Psychological Bulletin
1968, Vol. 69, No. 3, 161–182

MULTIPLE REGRESSION IN PSYCHOLOGICAL RESEARCH AND PRACTICE

RICHARD B. DARLINGTON [1]
Cornell University

A number of common practices and beliefs concerning multiple regression are criticized, and several paradoxical properties of the method are emphasized. Major topics discussed are the basic formulas; suppressor variables; measures of the "importance" of a predictor variable; inferring relative regression weights from relative validities; estimates of the true validity of population regression equations and of regression equations developed in samples; and statistical criteria for selecting predictor variables. The major points are presented in outline form in a final summary.

In recent years, electronic computers have made the multiple regression method readily available to psychologists and other scientists, while simultaneously making it unnecessary for them to study in full the cumbersome computational details of the method. Therefore, there is a need for a discussion of multiple regression which emphasizes some of the less obvious uses, limitations, and properties of the method. This article attempts to fill this need. It makes no attempt to cover thoroughly computational techniques or significance tests, both of which are discussed in such standard sources as McNemar (1962), Hays (1963), DuBois (1957), and Williams (1959). The discussion of significance tests by Williams is especially complete, as is the presentation of computing directions by DuBois. The latter source also contains many basic formulas of considerable interest. Anderson (1958) gives a very complete mathematical presentation of the exact sampling distributions of many of the statistics relevant to multiple regression. Elashoff and Afifi (1966) reviewed procedures applicable when some observations are missing. Beaton (1964) described a set of elegantly simple computer subroutines which a FORTRAN programmer can use to write quickly almost any standard or special-purpose regression program he may require.

[1] For critical comments on preliminary drafts, the author is indebted to J. Millman, P. C. Smith, and T. A. Ryan, and to his students J. T. Barsis, W. Buckwalter, H. Day, B. Goldwater, and G. F. Stauffer. He is especially grateful to his student C. S. Otterbein, whose editorial and substantive contributions amounted nearly to coauthorship.

Some of the points made herein are original, some have been derived independently by several workers in recent years, and some surprisingly little-known points were made in print 40 or more years ago.

In general, the dependent or criterion variable will be denoted by X_0, and the independent or predictor variables by X_1, X_2, $\cdots$, X_n. The score of person i on variable X_j is symbolized by x_{ij}. The population multiple correlation is denoted by $\bar{R}$, ordinary correlations by ρ, standard deviations by σ. Population regression weights are denoted by β, with β' denoting the corresponding weights when all variables have been adjusted to standard score form. Sample values of these parameters are denoted by R, r, s, b, and b'.

The purpose of the multiple regression method is to derive weights β_1, β_2, $\cdots$, β_n for the variables X_1, X_2, $\cdots$, X_n, and an additive constant α, such that the resulting weighted composite $\hat{X}_0$, which is defined by the multiple regression equation

$$\hat{X}_0 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \alpha \quad [1]$$

predicts a specified criterion variable X_0 with a minimum sum of squared errors; thus $\hat{X}_0$ correlates maximally with X_0. This paper deals directly only with the linear additive model, in which $\hat{X}_0$ is a linear function of the predictor variables. This restriction is more apparent than real, however, since if desired some of the variables in the equation can be curvilinear or configural (interactive) functions of other variables.

Basic Formulas [2]

As described in standard works on the subject, the n multiple regression weights β_1, β_2, β_3, $\cdots$, β_n are found by solving a set of n simultaneous linear equations in which these weights are the unknowns. These equations are the so-called *normal equations* of regression theory, although this term does not imply any dependence on an assumption that variables are normally distributed. The known quantities in the normal equations are the standard deviations of the n predictor variables and the criterion variable, and the intercorrelations among these $n + 1$ variables. A change in the standard deviation of one of the predictor variables will affect only the beta weight of that one variable, while a change in the correlation between any two variables will generally affect all the beta weights.

