beta}_j + \dot{\gamma}_{ij}$$

with

$$\sum_{i=1}^{a} \dot{\alpha}_i = 0, \quad \sum_{j=1}^{b} \dot{\beta}_j = 0, \quad \sum_{j=1}^{b} \dot{\gamma}_{ij} = 0 \;\forall\, i, \text{ and } \sum_{i=1}^{a} \dot{\gamma}_{ij} = 0 \;\forall\, j. \tag{10}$$

Then (9) is

$$\text{BLUE}\ (\dot{\alpha}_i - \dot{\alpha}_h) = \bar{y}_{i\cdot\cdot} - \bar{y}_{h\cdot\cdot}\ .$$

With this notation we can clearly state an important difference between the model (4) and the model (10): in (4) the individual parameters μ, α_i, β_j and γ_{ij} do not have BLUEs whereas in (10) their counterparts, $\dot{\mu}, \dot{\alpha}_i, \dot{\beta}_j$ and $\dot{\gamma}_{ij}$ do. One may therefore rightly ask what is the relationship between the two sets of parameters. One description is

$$\dot{\mu} = \mu + \bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot\cdot} \qquad \dot{\alpha}_i = \alpha_i - \bar{\alpha}_{\cdot} + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot\cdot}$$

$$\dot{\gamma}_{ij} = \gamma_{ij} - \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot j} + \bar{\gamma}_{\cdot\cdot} \qquad \dot{\beta}_j = \beta_j - \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot j} - \bar{\gamma}_{\cdot\cdot} \tag{11}$$

That every parameter in (10) has a BLUE means that so does every linear combination of these parameters - again a distinction of (10) from (4).

2.4 Analysis of variance F-tests

For balanced data the use of neither estimable functions nor the $\sum$-restrictions affects the analysis of variance except for description of the hypotheses that are tested by the F-statistics (based on the usual normality assumptions). For example, the F-statistic that has $\sum\sum\sum(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$ as its numerator sum of squares for the model (4) tests the hypothesis

$$H : \text{all}\ (\alpha_i + \bar{\gamma}_{i\cdot}) \text{ equal for } i=1,\ldots,a. \tag{12}$$

In terms of estimable functions this is equivalent to $H : \alpha_1 - \alpha_i + \bar{\gamma}_{1\cdot} - \bar{\gamma}_{i\cdot} = 0$ for $i=2,3,\ldots,a$; and in terms of a reparameterized model using the

$\sum$-restrictions (6) the equivalent hypothesis is, using the distinguishing notation, $H : \dot{\alpha}_1 - \dot{\alpha}_i = 0$ for i=2,3,...,a; and because in this model $\sum_{i=1}^{a} \dot{\alpha}_i = 0$ this is also equivalent to $H : \dot{\alpha}_i = 0$ for i=1,...,a. The distinguishing notation makes it clear that although the hypothesis is $H : \dot{\alpha}_i = 0$ it is not $H : \alpha_i = 0$.

3. UNBALANCED DATA

We here restrict the meaning of unbalanced data to that discussed in the final paragraph of Section 1. It does not include data that have just a few missing observations or that exhibit planned unbalancedness. Then, with this meaning, the difficulties caused by overparameterized models are by no means as easily solved for unbalanced data as they are for balanced data, for which using estimable functions or reparameterization with $\sum$-restrictions provides reasonable alternatives. It is here that the distinction between all-cells-filled data and some-cells-empty data is useful.

3.1 All-cells-filled data

Within the framework of analysis of variance, the most useful sums of squares for all-cells-filled data are those coming from Yates' weighted-squares-of-means analysis. For our rows-by-columns example, the sum of squares due to rows in this analysis is

$$SSA_w = \sum_{i=1}^{a} w_i(\tilde{y}_{i..} - \sum_{i=1}^{a} w_i \tilde{y}_{i..} / \sum_{i=1}^{a} w_i)^2 \qquad (13)$$

for

$$\tilde{y}_{i..} = \sum_{j=1}^{b} \bar{y}_{ij.}/b \text{ and } w_i = \sigma^2/v(\tilde{y}_{i..}) = b^2 / \sum_{j=1}^{b} \frac{1}{n_{ij}} \quad . \qquad (14)$$

With $SSE = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij.})^2$, the hypothesis tested by the F-stati-

stic (n.. - ab)SSA_w/(a-1)SSE is then (12). This is a useful hypothesis.

