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.

the values of the predictors is important, since an early prediction (e.g. of the size of a harvest) is more valuable than a later one of the same accuracy.

If we make the usual assumption[2] that the probability distribution of y is, for every set of values of the predictors, normal with a fixed variance σ^2 and an expectation that is a linear function of the predictors, we shall wish to minimize σ^2 subject to appropriate limitations, and this amounts to the same thing as maximizing the multiple correlation ρ of y with the predictors, since $1 - \rho^2$ is the ratio of σ^2 to the total variance of y, which is the same for all sets of predictors. The estimates s and R of σ and ρ obtained from the available sample are of course a different matter. But it is clear that the value of R provides a suitable criterion of choice under the following conditions: We are called upon to choose one among two or more sets, each consisting of a fixed number of predictors; for each predictor we have a known value corresponding to each of the values $y_1, \cdots, y_N$ observed for the predictand; and there is no basis for preferring one of these sets to another either in theory, in observations extraneous to those just specified, or in cost or time of availability. In particular, if just one predictor is to be used, that having the highest sample correlation with the predictand should under these conditions be the one adopted. But in making such a choice a test of its accuracy is required, to take account of the possibility that the wrong choice has been made because of chance fluctuations in the sample correlation coefficients.

There are innumerable economic variates available for prediction of business conditions, and most of these are highly correlated with each other. The selection of one business index instead of another for a particular purpose will involve the question which has exhibited the higher correlation with the quantity to be predicted, and consequently the question of the definiteness with which the difference between the calculated correlations can be regarded as significant.

Our problem evidently has a bearing on governmental policy in selecting among the numerous series of data those whose continuation will be most valuable. The high cost of assembling these statistics dictates a careful selection of a limited number of series having little correlation with each others' current values, but with correlations as great as possible with those things whose prediction or estimation is most important.

2. The Choice of one Predictor with Two Available. Let us take first the simplest case, which may be illustrated by a Michigan State College problem of

[2] We shall not here go into the question of the applicability of these standard assumptions to time series otherwise than to note that some transformations of observations ordered in time are usually necessary *and sufficient* to obtain quantities satisfying the assumptions so closely that deviations from them cannot be detected. Such transformations include replacing a variate by its logarithm, and eliminating trend and seasonal variations by least squares. In view of the satisfactory adjusted observations found empirically by these and similar methods, the usual objections to studying time series by exact methods seem much exaggerated.

which Dr. W. D. Baten has told me. The ultimate weight of a mature ox is estimated by means of his length at an early age. The question has been raised, however, whether a more accurate prediction might not be made by means of the calf's girth at his heart. Records were at hand of 13 oxen showing their lengths and girths as calves and also their weights when mature. A regression equation involving both length and girth would presumably give greater accuracy than either variate alone; but it appears that those who make the estimates desire a simple formula involving only one variate. Suppose, then, that in such a sample the correlation of weight with length is $r_1 = .7$, that the correlation of weight with girth is $r_2 = .5$, and that the correlation of girth with length is $r_0 = .4$. Is the difference $r_1 - r_2 = .2$ sufficiently great in relation to its sampling errors to warrant the inference that girth is really a better predictor than length, or must the question be left in abeyance until more observations can be accumulated?

A straightforward procedure which would have been used with little question before the advent of modern exact methods is to calculate the asymptotic approximation to the standard error of $r_1 - r_2$ by the differential method, assuming the three variates to have the trivariate normal distribution, and to regard the difference of the correlations as significant if it exceeds a multiple of this standard error determined by the tables of the normal distribution. The calculation of the asymptotic approximation $\sigma_{r_1-r_2}$ may be carried out in the following manner. Let ρ_1, ρ_2, and ρ_0 be the population values of r_1, r_2, and r_0 respectively. Then if σ_{ij} denote the population covariance of x_i and $x_j (i, j = 0, 1, 2)$, we have

