BU-307-M 11/71

ABSTRACT

ESTIMATION ASSOCIATED WITH LINEAR MODELS: A REVISITATION*

by

N. S. Urquhart, D. L. Weeks, and C. R. Henderson†

The association of linear models with the analysis of complex sets of data dates back to Gauss (about 1800). But linear models assumed a major role in statistics only after Fisher's colleagues introduced them in explaining the analysis of variance. Since then it has become a common practice to describe experimental situations by associated linear models. The emergence of the concept of estimability and its associated pedagogical difficulties accompanied this practice. This paper reconsiders the definition of a linear model with special reference to its association with the experimental context. The parameters of the resulting linear model all are estimated simply and without definitional ambiguity. These ideas are illustrated by considering the analysis of unbalanced cross classifications, a situation in which the definitional ambiguities of the usual linear models pose serious problems. Finally, the proposed model is compared to the usual less than full rank model.

* Presented to the session on Linear Models at the Joint Statistical Meetings in New York City, August 19, 1969.

† The authors are, respectively, Associate Professor of Biological Statistics, Biometrics Unit, Cornell University; Professor of Statistics, Statistical Unit, Department of Mathematics and Statistics, Oklahoma State University; Professor of Animal Breeding, Department of Animal Science, Cornell University. Paper No. BU-195 in the Biometrics Unit mimeograph series, Cornell University, Ithaca, New York 14850.

Also BU-307-M

ESTIMATION ASSOCIATED WITH LINEAR MODELS: A REVISITATION

by

N. S. Urquhart, D. L. Weeks, and C. R. Henderson[†]

We feel that substantial confusion surrounds the topic of estimation associated with linear models. Many contemporary presentations of linear models foster this confusion in students and others through statements such as: "the sum of the effects is zero", "the sum of the effects is assumed to be zero", "that function is estimable", "this function is not estimable", "putting these restraints on the solution of the normal equations", "putting these restrictions on the parameters of the model", etc. We believe that the resulting confusions can be eliminated if we keep in mind a few basic fundamental assumptions about how a linear model relates to its associated experimental context.

Thus we briefly review here the historical development of estimation associated with linear models and present an approach which eliminates much of the confusion now associated with this topic. The 'revisitation' part of the title of this paper reflects our view that the approach set out here has implicitly existed for some time, but that its relevance has not been appreciated. Our approach seems especially relevant in teaching about linear models and in thinking about their application to the 'messy data' problem.

We will assume that an experimenter has sampled m different populations for the purpose of studying relationships among the means of these populations. We acknowledge that some experiments cannot be properly treated from this point of

[†] The authors are, respectively, Associate Professor of Biological Statistics Biometrics Unit, Cornell University; Professor of Statistics, Statistical Unit, Department of Mathematics and Statistics, Oklahoma State University; Professor of Animal Breeding, Department of Animal Science, Cornell University. Paper No. BU-195 in the Biometrics Unit mimeograph series, Cornell University, Ithaca, New York 14850.

view, but this approach permits one to gain insight into many of the problems with which linear models now are associated. These m means, $\underset{\sim}{\mu}$, constitute a (vector-valued) parameter for the problem in the sense of Fraser [1957], namely, a function from the class of probability measures under consideration to m-dimensional Euclidean (parameter) space.

The parameterization of a given problem is, unfortunately, not unique. Suppose $\underset{\sim}{\theta}$, an s-component vector $(s > m)$, stands as a candidate to replace $\underset{\sim}{\mu}$. In the context of linear models this means that $\underset{\sim}{\mu} = \underset{\sim}{A}\underset{\sim}{\theta}$. The fundamental point in the subsequent discussion is that a parameterization in terms of $\underset{\sim}{\mu}$ will support the stated intent of examining relationships among population means. Parameterization in terms of $\underset{\sim}{\theta}$ leads to confusion as to the meaning of its elements. Even worse, it leads experimenters to abdicate the responsibility which belongs uniquely to them, namely the association of specific functions of $\underset{\sim}{\mu}$ with experimental interpretations.

Our principal motivation for formulating this approach comes from the analysis and interpretation of large sets of unbalanced data with numerous missing cells. Our examples are similarly motivated, but necessarily smaller than most actual messy data problems.

1. SOME RELEVANT HISTORY*

Statisticians tend to associate the birth of Statistics with the names of Francis Galton, Karl Pearson or R. A. Fisher, i.e. sometime between 1880 and 1920, depending upon their bias. These men certainly influenced Statistics greatly, but the estimation of parameters from data far predates this period. Astronomers and land surveyors wrote extensively upon this subject in the 1800's

* The early part of this history rests heavily upon Eisenhart's [1964] and Merriman's [1877] interpretation of 18th and 19th century writings to which we did not have access.

although it certainly concerned earlier scientists. For example, Eisenhart [1964] indicates that arithmetic means were in use by the late 1500's and that Cotes [1722] advocated the use of weighted arithmetic means. Simpson [1755] used probability and a germ of the idea of sampling from a population to argue that the mean of several observations provided a better estimate of a parameter than a single similar observation.

The concept of 'least squares' emerged about 1800 in the course of continuing attempts to get 'best' values for parameters. Although Gauss used least squares from 1795 onward and contributed heavily to its development [1809 and 1821], Legendre [1805] preceded him in publication. Least squares attracted immediate and sustained interest, a fact attested to by the size of Merriman's [1877] bibliography of writings about least squares between 1805 and 1875 (354 titles). The writers view this interest in least squares as significant because the simple forms of least squares cannot exist without at least the implicit existence of a linear model; actually they appear frequently in these early writings. Thus linear models have existed since about 1800, but the lack of the currently fashionable concept of estimability caused no pain at all. The early researchers framed their problems in terms of functionally independent parameters and got at least as many observations as there were parameters to be estimated; neither practice appeared to concern them. Even the people who might have had difficulty with estimability and associated problems, the land surveyors whose triangles had to close, avoided it by imposing, on the parameters of their models, the restriction present in the real world problem.

The interest in least squares has continued right up to the present, but for our purposes R. A. Fisher's introduction of the analysis of variance constitutes the next significant event. His writings reflect a maturing view of the analysis of variance between the time he coined this term and partitioned genetic variation [1918] and his publication of THE DESIGN OF EXPERIMENTS [1935]. Initially he associated it closely with intraclass correlation, perhaps because many of his contemporaries were familiar with correlation. His introduction to the analysis of variance in STATISTICAL METHODS [1925] supports this view. Eventually his interest in factorial sets of treatments led him to consider the analysis of variance as the partitioning of sums of squares into additive components each of which relates to a specific facet of an experiment, the view reflected when he introduced the analysis of variance in THE DESIGN OF EXPERIMENTS by discussing a randomized blocks field experiment. By current practices, however, one aspect of Fisher's treatment of the analysis of variance is surprising. He did not use linear models to explain the analyses of variance of designed experiments even though his writings on regression and correlation (both simple and multiple) lean toward linear models.

Fisher's colleagues at Rothamsted introduced the association of linear models with the analyses of variance of designed experiments in the early 1930's during their attempts to codify Fisher's ideas and explain what the analysis of variance was doing. This progressed through several stages. Allan and Wishart [1930], writing on estimating the yield of a missing plot, supplied the first stage by formalizing the additive contributions of several components to the true cell mean when they wrote, "Thus $Y = b_p + t_q$, where Y is the supposed true deviation from the mean yield of the plot having treatment q in block p."

Irwin [1931] supplied the next step by introducing the 'error term' and being rather precise about the underlying population structure. For example, he considered sample values x_{uv} from a two-way structure of populations having means m_{uv} and wrote (p. 293)

$$x_{uv} = m_{uv} + \xi_{uv}$$

where ξ_{uv} has expectation zero and variance σ^2. From this, he proceeded to evaluate the expected values of the analysis of variance sums of squares in terms of σ^2 and the m_{uv}. Subsequently (p. 295) he did write

$$x_{uv} = t_u + b_v + \xi_{uv} ,$$

but only in order to show that this sort of assumed structure eliminates the term $\Sigma(m_{uv} - m_{u\cdot} - m_{\cdot v} + m)^2$ from the expected value of the residual sum of squares. He carefully avoided estimating t_u or b_v by instead estimating $t_u + b_v$.

