CHAPTER 13

THE SELECTION OF PREDICTOR VARIABLES

13.1 Introduction

The typical approach to the problem of the prediction of job or academic performance employs multiple regression techniques. As we indicated in the previous chapter, standard nonlinear regression methods and some other prediction methods, such as those using moderator variables, can be reduced to problems in linear multiple regression.

In Chapters 6 and 12, we discussed some of the problems involved in comparing two competing tests. In this chapter, we shall be concerned with a slightly different problem. We assume that we are able initially to obtain some large number of potential predictors and wish to select some smaller number for use on a continuing basis. After discussing some relevant sampling problems in the next section, we describe two convenient procedures for selecting some smaller set of predictor variables from a larger set. In Section 13.4, we shall briefly discuss the very difficult problem of using data from a small sample to make predictions about a second sample, and we shall illustrate many of the techniques discussed in this and the previous chapter with a typical validity study.

In Section 13.5, we shall derive formulas for the effect of changes in test length on reliability, validity, and covariance matrices of predictor variables for the multipredictor case. These results are then used in conjunction with a very general problem in predictor variable selection. For this problem, we assume that the length (in time or number of items, as appropriate) of each predictor may be increased or decreased as required. We then determine the amount of time that, for optimal prediction, should be assigned to each predictor under the restriction that the total testing time is some fixed constant.

13.2 Some Sampling Problems

The development in the previous chapter involved population parameters, which we assumed to be known exactly. In practice, of course, this is never the case and often the sample data at hand are not of enormous size. This introduces further problems, some of which have not yet proved amenable to analytic solution. In this and the next section, we shall discuss some of these problems, the few analytic results which bear on them, and the inference procedures that have been adopted to partially overcome them.

If variables are selected for a prediction battery on the basis of sample correlations, the investigator will be selecting not only on the basis of true correlation but also on the basis of sampling errors. In Section 3.7, we showed that the regression model yields an expected true score that is less than the examinee's observed score for all observed scores greater than the mean population observed score. Similarly it is apparent that the regression estimate of true correlation is lower than the obtained sample correlation for observed correlations that are greater than the average of the observed correlation. Since selection typically involves choosing some small number of variables with the highest sample correlations, the effect of selection is usually to overestimate the true multiple correlation for the variables selected. We describe this by saying that there is a "capitalization on chance" when variables are selected in this way. Typically a "shrinkage" in the multiple correlation is generally found when these variables are used on a new sample.

Because the basic statistical problems in this area remain unsolved, an empirical approach to the problem is necessary. One commonly used procedure is to select variables on the basis of one sample, the *screening sample*, and then to estimate the multiple correlation and regression weights for these selected variables on the basis of a second sample, the *calibration sample*, for which the predictor variables have been prespecified. If variables are selected and regression weights estimated in a single sample, then it generally is advisable to apply these variables and weights in a new sample to see how valuable the specified composite is. This is called *cross validation*.

Even with a moderate sample size, the amount of shrinkage in the multiple correlation between the screening and calibration or cross validation samples may, in fact, be substantial if a very small number of predictor variables has been selected from a very large set of potential predictors. If, additionally, the sample size in the screening sample is small, the shrinkage can be nearly total.

If the reader has been left unworried by the above remarks, he should acquaint himself with the shrinkage encountered by Mosteller and Wallace (1964, Chapter 5) in their discriminant function analysis of the Federalist papers. Had they not had the foresight to reserve some of their data for use as a calibration sample, they would have overestimated the true "discrimination index" for their discriminators *by more than 50%*. Also, two cross validation studies are described in Section 13.5; in one case the shrinkage is far less drastic, in the other the shrinkage is almost total.