After the βs are found, the additive constant α is chosen so as to make the mean of the scores on $\hat{X}_0$ equal to the mean of the scores on X_0. The multiple correlation $\bar{R}$ can then be found in any of several ways, of which the simplest conceptually (though not computationally) is to compute each person's score on $\hat{X}_0$, and then to correlate these scores with X_0.

If all predictor variables are uncorrelated, then the above-mentioned procedures for computing beta weights and $\bar{R}$ reduce to the simple formulas

$$\beta_j = \rho_{0j} \frac{\sigma_0}{\sigma_j} \qquad [2]$$

and

$$\bar{R}^2 = \rho_{01}^2 + \rho_{02}^2 + \rho_{03}^2 + \cdots + \rho_{0n}^2 \qquad [3]$$

If we define the "usefulness" of predictor variable X_j as the amount $\bar{R}^2$ would drop if X_j were removed from the regression equation and the remaining variables appropriately reweighted, then Formula 3 shows that the usefulness of X_j equals ρ_{0j}^2 when predictor variables are uncorrelated.

Results analogous to some of those stated above can be derived for the case in which predictor variables are intercorrelated. Suppose we have a regression equation predicting X_0 from X_1, X_2, $\cdots$, X_n. Consider a second multiple regression equation in which one of the predictor variables X_j $(1 \leq j \leq n)$ is the dependent variable and the remaining $n - 1$ predictor variables X_1, X_2, $\cdots$, X_{j-1}, X_{j+1}, $\cdots$, X_n are the predictors. Let the residuals in this second equation (i.e., the set of scores obtained by subtracting each person's score on this regression function from his actual score on X_j) constitute the variable $X_{j(p)}$. The variable $X_{j(p)}$ is uncorrelated with all of the variables used to construct the regression equation predicting X_j. Following Rozeboom (1965) and others, $X_{j(p)}$ is termed the component of X_j orthogonal to the other predictor variables, or more simply the orthogonal component of X_j. (We shall later have occasion to denote the component of X_0 orthogonal to *all* the predictor variables—which component is the residual in the regression equation predicting X_0 from those predictor variables—as $X_{0(p)}$. The component of any predictor variable X_j orthogonal to the criterion variable X_0 will be denoted by $X_{j(c)}$. In the present terminology, the partial correlation between two variables X_j and X_k holding m other variables constant is the correlation between the components of X_j and X_k orthogonal to the other m variables.)[3]

The standard deviation of $X_{j(p)}$ is denoted by $\sigma_{j(p)}$, and the correlation of $X_{j(p)}$ with

[2] A number of points made in this paper are amplified and proven in a supplementary document by the author entitled "Proofs of Some Theorems on Multiple Regression." Statements in the present section are given as Theorems 4, 6, 10, 11, 12, and 13 of that document. Although the proofs are not original in any important sense of the word, the author has tried to simplify many of the standard proofs to a level readily grasped by students in an intermediate-level course in psychometric theory. The document will routinely be sent along with responses to reprint requests. It as also available from the American Documentation Institute. Order Document No. 9810 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $1.75 for microfilm or $2.50 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

[3] Dunlap and Cureton (1930) called the correlation between a variable and the orthogonal component of another variable a "semipartial correlation," and McNemar (1962, p. 167) called it a "part correlation." Both these terms emphasize its similarity to the partial correlation.

any variable X_h by $\rho_{h \cdot j(p)}$. Then we can write a formula for β_j which is very similar to Formula 2 but which applies whether the predictor variables are intercorrelated or not:

$$\beta_j = \rho_{0 \cdot j(p)} \frac{\sigma_0}{\sigma_{j(p)}} \qquad [4]$$

Further, just as the usefulness of X_j equals ρ_{0j}^2 when predictor variables are uncorrelated, so it equals $\rho_{0 \cdot j(p)}^2$ when predictor variables are intercorrelated. There is no formula quite so directly analogous to Formula 3; in general it is *not* true that

$$\bar{R}^2 = \rho_{0 \cdot 1(p)}^2 + \rho_{0 \cdot 2(p)}^2 + \cdots + \rho_{0 \cdot n(p)}^2$$