$$\rho_1 = \frac{\sigma_{01}}{\sqrt{\sigma_{00}\sigma_{11}}},$$

with similar formulae for ρ_2 and ρ_0. Likewise the sample estimates of these parameters are given by such expressions as

$$r_1 = \frac{s_{01}}{\sqrt{s_{00}s_{11}}}.$$

Taking the logarithm of this last expression, expanding about the population values, denoting by the operator δ the deviation of sample from population values of the covariances, and the resultant deviation in r_1, and dropping terms of order higher than the first, we have:

$$\delta r_1 = \rho_1\left(\frac{\delta s_{01}}{\sigma_{01}} - \frac{\delta s_{00}}{2\sigma_{00}} - \frac{\delta s_{11}}{2\sigma_{11}}\right).$$

In the same way

$$\delta r_2 = \rho_2\left(\frac{\delta s_{02}}{\sigma_{02}} - \frac{\delta s_{00}}{2\sigma_{00}} - \frac{\delta s_{22}}{2\sigma_{22}}\right).$$

The asymptotic value of the sampling covariance is obtained by multiplying these two expressions together and taking the expectation. The sampling covariance of two estimates of covariance of the usual kind (sum of products

divided by number of degrees of freedom) in the same sample, having n degrees of freedom (which ordinarily means that there are $n + 1$ individuals in the sample and that the means are eliminated), is given exactly by the formula[3]

$$E(\delta s_{ij}\delta s_{km}) = (\sigma_{ik}\sigma_{jm} + \sigma_{im}\sigma_{jk})/n,$$

in which the subscripts may have any values, equal or unequal. When this formula is applied to each of the nine terms of the product and the results are expressed in terms of the correlations ρ_i, there results the asymptotic expression for the covariance given by

$$nE(\delta r_1\delta r_2) = \tfrac{1}{2}\rho_1\rho_2(\rho_1^2 + \rho_2^2 + \rho_0^2 - 1) + \rho_0(1 - \rho_1^2 - \rho_2^2).$$

This method provides also one of the derivations of the familiar formula which may be written

$$n\sigma_{r_1}^2 = nE(\delta r_1)^2 = (1 - \rho_1^2)^2, \qquad n\sigma_{r_2}^2 = (1 - \rho_2^2)^2.$$

The variance of the difference of r_1 and r_2 is the sum of their variances minus twice their covariance. Hence

$$n\sigma_{r_1-r_2}^2 = (1 - \rho_1^2)^2 + (1 - \rho_2^2)^2 - \rho_1\rho_2(\rho_1^2 + \rho_2^2 + \rho_0^2 - 1) + 2\rho_0(\rho_1^2 + \rho_2^2 - 1).$$

We are testing the hypothesis that $\rho_1 = \rho_2$. If we put a common value ρ for them in the last expression and simplify, we obtain for the standard error of the difference,

$$\sigma_{r_1-r_2} = \sqrt{\frac{(1 - \rho_0)(2 - 3\rho^2 + \rho_0\rho^2)}{n}}.$$

The second factor in parentheses is always positive because of the inequalities limiting the correlations among three variates.

This formula contains two unknown parameters, ρ and ρ_0. The classical procedure would be substitute r_1, r_2 and r_0 respectively for ρ_1, ρ_2, and ρ_0 in the previous formula, and use the resulting standard error expression as if the ratio to it of $r_1 - r_2$ were normally distributed. A first modification, more in line with modern ideas, would be to use some kind of average of r_1 and r_2 as an estimate of both ρ_1 and ρ_2, since the null hypothesis tested is that these are equal. But whatever sample estimates we substitute for ρ and ρ_0, the formula remains unsatisfactory, since no suitable limits of error are available. If instead of the standard error we were to work out the exact distribution of $r_1 - r_2$ we should still not be free from the difficulty. This exact distribution clearly involves both ρ and ρ_0, since its variance does so. Neither can we escape from the trouble by using some function $z = f(r)$, such as the inverse hyperbolic tangent suggested by R. A. Fisher, and considering the standard error of $z_1 - z_2 =$

[3] I have given a derivation of this formula from the characteristic function of the multivariate normal distribution [1]. Numerous special cases appear in earlier literature. The derivation above is a simplification and improvement of several versions, appearing in the various early writings of Karl Pearson.