Actually, because of errors due to sampling of persons, the sample squared multiple correlation coefficient for *prespecified* variables has a positive bias as an estimator of the true multiple correlation. When the random variables have a multivariate normal distribution, however, a correction for *this* bias is possible. Olkin and Pratt (1958) have derived an unbiased estimator of the squared multiple correlation. For most practical use, the approximation

$$\widehat{\rho^2} = r^2 - \frac{n-2}{N-n-1}(1-r^2) - \frac{2(N-3)}{(N-n-1)(N-n+1)}(1-r^2)^2, \tag{13.2.1}$$

where $r^2 = r^2_{0\cdot 12\ldots n}$ is the sample multiple correlation, n is the number of predictors, and N is the sample size, will be satisfactory. An unfortunate feature of the unbiased estimator and the approximation to it (and many other unbiased estimators of positive quantities) is that they may sometimes take on negative values.

In contrast to the correlation coefficient, very simple unbiased estimates of the regression weights are available. If the theory of minimum mean squared error is applied in a sample, the resulting estimates (which in this case are just the sample regression weights) provide unbiased estimates of the corresponding population quantities. These estimates are called *least squares estimates.* The application of the method of least squares to the estimation of variances and partial variances leads to the unbiased estimate $[N/(N-n-1)]s^2$, where s^2 is the relevant sample partial variance, N is the sample size, and n is the number of predictor variables. For the zero-order variance ($n = 0$), this reduces to the familiar form $[N/(N-1)]s^2$.

Wherry (1940) has also provided a simple, alternative (though less accurate) approximate correction for the bias in the sample correlation coefficient; it also provides a reasonable working rule for deciding whether or not to include an additional variable or variables in a regression equation. The squared multiple correlation may be written as (12.5.5):

$$\rho^2_{0\cdot 12\ldots n} = 1 - \frac{\sigma^2_{0\cdot 12\ldots n}}{\sigma^2_0}. \tag{13.2.2}$$

If we denote the corresponding sample variance and partial variance by s^2_0 and $s^2_{0\cdot 12\ldots n}$, respectively, the sample multiple correlation will be given by

$$r^2_{0\cdot 12\ldots n} = 1 - \frac{s^2_{0\cdot 12\ldots n}}{s^2_0}. \tag{13.2.3}$$

Now, replacing the numerator and denominator of the second term on the right-hand side of (13.2.2) by their unbiased estimates $[N/(N-n-1)]s^2_{0\cdot 12\ldots n}$ and $[N/(N-1)]s^2_0$, we obtain the estimate

$$\begin{aligned}\widehat{\widehat{\rho^2}} = 1 - \frac{\left(\dfrac{N}{N-n-1}\right)s^2_{0\cdot 12\ldots n}}{\left(\dfrac{N}{N-1}\right)s^2_0} &= 1 - \left(\frac{N-1}{N-n-1}\right)(1 - r^2_{0\cdot 12\ldots n}) \\ &= \frac{(N-1)r^2_{0\cdot 12\ldots n} - n}{(N-n-1)},\end{aligned} \tag{13.2.4}$$

where N is the sample size, n is the number of predictor variables, and $N > n + 1$. This differs slightly from Wherry's original formula, which has the value N where $N - 1$ is found in the numerator and denominator of (13.2.4).

Formula (13.2.4) and the form originally given by Wherry have been called Wherry's *correction for shrinkage*. This terminology is confusing and undesirable because this "correction" has nothing at all to do with the capitalization on chance that occurs when a sample multiple correlation is obtained from *selected* variables, nor with the resulting shrinkage in the multiple correlation obtained in the calibration sample.

This correction can be justified from a different point of view. Under an assumption of multivariate normality, or otherwise asymptotically under a very broad assumption, it is true that for $n > i$,

$$\frac{(r_n^2 - r_i^2)/(n - i)}{(1 - r_n^2)/(N - n - 1)} \qquad (13.2.5)$$

is distributed as F with $n - i$ and $N - n - 1$ degrees of freedom, r_n^2 and r_i^2 being the sample squared multiple correlations based respectively on n and on any prespecified i of the given n variables (Kendall and Stuart, 1961). Now suppose $F = 1$. Then

$$\frac{(r_n^2 - r_i^2)}{n - i} = \frac{(1 - r_n^2)}{N - n - 1}. \qquad (13.2.6)$$

Suppose then that the Wherry correction formula is applied to r_n^2 and r_i^2, and that the two resulting corrected squared multiple correlations are equal; that is,

$$\frac{(N - 1)r_n^2 - n}{N - n - 1} = \frac{(N - 1)r_i^2 - i}{N - i - 1}. \qquad (13.2.7)$$

