Teaching Researchers to Create Linear Models to Solve Practical Problems

Joe H Ward Jr, Stanley B Polk and William E Alley - Texas, USA

1. Background

Over thirty years ago the need arose at the United States Air Force Personnel Research Laboratory to answer many research questions for which standard statistical analysis procedures were not appropriate. It was observed that many researchers had been trained to force research problems into standard "cookbook" procedures. Figure 1 depicts a typical researcher who would search through a collection of statistical procedures and attempt to fit the problem of interest into one of the available analysis methods. Occasionally, a researcher would change the research question to fit a familiar statistical "cookbook" model.

When it became apparent that many practical problems could not be solved by the standard procedures, a developmental effort began which combined the conceptual power of a prediction (e.g. regression or general linear) model approach with the computational power of high speed computers. A series of short courses was developed to introduce researchers to the new capabilities brought about through the combination of effective application of regression models with computing techniques.

After several experimental courses were introduced, the approach was documented in 1963 by Bottenberg and Ward in *Applied Multiple Linear Regression.* The approach allowed researchers to create statistical models to fit the questions of interest, rather than attempting to fit the problems into existing procedures. With the power to solve large-scale regression problems the researcher no longer is concerned with "matching experimental subjects", "equating cell frequencies"; and can deal systematically with "missing cells" and other problems that can arise from the use of "standard" procedures.

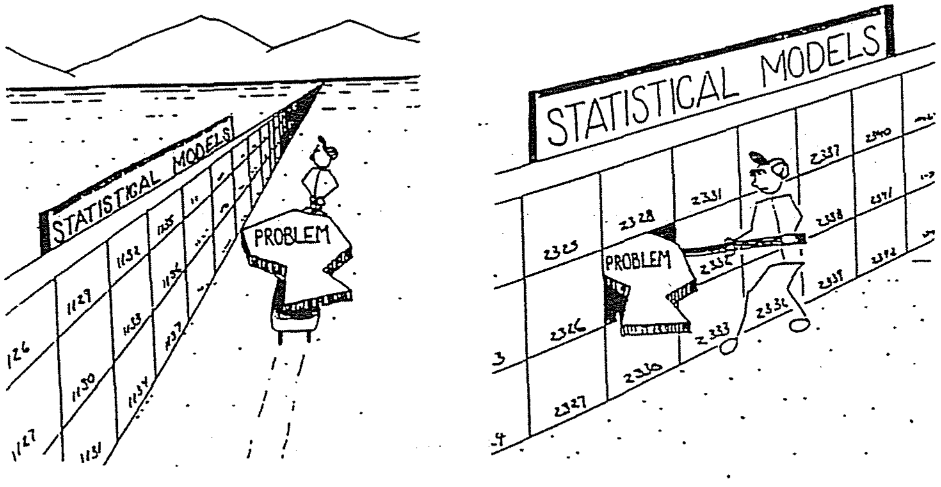

FIGURE 1
Forcing problems into cookbook models

2. An outline of the approach

First we list some of the main characteristics of the approach. These are:

(i) A minimum amount of technical vocabulary and computational formulas. Each problem is viewed as a prediction (i.e. regression or linear model) problem.
(ii) Practice in translating natural language research problems into statistical models with specific properties needed to answer questions of interest.
(iii) Practice in algebraic manipulation of models that is necessary to create an Assumed Model and to obtain a Restricted Model to test hypotheses about parameters of the Assumed Model. Section 3 contains simple examples of model development activities to illustrate a few objectives of the approach.
(iv) Many analysis procedures traditionally taught as unrelated topics are more efficiently acquired when conceptualised under one general approach.
(v) Researchers can acquire capability to solve a wider range of research questions than with a traditional approach.

The steps that are used in the analysis of problems are:

(i) State the research problem or question in natural language.
(ii) Translate the natural language statements into statements (i.e. hypotheses) about Expected Values (Means) in model-independent form.
(iii) Create an Assumed Model by expressing the Dependent Variable (i.e. Response or Outcome) as a function of Independent Variables plus Error.
(iv) Investigate the properties of the Assumed Model to be sure that it allows for answers to questions of interest.
(v) State the restrictions on the parameters of the Assumed Model that are implied by the hypotheses stated in Step (ii).

