CURVILINEAR REGRESSION

Charles L. Aird

*Department of Industrial Management, University of Petroleum and Minerals, Dhahran, Saudi Arabia**

الخلاصة

ان ابتكار الكومبيوتر من أجل تركيب معادلات التنبؤ قد فتح باباً للتعامل مع المغايرات الغير خطية تساوي بسهولتها التعامل مع المغايرات الخطية .

ان تحليل التراجع التدريجي على خط منحني يعتبر بمثابة عامل تغيير قوي للتحليل التراجعي المتضاعف الذي يوفر وسيلة لاختيار المغايرات المستقلة التي تركب بدورها أفضل الطرازات الممكنة لأجهزة التنبؤ بأقل عدد ممكن من التغايرات . لذلك يمكن تحقيق رصيد يجمع بين الإيجاز وجودة التنبؤ بالإضافة الى توفير وإدخار أكبر قدر من الجهود الممكنة .

ABSTRACT

The advent of the computer for the construction of predictor equations has made dealing with nonlinear variables as easy as dealing with linear variables. Stepwise curvilinear regression analysis is a powerful variation of multiple regression analysis which provides a means of choosing independent variables that construct the best predictor model possible with the least number of variables. Thus, a balance may be achieved among precision, predictive validity, and economy of effort.

**Present address: Supreme Court of Virginia, Fifth Floor, 11 South 12th Street, Richmond, Virginia 23219, U.S.A.*

CURVILINEAR REGRESSION

INTRODUCTION

One of the most frequently occuring situations in applied statistics is the analysis of data consisting of observed responses which are dependent upon a one-dimensional array of inputs. The problem is to construct a predictor model. The observed responses, y, can be represented by a regression equation and is a function of the independent responses, x. Most frequently the variables are related by the general multiple linear regression model:

$$y = a_1x_1 + a_1x_1 + \ldots + a_mx_m + e$$

where:

y is the one-dimensional array of observed or estimated values of some random variable and is referred to as the dependent or criterion variable.

x_i is a one-dimensional array of the known values of the independent or predictor variable.

a_i is a parameter of the system which is to be estimated from the data.

e is the one-dimensional array which has as elements differences or discrepancies between corresponding observed and estimated values in the dependent variable, y. This is often referred to as the error.

However, there are several considerations which may prove some or all of the independent variables to be nonlinear. A graph of the data may give an obvious picture of nonlinearity. Simple observations may prove the data to be periodic . Until the advent of the computer, nonlinear regression was difficult to work with and usually ignored.

Suppes [10] puts forth the following argument in favor of nonlinear regression analysis:

> ... it is almost as easy to deal with simple nonlinear models as linear ones. Exploring alternatives to linearity provides excellent insight into the nature of the relations between variables...
>
> If we think the effects of increase in x or y is proceeding at a faster than linear model, we can estimate the number of parameters in a quadratic model. On the other hand, if we think the nonlinear increase in y with increases in x is less than linear, we can test the logarithmic model.
>
> It is fantasy that we must always test for linear relations.

Thus, a general model in the form

$y = ax + e$ linear

may become

$y = ax^2 + e$ quadratic

$y = a \log x + e$ logarithmic

$y = a \sin x + e$ periodic

or some combination of these and other nonlinear relations. The multiple regression equation,

$$y = a_1x_1^2 + a_2 \log x_2 + a_3 \sin x_3 + e$$

is an example.

CURVILINEAR REGRESSION ANALYSIS

This article describes the use of the nonlinear technique of regression called curvilinear regression analysis. The most general form of a curvilinear regression equation is similar to the linear model. It is written as the linear combination,

$$y = a_1f_1(x_1) + a_2f_2(x_2) + \ldots + a_mf_m(x_m) + e$$

where:

y is the one-dimensional array of observed or estimated values of the dependent variable. x_i is one-dimensional array of the known values of the independent variable.

f_i is the function which provides the curvilinear transformation of the independent variable x_i.

a_i is a parameter of the system which is to be estimated from the data.

e is the residual array which has as elements differences or discrepancies between corresponding observed and estimated values in the dependent variable, y.

It is possible for some i and j that $f_i = f_j$ or $z_i = x_j$, but not both.

The use of such curvilinear regression models is appropriate in situations in which the simple linear forms of variables are inadequate and in the case where a single function is not capable of expressing the criterion variables as a combination of the independent variables.