Then we can easily see that (13.2.6) and (13.2.7) are equivalent (Exercise 13.3). Now consider a procedure that adds predictors one at a time in an order that maximizes the incremental validity at each step. The procedure stops adding variables when the corrected squared multiple for the larger set is less than the corrected squared multiple for the smaller set. Compare this with a procedure based on the rule to stop adding variables when the variance ratio (13.2.5) is less than one. Clearly the two procedures are equivalent. Thus the Wherry correction has a reasonable theoretical justification although this justification is not associated with the concept of shrinkage resulting from the selection of variables.

It should also be pointed out that $\sigma^2_{0\cdot 12\ldots n}$ is the variance of the errors made in the population when predicting X_0 from the known best linear combination of $x_1, x_2, \ldots, x_n$. In practice, however, the true regression weights are unknown, and instead a set of least squares estimates of these regression weights based on a prior sample must be used to define a linear prediction function. These estimated regression weights will almost never be the true regression weights; hence, when this linear prediction function is used, the variance of the errors of prediction in the population will almost always be greater than $\sigma^2_{0\cdot 12\ldots n}$, the variance of the errors of prediction when the linear regression function is used. Thus the usual estimate of the partial variance is not an unbiased estimate of the error variance associated with the use of this linear prediction function, but only an unbiased estimate of the error variance associated with the use of the true, but unknown, linear regression function. On the average, when the estimated regression function is used in a new sample, the error variance will be greater than the estimated residual variance for the reasons already given. Because of sampling variation, however, the residual variance in any particular second sample may in fact take the extreme value zero, on the one hand, or a value equal to the variance of the criterion, on the other hand.

13.3 Formal Procedures for Selecting Predictor Variables

Even if sampling problems could be ignored, the only way to be sure of obtaining the best n of N predictors would be to determine the multiple correlation for every such set. This *exhaustion procedure* can seldom be justified economically unless N is very small. There are two basic formal algorithms for selecting a "good" set of n predictor variables from a larger set of N possible predictor variables. The first of these, associated with the names of Wherry (1940), Dwyer (1945), and Summerfield and Lubin (1951), may be called the *forward selection procedure*. It involves a sequential selection of predictor variables such that the predictor variable selected at each stage is the one that provides the largest incremental validity, given all the predictor variables previously selected. In the first stage, this results in the selection of the variable having the highest zero-order correlation with the criterion. In the second stage, the variable selected is the one that has the largest partial correlation with the criterion when the first selected predictor is partialed out. This pair of variables gives the largest multiple correlation among all pairs of variables that include the variable selected in the first stage. However, it is possible to show that this pair of variables does not necessarily provide the highest multiple correlation over all possible pairs of predictor variables. We may show that in general the addition of the $(n + 1)$-variable with the highest incremental validity does not necessarily yield the best set of $(n + 1)$ predictor variables. So far as the present writers have been able to determine, no analytic results have ever been provided to show just how efficient the forward method is for typical problems.

Summerfield and Lubin (1951) have presented what appears to be the most reasonable approach to the problem of deciding when to stop adding variables to the predictor set. In their method, the F-statistic (13.2.5) is computed at each stage, with $n = n$ and $i = n - 1$, to determine whether or not the additional variable is indeed contributing to prediction. A second F-statistic is also computed, with n equal to the total number of variables in the pool and i equal to the number of variables so far selected, to determine whether or not

e variance of the
er than $\sigma^2_{0\cdot 12\ldots n}$,
ssion function is
nbiased estimate
liction function,
with the use of
rage, when the
or variance will
already given.
any particular
one hand, or a

e of obtaining
orrelation for
economically
or selecting a
ble predictor
nerry (1940),
the *forward*
riables such
provides the
sly selected.
the highest
he variable
terion when
s gives the
he variable
his pair of
on over all
e addition
not neces-
e present
been pro-
blems.
the most
variables
puted at
the ad-
tistic is
ol and i
or not

the remaining variables, in combination, contribute to prediction. The evaluation of this second F-ratio is designed to discover any errors in the forward method, errors possibly due to the existence of suppressor variables.

