SYNTHESIZING REGRESSION MODELS -
AN AID TO LEARNING EFFECTIVE PROBLEM ANALYSIS

by
Joe H. Ward, Jr.
Southwest Educational Development Laboratory

Regression models can be used to assist in the analysis of a wide variety of problems. However, the power of regression models is not widely utilized. There are two major reasons for the lack of use of general regression models. First, there have been too few attempts by teachers to develop the behaviors in students that are necessary to effectively create models appropriate to the particular problem of interest. Second, many of the models that should be utilized for a particular problem require the use of a computer, but many research workers do not have effective software systems to facilitate communication with the computer.

These two problems can be helped by 1) providing an instructional system that will develop in students the capability of defining regression models appropriate to their problems of interest; and 2) providing computational software that facilitates the analysis by a high speed computer.

Even though both of the above areas are important, the first - defining appropriate models - is the most important and difficult behavior to bring about in research workers. The following presentation will be devoted to the discussion of several aspects of this problem. First, a few general comments will be made about the general problem of teaching (and learning) techniques of model generation. This will be followed by a specific example of an instructional approach--a description of a synthesis of several different regression models.

The Generation of Models

Some of the intermediate behaviorial objectives that lead to effective model generation are:

1. A research worker should be able to define the vector of interest (i.e., dependent vector) appropriate to his problem. This is developed by extensive practice on problems with increasing difficulty.

2. A research worker must develop the capability of expressing his vector of interest (call it Y), as a linear combination of appropriately defined vectors (call them X(1), X(2), ..., X(k)) plus an error vector (call it E). Extensive practice in defining vectors is required to develop the desired capability. The research worker should think "I need to find 'another name for Y' so that the statements that I make about this 'other name' will be relevent to my problem." A student should have extensive practice in defining those vectors which are to be used in the "renaming " of Y.
3. After the vector of interest (Y) has been expressed as a linear combination of the new vectors (X(1), X(2), ..., X(K)) plus an error, the research worker can then make statements (or hypotheses) about "expected" or "predicted" values of Y. This involves the translation of the research question from natural language into the language of the mathematical representation. This translation process is sometimes quite difficult, and the student should have extensive practice, using simple problems in the beginning.
4. The translation of statements about the model leads to the imposition of restrictions on the model. The student should impose his restrictions on the model to determine the effects of the restrictions on the error. It is sometimes helpful to view these restrictions as the "giving up" of information.
5. After the student has imposed the restrictions it is important that the student verify that the restricted model actually does possess the properties imposed by the restrictions. This serves as a check for the student's substitution. It also provides frequent insights into previously unrecognized properties of various models.

A Synthesis of Regression Models

The following illustration is designed to show the idea that is common to four regression models that are often treated quite separately in our instructional programs. The basic problem of interest is the same in all four models; however, the models appear quite different due to the differences in the original assumptions that were made for the four models.

For our example, we consider four different research workers who are studying the effects that different amounts of practice have on typing proficiency. Furthermore, there is some concern by these research workers for the possible "contaminating" effect of the age of the students on the research results. Each research worker feels that something should be done to "hold age constant" or "take out the effects of age." However, a couple of the researchers aren't quite sure what they must do to "take out the effects."

Now the first research worker was located at a university (ANOVA U.) where there was strong emphasis on analysis of variance--with very little instruction in covariance analysis, or multiple correlation and regression. And this research worker was particularly fond of the "two-factor design." The second researcher was at a university (COVARIA U.) which to no one's surprise was really strong on covariance analysis. This school had a complete course in covariance analysis to stress its importance. The third worker received his education many years ago at a university (MULCOR U.) that had only taught multiple correlation and regression analysis. The analysis of variance and covariance course was started the year after he completed his statistics course.

The fourth researcher had attended a university (VARICO U.) that had stressed a slightly different approach which they described as "a sort of reverse covariance analysis" which they have named VARICO ANALYSIS.

All four of these research workers have conceptualized a common problem. First, they are all interested in studying the effect of practice on typing proficiency while "controlling" or "taking out the effects" of age. Furthermore since they are dealing with the age information they all wish to test for interaction since it may be that the effects of amount of practice is different for students of different ages. All four are interested in, first, testing for interaction, and then if there is no interaction they will test for the effect of amount of practice.