Yates [1933 and 1934] supplied the third major step in the early clarification of the structure of the analysis of variance. In both of these papers Yates considered nonorthogonal designs. In the first paper he introduced orthogonality and applied his ideas to the analysis of two moderately complex designed experiments. The second paper treated the analysis of multiple classifications with unequal numbers directly. In both of these papers he relied upon linear additive models to associate observations with parameters, but he was careful to relate these parameters to the associated population structure. Thus the parameters of his linear models satisfied certain restrictions which he used, without any apparent reservations, to assure uniqueness of parameter estimates. The following quote from his discussion of the no interaction case in the two-way cross classification appears to be typical:

> "It may happen, however, that the phenomena we are investigating are such that the A and B effects are additive, so that the hypothetical sub-class means are of the form
>
> $$\mu + \alpha_r + \beta_s; \quad r = 1, 2, \cdots, p; \quad s = 1, 2, \cdots, q;$$
>
> where μ may be called the hypothetical general mean and α's and β's the hypothetical deviations due to the treatments, these being subject to the relations
>
> $$\alpha_1 + \alpha_2 + \cdots + \alpha_p = 0$$
> $$\beta_1 + \beta_2 + \cdots + \beta_q = 0 ."$$

By 1950 linear models had become so closely identified with the analysis of variance of designed experiments and related problems that many discussions, both applied and theoretical, began with a linear model rather than with the associated experimental or sampling situation. The evolution to this point certainly was not uniform. For example, Rao [1945] credits Bose with discussing estimability in 1943, a concept which is unnecessary in the framework of Yates' original linear models. Thus less completely specified linear models were being considered by this time. Likewise Tukey [1949] exhibited an interest in discarding the zero-sum conditions present in the Yates viewpoint and Cornfield and Tukey [1956] reiterated this position by saying " ··· and a desire to treat contributions [of components of linear models] more as things with independent existence rather than as differences between certain averages." Yet as late as 1947, Eisenhart was still very careful to speak in essentially the same terms as Yates. Today, the situation is similar to what it was in the early 1950's even though there have been several efforts to be specific about what is meant by a linear model.

We can summarize this history and our intent as follows: (1) Least squares-type estimation has been around since about 1800; (2) Fisher introduced the analysis of variance and associated estimation about 1920; (3) linear models were introduced to explain what was going on in (2); (4) it has become common in many quarters to regard the linear model, rather than the experimental setting, as basic. We want to examine estimation in the simplest possible linear model, namely $Y = E(Y) + [Y - E(Y)] = \mu + \epsilon$, which is scarcely a linear model in the contemporary sense. This approach clarifies what common techniques are really estimating without the necessity of introducing estimability or imposing restrictions. These considerations seem especially relevant to the analysis of large sets of unbalanced data such as, for example, are encountered in the animal sciences and to the teaching of statistics at all levels. The authors and several colleagues have used the approach set out here in classes ranging from methodological courses to research topics seminars. This viewpoint has been well received by students familiar with as well as oblivious of the usual approach.

2. GENERAL FORMULATION

Here we formulate a rather general linear model. There does exist diversity of opinion as to what constitutes a linear model; witness, for example, the contrast between Graybill [1961, chapter 5] and Scheffé [1959, sections 1.2, 7.2, 8.1] or between Eisenhart [1947] and Cornfield and Tukey [1956]. Still there seems to be general agreement that a linear model consists of a model equation with allied assumptions. The model equation relates the observable random variables to underlying parameters and random variables in a linear fashion. The assumptions must specify the nature of the random components and should state whatever restrictions the parameters must satisfy.

What model should we choose? It should be strong enough to support examination of the parameters of interest, but weak enough to assure wide applicability. One thing seems obvious: _The study or experiment under consideration should motivate the model._ A large class of studies exhibit a common characteristic, namely that they seek to compare several populations by comparing responses from samples from each of the populations. Minimally, it is reasonable to assume that each population has a mean; in a wide variety of circumstances the comparison between populations is vested in comparisons among their means. It is this viewpoint, which is general enough to include multiple regression and the analysis of designed experiments, which motivates us to make the following set of assumptions:

1. Interest centers on some response, denoted here by Y, in m different populations. The populations are indexed by the elements of the set $S = \{\underset{\sim}{\alpha}\}$ which have an ordering $\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \cdots, \underset{\sim}{\alpha}_m$ chosen to meet the needs of the particular problem under consideration.

2. A random sample of size $n_{\underset{\sim}{\alpha}}$ is drawn from the population indexed by $\underset{\sim}{\alpha}$; define $n = \sum_{\underset{\sim}{\alpha} \in S} n_{\underset{\sim}{\alpha}}$.

3. $Y_{\underset{\sim}{\alpha} k}$ denotes the k^{th} observable random variable from the population indexed by $\underset{\sim}{\alpha}$, $k = 1, 2, \cdots, n_{\underset{\sim}{\alpha}}$. The random sampling assures that $Y_{\underset{\sim}{\alpha} 1}, \cdots, Y_{\underset{\sim}{\alpha} n_{\underset{\sim}{\alpha}}}$ all have the same probability distribution, and in particular that $E(Y_{\underset{\sim}{\alpha} k}) = \mu_{\underset{\sim}{\alpha}}$ for $k = 1, 2, \cdots, n_{\underset{\sim}{\alpha}}$.

4. $Y_{\underset{\sim}{\alpha} k} = E(Y_{\underset{\sim}{\alpha} k}) + [Y_{\underset{\sim}{\alpha} k} - E(Y_{\underset{\sim}{\alpha} k})] \equiv \mu_{\underset{\sim}{\alpha}} + \epsilon_{\underset{\sim}{\alpha} k}$. The 'errors'' $\epsilon_{\underset{\sim}{\alpha} k}$, are unobservable random variables which obviously have an expectation of zero; their covariance structure will be discussed later. (The historical association of the word 'errors' with $\epsilon_{\underset{\sim}{\alpha} k}$ is unfortunate because they rarely symbolize errors in the sense of mistakes; instead they usually symbolize the vagaries of random sampling.)

If we now introduce the following matrices:

(i) $\underset{m\times 1}{\underset{\sim}{\mu}} = \{\mu_{\underset{\sim}{\alpha}_1}\}$; (ii) $\underset{n\times 1}{\underset{\sim}{\epsilon}} = \{\underset{\sim}{\epsilon}_{\underset{\sim}{\alpha}_1}\}$ where $\underset{n_{\underset{\sim}{\alpha}_1}\times 1}{\underset{\sim}{\epsilon}_{\underset{\sim}{\alpha}_1}} = \{\epsilon_{\underset{\sim}{\alpha}_1 j}\}$; (iii) $\underset{n\times 1}{\underset{\sim}{y}} = \{\underset{\sim}{y}_{\underset{\sim}{\alpha}_1}\}$

where $\underset{n_{\underset{\sim}{\alpha}_1}\times 1}{\underset{\sim}{y}_{\underset{\sim}{\alpha}_1}} = \{Y_{\underset{\sim}{\alpha}_1 j}\}$; and (iv) $\underset{\sim}{W} = \begin{bmatrix} \underset{\sim}{j}^{n_{\underset{\sim}{\alpha}_1}} & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{j}^{n_{\underset{\sim}{\alpha}_2}} & \cdots & \underset{\sim}{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{j}^{n_{\underset{\sim}{\alpha}_m}} \end{bmatrix}$

where $\underset{\sim}{j}^a$ is an a × 1 vector of ones and $\underset{\sim}{0}$ is a vector of zeros, then the above assumptions enable us to decompose the observation vector as

$$\underset{\sim}{y} = \underset{\sim}{W}\underset{\sim}{\mu} + \underset{\sim}{\epsilon} .$$

This will serve as the model equation for a simple linear model, but appropriate assumptions about the parameters $\underset{\sim}{\mu}$ and the random variables $\underset{\sim}{\epsilon}$ are needed before we have a linear model.

