LACK OF FIT AND THE MULTIPLE REGRESSION OF ANOVA MODEL*

Dwight E. Spence, University of Pennsylvania

ABSTRACT

Suggested in this paper is a test for lack of fit and its implications for model change applicable to the general regression analysis model, especially when used to handle analysis of variance with qualitative independent variables. Draper and Smith (1967) have developed a procedure to test for lack of fit of the general regression analysis model in the case where the data includes repeat measurements of Y at the given X values. However, when the independent variable(s), X, is qualitative, as is more usually the case in multiple regression-ANOVA (MR/AV), Draper's estimate of "pure error" will be exactly that value given as the residual mean square estimate. Thus, any possibility for such a test of lack of fit, i.e., for finding "pure error" to be less than the residual mean square, will be obviated in this MR/AV case. An alternative is to obtain repeated treatment levels from which an estimate of pure error can be obtained. Use of this suggested alternative pure error term in a test for lack of fit is used as an initial indicator as to whether attempts should be made to (a) find dependent variable transformations and (b) try various possible independent variable product-variables.

FITTING THE REGRESSION ANALYSIS MODEL IN EXPERIMENTATION

INTRODUCTION

The following is not intended as an exhaustive presentation of the use of the regression analysis model vis-a-vis experimental design data. In part, it is a review of the major aspects of this subject; however, the emphasis is on testing the adequacy of a specific model against real data. It is with respect to this latter point that we have ventured beyond what is already available in the literature on this subject. The first section will examine the data coding techniques to be used in applying the regression model to experimental design data; the special concern here (as throughout this paper) is for the case of continuous dependent variables and categorical independent variables. Section II considers methods to be used in testing for lack of fit of a specific model to a given set of observations. Finally, the focus of Section III centers on the large spectrum of alternative regression models available for application in data analysis.

SECTION I

ONE-DIMENSIONAL ANOVA

If we are to use multiple regression analysis to do analysis of variance, the ANOVA data must be recoded prior to applying the multiple regression model (MR/AV). Thus, in this section we will consider the problem of recoding with regard to simple ANOVA, factorial designs with equal cell sizes, and factorial designs with unequal cell sizes.

First the case of simple ANOVA. The most direct approach to recoding in MR/AV is referred to as dummy coding, and has as its objective a vector representation of each data observation such that observations in different cells are distinctly represented in terms of a binary system of 1's and 0's (Kerlinger and Pedhazur, 1973). Recall that the general linear regression model

$$Y = \beta_0 Z_0 + \beta_1 Z_1 + \ldots + \beta_n Z_n + e \qquad (1)$$

where $Z_i = f(x_1, x_2, \ldots, x_m)$; $Z_0 = 1$

can also be expressed in the form

$$Y = \mu_Y + \beta_1(Z_1 - \mu_Z) + \ldots + \beta_n(Z_n - \mu_Z) + e \qquad (2)$$

where $\mu_Y - \beta_1\mu_Z - \ldots - \beta_n\mu_Z = \beta_0 Z_0$

which, when compared with the analysis of variance model

$$Y = \mu_Y + \alpha_i + \ldots + \delta_j + e \qquad (3)$$

reveals the equality (Hays, 1973)

$$\beta_1(Z_1 - \mu_Z) = \alpha_i \qquad (4)$$

However, in the present case of categorical independent variables where the variables are being recoded each level of each factor-variable is represented in the model. Consider a factor-variable with three levels, then observations at the first level are represented as the row vector

[1 1 0]

while observations at the second and third levels would have vector representations of

[1 0 1] and

[1 0 0] respectively. The first column is a unit vector; the second column has a "1" if an observation is a member of the first level and a "0" otherwise; the third column has a "1" where the observation belongs to the second level and a "0" otherwise. In contrast, the third level is indicated by the fact that its members belong to neither level one nor two, i.e., zeros in both columns two and three. Therefore, if we have three observations at each of the three levels, the result of recoding would be the following matrix:

X_0	X_1	X_2
1	1	0
1	1	0
1	1	0
1	0	1
1	0	1
1	0	1
1	0	0
1	0	0
1	0	0

Thus the number of predictor variables "increases" by one (and alas the number of Df decreases!); and the general linear regression model is

$$Y = \beta_0 Z_0 + \beta_1 Z_1 + \beta_2 Z_2 + e \qquad (1)$$

*Presented at the National Convention of the American Statistical Association, 1974

even though we are still concerned with only one factor with three levels. Computation of the square of the multiple correlation coefficient leads to an F test as follows:

$$F = \frac{R^2/P}{(1-R^2)/n-P-1}$$

where P = number of predictors;
n = number of observations.
The obtained F value with its respective number of degrees of freedom has the same level of probability as that which we could have computed using the typical ANOVA F-test