When $n = 2$, it can be shown that

$$\rho_{0 \cdot 1(p)} = \frac{\rho_{01} - \rho_{02}\rho_{12}}{\sqrt{1 - \rho_{12}^2}} \qquad [5]$$

and that

$$\sigma_{1(p)} = \sigma_1 \sqrt{1 - \rho_{12}^2} \qquad [6]$$

Interchanging the subscripts 1 and 2 in Formulas 5 and 6 gives $\rho_{0 \cdot 2(p)}$ and $\sigma_{2(p)}$. Substituting Formulas 5 and 6 in Formula 4 gives the familiar formula

$$\beta_1 = \frac{\rho_{01} - \rho_{02}\rho_{12}}{1 - \rho_{12}^2} \cdot \frac{\sigma_0}{\sigma_1} \qquad [7]$$

for the case in which $n = 2$. Analogously, when $n = 2$,

$$\beta_2 = \frac{\rho_{02} - \rho_{01}\rho_{12}}{1 - \rho_{12}^2} \cdot \frac{\sigma_0}{\sigma_2} \qquad [8]$$

When $n = 2$, the useful relation

$$\bar{R}^2 = \rho_{01}^2 + \rho_{0 \cdot 2(p)}^2 = \rho_{02}^2 + \rho_{0 \cdot 1(p)}^2$$

follows directly from the fact that $\rho_{0 \cdot j(p)}^2$ equals the usefulness of X_j. This result fits intuitively with Formula 3 since X_1 and $X_{2(p)}$ (or X_2 and $X_{1(p)}$) are uncorrelated variables which contain the same basic information (and thus yield the same prediction of X_0) as X_1 and X_2.

Suppressor Variables

This section describes suppressor variables and criticizes the common belief that dealing with them requires a modification of standard multiple regression procedures. For simplicity, the section assumes that each predictor variable with a nonzero correlation with X_0 is scored in the direction which makes that correlation positive in the population. The scoring direction of variables correlating zero with X_0 can be chosen arbitrarily.

Although various definitions have been given (cf. Guilford, 1954; Horst, 1941; McNemar, 1962), a suppressor variable is here defined as a variable which, when included in a regression equation in which the variables have been scored as described above, receives a negative weight when the equation is derived in the population. (This definition thus excludes variables whose negative weight is the result of sampling error.) Since a variable correlating zero with X_0 is allowed to be scored in either direction, and since it will receive a negative regression weight when scored in one of those directions if its weight is not zero, then by our definition any variable which correlates zero with X_0 but which receives a nonzero weight can be called a suppressor variable. Contrary to most previous definitions, by the present definition a suppressor variable need not have a low or zero validity, although in practice it usually does.

Since the multiple regression method chooses those weights, whether positive or negative, which maximize the multiple correlation, it necessarily follows that a suppressor variable improves prediction in the population when it is given a negative weight. A typical example in which prediction is improved by assigning a negative weight to a variable might be a situation in which a test of reading speed is used in conjunction with a speeded history achievement test to predict some external criterion of knowledge of history. Since the history test is contaminated by reading speed, assigning a negative weight to the reading-speed test would help to correct for the disadvantage suffered by a student with low reading speed who is competing with faster readers.

To understand more fully the functions of suppressor variables, it is helpful to examine the exact conditions under which negative weights appear. For simplicity, regression equations with only two predictor variables X_1 and X_2, and only the case in which X_2 receives a negative weight, will be considered.

Let $X_{1(c)}$ denote the component of X_1 orthogonal to the criterion variable. $X_{1(c)}$ can be considered to measure the sources of error in X_1, that is, those aspects of X_1 which prevent a perfect prediction of X_0 from X_1. For example, consider a hypothetical modification of the previous example in which reading speed correlates zero with the criterion variable, and is the only source of error in the history test. If the history test were X_1, then $X_{1(c)}$ would measure only reading speed. In real-life cases, $X_{1(c)}$ does not represent a single source of error, such as reading speed, but measures instead a composite of all the sources of error in X_1.