Insofar as estimation is concerned, since every cell contains data,

$$\text{BLUE } (\mu + \alpha_i + \beta_i + \alpha_{ij}) = \bar{y}_{ij}. \tag{15}$$

is true for every cell. Hence any linear combination of the terms ($\mu + \alpha_i + \beta_j + \gamma_{ij}$) can be estimated and, in particular, because every cell contains data,

$$\text{BLUE } (\alpha_i - \alpha_k + \bar{\gamma}_{i.} - \bar{\gamma}_{h.}) = \tilde{y}_{i..} - \tilde{y}_{h..} \tag{16}$$

for $\tilde{y}_{i..}$ defined in (14). Note that (16) differs from (8) in its use of $\tilde{y}_{i..}$ rather than $\bar{y}_{i..}$; but , of course, (16) reduces to (8) when the data are balanced, for then $\tilde{y}_{i..}$ and $\bar{y}_{i..}$ are the same.

3.2 Some-cells-empty data

We distinguish two cases: models without interactions, and those with interactions.

3.2a Models without interactions. Models having nested fixed effects are not considered; and models with no interactions are then main-effects-only models. Using such models on data that have empty cells demands ascertaining for whatever the main-effects-only model may be, that the data are connected. Unfortunately, the concept of connectedness is not particularly easy, and ascertaining whether data are connected or not is even less so. The essential idea is that connectedness is a property the presence of which ensures that* contrasts among all levels of each factor are estimable. For example, con-

sider Grid 1 where a check mark (✓) indicates the presence of data.

Grid 1

✓	✓	✓		
✓	✓	✓		
			✓	✓
			✓	✓

Expressions $\mu + \alpha_i + \beta_j$ are estimable for each of the filled cells but, for example, because all columns that have filled cells in rows 1 and 2 also have empty cells in rows 3 and 4, it is not possible to estimate any differences between rows 1 and 2 and rows 3 and 4; e.g., $\alpha_2 - \alpha_3$ has no BLUE.

These data are not connected. Further elementary ideas about connectedness can be found in Searle (1971), and for more mathematical considerations the reader is referred to Stewart and Wynn(1981). If suffices to say that most data sets are connected, expecially when the model consists of just a few (main effect) factors.

Given that some-cells-empty data are connected, their analysis using a main-effects-only model then provides (on the assumption of no interactions) useful results. Those results all come from using estimation by least squares and only in simple cases (e.g., the 1-way classification) can they be stated as simple functions of observed means. Otherwise they involve calculations that are best stated in terms of matrices and vectors, for which the reader is referred to such texts as Searle (1971), Rao (1973) and Guttman (1982). Nevertheless, description of the available results is easily given. (1) First, all contrasts among levels of each factor can be estimated, and tested . (2) Second, population cell means, even of empty cells, can be estimated

lity of all levels of each factor can be tested, using an F-statistic.

(4) For a model with f main-effect factors there are f! different ways to partition the total sum of squares. Hence there f! what-might-be-called analysis of variance tables, none of which is of any particular interest. The only F-statistics in those tables (just one from each table) that test useful hypotheses are those alluded to in (3).

3.2b Models with interactions. Consider analyzing data of the nature indicated in Grid 2.

Grid 2

Row 1	✓	✓	✓
Row 2	✓		✓
Row 3		✓	✓

Based on the model (4), the function $\mu + \alpha_i + \beta_i + \gamma_{ij}$ has a BLUE

$$\text{BLUE}\ (\mu + \alpha_i + \beta_j + \gamma_{ij}) = \bar{y}_{ij\cdot} \text{ for } n_{ij} > 0 \ .$$

Now row 1 has data in every column, so that there is a BLUE

$$\text{BLUE}(\mu + \alpha_1 + \tfrac{1}{3}[\beta_1 + \beta_2 + \beta_3] + \tfrac{1}{3}[\gamma_{11} + \gamma_{12} + \gamma_{13}]) = \tfrac{1}{3}(\bar{y}_{11\cdot} + \bar{y}_{12\cdot} + \bar{y}_{13\cdot}) \ .$$

Thus for row 1 there is a BLUE of a function that, along with $\mu + \alpha$, includes the mean of all the βs and γs for every column in row 1. But this latter characteristic cannot occur for row 2; because it has an empty cell in column 2, β_2 and γ_{22} cannot occur in a BLUE:

$$\text{BLUE}\ (\mu + \alpha_2 + \tfrac{1}{2}[\beta_1 + \beta_3] + \tfrac{1}{2}[\gamma_{21} + \gamma_{23}]) = \tfrac{1}{2}(\bar{y}_{21\cdot} + \bar{y}_{23\cdot}) \ .$$