$$F = \frac{MS_{Bet.}}{MS_{With.}}$$

In place of recoding through the use of dummy coding, we can use either "effect coding" or "orthogonal coding." Besides doing the same thing as the first technique, these latter two recoding techniques provide a convenient approach to the multiple comparison of means. Effect coding differs from dummy coding in that it uses "-1's" instead of "0's" to represent that group which was identified by the fact of non-membership in any other group. Thus, rather than just 1's and 0's, the coding entails 1's, 0's and -1's. Using this coding approach post hoc multiple comparisons of means can be carried out by what in effect is a statistical test of the significance of the difference between any two or more relevant regression coefficient estimates, $\hat{\beta}$. In terms of effect coding, the above matrix would take the form

X_0	X_1	X_2
1	1	0
1	1	0
1	1	0
1	0	1
1	0	1
1	0	1
1	-1	-1
1	-1	-1
1	-1	-1

In contrast, orthogonal coding leads to the opportunity to perform planned orthogonal comparisons among group means. To do this the column vectors must consist of terms such that each vector is orthogonal to all other vectors, with the exception of the initial unit vector which is omitted in many computer routines. Each column vector in turn expresses a separate orthogonal contrast. Again we can construct the matrix based on the above hypothetical ANOVA problem using orthogonal coding thus:

X_0	X_1	X_2
1	1	½
1	1	½
1	1	½
1	-1	½
1	-1	½
1	-1	½
1	0	-1
1	0	-1
1	0	-1

Column two gives a contrast between the means for group one and group two, while column three compares the average of group one and two with group three--orthogonal to the first comparison. A test of significance applied to a particular regression weight ($\hat{\beta}$) indicates whether the comparison specified by the corresponding column vector is significant or not, i.e., if $\hat{\beta}$ is significantly different from zero, the comparison of column vector X_1 is also significant with respect to the difference between group one and group two means.

Factorial Design:

For the case of multiple independent variables, each expressed in terms of the categorical level of measurement, the coding systems dealt with above can be shown to have direct applicability through generalizations of the principles used in the simple ANOVA situation. Suppose our analysis is of a 2 x 3 factorial design with two observations per cell. The resulting orthogonally coded matrix would take the form

(X_0)	(X_1)	(X_2)	(X_3)	(X_{12})	(X_{13})
1	1	1	1	1	1
1	1	1	1	1	1
1	1	1	1	1	1
1	1	1	1	1	1
1	1	1	1	1	1
1	1	1	1	1	1
1	1	-1	1	-1	1
1	1	-1	1	-1	1
1	1	-1	1	-1	1
1	-1	-1	1	1	-1
1	-1	-1	1	1	-1
1	-1	-1	1	1	-1
1	-1	0	-2	0	2
1	-1	0	-2	0	2
1	-1	0	-2	0	2
1	-1	0	-2	0	2
1	-1	0	-2	0	2
1	-1	0	-2	0	2

The X_0 column vector is again a vector of 1's and refers simply to the overall mean level of our regression model. Vector X_1 refers to factor A of two levels and uses a 1 to indicate membership in the first level and -1 for membership in the second level, contrasting these two levels. Columns X_2 and X_3 of course refer to independent variable B which has three levels. Thus X_2 provides a comparison between levels one and two of variable B; whereas, levels one and two are compared with level three in column X_3. Finally, X_{12} and X_{13} account for the possible interaction between the two independent variables for any given dependent variable. These last to vector are formed by simply taking the cross-products of the relevant main effects column, e.g., X_{12} consist of cross-products between the elements of vectors X_1 and X_2. The several F-ratios for main and interaction effects can be computed through the computation of R^2 values based on subsets of the orthogonal column vectors. For example, if vectors X_{12} and X_{13} are separately analyzed we get a R^2 value indicating the amount of variance accounted for by the interaction of the two (or more!) factors.

An added complication enters the picture if the cell sizes contain unequal numbers of observations. This stems from the fact that the main and interaction effects are no longer orthogonal to each other; and therefore, different results can be obtained depending on the order in which the total variance is partitioned into the variance of the separate components (Overall and Klett, 1973). Briefly, these are the three possible partitions. Simply consider the problem of calculating the variance

due to the main effect of variable A in the hypothetical 2 x 3 factorial design mentioned above.
a) R^2 can be computed for the column vector(s) of main effect A, and the variance thus accounted for can be determined by the product
Total Variance $\cdot R^2_{x_1}$ {Variance due to main effect A}

b) If R^2 for main effect B is subtracted from R^2 for both main effects A and B, then the variance accounted for by main effect A is given by the product
Total Variance $\cdot [R^2_{x_1,x_2,x_3} - R^2_{x_2,x_3}]$
{Variance due to main effect A}

c) Yet a third approach might entail taking the difference between R^2 computed for effects A, B and interaction and R^2 computed for effects B and interaction; in this instance, the variance due to main effect A would be given by the product
Total Variance $\cdot [R^2_{x_1,x_2,x_3,x_{12},x_{13}} - R^2_{x_2,x_3,x_{12},x_{13}}]$
{Variance due to main effect A}

In each of these three efforts to accounting for the variance due to main effect A different answers are obtainable given the situation of unequal cell sizes. It is the second strategy, (b), that corresponds to the ANOVA technique as used in experimentation (see Overall and Klett, 1972).