Ordinarily, the variable $X_{1(c)}$ is not directly measurable. However, suppose a second test X_2 were available which correlated perfectly with $X_{1(c)}$; in other words, suppose X_2 were a perfect measure of the sources of error in X_1. (For instance, in our hypothetical example, suppose X_2 was a perfect test of reading speed.) Then, by giving X_2 a negative weight in a regression equation in which X_1 had a positive weight, it would be possible to subtract out, or correct for, or "suppress," those sources of error.

Generally, of course, a second variable X_2 does not correlate perfectly with $X_{1(c)}$. The example should make clear, however, that X_2 can be used in either of two ways, depending on its characteristics: to measure X_0 directly, or to measure $X_{1(c)}$. X_2 should receive a positive weight if used the first way or a negative weight if used the second way. It will be shown that whether X_2 receives a positive or a negative weight, when used in a regression equation with X_1, depends upon the ratio between its abilities to perform these two different tasks; specifically, on the ratio $\rho_{2.1(c)} : \rho_{02}$.

In a regression equation with two predictor variables, Formula 8 shows that X_2 receives a negative weight if

$$\rho_{02} - \rho_{01}\rho_{12} < 0 \qquad [9]$$

Some algebra shows that Inequality 9 is equivalent to

$$\frac{\rho_{2.1(c)}}{\rho_{02}} > \frac{\rho_{1.1(c)}}{\rho_{01}} \qquad [10]$$

The algebra here consists of using Formulas 5 and 6 to set

$$\rho_{2.1(c)} = \frac{\rho_{12} - \rho_{01}\rho_{02}}{\sqrt{1 - \rho_{01}^2}}$$

and

$$\rho_{1.1(c)} = \sqrt{1 - \rho_{01}^2}$$

and then substituting these expressions in Inequality 10, which then simplifies to Inequality 9. These same two formulas can be used to derive an inequality which is similar to Inequality 10, but which some readers will find more meaningful. If both sides of Inequality 10 are squared, then the resulting quantities ρ_{01}^2, ρ_{02}^2, $\rho_{1.1(c)}$, and $\rho_{2.1(c)}$ equal the proportions of variance in X_1 and X_2 "accounted for" by X_0 and $X_{1(c)}$, respectively, so that the formula is in terms of proportions of variance rather than correlation coefficients. Some readers will prefer this alternative since ratios between proportions of variance are more familiar than ratios between correlation coefficients.

Inequality 10 provides the basis for a clear and simple statement of the algebraic nature of a suppressor variable. The left side of Inequality 10 is the ratio mentioned above, showing the ability of X_2 to measure $X_{1(c)}$ relative to its ability to measure X_0. The right side is an analogous ratio for X_1, showing the ability of X_1 to measure $X_{1(c)}$ relative to its ability to measure X_0. If the ratios on the left and right sides of Inequality 10 are equal, then X_2 cannot usefully supplement X_1 in measuring either X_0 or $X_{1(c)}$, so it receives a zero weight. Normally, X_1 is a better measure of its own sources of error than is X_2, so $\rho_{1.1(c)}$, the numerator of the right side, is normally larger than $\rho_{2.1(c)}$, the numerator of the left side. Hence the fraction on the right side is normally larger than the fraction on the left. If this occurs, then X_2 is more useful as a measure of X_0 than as a measure of $X_{1(c)}$, so it receives a positive weight. However, if X_2 correlates so highly with $X_{1(c)}$ that the left side of Inequality 10 is larger than the right (as a test of reading speed would correlate highly with the error in the history test in the above example), then X_2 is more useful as a measure of $X_{1(c)}$, and so receives a negative weight.

Inequality 9 shows that in regression equations with two predictor variables, β_2 is never negative if ρ_{02} is greater than or equal to ρ_{01} or if both predictor variables correlate positively with X_0 and negatively with each other. Further, β_2 is always negative if $\rho_{02} = 0$ and ρ_{12} and ρ_{01} are both positive.

