THE SELECTION OF VARIATES FOR USE IN PREDICTION WITH SOME COMMENTS ON THE GENERAL PROBLEM OF NUISANCE PARAMETERS

By Harold Hotelling

1. Maximum Correlation as a Test. For predicting or estimating a particular variate y there is frequently available an embarrassingly large number of other variates having some correlation with y. For example, in fitting demand functions by means of economic time series, the number of series of observations having some relation to the demand which is sought to be estimated is apt to be very large, whereas the number of good independent observations on each is quite small. The proper coefficients in the regression equation must ordinarily be determined from the observations, and must not exceed in number the observations on each variate. Furthermore, in order to have a measure of error that will make it possible to distinguish real effects from those due to chance, it is necessary that the number of predictors[1] shall be enough less than the number of observations on each variate so that the residual chance variance can be determined with an appropriate degree of accuracy. It is desirable to select a set of predictors yielding estimates of maximum but determinable accuracy, and at the same time to avoid the fallacies of selection among numerous results of that one which appears most significant and treating it as if it were the only one examined.

Considerations other than maximum and determinate accuracy are of practical importance. The labor of calculation by the method of least squares becomes a serious obstacle to the use of the theoretically optimum set of variates when these are very numerous, though the rapid current development of mechanical and electrical devices suitable for these computations offers a hope that the limits now set in practice in this way will soon be considerably increased. Furthermore, predictions or estimates must, as in speculative business or in military activity, be made from moment to moment, often in a rough manner by persons incapable of or averse to using complex formulae, and in such activities frequent revisions of the regression equations must be made to accord with altered conditions. Also, in temporal predictions, the time of availability of

[1] I use this term for what are often called the independent variates in a regression equation, since these ordinarily are not really independent in the probability sense. Similarly I shall call the "dependent" variate the *predictand*. By *prediction* I mean merely the use of regression equations to estimate some unknown variate by means of the values of related variates, without any necessary connotation of temporal order, though the most interesting applications seem for the most part to be those in which we pass from a knowledge of the past to an estimate of the future.