MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 11, NUMBER 3 SUMMER 1982

TEACHING APPLIED RESEARCHERS TO CREATE THEIR OWN STATISTICAL MODELS

Joe H. Ward, Jr.
Brooks Air Force Base, Texas

Earl Jennings
The University of Texas at Austin

The purpose of the following remarks is to give you something of the flavor of a novel approach to the teaching of statistical model building and manipulation. Historically, it evolved out of an applied environment in which many of the classical models appeared to be inadequate or at least deficient in one or more respects. Students in applied areas who have been exposed to the approach respond enthusiastically to it, and, in general, the more "traditional" work they have had, the greater their enthusiasm. The response of teachers has been mixed. Many of the critics make remarks similar to those criticisms that are directed at the "new math." It is certainly accurate to state that students of this approach get very little practice in arithmetic for even the most elementary models. In fact, the primary text [6] is almost totally devoid of computing formulae.

With respect to mathematical and statistical foundations, we rely very heavily on the theory of the classical fixed-x linear model, and the text bears some superficial resemblance to a typical text on linear models. However, a great deal of the material covered in a typical linear models text will be

Presented at AERA 1982, MLR Special Interest Group
Not refereed by editorial staff

found in ours only indirectly, if at all. Conversely, the concepts we identify and the skills we try to develop are only indirectly inferable from the typical text.

In general, our approach has the following characteristics:

1. A technical vocabulary of minimal length.
2. Very few special symbols and computational formulae. In those places where a new special symbol or formula would ordinarily be introduced, we make every effort to identify the concept as a special case of a more general concept and the formula as a special case of a more general formula. The cumulative effect of this is, we believe, a hierarchical structuring of the content that enhances learning. See Appendix A for an example of the way we summarize the models of one-way analysis of variance, a test for non-linearity, and simple regression analysis, and Appendix B for a summary of a two-factor problem. Students are assumed to have access to a computer, so very little arithmetic is required.
3. An emphasis on the idea that a model is a way of formalizing an argument.
4. Practice in translating natural language into models with unambiguously specified properties. The kind of skill required to do this is similar to the skill required to translate elementary algebra "word problems" into algebraic equations.
5. Extensive practice in the algebraic manipulation of models. This skill is frequently necessary to create an assumed model with specified properties and almost always required to produce a restricted model that can be used in tests of hypotheses about the parameters of the

assumed model. Although the amount of algebra required is burdensome for some models, the level of skill required is minimal.

Some of the features of the approach can best be understood by an example. Suppose we were interested in evaluating the differential effects of two different methods of teaching reading in the second grade. Students are randomly assigned to the two conditions. A measure of reading achievement is obtained before instruction begins, and another measure is obtained at the end of instruction. Because girls tend to read better at this age than boys, we can probably increase the precision of our estimations and the power of our tests by considering sex in the model. Moreover, there is a possibility that sex might interact with teaching method, initial performance, or both.

Ultimately, we are going to argue that if we can reject the hypotheses

$$E(1, \text{boy}, x) = E(2, \text{boy}, x)$$

$$E(1, \text{girl}, x) = E(2, \text{girl}, x)$$

we are in a position to conclude that the methods are not equally effective. Stated in prose, the hypothesis is that the expected posttest performance for a Method 1 boy with initial performance x is the same as the expected posttest performance for a Method 2 boy with the same initial performance, x. A similar statement is made for girls, and x takes on all possible values of initial performance. Suppose the potential range of x is 20 to 80. We seek a model that will produce 2 (methods) X 2 (sexes) X 61 (values of x) = 244 estimates of expected values. If we are not willing to make any simplifying assumptions about the relationships among the expected values, we need a model with 244 parameters, which we refer to as the mutually exclusive categorical model. Fortunately, in this problem, it seems reasonable to assume that the expected difference in posttest performance per unit difference in initial performance