When an equation contains more than two predictor variables, any variable X_j is a suppressor if Inequality 10 holds when X_2 is replaced by X_j, and X_1 is replaced by the variable formed by dropping the jth term from the multiple regression equation which uses all n predictor variables. This follows from the fact that if the variable formed in this way were used with X_j in a two-variable regression equation predicting X_0, then clearly the weight of X_j in this equation would equal the weight of X_j in the regression equation computed directly from all n variables. Thus β_j in the n-variable equation would be negative whenever β_j in the two-variable equation was negative, which would occur whenever Inequality 10 holds for that equation.

The multiple regression method considers suppressor relationships in that it chooses the weights, positive or negative, which give the highest multiple correlation. Hence, the observation of negative weights in a sample regression equation, indicating that suppressor variables may be present, does not alone imply that there should be a deviation from standard regression procedures. However, Gulliksen (1950) stated that negative weights "should lead to a careful scrutiny of the test and a consideration of the reasonableness of such a finding [p. 330]." Although such a scrutiny can often be attempted with confidence in regression equations with two predictor variables, it is difficult for an investigator to reach a conclusion about the reasonableness of a negative weight in a complex, multipredictor situation. For example, consider a three-predictor situation in which $\rho_{01} = \rho_{02} = .15$, $\rho_{03} = .2$, $\rho_{12} = 0$, and $\rho_{13} = \rho_{23} = .7$. Although X_3 is the most valid single predictor and would be assigned a positive weight when used in conjunction with either X_1 or X_2 alone, it can be shown that it is given a negative weight when X_1, X_2, and X_3 are all used together. Yet using X_3 in this way raises the multiple correlation for all three variables to .5, while the highest multiple correlation without negative weights is only .21, using X_1 and X_2.

Thus, suppressor relationships appear in situations in which a "reasonable" interpretation of the relationship is extremely difficult. Relationships among more than three predictor variables are even more complex. Therefore, even if the improvement resulting from using a negative weight were small, it is difficult to imagine an investigator with such faith in his ability to conceive of all possible suppressor situations that he would ignore the improved prediction resulting from the use of a negative weight.

Measures of the "Importance" of a Predictor Variable

The present section deals with five different measures of the "importance" of predictor variables; for variable X_j, the measures are ρ_{0j}^2; $\beta'_j{}^2$; $\rho_{0\cdot j(p)}{}^2$, which as we saw above equals the usefulness of X_j; $\beta'_j\rho_{0j}$; and a measure proposed by Englehart (1936). It will be recalled that β'_j was defined as the weight given to X_j when all variables have been adjusted to unit variance, and that the usefulness of X_j was defined as the amount $\bar{R}^2$ would drop if X_j were removed from the regression equation and the weights of the remaining predictor variables were then recalculated.

When all predictor variables are uncorrelated, all five of these measures are equivalent. The equivalence of the first four can be verified merely by inspection of Formulas 2 and 3; the fifth will be discussed later. If predictor variables are uncorrelated, each of the five measures also equals the difference (expressed as a proportion of the variance of X_0) between the original variance of X_0 and the variance of X_0 in a subpopulation whose members all have the same score on X_j. (If this latter variance varies across subpopulations, then an average is taken.) Further, if any of the five measures is summed across all the predictor variables in a regression equation, the total is $\bar{R}^2$. Thus it is meaningful and useful to consider $\bar{R}^2$ to be the sum of the proportions of variance in

the criterion variable "accounted for by," or "attributable to," or "contributed by" each of the predictor variables. The interpretation is completely analogous to the interpretation of results in analysis-of-variance designs.