Consider $\underset{\sim}{\mu}$ first; either we assume nothing further about it or we restrict it in some manner. Restrictions occupy two very different roles in the context of linear models. In overly parameterized situations, restrictions serve to restrict the parameters to some subspace of the parameter space for the purpose of introducing an element of uniqueness. Such restrictions serve no function in the present context because the assumption of random sampling guarantees that each of the parameters of interest, the $\mu_{\underset{\sim}{\alpha}_1}$, can always be consistently estimated (by $\left[\sum_{j=1}^{n_{\alpha_1}} Y_{\underset{\sim}{\alpha}_1 j}\right] / n_{\underset{\sim}{\alpha}_1}$). Further assumptions can lead to estimators with more desirable properties, for example efficient or sufficient ones. But in any event reasonable estimates of the $\mu_{\underset{\sim}{\alpha}_1}$ exist without restrictions.

The other kind of restriction arises very naturally in certain experimental contexts where the problem dictates that certain relations should exist among the parameters, the $\mu_{\underset{\sim}{\alpha}_1}$ here. For example in the analysis of cross-classified data, the knowledge that a certain interaction is zero forces cell means to satisfy specific relations. As a second example, multiple regression effectively requires that $\underset{\sim}{\mu} = \underset{\sim}{X}\underset{\sim}{\beta}$ where $\underset{\sim}{X}_{m \times k}$ is of rank $k \leq m$; in turn this requires $[\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}']\underset{\sim}{\mu}$ $= [\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}']\underset{\sim}{X}\underset{\sim}{\beta} = \underset{\sim}{0}$. A regression setting may produce additional restrictions of a different sort from these. Occasionally an experimental setting demands that a regression curve go through specified points other than the origin. Each of the restrictions discussed here and many more are covered by requiring $\underset{\sim}{\mu}$ to satisfy $\underset{\sim}{P}'\underset{\sim}{\mu} = \underset{\sim}{c}$ where $\underset{\sim}{P}'$ is $q \times m$ of rank q and $\underset{\sim}{c}$ is known. Thus we will assume for the remainder of this paper that these restrictions are part of the model. For those problems where this restriction is not desired, observe that it vanishes when $q = 0$. The known constants $\underset{\sim}{c}$ frequently will be zero, but in order to preserve applicability to the kinds of problems mentioned above, we will assume merely that $\underset{\sim}{c}$ is known.

Now consider the random deviations denoted by $\underset{\sim}{\epsilon}$. From their definition it is natural to assume that each of them has expectation of zero or collectively that $E(\underset{\sim}{\epsilon}) = \underset{\sim}{0}$. Further, we assume that $\underset{\sim}{V} = E(\underset{\sim}{\epsilon}\underset{\sim}{\epsilon}') = \mathrm{Cov}(\underset{\sim}{\epsilon})$ is known up to a multiplicative scalar. Further assumptions about the errors could be made but they are unnecessary for our purposes. Thus consider the linear model

$$\underset{\sim}{y} = \underset{\sim}{W}\underset{\sim}{\mu} + \underset{\sim}{\epsilon} \tag{1}$$

where

$$\underset{\sim}{P}'\underset{\sim}{\mu} = \underset{\sim}{c}, \quad E(\underset{\sim}{\epsilon}) = \underset{\sim}{0} \quad \text{and} \quad \mathrm{Cov}(\underset{\sim}{\epsilon}) = \underset{\sim}{V} \, .$$

Various choices of $\underset{\sim}{P}'$, $\underset{\sim}{c}$, $\underset{\sim}{V}$ produce linear models for the usual situations of multiple regression and designed experiments or cross-classified data, an illustration of its generality. It may not be immediately apparent that it also covers situations commonly described as 'mixed' and 'random' as well as 'fixed'. The 'mixed' and 'random' models differ from the conventional 'fixed' model (usually) only in the structure of the matrix $\underset{\sim}{V}$ and in the dimension of the parameter space to which $\underset{\sim}{\mu}$ is restricted. For example in the all-variance-component model, $\underset{\sim}{\mu}$ lies in the one dimensional subspace spanned by $\underset{\sim}{j}^n$ because $\underset{\sim}{\mu} = \mu\underset{\sim}{j}^n$ (or $(\underset{\sim}{I} - \frac{1}{n}\underset{\sim}{j}\,\underset{\sim}{j}')\underset{\sim}{\mu} = \underset{\sim}{0}$) and the elements of $\underset{\sim}{V}$ are linear functions of the variance components. In the usual 'mixed' model, with an appropriate ordering of the elements of $\underset{\sim}{y}$, $\underset{\sim}{\mu}$ lies in a subspace of the parameter space of dimension equal to at most the number levels (or combination of levels) of the fixed factor and $\underset{\sim}{V}$ displays a block diagonal form. For the usual 'fixed' model, $\underset{\sim}{V} = \sigma^2\underset{\sim}{I}$.

Other assumptions about $\underset{\sim}{\epsilon}$ are possible. For example normality assumptions about $\underset{\sim}{\epsilon}$ would lead to the kind of model to which the usual maximum likelihood methods are applied. However, other assumptions made above, (1), suffice to allow least squares estimation, the method we intend to use for reasons explained in the next section.

The model, (1), possesses a reasonable degree of generality, but does it support an examination of the experimentally relevant questions? The parameters $\mu_{\underset{\sim}{\alpha_1}}$ have a clear relation to the experimental context and the corresponding sample means provide at least consistent estimates of each of them individually. Subsequently we will examine their joint estimation subject to the model restrictions and again unique estimates will result. The capability of estimating $\underset{\sim}{\mu}$ assures the capability of estimating $\underset{\sim}{t}'\underset{\sim}{\mu}$ or $\underset{\sim}{T}'\underset{\sim}{\mu}$ by the corresponding function of the estimate of $\underset{\sim}{\mu}$; of course this produces estimates of $\underset{\sim}{t}'\underset{\sim}{\mu}$ or $\underset{\sim}{T}'\underset{\sim}{\mu}$ with as yet unspecified properties, but existence, not properties constitutes the essential issue for the final paragraph of this section.

An experimenter, who professes interest in how the means of his populations relate to each other, should be able to associate experimental interpretations with functions $\underset{\sim}{t}'\underset{\sim}{\mu}$. If he cannot, no linear model can help him interpret his experimental results because any estimable function, in the usual terminology of linear models, is merely estimating $\underset{\sim}{t}'\underset{\sim}{\mu}$ for some $\underset{\sim}{t}$. (This statement will be verified in section 5.) It seems apparent that an experimenter should have a much easier time interpreting $\underset{\sim}{t}'\underset{\sim}{\mu}$ when he picks $\underset{\sim}{t}$ so that $\underset{\sim}{t}'\underset{\sim}{\mu}$ has a fairly obvious relation to his experimental concerns than when he abdicates responsibility for picking $\underset{\sim}{t}$ to some arithmetic process which he does not understand.

3. ESTIMATION

We choose to estimate $\underset{\sim}{\mu}$ by the method of least squares. This is not a particularly restrictive approach because there are several viewpoints which lead to this method: (1) The minimization of the sum of squares of deviations between observation and prediction has been regarded as intuitively appealing ever since Legendre [1805] introduced it. The geometric analogue of this offers equal appeal. (2) Gauss [1809] arrived at least squares estimation by trying to get 'maximum probability of zero error of estimation'. This corresponds closely to maximum likelihood estimation under normal-theory distribution of errors. (3) Gauss [1821] again ended up with least squares estimation when he considered 'least mean squared error of estimation' which we now think of as best (minimum variance) linear unbiased estimation (BLUE). (4) Several decision theoretic approaches produce estimates which must satisfy the least squares criterion.

The resulting least squares estimates of $\boldsymbol{\mu}$ possess various properties, depending upon the viewpoint taken. We will not dwell upon these properties because instead we want to consider the relation of the parameterization and associated estimation to the experimental context. Since the principle of least squares generally produces consistent estimates which, in the present context, also possess the property of unbiasedness, it seems reasonable to regard them as estimates. The interested reader can pursue a consideration of properties elsewhere.