is constant (sometimes called the linearity assumption), although perhaps a different constant for each of the four groups. If this assumption is true, then the 244 expected values are expressible as a function of only eight parameters. In the text, we discuss ways of investigating the tenability of this assumption. Although there are an infinite number of ways of parameterizing a model to estimate the eight parameters, one with intuitive appeal is

$$Y = a_1B^{(1)} + a_2B^{(2)} + a_3G^{(1)} + a_4G^{(2)} +$$

$$c_1(X * B^{(1)}) + c_2(X * B^{(2)}) + c_3(X * G^{(1)}) +$$

$$c_4(X * G^{(2)}) + E^{(1)}$$

where

Y is a column vector of dimension n containing the observed posttest scores.

$B^{(i)}$ is a column vector of dimension n containing a one if the corresponding value in Y was observed on a boy in method i; zero otherwise. (i = 1,2)

$G^{(i)}$ is defined for girls similar to $B^{(i)}$ for boys.

X is a column vector of dimension n containing pretest scores arranged in the same order as Y.

The a's and c's are unknown scalars, and $E^{(1)}$ is an unknown column vector. A least squares solution to Model 1 might produce values that could be represented as in Figure 1.

The a's are the intercepts and the c's the slopes of the four separate straight lines. They are also estimates of the eight parameters which are assumed to yield the expected values. We could proceed to investigate our

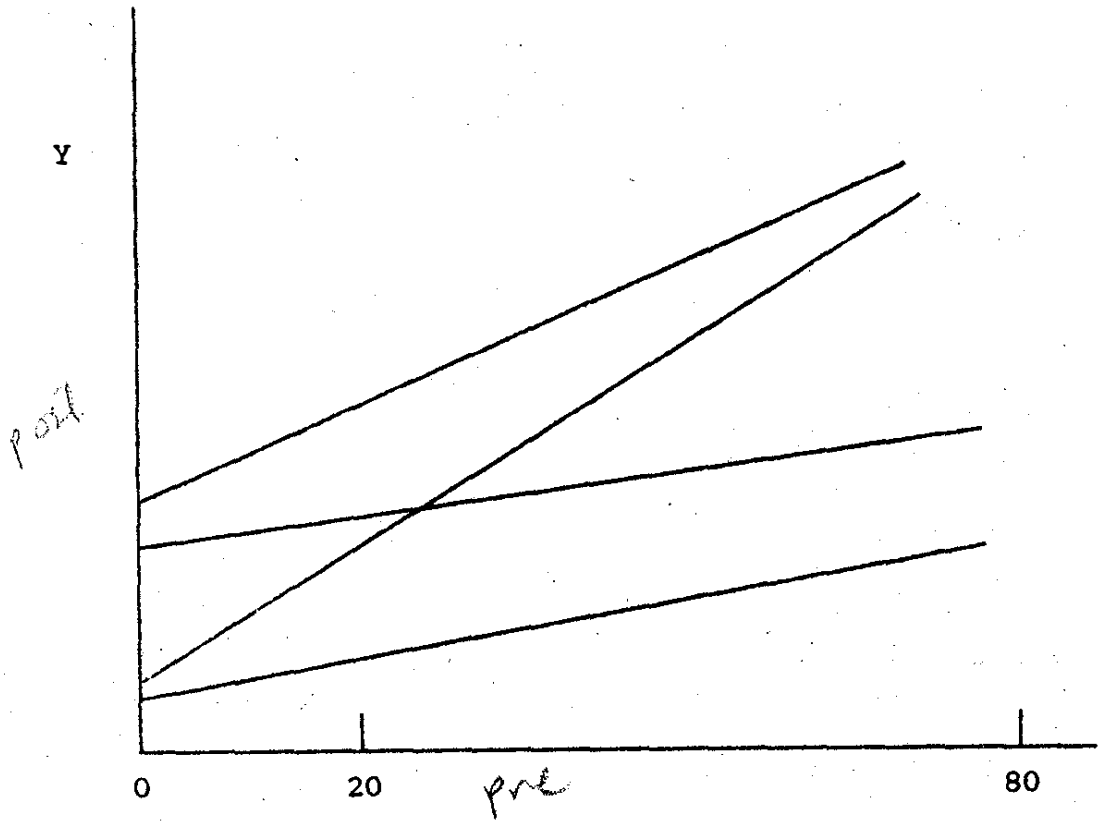

Figure 1. Possible results for Model 1.

ultimate hypothesis using Model 1 as an assumed model. However, such a test based on the F distribution would involve four degrees of freedom in the numerator and would not produce an unqualified recommendation with respect to method.