In analysis-of-variance designs, the complete independence of all the independent variables is assured by the requirement of equal or proportional cell frequencies (or by the requirement of statistical adjustments, such as those given by Federer and Zelen, 1966, designed to produce estimates of the same parameters as those estimated with equal cell frequencies). In multiple regression, however, there is no requirement that predictor variables be uncorrelated. This property gives regression analysis a substantial element of flexibility lacking in analysis of variance. When predictor variables are intercorrelated, however, the five measures of importance are no longer equivalent, so that the term "contribution to variance" suddenly becomes very ambiguous. The different measures of importance do not even necessarily rank order the variables in a regression equation in the same order. For example, consider the case in which $\rho_{01} = .4$, $\rho_{02} = .44$, $\rho_{03} = .3$, $\rho_{12} = .8$, $\rho_{13} = 0$, and $\rho_{23} = .4$. Given these values, standard formulas show that $\bar{R} = .5$. The three β' weights computed from these numbers are, respectively, .4, 0, and .3, and the three decreases in $\bar{R}^2$ used as measures of usefulness are, respectively, .038, 0, and .050. Thus X_1 has the highest β' weight, X_2 is most valid, and X_3 is most useful. Although in this example the variable with the lowest β' weight, X_2, is also the least useful, other examples can be constructed in which this is not true.

The rest of this section attempts to explain what meaning, if any, can be attached to each of the five measures of importance. Of the five, only $\beta'_j\rho_{0j}$ and Englehart's measure total to $\bar{R}^2$ when summed across the variables in a regression equation. Nevertheless, of the five, these two will be shown to be of least interest and value.

Squared Validity

Of the five measures, ρ_{0j}^2 or the squared validity, needs the least comment. It is the only one of the five measures unaffected by the choice of the other variables in the regression equation.

Beta Weights as Measures of the Importance of Causal Relationships

For many purposes, β'_j is of more interest than $\beta'_j{}^2$. In the present discussion, they will be considered equivalent, since either can be computed from the other (provided the sign of the weight is known) and since they rank variables in the same order of importance.

Previous sections showed that beta weights, like usefulness, are determined solely by the characteristics of the orthogonal component of the variable under consideration. They thus have little relation to validity and are heavily influenced by the nature of the other variables in the regression equation. Beta weights can even change in sign as variables are added to or removed from the equation; one example was given in the section on suppressors, another is given by Kendall (1957, p. 74).

It was shown above that β'_j (or $\beta'_j{}^2$) is not a measure of the usefulness of X_j when predictor variables are intercorrelated. The present section describes a particular case in which beta weights are nevertheless of considerable interest as a measure of the "importance" of a variable.

It is true that "correlation does not imply causation." In most cases, an investigator cannot determine whether an observed correlation between two variables X_0 and X_1 is due to the effect of X_0 on X_1, or to the effect of X_1 on X_0, or to some combination of effects which might include the effects of other outside variables on both X_0 and X_1. However, there are cases in which some of these alternatives can be ruled out by the nature of the variables involved; thus, if there is a correlation between snowfall and traffic accidents, it can be assumed that the traffic accidents did not cause the snowfall. If a large enough number of such causal hypotheses can be eliminated, then there are certain situations in which a multiple regression equation can be used to estimate the importance of the remaining causal relationships. Partly because this technique has been used in cases in

which it was not wholly appropriate, this section attempts to make explicit the assumptions necessary for the use of the technique.

Consider a situation in which (*a*) a given dependent variable is affected only by a specified set of measurable variables, (*b*) the effect of each of these variables on the dependent variable is linear, and (*c*) the dependent variable has no effect, either directly or indirectly, on any of the independent variables. In such a situation, consider a linear function of the causal variables in which the weight of each variable equals the causal importance of that variable; that is, if increasing X_j by 1 unit increases the dependent variable by g units, then g is the weight of X_j. This linear function will perfectly predict the dependent variable. Since the multiple regression method computes the weights which result in the best prediction of the dependent variable, in this situation a multiple regression equation computed in the population necessarily computes the true causal weights for the set of variables involved in the equation, since these are the only weights which result in perfect prediction. Further, if an investigator has inadvertently included among the predictors a variable which in fact has no effect on the dependent variable, then that predictor variable will receive a weight of zero.

