The Annals of Mathematical Statistics
1971, Vol. 42, No. 4, 1273–1284

MODEL BUILDING FOR PREDICTION IN REGRESSION BASED UPON REPEATED SIGNIFICANCE TESTS[1]

By W. J. Kennedy and T. A. Bancroft
Iowa State University

1. Introduction.

1.1. Regression analysis considered herein will deal with the fitting of linear models for the purpose of prediction. The method of least squares provides a well-defined mathematical procedure for obtaining a unique prediction equation whenever it can reasonably be assumed that the data arose from a situation which can adequately be represented by a linear model having one dependent variable and a definite number of independent variables. The usual additional assumption of independently normally distributed errors having zero mean and a constant variance allows application of additional statistical theory to test selected hypotheses of interest and to set confidence intervals.

In application of the theory of regression analysis to experimental data, uncertainty often arises as to the exact number of independent variates to include in the final model. Many situations are such that among the total set of independent variates only a small subset is of real value when attempting to predict the behavior of the dependent variable.

Several different procedures have been recommended for use in determining a suitable subset of independent variables for use in predicting the dependent variable of interest (see Abt [1], Draper and Smith [9], Efroymson [10], Gorman and Toman [12], Hocking and Leslie [13]). These procedures involve the use of repeated tests of significance and rely upon inferences based upon the outcome of such tests. The decision rules used in these procedures were, for the most part, selected for their intuitive appeal and little consideration has been given to the consequences, with respect to the fitted model, of the effect on subsequent inferences of such repeated testing.

The problem of model building in regression is one of the general class of problems called problems of incompletely specified models involving the use of repeated tests of significance. This classification serves to clarify the nature of this regression problem and to emphasize the need for more relevant theoretical development in this problem area. The development of model building techniques in general has been carried out under the assumption that no *a priori* information is available to the experimenter concerning which variables should remain in the final model.

1.2. *Objectives of the present study*. The present study will concern itself with two different model building procedures, called "Forward Selection" and "Sequential Deletion". We will consider these for use in model building when the experimenter believes that the usual error assumptions are appropriate in his full regression

Received March 19, 1970; revised January 4, 1971.

[1] This research was partially supported by the National Science Foundation Grant GP9046.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org