$f(r_1) - f(r_2)$; for this standard error will have as the first term in its expansion in a series of powers of n^{-1} simply the product of the expression above for $\sigma_{r_1-r_2}$ by $f'(\rho)$; and this must clearly involve both ρ_0 and ρ.

3. Nuisance Parameters. This is not by any means the only statistical problem in which unknown and undesired parameters enter into the distribution of the statistic which we should naturally use to test a hypothesis. Indeed, the early investigation which was perhaps most influential in setting the whole tone of modern statistical research was that [2] in which W. C. Gosset ("Student") arrived at the exact distribution of the ratio of a deviation in the mean to the *estimated* standard error. The previous practice (which unfortunately survives today in some quarters, and is even taught to students without explaining its approximate character) was to neglect the sampling errors in the estimate of the unknown variance σ^2 and to treat the ratio as normally distributed with unit variance. The rigorous derivation by Fisher [3] of the Student distribution makes clear the manner in which the nuisance parameter ϑ may in this, and in some other, problems be eradicated from the distribution through integration, after altering the original statistic (the deviation in the mean) by dividing it by another statistic. The new statistic, the Student ratio, vanishes whenever the old statistic, the deviation in the mean, does so, and the same hypothesis is tested by both. This then is one way to get rid of a nuisance parameter: when you have a statistic estimating a parameter whose vanishing is in question, but whose distribution involves another parameter, alter the statistic by multiplying or dividing by another statistic in such a way that the new function vanishes whenever the old one does so; and *do this in such a way that the new distribution will be independent of the nuisance parameter.* Unhappily, this method has been applied successfully only in particular cases, and no way to use it in the problem at hand has been found.

A second method is that of transformation employed by Fisher in dealing with such problems as testing the significance of the difference between the correlation coefficients in independent samples between the same two variates. The need for the transformation in this case is occasioned by the presence in the distribution of the difference of the sample correlations of the unknown true value, which is not directly relevant to the comparison. We have seen that this method also fails to solve our problem.

A third method of dealing with nuisance parameters is the use of fiducial probability by R. A. Fisher [4] and by Daisy M. Starkey [5] in testing the significance of the difference between the means of two samples when the variances may be unequal. Criticisms of these applications of fiducial probability have been made by M. S. Bartlett [6] and B. L. Welch [7], and the field of applicability of such methods is still in need of elucidation.

Some findings of J. Neyman [8] having a bearing on the general nuisance parameter problem should also be noted.

The only other class of methods for dealing with nuisance parameters of which

I am aware involves the comparison of the particular sample obtained, not with the whole population of samples with which a comparison might be made if we knew the value of the troublesome parameter, but with a sub-population selected with reference to the sample in such a way that the distribution, in this sub-population, of the statistic used does not involve any unknown parameter. An example is the testing of significance of a regression coefficient. Thus if we suppose that a sample of values of x and y is drawn from a bivariate normal population, and calculate the regression coefficient b of y on x in the sample, the distribution of b involves not only the population value β, but also the ratio α of the variances in the population. Since this second parameter is unknown, and can only be estimated from the sample, it is not possible to use the distribution of b in the whole population directly to test the significance of $b - \beta$. What we do is to find the place of this difference, not in the whole population of values in which both x and y are drawn at random, but in a sub-population for which the values of x are the same as in our sample. We may alternatively say that we limit the sub-population only to that for which the sum of the squares of the deviations of the values of x from their mean is the same as in our sample; the results are the same. The distribution in this sub-population of the ratio of $b - \beta$ to its estimated standard error is of the Student form, with no unknown parameters, and on this basis it is possible to make exact and satisfactory tests and to set up fiducial limits for b. Another example is that of contingency tables. The practice now accepted (after a controversy) for testing independence of two modes of classification, such as classification of persons according as they have or have not been vaccinated, and again according as they live through an epidemic or die, is to compare the observed contingency table, not with all possible contingency tables of the same numbers of rows and columns, but only with the possible contingency tables having exactly the same marginal totals as the observed table.