Minimization of $(\mathbf{y} - \mathbf{W}\boldsymbol{\mu})'V^{-1}(\mathbf{y} - \mathbf{W}\boldsymbol{\mu})$ over possible values of $\boldsymbol{\mu}$ produces

$$\hat{\boldsymbol{\mu}}_u = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}\mathbf{y} \qquad (2)$$

as the generalized least squares estimate of $\boldsymbol{\mu}$. The subscript of u on $\hat{\boldsymbol{\mu}}$ here should convey the unrestricted estimate of $\boldsymbol{\mu}$ in contrast to the restricted case when $\boldsymbol{\mu}$ must satisfy $\mathbf{P}'\boldsymbol{\mu} = \mathbf{c}$:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_r &= (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\{\mathbf{W}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{P}[\mathbf{P}'(\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{P}]^{-1}[\mathbf{P}'(\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{c}]\} \\ &= \hat{\boldsymbol{\mu}}_u - (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{P}[\mathbf{P}'(\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{P}]^{-1}[\mathbf{P}'\hat{\boldsymbol{\mu}}_u - \mathbf{c}] \ .\end{aligned} \qquad (3)$$

These estimates, in a sense, complete the task we set for ourselves, but they shed very little light on what happens in the usual cases. Thus the remainder of this paper will be devoted to elucidation; the usual fixed model will serve to communicate our remaining ideas. Thus we will assume for the subsequent discussion that $\mathbf{V} = \sigma^2\mathbf{I}$, the usual form for the fixed model. The simplicity of this assumption will not obscure the facts we wish to show.

Linear independence of the elements of $\boldsymbol{\mu}$ produces from Eq. (2) the well known result

$$\hat{\boldsymbol{\mu}}_u = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} \equiv \mathbf{D}^{-1}\mathbf{W}'\mathbf{y} \equiv \bar{\mathbf{y}} \, . \tag{4}$$

Here, $\mathbf{D}$ is a diagonal matrix with diagonal elements $n_{\boldsymbol{\alpha}_1}$ and $\bar{\mathbf{y}}$ is the m × 1 vector of observed means ordered the same as the elements of $\boldsymbol{\mu}$. If the elements of $\boldsymbol{\mu}$ satisfy $\mathbf{P}'\boldsymbol{\mu} = \mathbf{c}$ instead of being linearly independent, then Eq. (3) produces

$$\hat{\boldsymbol{\mu}}_r = \hat{\boldsymbol{\mu}}_u - \mathbf{A}(\mathbf{A}'\mathbf{D}\mathbf{A})^{-1}(\mathbf{P}'\hat{\boldsymbol{\mu}}_u - \mathbf{c}) \tag{5}$$

as the estimate of $\boldsymbol{\mu}$ where $\mathbf{A}' = \mathbf{P}'\mathbf{D}^{-1}$.

Although an experimenter wants to estimate $\boldsymbol{\mu}$, he usually also wants to estimate certain linear functions of the means. In either the restricted or unrestricted cases, the linear parametric function $\mathbf{t}'\boldsymbol{\mu}$ has an unbiased estimate since

$$E(\mathbf{t}'\hat{\boldsymbol{\mu}}_u) = \mathbf{t}'\boldsymbol{\mu} = E(\mathbf{t}'\hat{\boldsymbol{\mu}}_r) \, .$$

However, these two estimates may not have the same variance because

$$\operatorname{Cov}(\hat{\boldsymbol{\mu}}_u) = \sigma^2\mathbf{D}^{-1} \quad \text{while} \quad \operatorname{Cov}(\hat{\boldsymbol{\mu}}_r) = \sigma^2(\mathbf{D}^{-1} - \mathbf{A}(\mathbf{A}'\mathbf{D}\mathbf{A})^{-1}\mathbf{A}') \, ,$$

which implies that

$$\operatorname{Var}(\mathbf{t}'\hat{\boldsymbol{\mu}}_r) = \operatorname{Var}(\mathbf{t}'\hat{\boldsymbol{\mu}}_u) - \mathbf{t}'\mathbf{A}(\mathbf{A}'\mathbf{D}\mathbf{A})^{-1}\mathbf{A}'\mathbf{t} \, .$$

Since the subtractive term cannot be negative, $\operatorname{Var}(\mathbf{t}'\hat{\boldsymbol{\mu}}_r) \leq \operatorname{Var}(\mathbf{t}'\hat{\boldsymbol{\mu}}_u)$.

It seems appropriate at this juncture to make several observations. The selection of $\mathbf{T}'\boldsymbol{\mu}$ as the set of linear functions of interest, where this set contains fewer linearly independent functions (k) than m, does not automatically make it appropriate to use $\hat{\boldsymbol{\mu}}_r$ by assuming that a further set of m - k linearly independent functions $\mathbf{P}'\boldsymbol{\mu}$ is zero. The fact that m - k more linearly independent functions of the elements of $\boldsymbol{\mu}$ <u>could</u> have been chosen, but were not, does not entitle one to say this is equivalent to letting them be zero. This should go without saying, but one benefits if $\hat{\underset{\sim}{\mu}}_r$ can be used, a benefit being that some elements of $\hat{\underset{\sim}{\mu}}_r$ have less variance than the corresponding elements of $\hat{\underset{\sim}{\mu}}_u$. This benefit derives from some additional knowledge about linear relationships between the elements in $\underset{\sim}{\mu}$. Ignoring things which could be estimated but are not, does not entitle one this benefit. This paragraph was motivated by a practice frequently suggested for analyzing very messy cross classified data, namely, the practice of fitting a 'main effects only' model.

An obvious fact which has a tendency to be forgotten in some situations, particularly in N-way cross classifications with many missing cells, is this: The vector of observed means provides an estimate (at least unbiased and consistent) of the vector of means of all populations from which at least one observation was taken. In turn, estimates of interesting linear functions of the population means result from the same linear functions of the corresponding sample means. (Of course, estimates with a smaller variance may be available if something is known about $\underset{\sim}{\mu}$, namely $\underset{\sim}{P}'\underset{\sim}{\mu} = \underset{\sim}{c}$.)

The selection of an experimentally interesting set of linear functions of the vector of means poses a major problem to both the experimenter and the statistician. Left to his own devices the experimenter may say that he is interested in $\underset{\sim}{T}'\underset{\sim}{\mu}$. Whichever of $\underset{\sim}{T}'\hat{\underset{\sim}{\mu}}_u$ or $\underset{\sim}{T}'\hat{\underset{\sim}{\mu}}_r$ is appropriate in a given situation, it provides an estimate of $\underset{\sim}{T}'\underset{\sim}{\mu}$ without regard to the number of rows in $\underset{\sim}{T}$. With this approach, there is never any question as to the 'estimability' of certain linear functions of the cell means. It is our opinion that the question of estimability of certain linear functions of parameters in linear models is a result of the statisticians failure to be precise in assisting experimenters to pick $\underset{\sim}{T}$'s for their problems.

To bring this discussion back to 'home' base, the foregoing discussion and results point to the vector $\underset{\sim}{\mu}$ as the thing to be estimated, either by $\hat{\underset{\sim}{\mu}}_u$ or $\hat{\underset{\sim}{\mu}}_r$. Thereafter $\hat{\underset{\sim}{\theta}} = \underset{\sim}{T}'\hat{\underset{\sim}{\mu}}_u$ or $\underset{\sim}{T}'\hat{\underset{\sim}{\mu}}_r$ provides the experimenter with estimates of the linear functions of the means which interest him.

4. A NUMERICAL EXAMPLE

We intend for the example of this section to illustrate the definitions, model and estimation of the preceding two sections; it also serves to motivate some further comments.