A second procedure, the *backward elimination procedure* described by Horst and MacEwan (1960), begins with all N variables and then successively eliminates variables so that the decrease in the multiple correlation is minimized at each stage. The problem encountered with the forward method is also to be found in this procedure, for we have no guarantee that we indeed have the best combination of predictors at any stage past the first. Unless the number of variables to be selected is very near the total number in the predictor pool, the forward procedure involves less computation; and certainly the computation will be very much less if only a small percentage of predictors from the pool is to be retained. Since all computational methods in effect involve an inversion of the matrix of predictors, problems of ill-conditioning of this matrix (the matrix being too nearly singular for computational purposes) are more likely to occur with more rather than with fewer variables. If this happens, then the backward procedure cannot be used, but the forward procedure "will provide usable regression equations prior to degeneracy" (Efroymson, 1966).

For the data given in Section 12.5, X_1 is the best single predictor. If X_2 is taken in combination with X_1, a higher multiple correlation is obtained than if X_3 is taken. The multiple correlations are 0.467 and 0.405, respectively. Hence the forward selection procedure takes X_1 and X_2 as the best two-variable set. However, the optimal two-variable set is X_2 and X_3, for which the multiple correlation is 0.702, and this set would be the one selected by the backward procedure.

A refinement of the forward and backward procedures called *a stepwise procedure* has proved useful. For the forward procedure, briefly, this refinement is based on reevaluation of each member of the set of selected predictors every time a new predictor is added to the set. "A variable which may have been the best single variable to enter at an early stage may, at a later stage, be superfluous because of the relationships between it and other variables now in the regression" (Draper and Smith, 1966). If this is the case, this variable may be eliminated from the regression equation. A similar refinement is applicable to the backward procedure. Swoyer (1966) has used the backward stepwise procedure and obtained some encouraging results.

13.4 Prediction in Future Samples

In practice, regression weights are never known exactly; they must be estimated from a calibration sample. These estimates are then substituted for the true but unknown regression weights in the linear prediction model. The resulting linear prediction equation is then used to "predict" values of the criterion for other individuals, given measurements on the predictors. Even if this calibration sample is distinct from an initial screening sample, not all problems are solved. There will still be some tendency to capitalize on chance in estimating the regression weights for these variables.

If the true regression weights were known, an increase in the number of predictor variables could never result in a decrease in precision of prediction. However, if the true regression must be estimated, and particularly if the calibration sample is not substantially larger than the number of predictor variables, it can happen that an increase in the number of variables results in a decrease in the precision of prediction for individuals not in the calibration sample. Thus we may have

$$\mathscr{E}\left(Y - \sum_{i=1}^{M} \hat{\beta}_i X_i\right)^2 \quad \text{greater than or less than} \quad \mathscr{E}\left(Y - \sum_{i=1}^{M+N} \hat{\beta}_i X_i\right)^2,$$

depending (1) on the incremental validity of the last N predictor variables and (2) on the loss of precision of estimation due to the introduction of N additional parameters to be estimated. In the extreme, if a linear prediction function that has been determined from a very small calibration sample is used for prediction in a new sample, then it can happen that the expected variance error of prediction is larger than the variance of the criterion. In such cases, an investigator would do better to discard his predictors and use the sample mean value of the criterion as his predicted value.

The predictor-variable selection procedures described in the preceding section are often used as methods for deciding on the specification of variables to be included in the predictor set. Other approaches, advanced by Burket (1964), Elfving (1961), Elfving, Sitgreaves, and Solomon (1961), and Horst (1941), are based on the assumption of an underlying "factor" structure (see Chapter 24), which in effect involves a reduction in the rank of the prediction system (see Section 16.7). These methods seem promising for use in large scale studies. Very recent work of Fortier (1966a, b) should also be studied. Another interesting approach to this problem is that of Linhart (1960), who assumes a normal distribution of errors and then specifies a stopping rule for the forward selection procedure based on the criterion of the minimization of the confidence interval for a future observation. Papers of Stein (1960) and Nicholson (1960) are also pertinent. At present, however, no entirely satisfactory solution to this problem is available.