2 is to consider a BLUE that involves $\mu + \alpha_1$ and just the βs of columns 1 and 3, the columns that contain data in row 2:

$$\text{BLUE } (\mu + \alpha_1 + \frac{1}{2}[\beta_1 + \beta_3] + \frac{1}{2}(\gamma_{11} + \gamma_{13}]) = \frac{1}{2}(\bar{y}_{11\cdot} + \bar{y}_{13\cdot}) \ .$$

Hence

$$\text{BLUE } (\alpha_1 - \alpha_2 + \frac{1}{2}[\gamma_{11} + \gamma_{13}] - \frac{1}{2}[\gamma_{21} + \gamma_{23}]) = \frac{1}{2}(\bar{y}_{11\cdot} + \bar{y}_{13\cdot}) - \frac{1}{2}(\bar{y}_{21\cdot} + \bar{y}_{23\cdot}) \ . \qquad (1$$

Thus $\alpha_1 - \alpha_2$ can be estimated in the presence of a difference between average interactions that are in columns 1 and 3 - because these are the only columns wherein there are data in both rows 1 and 2. For similar reasons $\alpha_1 - \alpha_3$ can be estimated only over a different set of columns, columns 2 and 3. This arises solely from the pattern of empty cells. It means, in this example, that no comparison between rows can be made over all columns. Thus no contrast among row effects can be made that involves interaction effects averaged over all columns, i.e. there is no BLUE of $\alpha_i - \alpha_h + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{h\cdot}$ as there is with balanced data and with unbalanced all-cells-filled data.

This inability to make all comparisons between rows across the same set of columns resulting from there being empty cells, also makes it difficult to test hypotheses about row effects (and column effects) - indeed about main effects in general when there are interactions in the model and empty cells in the data. In fact, there is then no test that row effects are all equal, i.e., there is no test of H : α_i all equal as in (12) for balanced data, nor even of H : $\alpha_i + \bar{\gamma}_{i\cdot}$ all equal, as tested using SSA_w of (13) for all-cells-filled data. The big question, therefore, is what does one do with some-cells-empty data when wanting to be concerned about interactions? The answer is "use the cell means model".

4. CELL MEANS MODELS

Rather than use the overparameterized model (4) when dealing with some-cells-empty data, it is conceptually much easier to just think of the data in each filled cell as being a random sample from a population peculiar to that cell. Thus for cell i,j in which $n_{ij} > 0$ the data $y_{ij1},\ldots y_{ijn_{ij}}$ are taken as a random sample of size n_{ij} from a population having mean μ_{ij}. Then the cell means model is

$$E(y_{ijk}) = \mu_{ij} \quad \text{for} \quad n_{ij} \neq 0 \ . \tag{18}$$

Promotion of this kind of model since the late 1960's began with Speed (1969), and has continued in papers by Speed and co-workers (e.g., Speed et al 1976, 1978).

Although on comparing (18) with (4) it is obvious that μ_{ij} and $\mu + \alpha_i + \beta_j + \gamma_{ij}$ are equivalent, it is far easier with some-cells-empty data to concentrate attention on the parameters μ_{ij} of (18) than it is to deal with the μ, α_i, β_j and γ_{ij} terms of (4). It is not only easier through simply making estimation very easy (as it does) but it is also in keeping with a more natural way of looking at data and of providing us with a more straightforward way of seeking interpretation of data than does (4) in the presence of empty cells.

The first thing to notice about (18) is that there are exactly the same number of parameters to be estimated as there are observed cell means to estimate them from. Only for each fitted cell do we have a μ_{ij} in the model for the data, and corresponding thereto is an observed cell mean $\bar{y}_{ij.}$. Insofar as the data gatherer (whom we shall call an experimenter) is concerned, this

conceptof population cell means, especially compared to the overparamterized model, is right in keeping with one's manner of thinking about some-cells-empty data. For example, with data of Grid 2, no experimenter seeing the pattern of empty cells there would envisage trying to compare rows 1 and 2 over all three columns; nor rows 1 and 3 either. And clearly the only comparison possible between rows 2 and 3 can come from data in column 3. Any good experimenter understands this, simply by scrutinizing the occurrence of empty cells, or more particularly the pattern of occurrence of the filled cells. Statistical consultants who are asked to deal with such data should therefore go along with this understanding; the alternative is to force the experimenter into the forest of overparameterized models with all their mathematical entwinements of estimability, useless F-statistics, reparameterizations and restrictions. Cell means models have none of these. Furthermore, they are directly in line with an experimenter's way of thinking about some-cells-empty data. In addition, the statistician is able , by means of the cell means model, to give the experimenter help and advice that is easy to carry out and which is readily understood, something which cannot be said for the overparameterized model.