Even though these four research workers have a common problem, each one would probably perform quite different analysis because of the varied educational emphasis. Also, they might each argue that they are doing quite different analyses. These analyses appear even more different because the computational procedures appear quite different.

The four different approaches will be presented below in a form that will emphasize that there are basic ideas common to all.

Approach 1 - (ANOVA U.)

The research worker at ANOVA U. wishes to think of his problem as a "two-factor design."

We assume for all problems that there are observed 15 levels of practice (16,17,..., 30 hours) and that there are 5 ages (14,15,..., 18 years).

We define the following vectors:

Y = a vector containing the typing proficiency scores of the n students in the study.

X(i,j) = a set of vectors with elements defined as 1 if the corresponding element of Y is observed from a person with practice hours i, and age j; and 0 otherwise, (i = 16,17,...,30),(j = 14,15,...,18)

Notice that if some X(i,j) vectors are null they are not included in the model.

E = a residual vector

Then the full model is

$$Y = \sum_{i}\sum_{j} a_{ij} X(i,j) + E$$

or in extended form

$$(1) \qquad \begin{aligned} Y = {} & a_{16,14} X(16,14) + a_{16,15} X(16,15) + \ldots + a_{16,18} X(16,18) + \\ & a_{17,14} X(17,14) + a_{17,15} X(17,15) + \ldots + a_{17,18} X(17,18) \\ & \quad \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ & + a_{30,14} X(30,14) + a_{30,15} X(30,15) + \ldots + a_{30,18} X(30,18) + E \end{aligned}$$

In this discussion different symbols will not be used to distinguish between the unknown parameters and their least squares estimators. In the model above the symbols a_{ij} will be used to represent both the unknown parameters and the estimators.

Consider four different students having the following characteristics:

Student	Hours of practice	Age in years
1	r	p
2	s	p
3	r	q
4	s	q

Figure 1

Extended Form of Vectors of Model (1)

Y	X(16,14)	X(16,15) ...	X(30,14)	X(30,15) ...	X(30,18)
$y_{16,14,1}$	1	0	0	0	0
$y_{16,14,2}$	1	0	0	0	0
$y_{16,14,3}$	1	0	0	0	0
$y_{16,15,1}$	0	1	0	0	0
$y_{16,15,2}$	0	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
$y_{30,14,1}$	0	0	1	0	0
$y_{30,15,1}$	0	0	0	1	0
$y_{30,15,2}$	0	0	0	1	0
$y_{30,15,3}$	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
$y_{30,18,1}$	0	0	0	0	1
$y_{30,18,2}$	0	0	0	0	1
$y_{30,18,3}$	0	0	0	0	1

Predicted (or expected) value for an individual who practiced 16 hours and who is 15 years old.

$$E(16,15) = (a_{16,14} * 0) + (a_{16,15} * 1) + (a_{16,16} * 0) + \dots + (a_{30,18} * 0)$$

$$E(16,15) = a_{16,15}$$

The hypothesis of no interaction can be stated as follows:

The difference between expected (or predicted) typing performance of the two students at age p but with different practice levels r and s is equal to the differences between the expected (or predicted) typing performance of the two students at age q but with different practice levels r and s. This is hypothesized for all values of p, q, r, and s where $p \neq q$ and $r \neq s$.

Calling the four expected values E(rp), E(sp), E(rq), and E(sq) <u>the hypothesis of no interaction is</u>

$$E(rp) - E(sp) = E(rq) - E(sq)$$

Now in the model employed by the ANOVA U. research worker.

$$E(rp) = a_{rp}$$

$$E(sp) = a_{sp}$$

$$E(rq) = a_{rq}$$

$$E(sq) = a_{sq}$$

Then we see that the <u>hypothesis of no interaction is</u>

$$a_{rp} - a_{sp} = a_{rq} - a_{sq} \qquad \begin{array}{l} p \neq q \\ q = 14, \ldots, 18 \\ r \neq s \\ s = 16, \ldots, 30 \end{array}$$

When we impose these 56 restrictions on the model the restricted model can be written as

(2) $$Y = a_{16}\, X(16) + a_{17}\, X(17) + \ldots + a_{30}\, X(30)$$

$$+ b_{14}\, Z(14) + b_{15}\, Z(15) + \ldots + b_{17}\, Z(17) + R$$

where

X(i) = 1 if the corresponding element of Y is from a student who practiced i hours; 0 otherwise (i = 16,17,...,30)

Z(j) = 1 if the corresponding element of Y is from a student having age j.