Rydberg (1963) and others have suggested that another problem may arise when regression weights obtained in one sample are used for prediction in a new sample. Often the determination of the beta weights is made on a group preselected on the basis of some of the potential predictor variables. If no allowance is made for such preselection, then variables so used in selection will typically have drastically reduced beta weights with the reduction being greatest for the best variables. The application of such weights to an unselected group could produce unsatisfactory results.

We may illustrate many of the techniques described in this and the previous chapter with some data from a simple yet effective validity study from the

Table 13.4.1

Data from the 1959–60 independent school
SSAT prediction study

		1960 Correlations					Mean		Standard deviation	
		V	Q	T	PAS	GPA	1959	1960	1959	1960
1959 Correlations	V	–	.2339	.8681	.2493	.3859	294.0	294.4	13.0	12.6
	Q	.3231	–	.6670	.3059	.4019	318.1	313.8	13.6	11.6
	T	.8771	.7188	–	.3519	.5088	304.0	302.4	10.3	9.2
	PAS	.2212	.2564	.2706	–	.4502	88.2	88.1	5.7	6.2
	GPA	.3901	.4753	.5223	.4271	–	74.5	75.5	7.1	7.4

statistical report *Secondary School Admission Test (SSAT) Scores as Predictors of Ninth Grade Averages, 1959–60 and 1960–61* by Barbara Pitcher of Educational Testing Service. We shall discuss only a small portion of that study here.

A sample of 109 ninth-grade enrollees was obtained in an independent secondary school in 1959 and a second sample of 120 from the same school in 1960. The previous average school marks (PAS) and the Verbal (V), Quantitative (Q), and Total (T) scores, T = V — Q, were among the predictors available for each enrollee. The tests had been administered in the previous year. Although several performance criteria were available, we shall consider only the overall end-of-year ninth-grade grade point averages (GPA). The data that were obtained in the first and second samples have been summarized in Table 13.4.1.

It should be observed that in most instances the 1960 correlations (the above-diagonal entries) are very close to the corresponding 1959 correlations (the below-diagonal entries). Also it should be noted that the 1959 and 1960 means and standard deviations are remarkably close; this indicates that there was little difference in the quality of the two entering classes.

Test-score validities found in this school follow a pattern typically found in validity studies of this kind. These validities are quite satisfactorily high, considering that they were obtained from the selected rather than the applicant group (see Sections 6.8 through 6.10).

Multiple correlations and regression weights were computed for several combinations of predictors. These values are given in Table 13.4.2. The combination of previous average either with GSAT-T or with GSAT-V and GSAT-Q provides a cross-validated multiple correlation of 0.60 *in the selected group.*

The final column of this table shows a particularly interesting feature of the analysis. The regression weights obtained from the 1959 (1960) sample were used to predict criterion scores from the corresponding predictors in the 1960 (1959) sample, the computations being carried out according to (4.7.3). These computations yielded the cross-validated multiple correlations in the

Table 13.4.2

Regression weights, multiple correlations, and cross-validated multiple correlations for several combinations of predictor variables

Year	GSAT T	GSAT V	GSAT Q	Previous average	Multiple correlation	Cross-validated composite correlation
1959	0.3013			0.3804	0.6008	0.5852
1960	0.3220			0.3717	0.5855	0.6005
1959		0.1431	0.2021		0.5370	0.5001
1960		0.1818	0.2103		0.5017	0.5357
1959		0.1184	0.1706	0.3631	0.6062	0.5785
1960		0.1473	0.1582	0.3747	0.5802	0.6044