4. An Exact Solution. We shall solve the problem of the significance of the difference of r_1 and r_2 with the understanding that the meaning of significance is to be interpreted by reference to the sub-population of possible samples for which the predictors x_1 and x_2 have the same set of values as those observed in the particular sample available. This procedure, besides yielding an exact distribution without unknown parameters, has the advantage of relaxing the stringency of the requirement of a trivariate normal distribution. We now make only the assumptions customary in the method of least squares, that the predictand y has the univariate normal distribution for each set of values of x_1 and x_2, independently for the different sets, with a common variance σ^2, and with the expectation of y for a fixed pair of values of the predictors a linear function of these predictors. No assumption is involved regarding the distribution of the predictors, since we regard them as fixed in all the samples with which we compare our particular sample. The advantages of exactness and of freedom

from the somewhat special trivariate normal assumption are attained at the expense of sacrificing the precise applicability of the results to other sets of values of the predictors.

Since the correlational properties are unchanged by additive and multiplicative constants, we may suppose that

$$Sx_1 = 0 = Sx_2, \qquad Sx_1^2 = 1 = Sx_2^2, \tag{1}$$

where S stands for summation over a sample of N individuals. The notation may be made more explicit by the adjunction of an additional subscript α, varying from 1 to N, to denote the individual member of the sample, so that instead of Sx_1, for example, we might write $Sx_{1\alpha}$. The omission of this additional subscript is convenient and will usually leave no ambiguity when we deal with sums, but it will be convenient to retain it in connection with individual values. The correlation r_0 of x_1 with x_2 in all those samples we shall consider is, by (1)

$$r_0 = Sx_1x_2.$$

Now consider the new quantities

$$x'_\alpha = \frac{x_{1\alpha} - x_{2\alpha}}{\sqrt{2(1 - r_0)}}, \qquad x''_\alpha = \frac{x_{1\alpha} + x_{2\alpha}}{\sqrt{2(1 + r_0)}}. \tag{2}$$

Evidently, from (1) and (2),

$$Sx' = 0 = Sx'', \qquad Sx'^2 = 1 = Sx''^2, \qquad Sx'x'' = 0. \tag{3}$$

Since the mean value $E(y_\alpha)$ is a linear function of $x_{1\alpha}$ and $x_{2\alpha}$, y_α may, upon subtracting a constant from all these expectations, be written

$$y_\alpha = \beta_1 x_{1\alpha} + \beta_2 x_{2\alpha} + \Delta_\alpha, \tag{4}$$

where $\Delta_1, \cdots, \Delta_N$ are normally and independently distributed with variances all equal to σ^2 and expectations zero. The assumption that x_1 and x_2 are equally correlated with y in the population leads to the conclusion that $\beta_1 = \beta_2$; and putting $\beta = \beta_1\sqrt{2(1 + r_0)}$, we then have from (4) and (2):

$$y_\alpha = \beta x''_\alpha + \Delta_\alpha. \tag{5}$$

Consequently, by (3)

$$Sx'y = Sx'_\alpha y_\alpha = \beta Sx'x'' + Sx'\Delta = Sx'\Delta;$$

and this function has a normal distribution with zero mean and variance σ^2.