Aside: Wolf and Cartwright have recently (1974) proposed a quite different approach to this problem of coding in MR/AV. Their approach requires that an experimenter solve for the coded matrix (x) using the matrix solution to the normal equations $b = (x'x)^{-1}x'y$. Of course y is the N×1 vector of dependent variable observations. The J×1 (where J = the number of comparisons) b vector is obtained by the matrix formula $b = c(V'V)^{-1}V'y$ where V is an N×K (K = the number of categories) matrix of dummy coded 1's and 0's, indicating membership (1) and non-membership (0) in the several categories; and C is a J×K matrix of weights for the comparison of category means. An example of C provided by Wolf and Cartwright is

$$C = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -\frac{1}{4} & -\frac{1}{2} & 1 & 0 \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix}$$

which applies in one-dimensional ANOVA when the cell sizes are equal. Thus computing the vector b allows for the generation of a coded matrix (x) to be used in the multiple regression analysis of ANOVA (by solving for x in $b = (x'x)^{-1}x'y$)). The end product is a coded matrix which in appearance is considerably different from the matrices obtainable using either of the three coding techniques that we have considered up to this point. Likewise, Wolf and Cartwright state that the computed b, error and F values all differ from those gotten using a coding system consisting of 1's, 0's and -1's. However, they do reveal that the inferences made using these two differing approaches do not, themselves, differ. The statistical conclusions remained the same.

SECTION II

TESTING FOR LACK OF FIT

If we make the very important shift in reference away from simply testing the relationship between the independent and dependent variables vis-a-vis a specific test statistic to testing the adequacy of the test statistic model with respect to the data at hand, we will at least become aware of the presen-e of diverse models; and at most we will find independent/dependent variable relationships which would have otherwise been overlooked given the limitations of any one specific model. This point is singularly fundamental throughout this and the next section: whenever we test for a relationship within a set of observations, we do so in the context of a very specific model imposed on the data; and thus, must consider the possibility that an alternative model might prove more revealing (without simply capitalizing on chance characteristics of the data).

Draper and Smith (1966) published a technique to test for the lack of fit of a regression model. The key component is the computation of a "pure error mean squares" term (S^2_e) used to evaluate a lack of fit mean squares term (MS_L). To compute this pure error value the data must contain repeat observations on the dependent variable (Y) at several levels of the independent variable (X). Thus, pure error refers to the variability of Y within levels of X:

$$\text{Pure Error } (SS_e) \quad \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2 \quad .$$

$$\text{Lack of Fit } (SS_L) \quad \text{Residual (SS)} - \sum\sum (Y-\bar{Y})^2$$

This test for lack of fit follows the F distribution with

$$F = \frac{SS_L/Df_L}{SS_e/Df_e} = \frac{S^2_L}{S^2_e} = \frac{\text{Lack of Fit (MS)}}{\text{Pure Error (MS)}}$$

A significantly large F value indicates that the residual error term consists of more than simple error variability, i.e., the model being considered fails to account for some non-error variance.

Other questions might be raised at this point; however, the crucial one here is: What happens if the independent variables (X) are categorical rather than continuous? After all, this is more typically the case in social science data analysis using ANOVA. If we attempt to apply the above technique to MR/AV given categorical independent variables, the finding is that this approach invariably gives us a value for S^2_e exactly the same as the residual MS value. And thus we have no basis for ever finding a model inadequate. It is for this case that we would like to suggest an alternative approach.

Perhaps the pure error term (S^2_e) should be computed by splitting the data in half for each level of theindependent variable, computing a mean for each half cell of data and looking at the variability of half cell means within the several independent variable levels.

$$\text{Pure Error (SS)}' \quad \{\sum\sum(\bar{Y}-\bar{\bar{Y}}_i)^2 \frac{n_i}{2}\}$$

where $\bar{Y}$ is half cell mean;
$\bar{\bar{Y}}$ is a mean for the entire cell.