Suppose that a 2 X 3 factorial set of treatments has been run as a completely randomized experiment with the statistical layout of the data as follows:

		Factor B		
		Level 1	Level 2	Level 3
Factor A	Level 1	$y_{111} = 6$ $y_{112} = 8$	$y_{121} = 5$ $y_{122} = 4$ $y_{123} = 3$	$y_{131} = 12$
	Level 2	$y_{211} = 11$	$y_{221} = 5$ $y_{222} = 7$	$y_{231} = 16$

The underlying sampling situation consists of m = 6 populations which are related by a 2 X 3 factorial structure. Also n = 10 and S = {(1,1), (1,2), (1,3), (2,1), (2,2), (2,3)}; this choice of the $\underset{\sim}{\alpha}$'s as two-component vectors was motivated by the experimental situation. The following is a convenient ordering of the $\underset{\sim}{\alpha}_i$'s: $\underset{\sim}{\alpha}_1 = (1,1)$, $\underset{\sim}{\alpha}_2 = (1,2)$, $\underset{\sim}{\alpha}_3 = (1,3)$, $\underset{\sim}{\alpha}_4 = (2,1)$, $\underset{\sim}{\alpha}_5 = (2,2)$, $\underset{\sim}{\alpha}_6 = (2,3)$. If for brevity we denote $\mu_{\underset{\sim}{\alpha}_1} = \mu_{(1,1)}$ by μ_{11}, etc., then $\underset{1 \times 6}{\underset{\sim}{\mu}'} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})$,

$$\underset{\sim}{W}' = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\underset{\sim}{W}'\underset{\sim}{W} \equiv \underset{\sim}{D} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

Suppose the researcher has indicated interest in $\underset{\sim}{\theta} = \underset{\sim}{T}'\underset{\sim}{\mu}$:

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & 0 & -1 & 1 & 0 \\ -1 & -1 & 2 & -1 & -1 & 2 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 1 & -2 & -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix}$$

Then $\underset{\sim}{\hat{\mu}}_u' = (\underset{\sim}{D}^{-1}\underset{\sim}{W}'\underset{\sim}{y})' = \bar{\underset{\sim}{y}}' = (7,4,12,11,6,16)$ and $(\underset{\sim}{T}'\underset{\sim}{\hat{\mu}}_u)' = \underset{\sim}{\hat{\theta}}_u' = (56,10,-8,28,-2,2)$.

If the experimenter somehow knows that $\underset{\sim}{p}'\underset{\sim}{\mu} = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$, then

$$\underset{\sim}{\hat{\mu}}_r = \begin{bmatrix} 7 \\ 4 \\ 12 \\ 11 \\ 6 \\ 16 \end{bmatrix} - \frac{3}{42} \begin{bmatrix} 3 & -3 & 0 & -3 & 3 & 0 \\ -2 & 2 & 0 & 2 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -6 & 6 & 0 & 6 & -6 & 0 \\ 3 & -3 & 0 & -3 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 7 \\ 4 \\ 12 \\ 11 \\ 6 \\ 16 \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 52 \\ 26 \\ 84 \\ 71 \\ 45 \\ 112 \end{bmatrix}$$

and

$$\underset{\sim}{\hat{\theta}}_r = (\underset{\sim}{T}'\underset{\sim}{\hat{\mu}}_r)' = (55\tfrac{5}{7},\ 9\tfrac{3}{7},\ -7\tfrac{3}{7},\ 28\tfrac{2}{7},\ 0,\ 2\tfrac{4}{7}).$$

The covariance matrix of $\hat{\underset{\sim}{\theta}}_u = \underset{\sim}{T}'\hat{\underset{\sim}{\mu}}_u$, $\sigma^2\underset{\sim}{T}'(\underset{\sim}{W}'\underset{\sim}{W})^{-1}\underset{\sim}{T}$, is

$$\frac{\sigma^2}{21}\begin{bmatrix} 91 & 14 & -14 & 35 & -7 & -14 \\ 14 & 91 & -7 & -14 & -14 & 35 \\ -14 & -7 & 49 & 14 & 14 & 7 \\ 35 & -14 & 14 & 217 & 7 & 14 \\ -7 & -14 & 14 & 7 & 49 & 14 \\ -14 & 35 & 7 & 14 & 14 & 217 \end{bmatrix}$$

whereas the covariance matrix of $\hat{\underset{\sim}{\theta}}_r = \underset{\sim}{T}'\hat{\underset{\sim}{\mu}}_r$, $\sigma^2[\underset{\sim}{T}'(\underset{\sim}{W}'\underset{\sim}{W})^{-1}\underset{\sim}{T} - \underset{\sim}{T}'\underset{\sim}{D}^{-1}\underset{\sim}{P}(\underset{\sim}{P}'\underset{\sim}{D}^{-1}\underset{\sim}{P})^{-1}\underset{\sim}{P}'\underset{\sim}{D}^{-1}\underset{\sim}{T}]$, is

$$\frac{\sigma^2}{21}\begin{bmatrix} 90 & 12 & -12 & 36 & 0 & -12 \\ 12 & 87 & -3 & -12 & 0 & 39 \\ -12 & -3 & 45 & 12 & 0 & 3 \\ 36 & -12 & 12 & 216 & 0 & 12 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -12 & 39 & 3 & 12 & 0 & 213 \end{bmatrix}$$

In commenting upon this example, we should explain the rationale which led to the choice of $\underset{\sim}{T}$. We chose the conventional mean, main effects and interaction of the two main effects as a guide here, θ_1 being associated with the 'over-all mean', θ_2 with the 'main effect A', θ_3 and θ_4 with the 'main effect B', and θ_5 and θ_6 with the 'interaction' of the main effects. The conventional nature of these contrasts relates closely to assuming

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \qquad (6)$$

and proceeding. While these contrasts seem logical in some contexts, other logical ones exist. For example, suppose the experimenter knew that the middle level of Factor B would change the response. The contrasts

$$\begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 \\ 1 & -2 & 1 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix}$$

could be much more interesting in this case. If instead the six populations were strata of a larger population, the experimenter might choose coefficients related to the sizes of the various strata. This illustrates our basic point: the experimental setting should dictate the statistical analysis, not the opposite. The assumption of an overly specified linear model such as Eq. (6) obscures this basic point.

Experimenters often know certain things about their experimental setting. We took $\theta_5 = 0$ to illustrate the impact of this sort of knowledge. The incorporation of $\theta_5 = 0$ into the estimation of the other five linear functions of interest led to different estimates with smaller variances. The usual procedure of testing for no interaction involves (equivalently), testing both θ_5 and θ_6 simultaneously zero. If however, one decomposes the interaction into components and subsequently determines that a subset of <u>these</u> components of the interaction is zero, then one can derive the benefit of this information. The usual formulation does not explicitly state how this may be done.

5. SOME COMPARISONS WITH THE CONVENTIONAL APPROACH

The presentations of Scheffé [1958] and Graybill [1961] exemplify what we call the conventional approach to linear models. Specifically they begin by assuming that the responses satisfy a model equation of the form $\underset{\sim}{y} = \underset{\sim}{Q}\underset{\sim}{\psi} + \underset{\sim}{\delta}$ where $\underset{\sim}{Q}_{nXs}$ may have rank less than s, s ≤ n. In this section we will contrast this formulation to our approach set out in sections 2 and 3 by considering the senses in which the two approaches are equivalent and different. By doing so we hope to emphasize some assumptions on which the conventional approach rests; assumptions which are easily ignored or even forgotten although they may be of major consequence.

The major similarity in the two approaches lies in the random or error parts; the major difference in the fixed or mean parts. For the conventional approach, the $\underset{\sim}{\delta}$ in $\underset{\sim}{y} = \underset{\sim}{Q}\underset{\sim}{\psi} + \underset{\sim}{\delta}$ satisfies at least $E(\underset{\sim}{\delta}) = \underset{\sim}{0}$ and $\text{cov}(\underset{\sim}{\delta}) = \underset{\sim}{V}$, where $\underset{\sim}{V}$ is known up to a scalar multiplier, but otherwise remains undefined except implicitly by $\underset{\sim}{\delta} = \underset{\sim}{y} - \underset{\sim}{Q}\underset{\sim}{\psi}$. The $\underset{\sim}{\epsilon}$ in $\underset{\sim}{y} = \underset{\sim}{W}\underset{\sim}{\mu} + \underset{\sim}{\epsilon}$, though more clearly related to the sampling context, satisfies the same assumptions. While the $\underset{\sim}{\epsilon}$ and the $\underset{\sim}{\delta}$ differ in definitional clarity, they occupy essentially the same roles in the two approaches. Consequently the fixed or nonrandom parts must produce the same expectation for $\underset{\sim}{y}$ because $E(\underset{\sim}{\epsilon}) = \underset{\sim}{0} = E(\underset{\sim}{\delta})$ implies $\underset{\sim}{W}\underset{\sim}{\mu} = E(\underset{\sim}{y}) = \underset{\sim}{Q}\underset{\sim}{\psi}$. Past this the similarity ceases.