This kind of problem is frequently approached in standard methods by a factorial analysis of covariance in which the assumed model is a subspace of Model 1 incorporating the assumption that each c is an estimate of the same parameter. This assumption is frequently referred to as the homogeneity of regression assumption. If this assumption is true, then the 244 expected values are expressible in terms of only five parameters. A model to estimate these parameters is

$$Y = a_1 B^{(1)} + a_2 B^{(2)} + a_3 G^{(1)} + a_4 G^{(2)} + cX + E^{(2)}$$

A least squares solution to Model 2 might be represented as in Figure 2.

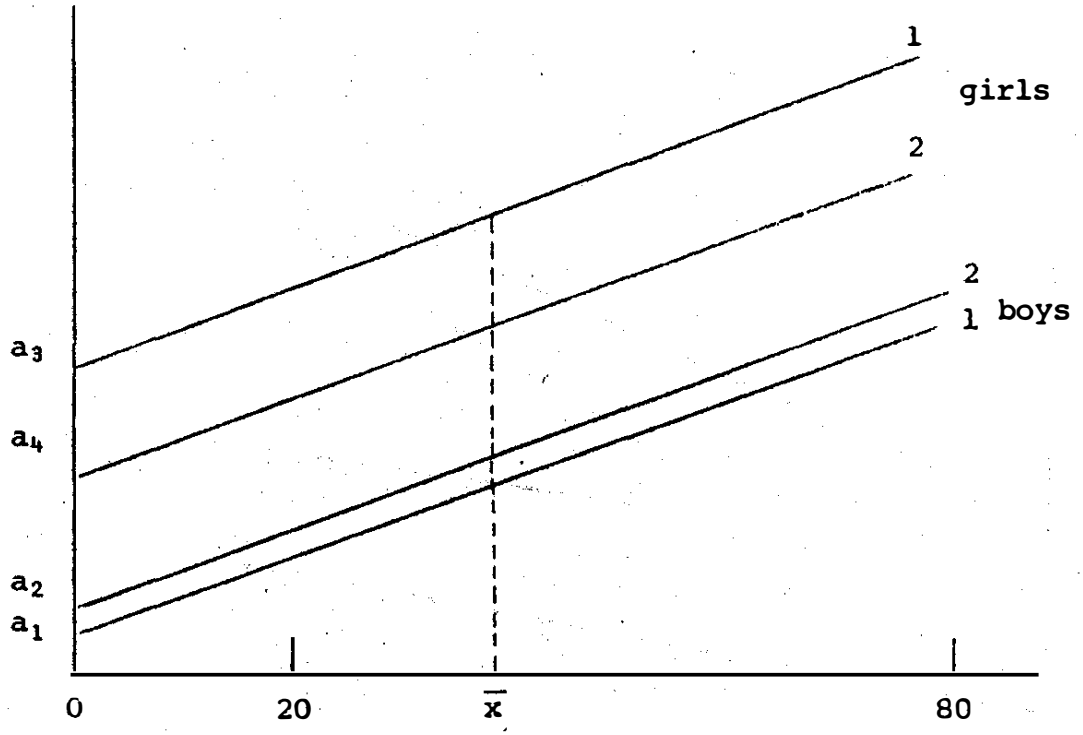

Figure 2. Possible results for Model 2.