4.1 Estimation

We have seen that the definition of μ_{ij} is straightforward: it is the population mean for cell i,j. And for every cell i,j that contains data, estimation of its μ_{ij} is easy:

$$\text{BLUE}\ (\mu_{ij}) = \bar{y}_{ij}. \quad \text{for } n_{ij} > 0; \tag{19}$$

i.e., a population cell mean, when that cell contains data, is estimated by the observed cell mean for that cell. Nothing could be easier.

Moreover, for every filled cell the μ_{ij} is estimated in this manner; and every linear combination of those μ_{ij}s is estimated by the same linear combination of the corresponding $\bar{y}_{ij}.$s . For example, for Grid 1,

$$\text{BLUE } (\mu_{11} + 3\mu_{12} - 5\mu_{13}) = \bar{y}_{11}. + 3\bar{y}_{12}. - 5\bar{y}_{13}. .$$

In general, for any values λ_{ij},

$$\text{BLUE } (\sum_{i,j} \lambda_{ij}\mu_{ij}) = \sum_{i,j} \lambda_{ij}\bar{y}_{ij}. \quad \text{for} \quad n_{ij} > 0 . \tag{20}$$

4.2 Hypothesis testing

The model for y_{ijk} based on (19) is

$$y_{ijk} = \mu_{ij} + e_{ijk}. \tag{21}$$

with e_{ijk} defined as $y_{ijk} - E(y_{ijk})$ just as in Section 1. On attributing the same variance-covariance properties to the e_{ijk} as there, namely that every e_{ijk} has variance σ^2 and all covariances are zero, the sampling variance of $\bar{y}_{ij}.$ is $v(\bar{y}_{ij}.) = \sigma^2/n_{ij}$ for $n_{ij} > 0$. Furthermore, the $\bar{y}_{ij}.$s are distributed independently of each other so that for (20)

$$v(\sum_{i,j} \lambda_{ij}\bar{y}_{ij}.) = \sigma^2 \sum_{i,j} \lambda_{ij}^2/n_{ij} \quad \text{for} \quad n_{ij} > 0 . \tag{22}$$

The variance σ^2 is estimated, as usual, by the pooled within-cell mean square

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij}.)^2}{n.. - s} \tag{23}$$

where s is the number of filled cells.

Testing linear hypotheses is done on the basis of customary normality assumptions. Thus the t-statistic (on $n.. - s$ degrees of freedom) for testing the simple linear hypothesis

$$H : \sum_{ij} \lambda_{ij}\mu_{ij} = m \quad \text{is} \quad t = \frac{\sum_{ij} \lambda_{ij}\bar{y}_{ij\cdot} - m}{(\hat{\sigma}^2 \sum \lambda_{ij}^2/n_{ij})^{\frac{1}{2}}} . \tag{24}$$

Testing a composite linear hypothesis is best described in terms of a few simple matrices and vectors, defined as follows:

$\underset{\sim}{\mu}$ = vector of μ_{ij}s for $n_{ij} > 0$, in lexicon order.

$\underset{\sim}{\bar{y}}$ = vector of $\bar{y}_{ij\cdot}$s for $n_{ij} > 0$, in lexicon order.

$\underset{\sim}{D}$ = diagonal matrix of terms $1/n_{ij}$ for $n_{ij} > 0$, in lexicon order.

$H: \underset{\sim}{K}' \underset{\sim}{\mu} = \underset{\sim}{m}$ is the hypothesis to be tested, where $\underset{\sim}{K}'$ has full row rank $r \leq s$.

Then the F-statistic on r and $n.. - s$ degrees of freedom for testing

$$H : \underset{\sim}{K}'\underset{\sim}{\mu} = \underset{\sim}{m} \quad \text{is} \quad F = (\underset{\sim}{K}'\underset{\sim}{\bar{y}} - \underset{\sim}{m})'(\underset{\sim}{K}'\underset{\sim}{D}\underset{\sim}{K})^{-1}(\underset{\sim}{K}'\underset{\sim}{\bar{y}} - \underset{\sim}{m})/r\hat{\sigma}^2 . \tag{25}$$

Essentially this is all there is to estimation and hypothesis testing in the cell means model. (Complexities such as restricted models and mixed models are mentioned in Section 5.) Estimation of means and linear combinations of them is given in (19) and (20), and of σ^2 in (23) and hypothesis testing is set out in (25), with (24) being a special case. Thus the mechanics of cell means models are patently simple. The problem is how, from this simplicity, does one use a cell means model to answer questions of interest to the experimenter?