Further comparisons between the conventional and proposed approaches require an examination of $\underset{\sim}{\mu}$ and $\underset{\sim}{\psi}$. Since the restrictions $\underset{\sim}{P}'\underset{\sim}{\mu} = \underset{\sim}{c}$ on $\underset{\sim}{\mu}$ somewhat cloud the comparison, we will first consider an unrestricted $\underset{\sim}{\mu}$. Let us next restate the essential characteristic of our formulation as follows: *A parameter space of dimension m* (the number of populations sampled in the study) *suffices to describe the results of the study*. The parameter vector $\underset{\sim}{\psi}$ of the conventional formulation lies in a s-dimensional parameter space. The equality $\underset{\sim}{W}\underset{\sim}{\mu} = \underset{\sim}{Q}\underset{\sim}{\psi}$ and the special structure of $\underset{\sim}{W}$ produce two relevant relations. The first $n_{\underset{\sim}{\alpha}_1}$ rows of $\underset{\sim}{W}$ are identical, so are the next $n_{\underset{\sim}{\alpha}_2}$, and so on. Pick one row from each of these sets of identical rows and let $\underset{\sim}{Q}_1$ denote the corresponding rows of $\underset{\sim}{Q}$. Thus $\underset{\sim}{\mu} = \underset{\sim}{Q}_1\underset{\sim}{\psi}$ and in turn because $\underset{\sim}{\mu}$ is unrestricted, $\underset{\sim}{Q}_1$ must be $m \times s$ of rank m and so $s \geq m$.

Is it plausible to restrict $\underset{\sim}{\psi}$ to some subspace of its s-dimensional parameter space? Recall that restrictions occupy two roles in the context of linear models: they may serve to introduce an element of uniqueness into an overly parameterized situation or they may be an attribute of the experimental context. For the present we have eliminated the second sort of restriction from consideration by assuming that $\underset{\sim}{\mu}$ does not satisfy $\underset{\sim}{P}'\underset{\sim}{\mu} = \underset{\sim}{c}$. As far as the uniqueness is concerned, subsequent considerations of $\underset{\sim}{\psi}$ will be general enough to allow any possible solution for $\hat{\underset{\sim}{\psi}}$. From these one could be picked satisfying a specified set of uniqueness criteria. Thus we will assume that $\underset{\sim}{\psi}$ lies in an unrestricted s-dimensional parameter space. This in turn requires that $\underset{\sim}{Q}$ have the same rank as $\underset{\sim}{Q}_1$, namely m.

When $s = m$, $\underset{\sim}{\mu}$ and $\underset{\sim}{\psi}$ are equivalent parameterizations because there is a one-to-one relationship between the points in the two parameter spaces. Hence, $\underset{\sim}{\psi}$ occupies a role essentially the same as the $\underset{\sim}{T}'\underset{\sim}{\mu}$, the set of 'experimentally interesting functions' discussed earlier. However, the conventional model rarely has $s = m$ in the present case, namely m populations with functionally unrelated means.

If instead, $s > m$, the parameterizations are very different. The example of the previous section will serve to illustrate this. Recall that it had a two-way cross classification with two rows, three columns, and no empty cells. The basic parameter $\underset{\sim}{\mu}' = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})$ lies in a six-dimensional parameter space which is clearly related to the experimental context. By contrast the conventional approach resting upon the scalar model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad i = 1, 2, \quad j = 1, 2, 3, \quad k = 1, 2, \cdots, n_{ij} > 0$$

is parameterized by the vector

$$\underset{\sim}{\psi}' = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, (\alpha\beta)_{11}, (\alpha\beta)_{12}, (\alpha\beta)_{13}, (\alpha\beta)_{21}, (\alpha\beta)_{22}, (\alpha\beta)_{23}).$$

The elements of this parameter vector generally remain undefined; attempts to define them almost always revert to $\underset{\sim}{\mu}$!

Certain computational relations exist in spite of the dissimilarity of the parameterizations. The least squares estimate for $\underset{\sim}{\mu}$ in this unrestricted case is the $\hat{\underset{\sim}{\mu}}_u$ which minimizes

$$(\underset{\sim}{y} - \underset{\sim}{W}\underset{\sim}{\mu})'(\underset{\sim}{y} - \underset{\sim}{W}\underset{\sim}{\mu})\Big|_{\underset{\sim}{\mu} = \hat{\underset{\sim}{\mu}}_u} ,$$

namely,

$$\hat{\underset{\sim}{\mu}}_u = (\underset{\sim}{W}'\underset{\sim}{W})^{-1}\underset{\sim}{W}'\underset{\sim}{y} = \underset{\sim}{D}^{-1}\underset{\sim}{W}'\underset{\sim}{y} = \bar{\underset{\sim}{y}} . \tag{7}$$

A least squares estimate for $\underset{\sim}{\psi}$ in this unrestricted case is any $\hat{\underset{\sim}{\psi}}_u$ which minimizes

$$(\underset{\sim}{y} - \underset{\sim}{Q}\underset{\sim}{\psi})'(\underset{\sim}{y} - \underset{\sim}{Q}\underset{\sim}{\psi})\Big|_{\underset{\sim}{\psi} = \hat{\underset{\sim}{\psi}}_u}$$

namely, all $\hat{\underset{\sim}{\psi}}_u$ which satisfy

$$\underset{\sim}{Q}'\underset{\sim}{Q}\hat{\underset{\sim}{\psi}}_u = \underset{\sim}{Q}'\underset{\sim}{y} . \tag{8}$$

If $\underset{\sim}{A}^-$ denotes a generalized inverse of $\underset{\sim}{A}$, namely any matrix which satisfies $\underset{\sim}{A}\underset{\sim}{A}^-\underset{\sim}{A} = \underset{\sim}{A}$, then Urquhart [1969] has shown that $\underset{\sim}{\psi} = \underset{\sim}{A}^-\underset{\sim}{b}$ generates all possible solutions to the consistent equations $\underset{\sim}{A}\underset{\sim}{\psi} = \underset{\sim}{b}$ provided $\underset{\sim}{b} \neq \underset{\sim}{0}$. Applied to (8), this yields

$$\hat{\underset{\sim}{\psi}}_u = (\underset{\sim}{Q}'\underset{\sim}{Q})^-\underset{\sim}{Q}'\underset{\sim}{y} . \tag{9}$$

In order to compare Eq. (7) and Eq. (9), recall that $\mu = Q_1\psi$ and $W\mu = Q\psi$ or $Q\psi = WQ_1\psi$ or equivalently that $Q = WQ_1$. The equality $\mu = Q_1\psi$ suggests $\tilde{\mu}_u = Q_1\hat{\psi}_u$ as a possible estimate of μ_u. The equivalence of $\hat{\mu}_u$ and $\tilde{\mu}_u$ rests upon the identity $Q_1(Q'Q)^-Q_1 = (W'W)^{-1}$ which is valid for any choice of a generalized inverse of $Q'Q$. To show this, consider the definition of $(Q'Q)^-$ as any matrix satisfying

$$Q'Q(Q'Q)^-Q'Q = Q'Q$$

or equivalently

$$Q_1'W'WQ_1(Q'Q)^-Q_1'W'WQ_1 = Q_1'W'WQ_1 \;.$$

Since both $W'W$ and Q_1Q_1' have inverses, premultiplication of this expression by $(W'W)^{-1}(Q_1Q_1')^{-1}Q_1$ and postmultiplication by $Q_1'(Q_1Q_1')^{-1}(W'W)^{-1}$ produces the identity. Thus in turn,

$$\tilde{\mu}_u = Q_1\hat{\psi}_u = Q_1(Q'Q)^-Q'y = Q_1(QQ')^-Q_1'W'y = (W'W)^{-1}W'y = \hat{\mu}_u \;. \qquad (10)$$

This result could have and should have been expected but was given to show how to obtain $\hat{\mu}_u$ from $\hat{\psi}_u$ once the latter has been calculated. Since $Q'Q$ is the coefficient matrix of the normal equations of the conventional 'linear model', if we obtain <u>any</u> solution to these equations and form $Q_1\tilde{\psi}_u$, this vector provides as an estimate of the vector of all means. When we have defined a vector of interesting linear functions of the cell means, say $T'\mu$, one then merely forms $T'(Q\hat{\psi}_u)$ to determine the vector of estimates of the functions of the parameters of interest.