In Model 2, the a's are the intercepts of the four lines in Figure 2, and c is the common slope. The test for "treatment effect" involves a comparison of what are called the "adjusted means," namely

$$\frac{(a_1 + c\bar{x}) + (a_3 + c\bar{x})}{2} = \frac{(a_2 + c\bar{x}) + (a_4 + c\bar{x})}{2}$$

which simplifies to $a_1 + a_3 = a_2 + a_4$.

A sufficiently large non-zero difference leads to a relatively large F, a rejection of the hypothesis, and the conclusion that the methods differ. Such a conclusion seems defensible, but we are still not in a position to make an unqualified recommendation with respect to method. In Figure 2, $a_1 + a_3$ is greater than $a_2 + a_4$, yet the available data seem to suggest that Method 1 is better for girls and Method 2 is better for boys.

A number of possibilities exist to reduce this ambiguity. The standard covariance sex by method interaction test is relevant information, but it does not directly address the issue. We could conduct pair-wise investigations

($a_1 = a_2$ and $a_3 = a_4$) and suffer the problems of an increased experimentwise Type I error rate or adopt some post hoc test and suffer the consequent loss of power.

An alternative is to consider an assumed model that avoids the ambiguity altogether. For example, if we are willing to assume the following relationships among the expected values

$$E(1, \text{boys}, x) - E(2, \text{boys}, x) =$$
$$E(1, \text{girls}, x) - E(2 \text{ girls}, x)$$

and

$$E(1, \text{boys}, x_1) - E(2, \text{boys}, x_1) =$$
$$E(1, \text{boys}, x_2) - E(2, \text{boys}, x_2)$$

and

$$E(1, \text{girls}, x_1) - E(2, \text{girls}, x_1) =$$
$$E(1, \text{girls}, x_2) - E(2, \text{girls}, x_2)$$

where

$$x, x_1, x_2 = 20, 21, \ldots 80 \qquad x_1 \neq x_2$$

the 244 expected values are expressible as a function of only five parameters as in Model 2, but because we are making different assumptions, the model we create will have different properties than Model 2. The skills required to create a model that incorporates the desired assumptions are identical to the skills required to test the assumptions. Involved is a simple substitution for the expected values above, their estimates in symbolic form from Model 1, and an algebraic simplification that results in three implied restrictions. Substituting the symbolic estimates from Model 1 for the expected values above,

$$a_1 + c_1x - a_2 - c_2x = a_3 + c_3x - a_4 - c_4x \qquad (1)$$

$$a_1 + c_1x_1 - a_2 - c_2x_1 = a_1 + c_1x_2 - a_2 - c_2x_2 \quad (2)$$

$$a_3 + c_3x_1 - a_4 - c_4x_1 = a_3 + c_3x_2 - a_4 - c_4x_2 \quad (3)$$

Equation (2) can be simplified to

$$(c_1 - c_2)(x_1 - x_2) = 0$$

Since $x_1 \neq x_2$, c_1 must equal c_2, and they can be given a common name.

$$c_1 = c_2 = b, \text{ a common value} \quad (4)$$

Similarly, Equation (3) can be simplified to

$$(c_3 - c_4)(x_1 - x_2) = 0$$

implying

$$c_3 = c_4 = g, \text{ a common value} \quad (5)$$

Substituting (4) and (5) into (1), we achieve

$$a_1 + bx - a_2 - bx = a_3 + gx - a_4 - gx$$

which can be written

$$a_1 - a_2 - a_3 + a_4 = 0 \quad (6)$$

a_1 through a_4 can be renamed so that they satisfy (6) as follows:

$$\begin{aligned} a_1 &= d_1 \\ a_2 &= d_1 + d_3 \\ a_3 &= d_2 \\ a_4 &= d_2 + d_3 \end{aligned} \quad (7)$$

In effect, we have renamed the eight parameter estimates in Model 1 in terms of only five names: d_1, d_2, d_3, b, and g.