There are situations when a graph of the known data and other considerations may indicate the non-

linear scope of unknown variables. For example, a set of responses may be periodic but with an unknown period. Another example may show a set of responses to be increasing to some unknown point and then decreasing. Such situations occur regularly in agricultural, engineering, and pharmaceutical studies.

Given a set of empirical data for which a curvilinear relationship is thought to exist among variables, the general equation is first generated. Lewis [7] feel that this is at best a haphazard process:

> It is apparent that curve fitting is largely a trial-and-error process. In a sense, it is an art; it cannot be reduced to a set of inflexible rules.There is always room for disagreement and for judicious decisions.

PROGRAM CONSTRUCTION

The exceptional computational speed and the data handling ability of the computer have made dealing with nonlinear analysis almost as easy as linear analysis. Thus, this author wrote a FORTRAN based computer program that would:

1. Tranform the linear variables using the FORTRAN functions defined by the system.
2. Construct a predictor or regression equation by stepwise methods.
3. Output the generated equation with a statistical analysis of the results.

The transformational functions include:

1. Applying a power function to variables: square and cube.
2. Applying trigonometric functions to appropriate variables: sine, cosine, and tangent.
3. Applying the exponential function with base e to appropriate variables.
4. Applying the natural and base 10 logarithmic functions to appropriate variables.

The program was adapted to the Control Data Corporation 6400 series computer. The program simultaneously analyses both categorical and continuous data, but does not transform categorical data. It also does not assume that the data is normally distributed for each variable.

Statistical analysis was done by using the predictor equation as a full model. Any hypothesis concerning relationships between independent variables and dependent variable places restrictions upon the full model. Thus, any new equation is formed taking these restrictions into account. This new equation is the restricted model.

The F statistic is generated to establish regions of acceptance or rejection of hypotheses. The equation

$$F = \frac{(R_1^2 - R_2^2) / df_1}{(1 - R_1^2) / df_2}$$

R_1^2 is the squared multiple correlation coefficient obtained from the full model;

R^2_2 is the squared multiple correlation coefficient obtained from the restricted model;

df_1 is degrees of freedom obtained by subtracting the number of linearly independent arrays of the restricted model from the full model;

df_2 is the degrees of freedom obtained by subtracting the linearly independent arrays of the full model from the observations of the dependent variable.

A number of arrays could be developed in any particular study. Additional arrays are generated by the program to investigate any relationships between independent variables. These arrays are formed as the direct product of independent variables under.

Stepwise regression is done on a combination of linear and curvilinear variables in order to achieve a 'best' model of the predictor equation. Stepwise regression is a powerful variation of multiple regression which provides a means of choosing independent variables that will provide the best prediction model possible with the smallest number of independent variables. In this way, a best fit is obtained from any set of empirical data. Stepwise regression analysis gives a quick and efficient method for establishing a near optimum solution.

The method is best described by relating its use in practice. The program recursively constructs a prediction model one independent variable at a time. The first step is to choose the single variable which gives the best prediction. The next variable added to the model is that which provided the best prediction in conjunction with the first variable. The method proceeds in this recursive fashion adding variables

step by step until either the desired number of independent variables is obtained from the model or until no other variable makes a significant contribution to the model. (It should be noted that the variables may be nontransformed or curvilinear.) For example, consider a specific model obtained from the data of a study. A precision model for each criterion was produced by specifying that variables be added to the model as long as there was an increase in the squared multiple correlation coefficient of at least 0.005. It is also possible to restrict the number of independent variables in the final equation. Thus, the exact choice is a user decision. This is essentially an arbitrary decision. However, Bormuth [4] has discussed the balance which must be achieved among precision, predictive validity, and economy of effort in determination and use of curvilinear regression equations. It appears that he is of the opinion that there are no general rules for determining the appropriate number of variables in the regression equation. At each step of the procedure, the program selects the optimum variable, given the other variables of the predictor model. The program also provides the following information: multiple correlation coefficient squared, increase in multiple correlation coefficient squared, beta weights, and the F-ratio and degrees of freedom of the full versus restricted models at each step.