4.3 Analysis of variance

First observe that there is no really useful analysis of variance table. The only table available from the model (21) would have but two sums of squares: $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij\cdot} - \bar{y}_{\cdots})^2$ and $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\cdot})^2$, and its only purpose would be as a summary of the arithmetic for testing the hypothesis H : all μ_{ij} (for $n_{ij} > 0$) equal. And clearly this is not a hypothesis of very much interest; in fact, in most cases it is of absolutely no interest at all. There is therefore little point in calculating an analysis of variance table for a cell means model.

4.4 Hypotheses of interest

Instead, on the basis of an experiment's (presumably expert) knowledge of the data at hand, the consulting statistician and the experimenter, working together, formulate what they think are interesting linear combinations of the cell means of the cells that contain data. These combinations can then be estimated using (20), the sampling variance of each estimate can be estimated from (22) and (23), and tests of hypotheses can be made about them by using (25). At the heart of this process is the experimenter's knowledge of the data; in the presence of empty cells the person whose data are being analzyed must contribute knowledge to deciding what combinations of means (of filled cells) are of interest. No longer can the automatic hypotheses like "equality of rows" be tested; considered thought must be given, under the spotlight of having empty cells, as to what combination of the filled cells are interesting. Knowledge of the data and the pattern of filled cells both have to be utilized. What one would like to consider, if all cells were fitted, has to be tempered by what can be considered in the light of certain cells being empty. Since the pattern of empty cells usually differs from one data set to another, the linear combinations of cell means considered

in one data set will not necessarily be the same as those in another. Yet in each case they must be combinations that are of interest, and they must also be combinations of filled cells. Different data sets from similar studies might therefore (probably will) have different analyses - depending on which cells in the data grid are empty in each case.

For some examples, consider Grid 2. Although row effects, such as α_i in the overparameterized model (4), are not specifically part of a cell means model, row means can be defined in terms of cell means (as linear combinations of them) and hence they can be studied. Thus for Grid 2

$$\bar{\mu}_{1\cdot} = (\mu_{11} + \mu_{12} + \mu_{13})/3$$

represents a row mean, with estimator

$$\text{BLUE } (\bar{\mu}_{1\cdot}) = (\bar{y}_{11\cdot} + \bar{y}_{12\cdot} + \bar{y}_{13\cdot})/3 \ .$$

But for comparing rows 1 and 2, $\bar{\mu}_{1\cdot}$ is not suitable because row 2 has no data in column 2. This comparison is confined to columns 1 and 3, which contain data in rows 1 and 2. We therefore consider $\frac{1}{2}(\mu_{11} + \mu_{13}) - \frac{1}{2}(\mu_{21} + \mu_{23})$ with

$$\text{BLUE } [\tfrac{1}{2}(\mu_{11} + \mu_{13}) - \tfrac{1}{2}(\mu_{21} + \mu_{23})] = \tfrac{1}{2}(\bar{y}_{11\cdot} + \bar{y}_{13\cdot}) - \tfrac{1}{2}(\bar{y}_{21\cdot} + \bar{y}_{23\cdot}) \ .$$

This is, of course, the same estimator as (17), but deriving it and understanding it is much easier in terms of cell means μ_{ij} than in terms of the too many parameters of the overparameterized model. This gain in understanding may not be as evident in this small example as it would be in a case of several rows and columns and numerous empty cells.

With Grid 2, every pair of rows has to be compared across a different set of columns. Nevertheless, it is precisely by scrutinizing a pattern of empty cells with a view to comparing means of those cells, that we find it possible to make comparisons that may be of interest. Further illustration of this in Grid 2 is that the interactions $\mu_{11} - \mu_{13} - \mu_{21} + \mu_{23}$ and $\mu_{12} - \mu_{13} - \mu_{32} + \mu_{33}$ can be estimated; and a test of the hypothesis that they are zero can be derived from (25).

4.5 Subset analyses

Analyses of the nature just described (which we call subset analyses) are certainly not as informative overall as either the analysis of variance of balanced data, or the weighted squares of means analysis of unbalanced but all-cells-filled data. But for unbalanced and some-cells-empty data they are much more useful than the analyses of variance of unbalanced data that come from fitting an overparameterized model (4) and its submodels - see, for example, Searle (1971, Chapter 7). Furthermore, these subset analyses for some-cells-empty data are vastly easier to understand, to interpret and to explain to decision-makers than are the analyses using an overparameterized model with interactions. Not only are the latter difficult to interpret, but in most cases of some-cells-empty data they are of no real interest.