If in the sample we work out a regression equation

$$Y = a + b'x' + b''x'',$$

the normal equations for determining b' and b'' must by (3) take the simple forms

$$a = \bar{y}, \qquad b' = Sx'y, \qquad b'' = Sx''y.$$

From the general theory of least squares it is known that the sum of squares of residuals is

$$Sv^2 = S(y - Y)^2 = Sy^2 - \bar{y}Sy - (Sx'y)^2 - S(x''y)^2,$$

and that Sv^2/σ^2 has the χ^2 distribution with $n = N - 3$ degrees of freedom, independently both of $Sx'y$ and of $Sx''y$. From these facts it follows that

$$t = Sx'y\sqrt{\frac{n}{Sv^2}} \tag{6}$$

has the Student distribution with n degrees of freedom. Since in accordance with the foregoing definitions and (1) we have

$$Sx'y = (r_1 - r_2)\sqrt{\frac{S(y - \bar{y})^2}{2(1 - r_0)}},$$

and since also it is known that

$$Sv^2 = S(y - \bar{y})^2 \frac{D}{1 - r_0^2},$$

where

$$D = \begin{vmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_0 \\ r_2 & r_0 & 1 \end{vmatrix},$$

(6) may be written

$$t = (r_1 - r_2)\sqrt{\frac{n(1 + r_0)}{2D}}. \tag{7}$$

The probability of a greater value of $|t|$ is given by tables of the Student distribution with $n = N - 3$. If this probability is sufficiently small (which conventionally means less than .05, or sometimes .01) we have a corresponding degree of confidence that the variate chosen because of a higher correlation in the sample has actually a higher correlation than the other in the population.

5. The Selection of One Variate from Among Three or More. Suppose that we are to choose one of the variates $x_1, \cdots, x_p$ in order to predict y. $(p < N - 1)$ We choose the one having highest correlation, and wonder how much confidence to place in this choice. We shall now determine the distribution of a function suitable for testing the hypothesis that there is no real difference between any pair of the correlations of $x_1, \cdots, x_p$ with y. Again we shall assume the values of these predictors fixed, and look for the place of our particular sample among all samples having these values, with only y free to vary normally by chance.

Let $a_{ij} = S(x_i - \bar{x}_i)(x_j - \bar{x}_j)$, and let c_{ij} be the cofactor of a_{ij} in the determinant a of these quantities, divided by a. Then

$$\sum a_{ij}c_{ik} = \delta_{jk} = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases} \tag{8}$$

Here Σ stands for summation from 1 to p. Let

$$w_i = \frac{\sum_j c_{ij}}{\sum\sum c_{ij}}, \tag{9}$$

$$l_i = S(x_i - \bar{x})y, \tag{10}$$

$$l = \Sigma w_i l_i. \tag{11}$$

From (9) it follows that

$$\Sigma w_i = 1. \tag{12}$$

From the hypothesis that y is in the population equally correlated with all the x_i it follows that $l_1, \cdots, l_p$ have equal expectations, which we may denote by λ; and from (11) and (12) it follows that also $E(l) = \lambda$. Obviously

$$E(l_i - \lambda)(l_j - \lambda) = \sigma^2 a_{ij}, \tag{13}$$

where σ^2 is the variance of those values of y corresponding to a fixed set of values of the x's. From (11), (13) and (9) we obtain

$$E(l - \lambda)^2 = \frac{\sigma^2}{\Sigma\Sigma c_{ij}}. \tag{14}$$

Since the l_i are linear functions of the y's, they have the multivariate normal distribution. From the theory of this distribution and the values (13) of the covariances it follows that the distribution has the form

$$(2\pi)^{-\frac{1}{2}p} a^{-\frac{1}{2}} \sigma^{-p} e^{-T/2\sigma^2}\, dl_1 \cdots dl_p,$$