PRECISION vs. VALIDITY

When an equation is obtained from a set of empirical data, how well does the equation represent that data? If the squared multiple correlation coefficient is quite high, it may be concluded that a major portion of the variance of the dependent variable is attributable to changes in the independent variables and that only a small portion is due to other factors. However, the results of work by Bormuth cast doubt on whether it is possible for a regression equation to simultaneously exhibit high precision and predictive accuracy For an equation to have high precision, it must contain a relatively large number of variables. But if many variables are included within the equation, the equation would almost certainly lack accuracy of prediction. Adding variables to a regression equation also adds to the error normally associated with the estimation of the beta coefficients. At some fairly early point, the error added by each new variable begins to exceed whatever predictive validity the variable may have added. Bormuth [4] made the following statement concerning this difficulty:

> Adding enough variables to obtain a formula having high precision will result in a formula having low predictive validity. Obviously, some sort of compromise has to be reached in a way which is not entirely clear.

The predictive validity of any model can be determined by a replication of investigation and construction of a new regression model. If there is homogeneity of the equations, then it is possible to use either model for predictive purposes and validity is assured.

TESTING THE PROCEDURE

This procedure was used in three major studies to evaluate those studies and to determine the effectiveness of the computer program and the particular method of statistical analysis. The three studies were:

1. An investigation of computer based instruction at the University of Virginia School of Engineering [1].
2. The construction of predictor models for admissions at T.C. Wiliiams Law School of The University of Richmond [2].
3. A study of the use and evaluation of remedial mathematics of Piedmont Virginia Community College.

The first study was conducted to determine the effectiveness of computer based instruction in a solids of mechanics course. The original statistical analysis was done using linear regression. The results were inconclusive. The computer program which carried out the curvilinear regression analysis was then constructed. The data was run with the new method of analysis with successful results. That is, the method of instruction was effective. Predictor equations were established and a replication of the investigation validated these equations.

The second study was primarily concerned with the development of predictor models for use with admissions procedures. The data was collected for all students and equations established. To test the predictive validity of the equations, a repeated study wss conducted. The results again proved the strength of curvilinear regression in establishing both precision and validity in models as well as evaluation of educational programs.

The last study was an attempt to evaluate developmental mathematics at the community college level. The method of analysis was by curvilinear regression. The results showed that a multi-media approach to teaching remedial mathematics was not as effective as traditional instruction. However, the statistical analysis and the predictor equations were shown to be valid.

A full account of the background and computational aspects underlying these studies is too detailed to present here. In all three cases the results proves satisfactory. Replications of studies 1 and 2 developed new predictor models, and there was homogeneity of the curvilinear regression equations. In all cases high squared multiple correlation coefficients were obtained. That is, the models developed have high precision, but according to Bormuth [4] may have low predictive validity. However, with the homogeneity of equations determined through follow up studies, these models are generalizable and have both high predictive validity and precision.

CONCLUSIONS

The results of the above studies indicate that stepwise curvilinear regression analysis and the particular procedure of the program are effective, efficient, and economical means of statistical analysis. However, there are broad gaps in the research literature on nonlinear regression. Curvilinear analysis should become an area of extensive concern to the applied statistician.

REFERENCES

[1] C.L. Aird, "An Investigation of Self-Study Computer-Based Instruction in Engineering", *Proceedings of* 1974 *AERA National Meeting*, April, 1974.

[2] C.L. Aird, "Development of Predictor Equations for Law School Admissions", *unpublished report*, June, 1974.

[3] C.L. Aird, "Computerized Techniques of Stepwise Curvilinear Regression Analysis", *Proceedings of ASA SIG on Computational Statistics*, November, 1975.

[4] J.R. Bormuth, *Development of Readability Analysis*, Chicago: University of Chicago, 1969.

[5] R.A. Bottenberg, and J.H. Ward, *Applied Multiple Linear Regression*, Texas: U.S. Air Force, 1963.

[6] A.R. Gallant, "Nonlinear Regression", *The American Statistician*, May, 1975.

[7] D. Lewis, *Quantitative Methods in Psychology*, New York: McGraw-Hill, 1960.

[8] D.W. Marquardt, and R.D. Snee, "Ridge Regression in Practice", *The American Statistician*, February, 1975.

[9] J.E. Matheson, *Optimum Teaching Procedures Derived from Mathematical Models, Stanford*, California: Stanford University, 1964.

[10] P. Suppes, "Facts and Fantasies of Education", *Technical Report No.* 193 *of Institute for Mathematical Studies*, October, 1972.

[11] J.H. Ward, "Multiple Linear Regression Models", *Computer Applications in the Behavioral Sciences*, New Jersey: Prentice Hall, 1962.

Reference Code for AJSE Information Retrival HA 3177 AI1.
Paper Received 25 October 1977.