4.5a Models with interactions. The analysis of models without interactions (i.e., main-effects-only models) is easy, as has already been described in Section 3.2a. No matter how many factors one has, the use of a main-effects-only model provides, for each factor, a test of equality of the effects of its different levels on the response variable. But such a model implies no interactions and this begs the question "when are we ever in a situation where we know there are no interactions?". The answer is probably "never".

However, in even the simplest case of only two factors, the use of a with-interaction overparameterized model for some-cells-empty data yields very little (if anything) that is either generally useful or readily interpretable. But the desire and/or need to investigate the occurrence of interactions is often very strong, and whereas the overparameterized model is of little help in this connection when dealing with some-cells-empty data, the use of a cell means model can be very helpful indeed. This is achieved by looking at the data from the cell means model perspective and scrutinizing the data grid to see which subsets of filled cells suggest themselves as possibilities for subset analyses that might yield information about interactions. The information so obtained may not be as far-reaching as when all cells are filled, but it will be better than nothing (which is only what a main-effects-only model can yield, insofar as interactions are concerned) and it will nearly always be better than an overparameterized analysis, because that analysis can, in the face of empty cells, be very difficult to interpret.

The crux of the procedure when wishing to consider interactions with some-cells-empty data is therefore to view the data from the perspective of a cell means model and to seek subsets of data that can yield information about interactions. Then, as consultant to a client who insists (as do some clients) on considering interactions when data have empty cells, the statistician, instead of having to complicate (at least from a client's, presumably non-mathematical viewpoint) the analysis of such data by introducing ideas of estimability and/or restricted models, can offer clarification in the form of helping the client decide which subsets of data might provide analyses of interest.

4.5b Examples. Suppose data occurred as indicated in Grid 3.

Grid 3

	1	2	3	4
I	✓	✓		✓
II	✓	✓		✓
III		✓	✓	✓

An overparameterized analysis would yield a sum of squares for interaction having three degrees of freedom. But in trying to ascertain which interactions appear to be significant, and to understand their occurence in the data with more depth than that sum of squares provides, scrutinizing the pattern of filled cells in the data reveals subsets of the data that are easier to interpret; e.g., for Grid 3, the following subsets:

Grid 3a

	1	2	4
I	✓	✓	✓
II	✓	✓	✓

Grid 3b

	2	4
I	✓	✓
II	✓	✓
III	✓	✓

True, analyses of Grids 3a and 3b are not independent of one another; but the analyses are simple and interpretation of each is straightforward.

The preceding example is so simple that it may fail to emphasize just how difficult the search for informative subsets of data can be. But consider Grid 4, of six rows and eight columns, with 19 of the 48 cells containing data.

Grid 4

	1	2	3	4	5	6	7	8
I	✓			✓		✓	✓	
II				✓	✓			✓
III				✓				✓
IV	✓	✓		✓		✓	✓	✓
V								✓
VI			✓	✓				✓

As an example, it illustrates how we can lead ourselves, through "feeling our way", towards analyses that may provide more useful interpretation than does analyzing the full data set "warts and all", in this case the warts being the large number of empty cells: 60 % are empty. In this regard the very grid itself is useful, because it provides opportunity to scrutinize just which cells have data and whether or not any of them form subsets of the data that may be open to straightforward analysis. Such scrutiny reveals that columns 2, 3 and 5, and row V each have but a single filled cell. Setting these data aside leaves Grid 4a:

Grid 4a

	1	4	6	7	8
I	✓	✓	✓	✓	
II		✓			✓
III		✓			✓
IV	✓	✓	✓	✓	✓
VI		✓			✓

This, we can easily see, falls into two subsets of data: rows I and IV in columns 1, 4, 6 and 7, and rows II, III, IV and VI in columns 4 and 8. These subsets have but one cell in common and account for all six of the degrees of freedom for interaction available in the analysis of the full data set. But now, in directing attention to these two subsets, we are able to understand, very straightforwardly, what interaction effects are being considered.