R = the residual vector

a_i and b_i = unknown coefficients

Notice that Z(18) is not included in this model since it is a linear combination of the other vectors.

Let

$q_1 = \sum_i E_i^2$, the sum of squares of errors in the full model

$q_2 = \sum_i R_i^2$, the sum of squares of the error in the restricted model

Then if the F statistic is desired to test the hypothesis we have

$$F = \frac{(q_2 - q_1)/(75-19)}{q_1/(n - 75)}$$

Now we will consider the situation in which the no-interaction hypothesis has been accepted as true.

Then we use the model

$$Y = a_{16}\ X(16) + a_{17}\ X(17) + \ldots + a_{30}\ X(30)$$
$$+ b_{14}\ Z(14) + b_{15}\ Z(15) + \ldots + b_{17}\ Z(17) + R$$

The next hypothesis (the effects of practice) is that the difference between the expected typing performance for two students at the same age p but who have practiced different amounts r and s is equal to zero. This must be true for all ages.

Then we consider the two expected values E(rp) and E(sp).

The hypotheses is

$$E(rp) - E(sp) = 0$$

Now in the above model we find that

$$E(rp) = a_r + b_p$$
$$E(sp) = a_s + b_p$$

Then we see that <u>the hypothesis is</u>

$$(a_r + b_p) - (a_s + b_p) = 0$$
$$a_r - a_s = 0 \qquad r \neq s$$
$$r,s = 16,17,\ldots,30$$

When we impose these 14 restrictions on the model the new restricted model can be written as

(3) $$Y = b_{14}\ Z(14) + b_{15}\ Z(15) + \ldots + b_{18}\ Z(18) + G$$

Notice that this restricted model has no information to distinguish amounts of practice; i.e., we have given up the information about differences in amounts of practice.

Let $q_3 = \Sigma G_i^2$, the sum of squares of the error in the new restricted model.

Then if desired we have

$$F = \frac{(q_3 - q_2) / (19-5)}{q_2 / (n-19)}$$

Approach 2 - (COVARIA U.)

Since the research worker from COVARIA U. likes to do covariance analysis it is necessary to have his "contaminating" or covariable in "continuous" form. Therefore, it is necessary to accept a certain hypothesis about the model used in the ANOVA approach (model 1) above. Before beginning his analysis this research worker must make the following assumptions:

$$a_{16,14} = c_{16} + d_{16} * 14$$
$$a_{16,15} = c_{16} + d_{16} * 15$$
$$\vdots$$
$$a_{16,18} = c_{16} + d_{16} * 18$$
$$a_{17,14} = c_{17} + d_{17} * 14$$
$$a_{17,15} = c_{17} + d_{17} * 15$$
$$\vdots$$
$$a_{17,18} = c_{17} + d_{17} * 18$$
$$\vdots$$
$$a_{30,18} = c_{30} + d_{30} * 18$$

or

$$a_{ij} = c_i + d_i * j \qquad i = 17,18,\ldots,30; \quad j = 14,15,\ldots,18$$

where

c_i and d_i are unknown parameters to be estimated by the least squares method.

These assumptions then lead to the acceptance of the following model:

(4) $$Y = c_{16}\, X(16) + d_{16}\, A(16) + c_{17}\, X(17) + d_{17}\, A(17) + \ldots + c_{30}\, X(30) + d_{30}\, A(30) + E$$

where

X(i) = 1 if the corresponding element of Y is from a student who practiced i hours; 0 otherwise

A(i) = the age of the student if the corresponding element of Y is from a student who practiced i hours; 0 otherwise.

Now, we emphasize the basic common element between the two approaches. The COVARIA approach is now stated exactly the same as the ANOVA approach, i.e., the hypothesis of no interaction is—the difference between the expected (or predicted) typing performance of the two students at age p but with different practice levels r and s is equal to the difference between the expected (or predicted) typing performance of the two students at age q but with different practice levels r and s.