(vi) Impose the restrictions identified in Step (v) on the Assumed Model created in Step (iii) to produce a Restricted Model.

(vii) Use statistical computing software to obtain the appropriate numerical results needed to answer the questions of interest (e.g. compare error sums of squares of Assumed and Restricted Models to test hypotheses and obtain confidence intervals).

(viii) State the conclusions in natural language.

The instructional approach has been elaborated with applications to numerous problems by Ward and Jennings (1973) in *Introduction to Linear Models*. The following instructional sequence has been used to develop skills in creating and manipulating models:

(i) Introduction to the use of "ordered list of numbers" (vectors) for creating prediction models of the form

Y (dependent vector) = Linear combination of X's (independent vectors)
+ E (an Error vector)

Using the prediction (regression, linear model) approach to investigate questions about:

(ii) One or more means from one categorically-coded predictor attribute. These problems interrelate one-factor ANOVA with t-tests about one or two means.

(iii) Several means created from two or more categorically-coded predictor attributes. These problems interrelate multi-factor ANOVA ("main-effects", "interaction effects") and illustrate analyses involving "missing cells" and other "non-standard" problems.

(iv) Problems in which the means might be related in special ways such as "linear" or "higher-degree polynomial" relationships. These problems consider tests for "non-linearity", "discontinuity", "non-linear covariance analysis", and other examples involving unusual relationships among means created from several attributes.

Over the years, researchers educated as Research Psychologists, Education and Training Researchers, Operations Researchers, Econometricians, Statisticians, and Computer Scientists, have worked at the Personnel Research Laboratory (now called the Human Resources Laboratory). It has been observed that there has been little change in the capability of these researchers to effectively combine the power of a general regression model approach and statistical computing procedures. Consequently, as new personnel arrive they are introduced to the more general approach. This provides researchers powerful capability to attack new problems and to effectively communicate with each other about their research projects.

Short courses have continued at the Human Resources Laboratory, at other research locations in the Air Force, at pre-sessions for the American Educational Research Association, and at several universities throughout the United States.

Extensive use of general regression models and statistical computing procedures has extended into the operations research programme of the Strategic Plans and Analysis

Division of the United States Air Force Commissary Service Headquarters. This capability brings a new dimension to the efforts of the Headquarters to increase the contribution of Commissaries to the Air Force mission.

Figure 2 depicts the "born-again" researcher who has the capability to create and analyse a model appropriate to the problem of interest.

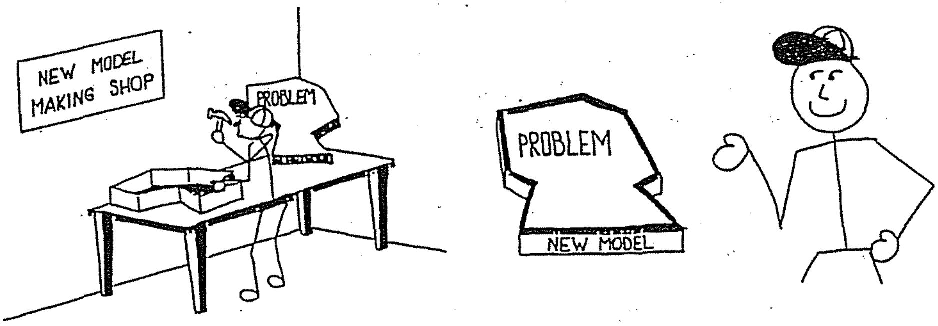

FIGURE 2
Building a new model to fit the problem

3. Examples of model development to test hypotheses

The following simple examples illustrate instructional objectives that allow applied researchers to create and manipulate models. Note that in this presentation the same symbols are used to represent the unknown parameters of a statistical model and the least squares estimates of the parameters. The hypothesis statements refer to the unknown parameters, which are frequently represented by Greek letters.

Consider a problem involving the typing performance of students who are in grades 9, 10, 11, and 12. We will create models expressing typing performance as a function of grade in school using the following vector definitions:

Y = a vector of dimension n in which each element is a typing-test score from a student;
X_9 = 1 if corresponding element of Y is from a 9th grader; 0 otherwise;
X_{10} = 1 if corresponding element of Y is from a 10th grader; 0 otherwise;
X_{11} = 1 if corresponding element of Y is from a 11th grader; 0 otherwise;
X_{12} = 1 if corresponding element of Y is from a 12th grader; 0 otherwise;
U = 1 for each element;
G = student's grade (9, 10, 11, 12);
E_i = an error vector for model i.