4.5c Difficulties. Two difficulties with subset analyses are readily envisioned. One is that a data set may not always yield subsets that are useful in the way that those of the preceding examples appear to be. For example, consider Grid 5:

Grid 5

	1	2	3	4
1	✓			✓
2		✓		✓
3	✓	✓	✓	

In analyzing the complete data set using an overparameterized model, there will be a 1-degree-of-freedom sum of squares for interaction that will be testing the hypothesis stated, as follows, in three equivalent ways:

$$H : \mu_{11} - \mu_{14} - \mu_{21} + \mu_{24} + (\mu_{21} - \mu_{22} - \mu_{31} + \mu_{32}) = 0$$

$$H : \mu_{11} - \mu_{14} - \mu_{31} + \mu_{34} - (\mu_{22} - \mu_{24} - \mu_{32} + \mu_{34}) = 0$$

$$H : \mu_{11} + \mu_{24} + \mu_{32} - (\mu_{14} + \mu_{22} + \mu_{31}) = 0 .$$

The first of these three statements involves the sum of two interactions whereas the second involves the difference between two such interactions. Whatever the utility of these may be (if any), scrutiny of Grid 5 reveals that no subsets of the data manifest themselves as being candidates for in-

formative analyses. When this kind of situation occurs the statistician can do little more than persuade the experimenter that this is so, and fall back on the main-effects-only model. Of course with data sets larger than that of Grid 5, coming to the conclusion of no useful subsets may not be as easy as it is with Grid 5, and much resourcefulness might be needed before such a conclusion can be firmly established.

A second and obvious difficulty inherent in subset analyses is that a data set might well be divisible into two or more different subset analyses. For example, Grid 6 can easily be divided into two subsets in two different ways: one way consists of rows I and II, and row III; and the other is columns 1 and 4, and columns 2 and 3.

Grid 6

	1	2	3	4
I	✓	✓	✓	✓
II	✓	✓	✓	✓
III	✓			✓

This situation emphasizes what is so important about analyzing unbalanced data, especially some-cells-empty data: there is seldom just a single, correct way of doing a statistical analysis. Therefore the first responsibility of a consulting statistician to those who have garnered such data is to impress upon them that analyzing those data has no single, easy, umbrella of interpretation. Within that umbra, the statistician can certainly provide advice as to what analyses might be helpful, and two different statisticians might well have two different lines of advice for analyzing the same data set. As with lawyers, the advice of statisticians is not necessarily uniform, let alone uniformly right or uniformly wrong.

4.5d The data gatherer. Naturally, the person whose data are to be analyzed (the data gatherer) must contribute to deciding on possible divisibility of the total data set into subset analyses. In the preceding examples the pattern of empty cells has been the sole criterion for suggesting subsets. That is always a useful criterion because it can guide us to analyses that are interpretable. But it must not be the sole criterion; data gatherers must be urged, from their prior knowledge of similar data and of the context from which the present data have come, to decide what specific levels of the factors (or pooled combinations thereof) are of prime interest, especially in the context of interactions. Indeed, as an alternative to the statistician's advice being in terms of estimability and estimable functions (as it must for overparameterized models), it seems that that advice should be in terms of helping data gatherers, nay even cajoling them, perhaps, into deciding what specific subset of filled cells might be of real interest. One would hope that when faced with empty cells, the combined efforts of statistician and data gatherer would usually reveal some data subsets that provide both easy analysis and straightforward interpretation through the use of the cell means model.

5. EXTENSIONS

5.1 More than two factors

The examples we have used are solely in terms of the 2-way classification, rows and columns. But the cell means model extends very easily and directly to more than two factors: the population mean of a cell defined by one level of each factor is, when that cell contains data, estimated by the mean of those data in the cell. This, for the 2-way classification is precisely the result (19). Its formal extension to more than two factors simply involves having three or more subscripts in place of i,j in y_{ijk} and μ_{ij}. And its

conceptual extension to means of sub-most cells of the data classification is, as already stated, very easy.

Difficulties of overparameterized models for the 2-way classification are simply aggravated when we come to deal with 3-, 4- and more-factor models with empty cells in the data. Although the desire for including interactions may be strong, one must never forget that a main-effects-only model provides straightforward tests about levels of each main effect - as has already been mentioned. Moreover, an alarming feature of interactions is that in a multi-factor situation they can be totally overwhelming, simply by virtue of the sheer number of possible interactions. The example in Searle (1971, Section 8.1) illustrates this. It is taken from the Bureau of Labor Statistics Survey of Consumer Expenditures in which family investment patterns are classified by a total of 56 levels of nine different factors: the numbers of levels of the factors are 12, 11, 4, 3, 4, 6, 6, 6 and 4. This 9-way classification has 5,474,304 possible cells, with 1,354 possible 2-factor interactions and 18,538 possible 3-factor interactions. Yet there were only 8,577 observations in the study! Clearly, then, any desire to "study interactions" in large data sets has to be tempered by the paucity of data relative to the number of possible interactions. Patterns of filled cells and the researcher's knowledge of the data have to be used with great perspicacity to hopefully elucidate subset analyses that provide information about interactions.