If the new names are substituted in Model 1, we get

$$Y = d_1B^{(1)} + (d_1 + d_3)B^{(2)} + d_2G^{(1)} + (d_2 + d_3)G^{(2)} +$$

$$b(X * B^{(1)}) + b(X * B^{(2)}) + g(X * G^{(1)}) +$$

$$g(X * G^{(2)}) + E^{(3)}$$

Expanding and simplifying yields

$$Y = d_1(B^{(1)} + B^{(2)}) + d_2(G^{(1)} + G^{(2)}) + d_3(B^{(2)} + G^{(2)}) +$$

$$b(X * B^{(1)} + X * B^{(2)}) + g(X * G^{(1)} + X * G^{(2)}) + E^{(3)}$$

A least squares solution to Model 3 might appear as in Figure 3.

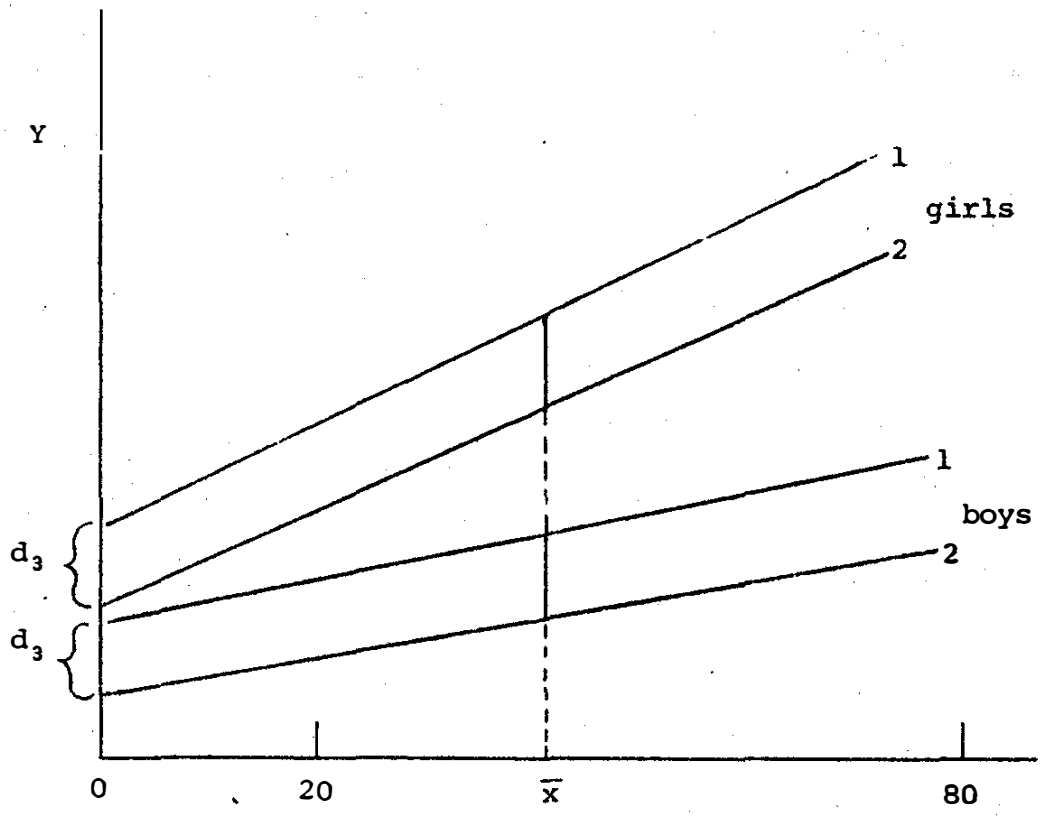

Figure 3. Possible results for Model 3.

The essential property of Model 3 for our purpose is that the expected difference between any pair of persons having the same sex and initial performance, differing only in the method of instruction, is estimated by the same constant, namely d_3. When the properties of a model are not immediately obvious by inspection, we encourage the practice of verifying that the model has the claimed properties. This involves writing the symbolic expressions that estimate the expected values and verifying that the symbolic expressions are related as the expected values are assumed to be, as shown in Table 1.