The restricted case, when $P'\mu = c$, presents one major difference from the unrestricted case, namely the two approaches must be subject to equivalent restrictions. The restrictions $P'\mu = c$ grow out of the experimental context; they are unnecessary to produce uniqueness of estimates. Consequently the experimental context should equivalently restrict ψ to satisfy a set of restrictions $R\psi = b$. Again μ has a direct estimate through Eq. (3) and an indirect estimate through ψ.

The first gives

$$\hat{\boldsymbol{\mu}}_r = \{\mathbf{I} - (\mathbf{W}'\mathbf{W})^{-1}\mathbf{P}[\mathbf{P}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{P}]^{-1}\mathbf{P}'\}\,(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$$
$$+ (\mathbf{W}'\mathbf{W})^{-1}\mathbf{P}[\mathbf{P}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{P}]^{-1}\mathbf{c}\ . \qquad (11)$$

The class of least squares estimates of $\boldsymbol{\psi}$, obtained in the same manner as Eq. (3), is given by

$$\hat{\boldsymbol{\psi}}_r = \{\mathbf{I} - (\mathbf{Q}'\mathbf{Q})^{-}\mathbf{R}[\mathbf{R}'(\mathbf{Q}'\mathbf{Q})^{-}\mathbf{R}]^{-}\mathbf{R}'\}\,(\mathbf{Q}'\mathbf{Q})^{-}\mathbf{Q}'\mathbf{y}$$
$$+ (\mathbf{Q}'\mathbf{Q})^{-}\mathbf{R}[\mathbf{R}'(\mathbf{Q}'\mathbf{Q})^{-}\mathbf{R}]^{-}\mathbf{b}\ . \qquad (12)$$

The equivalence of $\hat{\boldsymbol{\mu}}_r$ and $\tilde{\boldsymbol{\mu}}_r = \mathbf{Q}_1\hat{\boldsymbol{\psi}}_r$ rests upon the equivalence of the restrictions $\mathbf{P}'\boldsymbol{\mu} = \mathbf{c}$ and $\mathbf{R}'\boldsymbol{\psi} = \mathbf{b}$. Since $\boldsymbol{\mu} = \mathbf{Q}_1\boldsymbol{\psi}$, the restrictions on $\boldsymbol{\mu}$ become, in the $\boldsymbol{\psi}$ parameter space, $\mathbf{c} = \mathbf{P}'\boldsymbol{\mu} = \mathbf{P}'\mathbf{Q}_1\boldsymbol{\psi}$. Thus equivalence of the two sets of restrictions requires that $\{\boldsymbol{\psi}\colon\ \mathbf{R}'\boldsymbol{\psi} = \mathbf{b}\} = \{\boldsymbol{\psi}\colon\ \mathbf{P}'\mathbf{Q}_1\boldsymbol{\psi} = \mathbf{c}\}$. Since $\mathbf{R}'$, $\mathbf{P}'$, and $\mathbf{Q}_1$ all have full rank, the coefficient matrices and the constant vectors can differ by at most a nonsingular premultiplier, i.e. $\mathbf{R}' = \mathbf{M}\mathbf{P}'\mathbf{Q}_1$ and $\mathbf{b} = \mathbf{M}\mathbf{c}$ where $\mathbf{M}^{-1}$ exists. The utilization of these two relations with $\mathbf{Q}_1(\mathbf{Q}'\mathbf{Q})^{-}\mathbf{Q}_1' = (\mathbf{W}'\mathbf{W})^{-1}$ in $\tilde{\boldsymbol{\mu}}_r = \mathbf{Q}_1\hat{\boldsymbol{\psi}}_r$ shows that $\hat{\boldsymbol{\mu}}_r = \tilde{\boldsymbol{\mu}}_r$ as with the unrestricted case.

The correspondence between restrictions in the two approaches has one additional aspect which warrants consideration. We have implicitly assumed that the parameterization in terms of $\boldsymbol{\psi}$ was complete enough to support comparisons with $\boldsymbol{\mu}$. There exist experimental situations where this assumption might not appear true. The analysis of large cross-classified sets of data with many empty cells often proceeds under the assumption that many or even all interactions do not exist. This still fits into the preceding framework; include all such interactions in $\boldsymbol{\psi}$ and include enough restrictions in $\mathbf{R}'\boldsymbol{\psi} = \mathbf{b}$ to make them equal zero. Since

$\underset{\sim}{\mu} = \underset{\sim}{Q}_1\underset{\sim}{\psi} = (\underset{\sim}{Q}_{11}, \underset{\sim}{Q}_{12})\begin{pmatrix}\underset{\sim}{\psi}_1 \\ \underset{\sim}{0}\end{pmatrix} = \underset{\sim}{Q}_{11}\underset{\sim}{\psi}_1$ where $\underset{\sim}{\psi}_2 = \underset{\sim}{0}$ by restrictions or omission (the partitions are conformable), the restriction $\underset{\sim}{\psi}_2 = \underset{\sim}{0}$ will usually restrict $\underset{\sim}{\mu}$ to some subspace of its m-dimensional parameter space, or equivalently $\underset{\sim}{P}'\underset{\sim}{\mu} = \underset{\sim}{c}$. Regardless of how we say it, the omission of a set of interaction terms from $\underset{\sim}{\psi}$ forces $\underset{\sim}{\mu}$ to satisfy restrictions, ones which might be quite objectionable to an experimenter if he knew they were present.

To illustrate the foregoing, we give another example. Consider a 3 X 3 structure of populations with one observation on eight of the populations and no observations on one population. (This choice of one observation per filled cell simplifies arithmetic, but essential features of the example do not depend upon it .) Label the corresponding means as follows:

		Factor B		
		1	2	3
	1	μ_{11}	μ_{12}	μ_{13}
Factor A	2	μ_{21}	μ_{22}	μ_{23}
	3	μ_{31}	μ_{32}	

First consider our model and suppose we assume $\mu_{ij} = \alpha_i + \beta_j$, that is, the cell mean in the i^{th} row and j^{th} column is expressible in this form. Writing μ_{ij} in this form means, in the conventional sense, we assume 'no interaction', a point we will illustrate with the conventional model shortly. Consider the following three linearly independent (not necessarily orthogonal) restrictions on $\underset{\sim}{\mu}$:

$$\begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \mu_{31} \\ \mu_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} .$$

They force the μ_{ij} to have the specified additive form. Thus

$$\hat{\underset{\sim}{\mu}}_r = \frac{1}{12} \begin{bmatrix} 7 & 3 & 2 & 3 & -1 & -2 & 2 & -2 \\ 3 & 7 & 2 & -1 & 3 & -2 & -2 & 2 \\ 2 & 2 & 8 & -2 & -2 & 4 & 0 & 0 \\ 3 & -1 & -2 & 7 & 3 & 2 & 2 & -2 \\ -1 & 3 & -2 & 3 & 7 & 2 & -2 & 2 \\ -2 & -2 & 4 & 2 & 2 & 8 & 0 & 0 \\ 2 & -2 & 0 & 2 & -2 & 0 & 8 & 4 \\ -2 & 2 & 0 & -2 & 2 & 0 & 4 & 8 \end{bmatrix} \begin{bmatrix} y_{111} \\ y_{121} \\ y_{131} \\ y_{211} \\ y_{221} \\ y_{231} \\ y_{311} \\ y_{321} \end{bmatrix} .$$

If the experimenter thought that $\mu_{ij} = \alpha_i + \beta_j$ for all cells, including the unfilled one, he might be motivated to define μ_{33} as $- \mu_{22} + \mu_{23} + \mu_{32}$. Thus if $\underset{\sim}{t}' = (0, 0, 0, 0, -1, 1, 0, 1)$, $\hat{\underset{\sim}{\mu}}_{33} = \underset{\sim}{t}'\hat{\underset{\sim}{\mu}}_r = (-3, -3, 6, -3, -3, 6, 6, 6)\underset{\sim}{y}$.

