DOCUMENT RESUME

ED 081 205 EM 011 376

AUTHOR Utter, Merlin; Wilkinson, John W.
TITLE Some Classroom Experiences in the Teaching of Empirical Model Building and Regression Analysis.
PUB DATE Jun 73
NOTE 3p.; Paper presented at the Conference on Computers in the Undergraduate Curricula (Claremont, California, June 18-20, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Calculation; College Mathematics; *Computer Assisted Instruction; Computer Programs; Digital Computers; Higher Education; *Mathematical Models; *Mathematics Instruction; *Multiple Regression Analysis; Program Descriptions; *Statistics; Time Sharing; Undergraduate Study
IDENTIFIERS BMD; CAI; LINREG; RPIREG; STEPREG

ABSTRACT

The use of the digital computer for the presentation of the topics of empirical model building and regression analysis is discussed. The author concentrates upon a description of computing exercises which are employed to provide the students with experience in model building and evaluation in a controlled situation. The types of exercises given are treated, followed by a discussion of the relative merits and dysfunctional aspects of the time-sharing and batch modes of operation. Details are presented concerning the main programs accessed by the students--the BMD multiple and stepwise regression programs, RPIREG, STEPREG, and LINREG. Finally, there is consideration of the strengths and weakness of the computer-assisted instructional (CAI) approach to these topics. (PB)

SOME CLASSROOM EXPERIENCES IN THE TEACHING OF EMPIRICAL MODEL BUILDING AND REGRESSION ANALYSIS

Merlin Utter and John W. Wilkinson*
Operations Research and Statistics
Rensselaer Polytechnic Institute
Troy, New York
(518) 270-6565

Without the computer the presentation of a topic such as empirical model building and regression analysis would be quite sterile. Hence, in a course involving a heterogeneous mixture of undergraduate and graduate students, involvement of the digital computer has been invaluable. In presenting the material, the following three vehicles have helped to immerse the student in the subject: (a) problem sets, which provide a familiarization with the available computing power as well as instruction in the topics of regression analysis; (b) projects, which introduce the student to the importance and difficulties of problem formulation as well as the problems encountered with the gathering and handling of real data; and (c) computing exercises, which provide an experience in model building and evaluation in a controlled situation. This paper will concentrate on the computing exercises portion of the course.

For the computer exercises, the students are provided with sets of data which have been artificially generated. Rather than fit a specific model to the data, as is commonly done in problem sets, the students try to estimate the regression model from which the data were generated. In the computing exercises, unlike the projects, the true model from which the data were generated is known, and it is felt that this information can eventually be used by the student to provide a capacity of comparison of his model with the true situation. This would allow feedback that might provide insight into what moves were important in making either the right or wrong conclusions in the modeling process. The data are generated by the instructor in a simple format. Once the model is decided upon, the response or dependent variable is generated without error as a known function of one or more independent or predictor factors. These independent factors are either chosen randomly over some preassigned factor space or according to some designed experiment. Once the error of some prescribed form has been generated, addition to the previously obtained errorless response value provides the "observed" value for the dependent variable. These corresponding values of the predictor factors and response variables are generated by the computer, requiring as input only the model to be used, the form of error desired, and the number and location of the observations to be taken.

The error chosen can be of such a magnitude that it is either very difficult or much too easy, to adequately fit a model to the generated data. For instance, the first set of data given to the student was generated from a low-order polynomial with extremely little error. The result was an unchallenging and most uninteresting rapid convergence to the correct model for the student. On the other hand, students should not by provided with data involving such a large random error that they are lulled into thinking that nothing but a n-1st order polynomial can provide an adequate fit. After experimentation with the magnitude of error relative to the range of the response variable, appropriate values were found for the students' model building and evaluation experiments.

Various types of exercises are given to the students. The first exercise introduces them to the techniques involved in solving a simple polynomial of either one or two variables of second order or less; and usually missing one of the terms, such as the cross-product term. Subsequent data sets are generated from models involving more complicated functions of the independent variables, such as $\sqrt{x}$, sin(x) and 1/x. Models involving transformations on the dependent variable, such as 1/y, $\sqrt{y}$ or ln y, have added spice to the model building game. Data generated from these latter types of models has created valuable learning experiences, due to the strange behavior of the residuals and other statistics obtained when fitting the wrong model. Also it has been interesting to occasionally add a factor which is merely random noise and actually has no influence on the true model. This is important because the student would soon find out if one always presented significant factors and this would considerably influence his model building. In other exercises, even though data have been generated from a model with two factors, the students are provided only one of the independent factors along with the dependent variable generated for the complete model. Such an exercise has provided an excellent introduction to the effects of missing variables as well as an awareness of the possible need to search for additional explanations of a dependent response. In the exercise, the students were initially perplexed when they obtained highly significant parameters and regression sum of squares but unusual residual plots. To complete the exercise, the "lost variable" was provided to give the students the

opportunity to re-evaluate and modify their model based on this new and more complete information.