where a is the determinant of the a_{ij}'s, and

$$T = \Sigma\Sigma c_{ij}(l_i - \lambda)(l_j - \lambda).$$

We may introduce linear functions $l_1', \cdots, l_p'$ of $l_1 - \lambda, \cdots, l_p - \lambda$ such that $T = l_1'^2 + \cdots + l_p'^2$, and such that $l_p'^2 = (l - \lambda)^2 \Sigma\Sigma c_{ij}$. Now $\dfrac{l_1'^2 + \cdots + l_{p-1}'^2}{\sigma^2}$ has the χ^2 distribution with $p - 1$ degrees of freedom. The numerator of this expression equals

$$\begin{aligned} T - l_p'^2 &= \Sigma\Sigma c_{ij}(l_i - \lambda)(l_j - \lambda) - (l - \lambda)^2 \Sigma\Sigma c_{ij} \\ &= \Sigma\Sigma c_{ij} l_i l_j - l^2 \Sigma\Sigma c_{ij} \\ &= \Sigma\Sigma c_{ij}(l_i - l)(l_j - l). \end{aligned}$$

The penultimate form shows that this function is independent of λ; the last, as a positive definite form in the deviations of the l's from their weighted mean, shows that sufficiently large values of the expression will reveal with definiteness the inequality of the predicting powers of the p variates when this exists.

It is well known that the regression coefficients of y upon the set of variates $x_1, \cdots, x_p$ are completely independent of the sum of squares Sv^2 of residuals from the regression equation. Since the l's are linear functions of these regression coefficients, (namely the linear functions appearing in the normal equations), they also are independent of Sv^2. Hence, if we put

$$s_1^2 = \frac{\Sigma\Sigma c_{ij} l_i l_j - l^2 \Sigma\Sigma c_{ij}}{p - 1},$$

$$s_2^2 = \frac{Sv^2}{N - p - 1},$$

the ratio $F = s_1^2/s_2^2$ will, in case of equality of the correlations of the various x's with y, have the variance ratio distribution with $n_1 = p - 1$ and $n_2 = N - p - 1$ degrees of freedom. When $p = 2$ this test reduces exactly to (7), as it should, and $F = t^2$.

In the numerical application of this method, the regression coefficients b_i of y on $x_1, \cdots, x_p$ should first be worked out by the inverse matrix method. The right-hand members of the normal equations are $l_1, \cdots, l_p$, the coefficients in these equations are the a_{ij}, and the calculation of s_1^2 is simplified with the help of the identity

$$\Sigma\Sigma c_{ij} l_i l_j = \Sigma b_i l_i .$$

6. Selection of Additional Variates When Some Have Been Chosen. Suppose now that q predictors have been included definitely in the regression equation, and that one more is to be selected for inclusion among p additional predictors that are available. The criterion now is that that one should be chosen tentatively which has the highest partial correlation with the predictand, eliminating those already definitely chosen; but the confidence to be placed in the choice is to be judged by an adaptation of the criterion of the preceding section. It is only necessary to consider the a_{ij}, l_i, c_{ij} and b_i $(i, j = 1, \cdots, p)$ as calculated from the new predictors and the deviations of y from the regression equation on the predictors already adopted. Formulae may easily be derived for the values of these quantities in terms of those already found and the sums of products, so as to simplify the calculations. Sv^2 will now stand for the sum of squares of residuals from the regression equation involving all the $p + q$ predictors. It is to be divided by $N - p - q - 1$ to obtain s_2^2. The numbers of degrees of freedom with respect to which F is to be judged are now $n_1 = p - 1$ and $n_2 = N - p - q - 1$. When $p = 2$ this test, like that of the preceding section, reduces to the use of the t-distribution of (7), with $n = N - q - 3$, and the correlations standing for partial correlations eliminating the predictors already definitely chosen.

A special instance in which this procedure is applicable is in economic time series, in which time, in the form of orthogonal polynomials, must ordinarily be "partialled out" in order that tests of significance may be sound.