If we now set up the normal equations for the conventional model equation,

$$y_{ijk} = \alpha_i + \beta_j + \epsilon_{ijk} \ ,$$

we get

$$\begin{bmatrix} 3 & 0 & 0 & 1 & 1 & 1 \\ 0 & 3 & 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 1 & 1 & 0 \\ 1 & 1 & 1 & 3 & 0 & 0 \\ 1 & 1 & 1 & 0 & 3 & 0 \\ 1 & 1 & 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} a_1 = y_{111} + y_{121} + y_{131} \\ a_2 = y_{211} + y_{221} + y_{231} \\ a_3 = y_{311} + y_{321} \\ b_1 = y_{111} + y_{211} + y_{311} \\ b_2 = y_{121} + y_{221} + y_{321} \\ b_3 = y_{131} + y_{231} \end{bmatrix}$$

A solution to this system of equations is

$$\begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{3}(a_1 - \hat{\beta}_1 - \hat{\beta}_2) \\ \frac{1}{3}(a_2 - \hat{\beta}_1 - \hat{\beta}_2) \\ \frac{1}{2}(a_3 - \hat{\beta}_1 - \hat{\beta}_2) \\ \frac{1}{12}(11b_1^* + 7b_2^*) \\ \frac{1}{12}(7b_1^* + 11b_2^*) \\ 0 \end{bmatrix}$$

where $b_1^* = b_1 - a_1/3 - a_2/3 - a_3/2$ and $b_2^* = b_2 - a_1/3 - a_2/3 - a_3/2$.

We form only the $(1,1)^{th}$ element of $\tilde{\mu}_r$, $\tilde{\mu}_{11} = \hat{\alpha}_1 + \hat{\beta}_1$, for illustration:

$$\tilde{\mu}_{11} = \frac{2a_1 - 2a_2 - 3a_3 + 5b_1 + b_2}{12} = \frac{1}{12}(7y_{111} + 3y_{121} + 2y_{131} + 3y_{211} - y_{221} - 2y_{231}$$

$$+ 2y_{311} - 2y_{321})$$

which is the same estimate obtained using the procedure of this paper to obtain the $(1,1)^{th}$ element of $\hat{\underset{\sim}{\mu}}_r$ shortly before.

This example provides a means of exhibiting some of the peculiarities which occur when no data appears in at least one cell. In the example, the expectation of $\hat{\alpha}_1 - \hat{\alpha}_3$ in terms of the μ_{ij}'s is

$$\frac{1}{12}(5\mu_{11} + 5\mu_{12} + 2\mu_{13} + \mu_{21} + \mu_{22} - 2\mu_{23} - 6\mu_{31} - 6\mu_{32}).$$

This reduces to $\alpha_1 - \alpha_3$ when $\mu_{ij} = \alpha_i + \beta_j$ but otherwise this linear function of the true cell means appears very odd. It seems very unlikely indeed that a researcher, faced with the problem of deciding what linear functions of the cell means would be of interest to him, would come up with this linear function.

This simple example gives some insight into what might be going on in more complicated settings. Analysis of the data from a multi-way cross classification problem involving as many as ten factors sometimes is done with a 'main-effects only' type model, primarily because even this 'simple model' taxes the capacity of most computers. In problems of this size, it is not uncommon to have ten to fifteen thousand observations with at most five percent of the cells having at least one observation. In such a situation, we can only imagine what linear function of the cell means (where at least one observation was seen) '$\hat{\alpha}_1 - \hat{\alpha}_3$' estimates.

While a matrix model allows for a compact expression of just what is being estimated under different model situations, it appears to us that this simple example reveals a great deal about the mess caused by empty cells. The way out of the dilemma seems to rest upon choosing from a 'list' of occupied cells (only) linear functions of those means of possible interest. In 'balanced' experiments this presents no difficulties since the statistician can easily describe several sets of linear functions of the means which might interest the experimenter. The question of whether the statistician should do this alone, or whether he should assist the researcher in selecting his own set, may still be debated; nevertheless it appears that in 'messy data' problems, this may be the only approach which offers interpretable results.

6. SUMMARY

Linear models have become closely identified with several of the standard statistical analyses. They were introduced initially to explain the analyses; more recently they have been elevated to the role of completely describing the statistically interesting features of the experiment, thereby exerting substantial effect on the analysis itself. As linear models have become more widely utilized, they have simultaneously become less well specified. Specifically they have become 'overly parameterized' in the sense that they contain more parameters than necessary to describe the experimental context. In turn the experiment will not support their estimation. The idea of estimability was introduced to circumvent this problem.

We have proposed an alternate kind of linear model which is closely identified with the experimental context. The means of the populations sampled in the experimental context serve as its parameters. They are estimated simply and without definitional ambiguity; it supports the examination of any kind of 'mean-type' parametric characteristic. We present examples and make comparisons with the conventional approach in order to illustrate its simplicity and generality.

7. ACKNOWLEDGMENT

The views set forth here have benefited from many provocative discussions with our colleagues in the Biometrics Unit at Cornell University. Among these

L. M. Male, S. R. Searle, and E. C. Townsend were especially helpful. In acknowledging our appreciation to these people we, of course, assume full responsibility for the views set out here.

One of us (Weeks), was supported by an NSF Science Faculty Fellowship during the year (1967-68) he spent with the Biometrics Unit and worked on this material. This support is gratefully acknowledged.

8. REFERENCES

Allan, F. E. and J. Wishart [1930]. A method of estimating the yield of a missing plot in experimental work. J. Agri. Sci. 20:399-406.

Cornfield, J. and J. W. Tukey [1956]. Average values of mean squares in factorials. Ann. Math. Stat. 27:907-949.

Cotes, R. [1722]. Aestimatia errorum in mixta mathesi, per variationes partium trianguli plani et spherici. Opera Miscellania (appended to his Harmonia Mensurarum, Cantabrigiae, 1722), pp. 1-22.

Eisenhart, C. [1947]. The assumptions underlying the analysis of variance. Biometrics 3:1-21.

Eisenhart, C. [1964]. The meaning of "least" in least squares. Wash. Acad. Sci. 54:24-33.

Fisher, R. A. [1918]. The correlation between relatives on the supposition of Mendelian inheritance. Trans. Royal Soc. Edin. 52:399-433.

Fisher, R. A. [1925]. Statistical Methods for Research Workers. Oliver and Boyd, London.

Fisher, R. A. [1935]. The Design of Experiments. Oliver and Boyd, London.

Fraser, D. A. S. [1957]. Nonparametric Methods in Statistics. Wiley, New York.

Gauss, C. F. [1809]. Theory of motion of the heavenly bodies moving about the sun in conic sections. Translation of original Latin published by Dover, New York, 1963.

Gauss, C. F. [1821]. On the theory of least squares. Translation of original Latin appears in Technical Report No. 5 of Statistical Techniques Research Group, Princeton, 1957.

Graybill, F. A. [1961]. Linear Statistical Models. Vol. 1. McGraw-Hill, New York.

Irwin, J. O. [1931]. Mathematical theorems involved in the analysis of variance. J. Roy. Stat. Soc. 94:284-300.

Legendre, A. M. [1805]. On the method of least squares. Translation of original French appears in D. E. Smith (1959). A Source Book in Mathematics. Dover, New York.

Merriman, M. [1877]. Writings on least squares and related topics. Trans. Conn. Acad. Arts and Sci. 4:151-232.

Rao, C. R. [1945]. Generalization of Markoff's theorem and tests of linear hypothesis. Sankhyā 7:9-16.

Scheffé, H. [1959]. The Analysis of Variance. Wiley, New York.

Simpson, T. [1755]. On the advantage of taking the mean of a number of observations, in practical astronomy. Phil. Trans. Roy. Soc. London 49:82-93.

Tukey, J. W. [1949]. Dyadic anova, an analysis of variance for vectors. Human Biology 21:65-110.

Urquhart, N. S. [1969]. The nature of the lack of uniqueness of generalized inverse matrices. SIAM Review 11:268-271.

Yates, F. [1933]. The principle of orthogonality and confounding in replicated experiments. J. Agri. Sci. 23:108-145.

Yates, F. [1934]. The analysis of multiple classifications with unequal numbers in the different classes. J. Amer. Stat. Assoc. 29:51-66.