Some exercises have dealt with different sets of data generated for the same model, but under various designs, thus providing a comparison of their respective powers for evaluating "goodness of fit." Such examples have been: (a) cyclic and factorial designs to fit a second-order, two-variable polynomial; (b) designs with $n/3$ replicates at each of three equally spaced points, $n/6$ replicates at each of six equally spaced points and n equally spaced points to fit a second-order, one-variable polynomial; and (c) designs with $n/6$ replicates at each of six equally spaced points and n equally spaced points to fit a third-order, one-variable polynomial. From exercises like these, there are often some side benefits that make significant contributions to the learning process. For instance, when fitting a second-order polynomial to data from a third-order, one-factor polynomial, a higher R^2 value (R representing the multiple correlation coefficient) was observed than when the correct model was fit to data generated from a second-order polynomial (obtained by eliminating the cubic term from the model above). This apparent anomaly is due to the larger sum of squares involved in the first situation. However, it provided a very sobering message as far as creating some impressions relating magnitude of R^2 to the goodness of the model.

These computer exercises would be carried out in either time-sharing or batch modes of operation. The main programs accessed by the students were the BMD multiple and stepwise regression programs, RPIREG, STEPREG and LINREG, the latter being specially written with the computing exercises in mind.

All the programs provide the standard correlation matrix, variance-covariance matrix, parameter estimates with their associated standard deviations and t-statistic values, ANOVA table, R^2 value and various printer plots of residuals. The stepwise programs also provide the partial correlations with the response variable of those factors not yet in the regression. In addition, LINREG allows inequality constraints on the parameters as well as the ability to test hypotheses of linear combinations of the parameters. Another side effect of the computer exercises has been their indirect effect on the refinement of applicable computer programs.

At the onset, it was thought that these computer exercises would best be done in the time-sharing mode and thereby fully utilize the benefits of such an interactive system, where model after model could be sequentially run in a logical fashion, leading toward a "good" model. In fact, the LINREG program allows one to choose each variable to be entered or deleted in the stepwise procedure manually in a true interactive manner. However, time-sharing is not crucial for this type of work and its use was not insisted upon. The result has been that this mode has not received utilization to the extent expected, due to many reasons, some involving program sophistication and others related to computer system utilization. Because of an overloaded computer system, elapsed time at the remote terminal has been too long for the amount of actual computing performed. This has been the main reason for students "giving up" on the time-sharing mode and going to a batch mode of operation. Another contributing factor has been the necessity during the day to sign up for terminal use, requiring the student to adapt his schedule to terminal availability. Also a computer system change early in the course resulted in a decline in the reliability of the time-sharing mode. Often the student would experience system crashes essentially requiring him to start over again. Also, when the system was working well, there was the temptation (often taken) to use the "shotgun" approach and to try as many models as possible without much thought other than to run as many as possible in the time the student had been assigned to the terminal. This tended to defeat any advantage that the interactive "instant turnaround" time-sharing mode was supposed to offer.

To cut down on the load added to the system and to reduce the computer costs for the course, groups of from 2 to 4 students were formed to jointly work on the computer exercises rather than each individual doing each exercise independently. Although this approach posed the danger of potentially allowing some students to coast, it had the advantage of encouraging interaction with each other which aided the model building process. Each group received a different set of data, often from different models. This encouraged independent work and also provided the class with a variety of experiences for later class discussion.

The use of the batch mode also had some problems. At certain times during the semester, the turnaround was quite slow, again pressuring the student to consider the "shotgun" approach to run many models with the hopes that if you try enough you might be lucky and pick a winner. To help avoid this, the students were given longer to complete the exercises so they would feel no real time constraint. In addition, the form of the written report required for each computer exercise was altered. At first, very few instructions were provided concerning the form that the report should take or the technique used to obtain the model that the group felt best fit the data. As might have been expected, the result was a

barrage of computer printouts, the output from all the models each group considered worth running. Besides requiring unnecessary amounts of computer time, this shotgun approach resulted in a minimal gain in knowledge of model building. Many students were content to let the stepwise program do the work and merely fit that polynomial which best fit the data, regardless of the true model. To discourage these practices, a step-by-step procedure to obtain the resulting model was required. It was stressed that the students use the data and any previously run models to determine the next model to be tried. Each step of this logical procedure involving how they decided the next model to try, as well as the pertinent information derived from each model attempted, was to be documented in the write-up. As part of the analysis, the students were to discuss the goodness of the fit, the precision of the estimates, the examination of residuals by both graphical procedures as well as by the use of various statistical tests and any possible signs that the error was not completely random. The results were clearly better.

Although the results of our computer exercises have been extremely valuable in presenting the concepts of model building and impressing the students with the effect of factor space coverage on this process, there are many improvements to be made. Students still run many more models than they need, and seem willing to substitute a little more keypunching of new models for a little less thinking and inspection of the results already obtained. Because of the format of the write-up, some students report on only those models which appear good. One possible solution is to monitor the amount of computer time used by each group, and use this time as a measure of their efficiency in the modeling process. In the past, only a typed copy of the data has been provided, forcing the students to type in the data themselves each time the terminal is to be used. To alleviate this situation, and to permit more variety in sample size it is planned to store the data on files in the computer for easy access by the student. Also planned is the development of other types of exercises which, among other things, will allow experimentation examining the violation of the various assumptions relating to such features as common variance and additive error.

*John W. Wilkinson will handle correspondence.