7. Further Problems. It is natural to ask whether the foregoing work can be extended to examine the soundness of the selection, on the basis of a greater multiple correlation, of a particular set of two or more variates, chosen from among several such sets. The simplest such problem that goes beyond what has been done above deals with two sets, each of two predictors, having in a sample multiple correlations R and R' with the predictand. The question is whether the difference $R - R'$ is significant.

Suppose that, in the interests of simplicity and the hope of attaining a solution satisfactorily free from unknown parameters, we assume as before that the predictors have a fixed set of values, the same in all samples. Since multiple correlations are invariant under linear transformations of predictors, we may without loss of generality assume that the predictors in each set are mutually uncorrelated and have sums of squares equal to unity. Indeed, we may go somewhat further in standardizing the sets of values to which consideration can be confined without loss of generality, with the help of some ideas introduced in the paper [1]. In the terminology of that paper, the variates in each set may be considered *canonical* with respect to the relationship between the sets. This means that linear functions x_1 and x_2 of the two variates in one set, and linear functions x_1' and x_2' of those in the other set, can be chosen so as to satisfy not only the conditions

$$\begin{aligned} Sx_1 = Sx_2 = Sx_1' = Sx_2' &= 0 \\ Sx_1^2 = Sx_2^2 = Sx_1'^2 = Sx_2'^2 &= 1 \\ Sx_1x_2 = 0 = Sx_1'x_2' &, \end{aligned} \tag{15}$$

but also the further conditions

$$Sx_1x_2' = 0 = Sx_2x_1' . \tag{16}$$

This means that, for all the purposes in view, the two sets of predictors can be characterized as to their mutual relationships by the values of the remaining two sums of products, namely

$$c_1 = Sx_1x_1' , \qquad c_2 = Sx_2x_2' .$$

In view of the conditions assumed earlier, c_1 and c_2 are what have been called the *canonical correlations* between the two sets.

To the sets thus standardized, the predictand y is related in a manner expressed by the population regression coefficients β_1 and β_2 of y on the first set, and β_1' and β_2' on the second. If we take y as having unit variance in the population, the squared multiple correlation coefficients in the two cases will be

$$\rho^2 = \beta_1^2 + \beta_2^2 , \qquad \rho'^2 = \beta_1'^2 + \beta_2'^2 .$$

The hypothesis to be tested is that $\rho = \rho'$. If b_1, b_2, b_1', b_2' denote the sample estimates of the regression coefficients, the statistic appropriate for the test would appear necessarily to be proportional to

$$w = \tfrac{1}{2}(b_1^2 + b_2^2 - b_1'^2 - b_2'^2).$$

The sample regression coefficients are normally distributed, with population correlations equal to the sample correlations among the corresponding predictors. The variance of each is σ^2. Thus their joint distribution may be written down at once, in a rather simple form in view of (15) and (16). From this it is possible to determine directly the characteristic function $M(t) = Ee^{tw}$ of w. If we write $K(t) = \log M(t)$ we obtain:

$$2K(t) = \Sigma\{(\beta_j^2 - 2c_j\beta_j\beta_j' + \beta_j'^2)t^2 + (\beta_j^2 - \beta_j'^2)t\}\{1 - (1 - c_j^2)t^2\}^{-1} - \Sigma \log\{1 - (1 - c_j^2)t^2\}.$$

Here the summations are with respect to j over the values 1 and 2. If each set of predictors had had s members, the same result would hold for $K(t)$ except that the summations with respect to j would then extend from 1 to s.

This is a very disappointing result because it contains so many parameters. The distribution of w must contain the same parameters as its characteristic function. All the four parameters β_j , β_j' appear in the expression above, though their effective number is reduced to three by the condition that the two sums of squares shall be equal which constitutes the hypothesis under test. The distribution of w thus contains at least three unknown parameters besides σ.