This is hypothesized for all values of p, q, r, and s where $p \neq q$ and $r \neq s$.

Exactly as in approach 1 <u>the hypothesis of no interaction is</u>

$$E(rp) - E(sp) = E(rq) - E(sq)$$

Notice that the two research workers are thinking about the problem in the same way.

Now we proceed to find that the expected values in the COVARIA approach are

$$E(rp) = c_r + d_r * p$$

$$E(sp) = c_s + d_s * p$$

$$E(rq) = c_r + d_r * q$$

$$E(sq) = c_s + d_s * q$$

Then the hypothesis is

$$(c_r + d_r * p) - (c_s + d_s * p) = (c_r + d_r * p) - (c_s + d_s * q)$$

$$(d_r - d_s) * (p - q) = 0$$

But since $p \neq q$

Then it is necessary that

$$d_r - d_s = 0$$

or $$d_r = d_s \qquad r \neq s \quad r,s = (16,17,\ldots,30)$$

Then imposing the restrictions on the full model (4) we have:

(5) $$Y = c_{16}\,X(16) + c_{17}\,X(17) + \ldots + c_{30}\,X(30) + d_0\,A + R$$

where

d_0 = a new unknown parameter which represents the coefficient common to <u>all</u> practice categories.

A = a vector containing the ages associated with the elements in Y.

Let $q_1 = \sum_i E_i^2$

$q_2 = \sum_i R_i^2$

Then the F statistic

$$F = \frac{(q_2 - q_1) / (30 - 16)}{q_1 / (n - 30)}$$

can be computed as a test for the interaction.

As before we consider the case where the research worker accepts the above hypothesis.

The hypotheses of the effects of practices is thought of in the same way as in the previous approach. The difference between the expected typing performance for two students at the same age p but who have practiced different amounts r and s is equal to zero.

Then the hypothesis is written exactly as in approach 1.

$$E(rp) - E(sp) = 0$$

But in this COVARIA MODEL WE HAVE

$$E(rp) = c_r + d_0 p$$

$$E(sp) = c_s + d_0 p$$

Then by substitution the hypothesis is

$$(c_r + d_0 p) - (c_s + d_0 p) = 0$$

or $c_r = c_s$ $\quad r \neq s$, $\quad r,s = (16,17,\ldots,30)$

Then imposing the restrictions on model (5) we have

(6) $$Y = c_0 U + d_0 A + G$$

where U = the unit vector of all 1's

c_0 = an unknown coefficient associated with the unit vector

Notice that in this model all information about practice has been eliminated.

Then

$$q_3 = \sum G_i^2$$

and

$$F = \frac{(q_3 - q_2) / (16 - 2)}{q_2 / (n - 16)}$$

Approach 3 - (MULCOR U.)

Now the research worker who was trained at MULCOR U. needs to have all his information in a continuous form since that's what is required in his approach.

Then before the MULCOR man can start he assumes not only the restrictions represented by the COVARIA worker (see model (4)) but in addition he must assume in model (4) that the following are true:

$$c_{16} = t_0 + t_1 * 16$$
$$d_{16} = w_0 + w_1 * 16$$
$$c_{17} = t_0 + t_1 * 17$$
$$d_{17} = w_0 + w_1 * 17$$
$$\vdots$$
$$c_{30} = t_0 + t_1 * 30$$
$$d_{30} = w_0 + w_1 * 30$$

or

$$c_i = t_0 + t_1 * i \qquad (i = 16,17,\ldots,30)$$

Imposing these assumptions on model (4) we develop a starting model as follows:

$$Y = (t_0 + t_1 * 16)\ X(16) + (w_0 + w_1 * 16)A(16) + \ldots + (t_0 + t_1 * 30)\ X(30) + (w_0 + w_1 * 30)\ A(30) + E$$

or

$$Y = t_0 \left[X(16) + \ldots + X(30)\right] + t_1 \left[16 * X(16) + \ldots + 30 * X(30)\right] + w_0 \left[A(16) + \ldots + A(30)\right] + w_1 \left[16 * A(16) + \ldots + 30 * A(30)\right] + E$$

Then

(7) $$Y = t_0\ U + t_1\ P + w_0\ A + w_1\ (P*A) + E$$

Where U = the unit vector of all 1's

P = a vector containing hours of practice

A = a vector containing ages

P * A = a vector whose elements are the product of the corresponding elements of P and A. This is called the direct product of P and A.

