# ON THE TEACHING OF LINEAR MODELS AND STATISTICAL DESIGN

W. T. Federer

BU-455-M March, 1973

## Abstract

The relations of linear model theory and statistical design theory are considered via a number of examples. It is concluded that the basic ingredients for obtaining appropriate statistical analyses are the nature of the investigation, of the responses, of the various sources of variation inherent in the investigation, and of the relationships among these sources; the basic ingredients are not necessarily the linear model or the statistical design.

# ON THE TEACHING OF LINEAR MODELS AND STATISTICAL DESIGN

W. T. Federer

BU-455-M March, 1973

In some quarters linear models have become so closely identified with analyses of variance of designed experiments and surveys and for related problems that many applied and theoretical discussions begin with a linear model rather than the associated experimental or sampling situation. These quarters consider the linear model rather than the experimental or survey set-up as basic. This approach may survive chalkboard discussions but will not be universally applicable for real world experiments and surveys. It is absolutely essential to make the distinction between chalkboard and real world situations in the training of real world statisticians. However, if we wish only to train statisticians who always hide within the confines of a statistics department, it is not necessary to make this distinction.

In teaching a course on linear models or on statistical design, the instructor should be thoroughly familiar with both fields. It is not sufficient to know some theory in one or the other field. Too often statisticians become too involved with the combinatorics or with the matrix manipulations to portray any connections that might exist between linear models theory and statistical design theory in real life situations. Their chalkboard examples envelope them and blind them to the realities that an experimenter faces. Statements by statisticians who profess to be linear models theorists to the effect that "the linear model is basic so why bother with statistical design" or by those who profess to be statistical design theorists to the effect that "the statistical design is

---

Paper Number BU-455-M in the Mimeograph Series of the Biometrics Unit, Cornell University.

basic so why bother about linear model theory", are grossly misleading and anti-scholarly. A scholar is an open-minded individual who is eager to learn all about his subject and not a selected subset of it.

To dispell the notion that one or the other field contains sufficient knowledge to handle all situations consider the following examples. A situation for which the intersection of linear model theory and statistical design theory is the null set is considered first. Suppose that a balanced incomplete block design has been constructed for an experiment such that pairwise balance of the treatments has been accomplished. Then, appropriate randomization procedures have been followed in laying out the experiment. Furthermore, suppose that the response of the $i^{th}$ treatment in the $j^{th}$ block is nonlinear in the effects. Since no linear model exists here the intersection of the two fields is the null set. In fact, many statistical designs can be constructed using only combinatorial ideas and if a nonlinear response model is appropriate, then the intersection set is again the null set.

Likewise, knowing the statistical design does not imply a linear model. For example, consider the two-way nested design for which

(i) p populations were randomly selected from P populations,

(ii) $s_h$ subpopulations were randomly selected for the $h^{th}$ population, and

(iii) $n_{hi}$ individuals were randomly selected from the $i^{th}$ subpopulation from the $h^{th}$ population.

The sampling procedure is specified but what about a linear model given that one is applicable? For the above sampling situations there are at least 18 linear models for various situations. Suppose that the yield equation is of the form $Y_{hij} = \mu + \delta_h + \pi_{hi} + \epsilon_{hij}$, for $h=1,2,\cdots,p$, $i=1,2,\cdots,s_h$, $j=1,2,\cdots,n_{hi}$, $\mu + \delta_h$ = the mean of the $h^{th}$ population, for $\mu + \delta_h + \pi_{hi}$ = the mean of the $i^{th}$ subpopulation from the $h^{th}$ population, and where $\mu$, $\delta_h$, $\pi_{hi}$, and $\epsilon_{hij}$ are independent effects. The 18 models are:

| | $E(\delta_h) = \delta_h$, $E(\delta_h^2) = (\delta_h^2)$ | | | $\delta_h$ are $IID(0,\sigma_h^2)$ | | |
|---|---|---|---|---|---|---|
| | $\epsilon_{hij}$ are: | | | $\epsilon_{hij}$ are: | | |
| | $IID(0,\sigma_\epsilon^2)$ | $IID(0,\sigma_{\epsilon h}^2)$ | $IID(0,\sigma_{\epsilon hi}^2)$ | $IID(0,\sigma_\epsilon^2)$ | $IID(0,\sigma_{\epsilon h}^2)$ | $IID(0,\sigma_{\epsilon hi}^2)$ |
| $E[\pi_{hi}] = \pi_{hi}$<br>$E[\pi_{hi}^2] = \pi_{hi}^2$ | | | | | | |
| $\pi_{hi}$ are $IID(0,\sigma_\pi^2)$ | | | | | | |
| $\pi_{hi}$ are $IID(0,\sigma_{\pi h}^2)$ | | | | | | |

From the preceding two examples it should be clear that neither the sampling procedure nor the linear model is sufficient to determine the appropriate analysis for the general experiment. One must know the nature of the sampling or design procedures, the nature of the responses, and the nature of all types of variation associated with the experiment before it is appropriate to write a model and obtain the associated statistical analyses. This would appear to be a reasonable approach but evidently textbook writers do not or they would not write statistical methods, linear models, or statistical design textbooks in the manner they are written. All too often one observes students in statistics and statisticians themselves denoting an equation as a model. It is not infrequent to observe people saying that the model for the randomized complete block design is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \text{ !}$$

In other quarters the following would qualify for a "definition" of a "general linear model" for a two-way crossed classification. Let the yield of the

$ij^{th}$ observation be

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where $\mu$ is an effect common to all observations, $\alpha_i$ is the effect common to the $i^{th}$ treatment, $\beta_j$ is the $j^{th}$ block effect which is independent of other effects and the $\beta_j$ are $IID(0,\sigma_\beta^2)$, $\epsilon_{ij}$ is a random effect which is $IID(0,\sigma_\epsilon^2)$, and $E[Y_{ij}] = \mu + \alpha_i$ . As stated above there is no indication as to which one of a very large number of experimental designs the above "model" is associated.