The estimate of variance s^2 obtained from the residuals from the grand regression equation of y on x_1 , x_2 , x_1' , and x_2' is independent of w. Its distribution is of the usual form and involves a parameter, the population variance, which is a function of β_1 , β_2 , β_1' , and β_2' . We could therefore pass by a single integration from the distribution of w to that of the statistic w/s^2, which vanishes with w, and which on this account, and on grounds of physical dimensionality, might be considered appropriate to test the hypothesis that $\rho = \rho'$. The question may be raised whether the distribution of this ratio might not be free from parameters. The answer unfortunately is in the negative, as appears from an examination of the characteristic function of the ratio. Even in the simplified case in which all the c_j are equal, a troublesome parameter persists in the distribution.

Thus we meet again the problem of nuisance parameters, and this time no escape is visible. Perhaps some such artifice as those enumerated in paragraph 3 (for example, some further limitation of the sub-population within which we should seek the place of our particular sample) is capable of yielding an exact, or "studentized" distribution, but this has not yet been found. The problem is of considerable interest, not only because of its practical importance, but because of its suggestiveness in connection with general theory.

Numerous other problems having both practical importance and general theoretical interest are associated with the selection of predictors. For example, we have not dealt at all with the problem of the *number* of predictors that should be used when maximum accuracy in prediction, or in evaluation of the regression coefficients, is the sole criterion. A particular case is the determination of the degree of the regression polynomial which should be fitted to obtain

maximum accuracy, for example of the number of orthogonal polynomials in fitting a trend. Such customary criteria as minimizing the *estimated* variance of deviations, in which the sum of squares which is the numerator and the number of degrees of freedom which is the denominator both diminish to zero as the number of variates is increased, do not rest upon any satisfactory general theory.

Another related set of problems is concerned with variates more numerous than the observations on each. It is clear that there is real information inherent in data of this kind, but existing theory and methods, including those of the present paper, are not adequate to utilize it in a thoroughly efficient manner. A recent paper of P. L. Hsu [9] is unique in not excluding the case in which the variates outnumber the observations.

8. Summary. A criterion has been obtained for judging the definiteness of the selection of a particular variate, from among several available for prediction, on the basis of its having the maximum sample correlation with the predictand. A variation of this criterion is applied in paragraph 6 to the problem of extending the list of variates to be used in a regression formula.

Some of the problems of "nuisance parameters" which affect general theory are illustrated in this problem. Some outstanding unsolved problems related to these questions are discussed in paragraph 7.

REFERENCES

[1] Harold Hotelling, "Relations Between Two Sets of Variates," *Biometrika*, Vol. 28 (1936), pp. 321–377.

[2] "Student," "The Probable Error of a Mean," *Biometrika*, Vol. 6 (1908), pp. 1–25.

[3] R. A. Fisher, "Applications of 'Student's' Distribution," *Metron*, Vol. 5 (1925), pp. 90–104.

[4] R. A. Fisher, "The Fiducial Argument in Statistical Inference," *Annals of Eugenics*, Vol. 6 (1935), pp. 391–398. See also Fisher's answer to Bartlett in the *Annals of Math. Stat.*, Vol. 10 (1939), pp. 383–388 and the references there given.

[5] Daisy M. Starkey, "A Test of the Significance of the Difference Between Means of Samples from Two Normal Populations Without Assuming Equal Variances," *Annals of Math. Stat.*, Vol. 9 (1938), pp. 201–213.

[6] M. S. Bartlett, "The Information Available in Small Samples," *Proc. Camb. Phil. Soc.*, Vol. 32 (1936), pp. 560–566.

[7] B. L. Welch, "On Confidence Limits and Sufficiency, with Particular Reference to Parameters of Location," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 58–69.

[8] *Statistical Research Memoirs*, Vol. 2 (1938), pp. 58–59.

[9] P. L. Hsu, "On the Distribution of Roots of Certain Determinantal Equations," *Annals of Eugenics*, Vol. 9 (1939), pp. 250–258.

COLUMBIA UNIVERSITY,
NEW YORK, N. Y.