Again we emphasize that the MULCOR research worker thinks about the problem in the same manner as the previous two. His hypothesis of no interaction is as before

$$E(rp) - E(sp) = E(rq) - E(sq)$$

Now we determine the expected values in the model assumed by the MULCOR researcher. Looking at model (7) we find

$$E(rp) = t_0 + t_1 * r + w_0 * p + w_1 * (r*p)$$

$$E(sp) = t_0 + t_1 * s + w_0 * p + w_1 * (s*p)$$

$$E(rq) = t_0 + t_1 * r + w_0 * q + w_1 * (r*q)$$

$$E(sq) = t_0 + t_1 * s + w_0 * q + w_1 * (s*q)$$

Then the hypothesis becomes

$$\left[t_0 + t_1 * r + w_0 * p + w_1 * (r*p)\right] - \left[t_0 + t_1 * s + w_0 * p + w_1 * (s*p)\right]$$

$$= \left[t_0 + t_1 * r + w_0 * q + w_1 * (r*q)\right] - \left[t_0 + t_1 * s + w_0 * q + w_1 * (s*q)\right]$$

or

$$w_1\left[p - q\right] \left[r - s\right] = 0$$

Then for this to be always true the hypothesis is

$$w_1 = 0$$

Imposing this restriction on the assumed model (7) we obtain the restricted model (8).

(8) $$Y = t_0 U + t_1 P + w_0 A + R$$

Then we can compute

$$q_1 = \sum_i E_i^2$$

$$q_2 = \sum_i R_i^2$$

and the F statistic is

$$F = \frac{(q_2 - q_1) / (4 - 3)}{q_1 / (n - 4)}$$

Now if model (8) is accepted as true then we proceed to test the effects of practice as in the previous two approaches. The hypothesis is as before

$$E(rp) - E(sp) = 0$$

Now in model (8) we have

$$E(rp) = t_0 + t_1 * r + w_0 * p$$

$$E(sp) = t_0 + t_1 * s + w_0 * p$$

And our hypothesis is

$$(t_0 + t_1 * r + w_0 * p) - (t_0 + t_1 * s + w_0 * p) = 0$$

$$t_1 (r-s) = 0$$

and since $r \neq s$ then the hypothesis is

$$t_1 = 0$$

Imposing this restriction on model (8) we have

(9) $$Y = t_0 U + t_1 P + G$$

Then if

$$q_3 = \sum_i G_i^2$$

we have

$$F = \frac{(q_3 - q_2) / (3-2)}{q_2 / (n-3)}$$

Approach 4 - (VARICO U.)

The research worker for VARICO U. has always preferred to have his data in a different form from the others. He wishes to assume that model (1) of the ANOVA approach has the following restrictions:

$$
\begin{aligned}
a_{16,14} &= k_{14} + m_{14} * 16 \\
a_{17,14} &= k_{14} + m_{14} * 17 \\
&\vdots \\
a_{30,14} &= k_{14} + m_{14} * 30 \\
&\vdots \\
a_{30,18} &= k_{18} + m_{18} * 30
\end{aligned}
$$

or

$$a_{ij} = k_j + m_j * i \qquad \begin{array}{l} i = 17,18,\ldots,30 \\ j = 14,15,\ldots,18 \end{array}$$

where k_j and m_j are unknown parameters to be estimated by the least squares method. These assumptions lead to the acceptance of the following model:

(10) $$Y = k_{14}\, Z(14) + m_{14}\, P(14) + k_{15}\, Z(15) + m_{15}\, P(15) + \ldots + k_{18}\, Z(18) + m_{18}\, P(18) + E$$

where

$Z(i)$ = 1 if the corresponding element of Y is from a student who is i years of age; 0 otherwise

$P(i)$ = the hours of practice of the student if the corresponding element of Y is from a student who is i years of age; 0 otherwise

Again, we emphasize the idea that is common to all four approaches.