This presumably would be comparable to a treatment (SS) value under the condition of no systematic relationships in a set of data. Thus, the F distribution would take the form

$$F = \frac{\text{Pure Error)(MS)}'}{\text{Residual (MS)}'} = \frac{\sum\sum \frac{n_i}{2} (\bar{Y}-\bar{\bar{Y}})^2/Df_e}{\text{Residual (SS)}'/Df_{res}}$$

and a significantly small F value would indicate that the residual term is inflated for the given

regression model. Of course this requires reasonably large sample sizes.

SECTION III

ALTERNATIVE REGRESSION ANALYSIS MODELS

A test for lack of fit is here intended to imply both an entire system of candidate models and a willingness to explore the possible usefulness of any one of these models. Again the context to be assumed is that of continuous dependent variables and categorical independent variables. Two major types of alternative models are to be dealt with. The maneuver distinguishing the first type will be the gamut of variable transformations which may be applied to the dependent variable, whereas the strategy of the second class of alternative models revolves around the generation of "new" independent variables through taking the cross-products of the original independent variables producing what will be called "interaction variables." Of course there exists no compelling reason why both of the above manipulations cannot be applied simultaneously.

There is quite a good deal of literature devoted to the subject of data variable transformations (Bartlett, 1947; Box, 1962; Tukey, 1957). One rationale for data transformations is that the transformations make for a more precise analysis by bringing real data into greater conformity with the assumptions of the analysis, e.g., homogeneity of variance, normality, additivity, etc. However, clear and convincing indicators as to where, for example, a square root or a reciprocal, a log or an arcsin transformation might most profitably be applied does not seem evident from the literature. The user must resort largely to trial and error, and testing for lack of fit. Some general guidelines have been put forward (Myers, 1972). Given a situation where the variance is proportional to the means (i.e., $\sigma^2/\mu_i = K$) the square root function should be tried to achieve variance constancy. Application of a log function to observed data might be attempted in the case where the standard deviation is directly proportional to the mean (i.e., $(\sigma/\mu)^2 = K$) and the distribution is considerably skewed. Use of the arcsin function has resulted in reports of success in the case where the data observations are proportions (or percentages) with $\sigma^2 = \mu(1-\mu)$.

General Regression Model

$$Y = \beta_0 Z_0 + \beta_1 Z_1 + \ldots + \beta_n Z_n + e = \sum_{i=0}^{n} \beta_i Z_i + e \qquad (1)$$

where $Z_i = f(X_1, X_1, \ldots, X_m)$

Regression Model with Dependent Variable Transformation

$$ln Y = \sum_{i=0}^{n} \beta_i Z_i + e \qquad (5)$$

This should be enough to make it clear that the possibilities with respect to use of dependent variable transofrmations are virtually limitless.

Compared to data transformations on continuous variables, there appears to be little published material concerned with the generation of new independent variables from a set of data where the independent variables are measured at the categorical level (Cohen, 1968). In general, this second group of alternative models can be characterized by the fact that they are the result of forming new independent variables by taking the cross-products of the basid set of independent variables in some cases referred to as moderate variables. Thus, for example,

General Regression Model $$Y = \sum_{i=0}^{n} \beta_i Z_i + e \qquad (1)$$

where $Z_i = f(X_0, X_1, \ldots, X_m)$

and X_0 through X_m make up the original data base. Typically we would simply have the equality $Z_i = X_i$; however, in the case of cross-products we might have

Interaction (Variable) Between X_{i-1} and X_i $\quad Z_i = X_{i-1} \cdot X_i \qquad (6)$

Given as few as, say, three original data variables there are, as can be quite easily shown, several interaction variables which can be computed (which is not to say that they should be!). Whether these interaction variables account for any variance of consequence can be tested by the usual multiple R^2 ratio following the F distribution:

$$F = \frac{(R^2_{y \cdot 12} - R^2_{y \cdot 1})/Df}{(1-R^2_{y \cdot 12})/Df} \qquad (7)$$

REFERENCES

Bartlett, M. S. The use of transformations. *Biometrics*, 3, 1957.

Box, G. E. P., and Tidwell, P. W. Transformation of the independent variables. *Technometrics*, 4, 1962

Cohen, J. Multiple regression as a general data analytic system. *Psychological Bulletin*, 1968, 70.

Draper, N. R., and Smith, H. *Applied Regression Analysis*. New York: Wiley, 1966.

Hayes, W. *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston, 1973.

Kerlinger, F., and Pedhazur, E. *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart and Winston, 1973.

Myers, J. *Fundamentals of Experimental Design* (2nd ed.). Boston: Allyn and Bacon, Inc., 1972.

Overall, J., and Klett, C. *Applied Multivariate Analysis*. New York: McGraw-Hill, 1972.

Tukey, J. W. On the comparative anatomy of transformations. *Ann. Math. Statistics*, 28, 1957.

Wolf, G., and Cartwright, B. Rules for coding dummy variables in multiple regression. *Psychological Bulletin*, 81, 1974.