The four different patterns of the elements of X_9, X_{10}, X_{11}, X_{12}, U and G are shown below:

X_9	X_{10}	X_{11}	X_{12}	U	G
1	0	0	0	1	9
0	1	0	0	1	10
0	0	1	0	1	11
0	0	0	1	1	12

Problem 1. Consider the hypothesis that there are no differences between the expected (mean) typing performance for the four grades. The "model independent form" of this hypothesis may be written as:

Expected value of Y for 9th graders = Expected value of Y for 10th graders
= Expected value of Y for 11th graders = Expected value of Y for 12th graders

Assumed Model 1.1A. A simple equation that can be used to express this hypothesis and to obtain least-squares estimates of the parameters is:

$$Y = a_9X_9 + a_{10}X_{10} + a_{11}X_{11} + a_{12}X_{12} + E_{1.1A}$$

and the hypothesis can be expressed in terms of Model 1.1A as:

$$a_9 = a_{10} = a_{11} = a_{12} = \text{a common value } a_0.$$

Substituting the value a_0 in the Assumed Model 1.1A gives

Restricted Model 1.1R

$$Y = a_0U + E_{1.1R}$$

and the Sum of Squares of Errors (SSE1.1A and SSE1.1R) can be compared to test the hypothesis using the F statistic:

$$F = \frac{(\text{SSE1.1R - SSE1.1A})/(4-1)}{(\text{SSE1.1A})/(n-4)}.$$

Some statistical software systems can directly impose these restrictions on the assumed model and perform the test, relieving the analyst from creating the restricted model and comparing the error sums of squares. Alternatively, it may be necessary to reformulate the model so that the hypothesis can be tested by "setting coefficients equal to zero". This can be done by rewriting the model as

$$Y = b_0U + b_9X_9 + b_{10}X_{10} + b_{11}X_{11} + E_{1.2A} \quad [E_{1.2A} = E_{1.1A}]$$

with the hypothesis $b_9 = b_{10} = b_{11} = 0$. Note that this step can be done only *after* the researcher's hypothesis is known.

Problem 2. Consider the hypothesis that there are *constant* differences between the expected (mean) typing performance for adjacent grades. That is, the difference between the expected performance of 9th and 10th graders is equal to the difference between the expected performance of 10th and 11th graders and is equal to the difference between the expected performance of 11th and 12th graders. This is frequently referred to as a test for "non-linearity". The "model independent form" of this hypothesis may be written as:

(Expected Y for 9th graders) - (Expected Y for 10th graders) =
(Expected Y for 10th graders) - (Expected Y for 11th graders) =
(Expected Y for 11th graders) - (Expected Y for 12th graders)

The most direct approach is to use *Assumed Model 2.1A* which is identical to *Assumed Model 1.1A*. The hypothesis can be expressed in terms of Model 1.1A as:

$$a_9 - a_{10} = a_{10} - a_{11} = a_{11} - a_{12} = \text{a common value } w_1;$$

and imposing the restrictions on Assumed Model 2.1A gives

Restricted Model 2.1R

$$Y = w_0 U + w_1 G + E_{2.1R}$$

and the Sum of Squares of Errors (SSE2.1A and SSE2.1R) can be compared to test the hypothesis using the F statistic.

As before, we may have statistical software systems that can impose the parameter restrictions on the assumed model and perform the test, or we may have to reformulate the model so that the restricted hypothesis corresponds to setting some parameters equal to zero, a task which in this case we leave to the reader.

In summary, it is important for an applied researcher to work with an assumed model that has properties that simplify expression of the hypothesis of interest in terms of the parameters of the assumed model. Once the researcher has identified the parameter restrictions, then the restricted model can be generated and the restrictions can be imposed by an appropriate statistical software system.

References

Bottenberg, Robert A and Ward, Joe H Jr (1963) *Applied Multiple Linear Regression.* PRL-TDR-63-6, AD-413 128, Personnel Research Laboratory, Lackland AFB, TX.

Ward, Joe H Jr and Jennings, Earl (1973) *Introduction to Linear Models.* Prentice-Hall, Englewood Cliffs, NJ.