final column. It should be noted that any shrinkage found here arises only from variations of weights and not from selection of variables during the study; on the contrary, the variables were chosen ahead of time on the basis of a wide background of prior experience. Indeed, in this study the amount of "shrinkage" in every case proved to be at most relatively negligible, and in some cases there was an actual increase in the composite-predictor correlation with criterion in the second sample. This contrasts sharply with the substantial shrinkage obtained in the Mosteller and Wallace (1964) study. The reason for this difference is that Mosteller and Wallace were forced to select a small number of predictor variables *during* the study from a much larger set of potential predictors. Thus they capitalized on chance in their selection. Only by cross-validating this selection were they able to obtain an accurate appraisal of the true predictability of their criterion. Their one outstanding predictor variable, however, actually improved on cross validation. This also is not atypical, for if one variable is an outstanding predictor, then it is chosen on its true merit rather than for its error, and hence it can yield either a lower or higher value on cross validation. It is when many variables have uniformly low true correlations that cross validation shrinkage is large. An even more drastic shrinkage occurred in a vintage study reported by Guttman (1941): In this case, the use of 84 regression coefficients in a sample of 136 produced a multiple correlation of 0.73, but when these same weights were used in a second sample of 140, the multiple correlation was 0.04.

Considering each of these groups as a sample from some larger (hypothetical) population, it is clear that the weights obtained in either year can only be approximations to the optimal weights. Hence in using such weights we are not using the true linear regression weights. However, the results of this study (and other studies) suggest that an approximate optimal linear combination of predictor variables often performs nearly as well as the true optimal com-

Cross-validated composite correlation
0.5852
0.6005
0.5001
0.5357
0.5785
0.6044

e arises only
ng the study;
asis of a wide
f "shrinkage"
ne cases there
h criterion in
hrinkage ob-
his difference
of predictor
ictors. Thus
lidating this
true predict-
le, however,
l, for if one
merit rather
alue on cross
elations that
ge occurred
ise of 84 re-
tion of 0.73,
he multiple

ypothetical)
an only be
ghts we are
f this study
ombination
timal com-

bination. Geometrically we would say that the composite variable correlation surface is reasonably flat in the region of the point determined by the linear regression weights.

13.5 The Effect of Relative Test Lengths on Reliability and Validity: The Multiple Predictor Case*

As we indicated in Chapter 5, the validity coefficient of any test containing errors of measurement can be increased by increasing the length of the test. Formula (5.11.2) gives the validity of a test at length k with respect to a fixed criterion in terms of its validity at unit length, its reliability at unit length, and the value k. Formula (5.10.1) gives the reliability of a test of length k in terms of its reliability at unit length and the value k.

Since the multiple correlation coefficient is, in fact, the zero-order correlation between the best linear combination of the predictor variables $X_1, X_2, \ldots, X_n$ and the criterion X_0, we might suppose the multiple correlation coefficient varies as the lengths of the various predictors are altered. In this section, we shall develop formulas in matrix notation for the effects of changes in test length on reliability, validity, predictor variable intercorrelation, partial regression weights, and multiple correlation. Among other things, we shall show that the partial regression weights depend on the lengths of the various tests, and indeed that the desirability of including a particular variable in a regression equation may depend on the total available testing time.

Let $X_1, X_2, \ldots, X_n$ be a set of n predictor variables and X_0 be a criterion variable. Let

$\mathbf{D}_a$ be a diagonal matrix whose diagonal elements are the lengths of the predictors $X_1, X_2, \ldots, X_n$,

$\boldsymbol{\rho}$ be the vector of validity coefficients of $X_1, X_2, \ldots, X_n$ with X_0,

$\mathbf{P}$ (upper case rho) be the matrix of intercorrelations of the predictors,

$\mathbf{D}_r$ be the diagonal matrix whose diagonal elements are the reliabilities of the predictors $X_1, X_2, \ldots, X_n$,

$\boldsymbol{\beta}$ be the vector of partial regression weights of $X_1, X_2, \ldots, X_n$ with X_0, and

R_n^2 be the multiple correlation of $X_1, X_2, \ldots, X_n$ with X_0.

We assume that each of the above quantities is known. Now suppose the length of each of the predictors is altered and the new predictors are denoted by $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$. The length of the criterion is assumed to remain unchanged. Let

$\mathbf{D}_b$ be a diagonal matrix whose diagonal elements are the new lengths of the predictors $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n$.

[*] Reading of this and the following section may be omitted without loss of continuity.