The hypothesis of no interaction is still stated as

$$E(rp) - E(sp) = E(rq) - E(sq)$$

Then we obtain these expected values from model (10).

$$E(rp) = k_p + m_p * r$$

$$E(sp) = k_p + m_p * s$$

$$E(rq) = k_q + m_q * r$$

$$E(sq) = k_q + m_q * s$$

Then the hypothesis is

$$(k_p + m_p * r) - (k_p + m_p * s) = (k_q + m_q * r) - (k_q + m_q * s)$$

$$(m_p - m_q) * (r - s) = 0$$

But since $r \neq s$ then the hypothesis must be

$$m_p - m_q = 0$$

or $m_p = m_q$ $\qquad p \neq q$, $p,q = (14,15,\ldots 18)$

Imposing these restrictions on model (10) we obtain the restricted model

(11) $$Y = k_{14}\ Z(14) + k_{15}\ Z(15) + \ldots + k_{18}\ Z(18) + m_0\ P + R$$

where

m_0 = a new unknown parameter which is common to all ages.

P = a vector containing the practice hours associated with the elements in Y.

Then we can compute

$$q_1 = \Sigma E_i^2$$

$$q_2 = \Sigma R_i^2$$

and

$$F = \frac{(q_2 - q_1) / (10 - 6)}{q_1 / (n - 10)}$$

can be computed to test the hypothesis.

As in the three previous approaches we next explore the case of no-interactions and hypothesize that the difference in typing performance for two students at the same age p but who have practiced different amounts r and s is equal to zero.

Again, the hypothesis is the same in all three previous approaches.

$$E(rp) - E(sp) = 0$$

Then we obtain these expected values in our VARICO model.

$$E(rp) = k_p + m_0 r$$

$$E(sp) = k_p + m_0 s$$

Then the hypothesis is

$$(k_p + m_0 r) - (k_p + m_0 s) = 0$$

or

$$m_0 = 0$$

Then imposing this restriction on model (11) we have

$$Y = k_{14} Z(14) + k_{15} Z(15) + \ldots + k_{18} Z(18) + G$$

computing

$$q_3 = \Sigma G_i^2$$

we can determine

$$F = \frac{(q_3 - q_2) / (6 - 5)}{q_2 / (n - 6)}$$

<u>Summary</u>

The ideas that were emphasized above are:

1. In all four approaches the statement of the hypothesis of no interaction was the same in the original thinking about the problem. Not until the specific assumed model was introduced did the approaches appear different.
2. In all four approaches the hypothesis testing the effects of practice (which followed the acceptance of no interaction) was the same. Not until the specific model was introduced did the approaches appear different.
3. The assumed models in all four approaches were obtained by accepting assumptions about the first approach (ANOVA).

If desired the research worker from COVARIA, MULCOR, and VARICO could test their assumed models to determine if these starting models are appropriate.

4. Even though the computational aspects were not emphasized, it can be observed that computing procedures required in all four approaches are quite similar.

It is interesting to notice that the original model of approach number one was basic to all others, and that the last three research workers chose to accept assumptions about the first model. Now the predictor vectors in this basic model that was the originator (or parent) of all others are binary coded, mutually exclusive vectors. Sometimes these basic vectors are called dummy vectors. This seems to imply that there is something "not quite right" or "bad" about these vectors. These binary vectors are really the parents of the other vectors and are in effect the most "brilliant" of them all. I would think that they should be called the "bright" vectors, and the other vectors might be called "dummy."

My guess is that since the binary (parent) vectors were recognized much later than their offspring, there was some attempt to apologize for the introduction of the parent.

Also, since many early studies were thought of in a multivariate normal setting there existed more need for users of these binary vectors to apologize for their use since they were not multivariate normal.

The first three approaches, ANOVA, COVARIA, AND MULCOR are frequently treated quite separately in the education of research workers. The fourth approach VARICO is not likely to appear at all.

I urge those teachers who are interested in developing in their students the capability of effective research analysis to consider carefully the objectives presented in the earlier part of this paper on page one. Then I suggest that the specific synthesis of models that has been presented will contribute to the development of the research capabilities that are desired of research workers.

jc/
6/8/67