An area in which linear model theorists get into trouble in statistical analyses is when they fail to distinguish between the following two situations:

(i) Two-way nested classification where the yield equation is said to be

$$Y_{hij} = \mu + \alpha_h + \pi_{hi} + \epsilon_{hij}$$

for $h=1,2,\cdots,a$; $i=1,2,\cdots,b_h$; and $j=1,2,\cdots,n_{hi}$ .

(ii) Two-way crossed classification in which one set of effects is either equal to zero or is set equal to zero for the following yield equation:

$$Y_{hij} = \mu + \alpha_h + \beta_i + \delta_{hi} + \epsilon_{hij}$$

for $h=1,2,\cdots,a$; $i=1,2,\cdots,b$; $j=0,1,\cdots,n_{hij}$ and for $\delta_{hi}$ being the interaction effect in a two-factor model.

The trouble arises when one considers the reduction in sums of squares of the form:

$$SS(\mu,\alpha,\beta,\delta) - SS(\mu,\alpha,\delta)$$

If no $n_{hi}$ are zero, the first sum of squares has ab degrees of freedom and the second has ab - (b-1) degrees of freedom. Some linear model theorists have observed case (ii) above, have assumed that they were in case (i), and have concluded that the above sum of squares is always zero. Their manner of computing the sum of squares may result in a zero sum of squares but the method of computing is incorrect.

In a similar situation it is sometimes assumed by statisticians and users of statistics that the estimators of effects are invariant under deletion of parameters in a linear model. For example, consider a two-way nested situation wherein there are p populations with $s_h$ subpopulations in the $h^{th}$ population from which a randomly selected sample of $n_{hi}$ observations are taken from the $i^{th}$ subpopulation of the $h^{th}$ population. For the responses obtained suppose the following model holds:

$$Y_{hij} = \mu_h + \pi_{hi} + \epsilon_{hij}$$

where $E[Y_{hij}] = \mu_h + \pi_{hi}$ and $\epsilon_{hij}$ are $IID(0,\sigma_\epsilon^2)$ . The solutions for $\mu_h$ that minimize the sum of squares

$$\sum_{h=1}^{k} \sum_{i=1}^{s_h} \sum_{j=1}^{n_{hi}} \left(Y_{hij} - \mu_h\right)^2$$

are not the same as those that minimize the following sum of squares

$$\sum_{h=1}^{p} \sum_{i=1}^{s_h} \sum_{j=1}^{n_{hi}} \left(Y_{hij} - \mu_h - \pi_{hi}\right)^2 ,$$

unless $n_{hi}$ = a constant and the $s_i = s$ . When the $n_{hi}$ and the $s_i$ are unequal, the estimators for the difference $\hat{\pi}_h - \hat{\pi}_{h'}$, $h \neq h'$, will be different from the two sets of minimizations.

Probably the area in which statisticians and users of statistics fail most often to understand design concepts is for the split-plot and split-block designs. Failure to comprehend design principles and techniques leads to incorrect statistical analyses. For example, consider the usual split-plot example wherein the levels of one factor form the a whole plot treatments and the levels of the second factor form the b split-plot treatments. For the whole plot treatments

designed in a randomized complete block design with r blocks and for the split-plot treatments randomly allocated to the experimental units within each whole plot, there are r randomizations on the a whole plot treatments and ra randomizations on the split plot treatments. For experiments designed in this fashion, the whole plot and split plot treatments are usually subject to different error variances. A common mistake is to consider the experiment as a three-factor (blocks, whole plot treatments, and split-plot treatments) factorial and proceed as if all contrasts were subject to the same error variance.

A second situation wherein the confounding aspects of the split plot design are misunderstood is when repeated measurements are made on an experiment designed as a randomized complete block (or other design) design. The repeated measurement may be made through t time periods, by j judges, by m methods, etc. As far as the judges are concerned there in only one randomization if one judge scores the entire experiment and the remaining judges do likewise. Frequently, statistical analysts consider the judges as the split plot treatments since they may tabulate the results in the manner they would for a split plot treatment. They follow a similar incorrect procedure for measurements taken over several time periods. Since the time periods are unreplicated, how could they possibly be considered as split plot treatments?

An area in which statistical design people have closed their eyes to other than a single yield equation is for incomplete block designs both with and without recovery of interblock information. In many experimental situations there is the possibility of a block-treatment interaction. Who treats this in general? A paper submitted on this topic several years ago was first accepted and then later rejected because it "lacked enough engineering content". Was the real reason for rejection the fact that this subject might rock the boat for experimental design analysts?

In summary, it is considered that neither the linear model nor the statistical design is the basic ingredient for statistical analyses. Instead the nature of the investigation of responses, and of the various types of variation are considered to be the basic ingredients in selecting appropriate statistical analyses for experimental data. Statistics needs to be more realistic with much less emphasis on the chalkboard variety of teaching statistics. If statisticians are to take the lead in model building, and I believe they should, it will be necessary to leave the chalkboard and return to basic facts about the phenomenon under study.