5.2 Models without interactions

Cell means models by their very nature implicitly include interactions in the population means of the sub-most cells. In the 2-way classification, the equivalence of μ_{ij} of the cell means model (18) to $\mu + \alpha_i + \beta_j + \gamma_{ij}$ of the overparameterized model (4) has already been noted; and in models

for more than two factors the cell means model includes implicitly all interactions of all orders. Contrasted to this are the main-effects-only models that have no interactions. Sometimes a middle ground may be suitable with some but not all interactions wanted in a model. This can be done with a cell means model by defining in terms of the population cell means those interactions which are not wanted, and then excluding them from the model by having, as part of the model, restrictions that put those interactions equal to zero. For example, in the case of just two rows and two columns the no-interaction form of the cell means model is

$$E(y_{ijk}) = \mu_{ij} \quad \text{for} \quad i,j = 1,2 \quad \text{and} \quad \mu_{11} - \mu_{12} - \mu_{21} + \mu_{11} = 0 \ . \qquad (26)$$

The standard procedure of restricted least squares can then be used to estimate the μ_{ij}s, subject to the no-interaction restriction $\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$. And, of course, this use of a restricted model such as (26) generalizes very directly to more complicated cases.

In a case like (26), where all interactions are to be excluded, the model is of course equivalent to a main-effects-only model. This is also true when, in cases of more than two factors, all interactions of all orders are to be excluded; but when in those many-factor cases only some interactions are to be excluded, then the estimability of linear combinations of cell means is affected both by the pattern of empty cells and by which interactions are excluded. Searle and Speed (1982) have a theorem that gives details of this situation.

5.3 Mixed models

The preceding discussion has been based on assuming that observations all

have the same variance, σ^2, and that covariances between all pairs of observations are zero. This is conventional for fixed effects models. In contrast, a mixed model consists of both fixed effects and random effects. So far as estimation of fixed effects is concerned, the consequence of also having random effects in the model is that they contribute structure to the pattern of variances and covariances among the observations. Using this structure, estimation of the fixed effects can nevertheless be dealt with by generalized least squares (equivalent to BLU estimation). Thus can cell means models be extended to mixed models: population cell means are defined for the submost cells defined by the levels of the fixed effects factors. Those cell means are then estimated by generalized least squares. Searle (1984a,b) shows details, including two useful results: (i) in using mixed models with balanced data, the BLUE of any contrast is the same as the ordinary least squares estimator. (ii) analytic expressions have been derived for the BLUE of treatment means in randomized complete blocks having unequal numbers of observations on the treatments in the blocks. These expressions simplify for balanced incomplete blocks to be equivalent to the results of Scheffé (1959).

REFERENCES

Guttman, I. (1982) Linear Models: An Introduction, Wiley, New York.

Meredith, M.P. and Cady, F.P. (1984) Methodology analysis of treatment means from factorial experiments with unequal replication. Technical Report BU-859-M, Biometrics Unit, Cornell University.

Rao, C.R. (1973) Linear Statistical Inference and its Applications, 2nd Edition, Wiley, New York.

Scheffé, H. (1959) The Analysis of Variance, Wiley, New York.

Searle, S.R. (1971) Linear Models, Wiley, New York.

Searle, S.R. (1984a) Cell means formulation of mixed models in the analysis of variance, Technical Report BU-862-M, Biometrics Unit, Cornell University.

Searle, S.R. (1984b) Best linear unbiased estimation in mixed models of the analysis of variance, Technical Report BU-864-M, Biometrics Unit, Cornell University.

Searle, S.R. and Speed, F.M. (1982) Estimability in the cell means general linear model, Technical Report BU-730-M, Biometrics Unit, Cornell University.

Speed, F.M. (1969) A new approach to the analysis of linear models. Ph. D. Thesis, Texas A. and M. University, College Station, Texas.

Speed, F.M. and Hocking, R.R. (1976) The use of the R() notation with unbalanced data, The American Statistician 30,30-34.

Speed, F.M., Hocking, R.R., and Hackney, O.P. (1973) Methods of analysis of linear models with unbalanced data, J. American Statistical Association 73,105-112.

Steel, R.G.D. and Torrie, J.H. (1980) Principles and Procedures of Statistics: A Biometrical Approach, 2nd Edition, McGraw-Hill, New York.

Stewart, I. and Wynn, H.P. (1981) The estimability structure of linear models and submodels, J. Roy. Stat. Soc. (B) 43,197-207.

Urquhart, N.S. and Weeks, D.L. (1978) Linear models in messy data: some problems and alternatives, Biometrics 34,696-705.