The Annals of Mathematical Statistics
1972, Vol. 43, No. 4, 1076-1088

MULTIPLE COMPARISON OF REGRESSION FUNCTIONS[1]

BY EMIL SPJØTVOLL

University of Oslo

1. Introduction. The situation discussed under the heading "Regression Analysis", treated in many textbooks, is that we have an $n \times 1$ random variable y which is $N(X'\beta, \sigma^2 I)$, where X' is a known $n \times p$ matrix, β is an unknown $p \times 1$ vector parameter and σ^2 is an unknown variance. The statistical problems are to estimate parameters and to test hypotheses concerning the parameters.

In practice, however, the situation is not so simple in most cases. Often the true form of the expectation of y is not known, but one has some variables which one suspects contribute to Ey in some way. Let these variables be the columns of the matrix X'. One then tries to describe Ey by $X'\beta$ for some β. But one cannot be sure, even with a large number p of variables that the statement "Ey is equal to $X'\beta$ for some β" is true. If y and the p variables of X' have a joint multinormal distribution, the above statement will be true conditionally given the p variables of X'. This argument cannot be used in many situations. In some, it can easily be seen that some of the variables in X' do not have a normal distribution, or maybe both a transform of a variable and the variable itself occur in X'. It is, however, possible (as remarked by a referee) that conditionally $Ey = X'\beta$ even if y and the variables in X' do not have a joint multinormal distribution.

Since usually too many variables are included in X', procedures have been developed for excluding variables which do not contribute to Ey; see, for example, Beale, Kendall and Mann (1967), and Draper and Smith (1966). An important class of such procedures are so-called stepwise regression methods. A description of some of these can be found in Draper and Smith (1966). It seems to be commonly accepted that stepwise regression methods lack justification by statistical theory. In particular the fact that different stepwise regression methods often give different results is confusing, see Hamaker (1962), and Draper and Smith (1966).

What one often ends up with is several regression functions which seem to be good candidates for the regression function to be used. Some of these regression functions might have been obtained by use of a stepwise procedure, some might have been obtained because the statistician has looked at some particular combination of variables. Lately, efficient computer procedures have been developed for computing all possible regression functions or certain subsets containing the "best" regression function; see [8] and [10]. The statistician, therefore, is faced with the problem of choosing among a (usually) large number

Received May 1, 1969; revised October 18, 1971.

[1] This paper was written while the author was visiting at the University of California, Berkeley, and revised at the University of Wisconsin and the University of Oslo.