Suppose now that the dependent variable is determined partly by chance factors or by nonchance factors which are uncorrelated with all of the predictors which the investigator uses. It can be shown that the weights in a multiple regression equation are unchanged by the addition of a new predictor variable which is uncorrelated with all the other predictors.[4] Therefore, the best possible prediction of the dependent variable from the causal measures used is still obtained when the weight of each variable equals the true causal effect of that variable on the dependent variable. Hence the population multiple regression weights still equal the true causal weights, although the multiple correlation is less than unity. And since sample beta weights are unbiased estimates of the population beta weights, they can be employed as unbiased estimates of the true causal weights. The method can also be extended to handle curvilinear or interactive effects by including such terms in the regression equation.

Thus, the method assumes:

1. All variables which might affect the dependent variable are either included in the regression equation or are uncorrelated with the variables which are included.
2. Terms are included in the regression equation to handle any curvilinear or interactive effects.[5]
3. The dependent variable has no effect on the independent variables.

Since these assumptions are rarely all fully met, the technique should be used with caution. Nevertheless, when they are met, it provides a technique for rationally inferring causal relationships in complex situations even though experimental manipulation of the independent variables is impossible.

The technique is actually a variant of the method of computing a partial correlation between the dependent variable and each of the independent variables. In the regression technique, however, the emphasis is on regression weights rather than correlation coefficients. The advantage is that the final conclusions are in the form, "Increasing X_j by 1 unit increases the dependent variable by β_j units"; for example, "Every inch of snowfall causes, on the average, 15 additional traffic accidents." This is the most useful form of a statement when the emphasis is on cause and effect.

In any attempt to illustrate the method by an example, valid questions can be raised concerning the applicability of the assumptions listed above to that specific example. However, as an illustration of the technique, consider a study of the effects of different weather conditions on the frequency of traffic accidents. Suppose that each day, in a large city, several measures of weather conditions were recorded, and that the number of traffic

[4] See Theorem 5 of the document cited in Footnote 2.

[5] Configural and curvilinear terms, however, can produce complications in the interpretation of linear terms. See Darlington and Paulus (1966) for a more complete discussion.

accidents in the city each day was also recorded. Suppose then that a multiple regression equation was constructed to predict the number of traffic accidents in a day from the various measures of weather conditions that day. Despite the fact that weather conditions cannot be manipulated at will, and despite the fact that, say, humidity may be correlated with temperature, the beta weights in this regression equation would give information on the causal importance of each aspect of the weather.

The questions which arise in connection with this example illustrate the types of questions which must be considered in any use of the technique. For example, in connection with Assumption 1 above, we must ask: (*a*) "Does temperature have a positive beta weight because vacations come in the summer, and people drive more during vacations?" (This could be handled by, say, using number of accidents per vehicle mile as the dependent variable.) (*b*) "Does an aspect of the weather which has not been recorded, but which correlates with some measures which were recorded, affect accidents?" (This would result in spuriously high beta weights for these recorded measures.) Similar questions arise concerning the appropriateness of Assumption 2, although Assumption 3 seems to hold for this example.

Whenever a causal relationship is established in any branch of science, there is always the possibility of investigating the causal relationships which mediate those relationships found. This is true of the present technique. Thus, if hot weather is found to increase accidents, there remains for future investigators the task of discovering whether this is mediated by the effect of heat on the alertness of drivers, on the reliability of brakes, or on other factors. This consideration, however, does not lessen the value of the original finding.

The method has been developed far beyond the limits indicated here. More complete discussions by social scientists of this and related techniques can be found in Simon (1957), Blalock (1964), Monroe and Stuit (1935), Dunlap and Cureton (1930), and Burks (1926). The method was first developed by Wright (1921) in the biological sciences, where it is known as "path analysis." Recent general discussions of the method were given by Wright (1954) and by Turner and Stevens (1959). At least one of these should be read by anyone planning to use the technique. Detailed recent discussions of particular aspects of path analysis have been given by Wright (1960a, 1960b) and by Turner, Monroe, and Lucas (1961). These and the more general articles also give references to further literature in the area. They also discuss in detail techniques applicable when some of the independent variables are themselves affected by other independent variables. The simple technique outlined above still applies in this situation, but the weight given to each independent variable measures only the direct causal effect which that variable has on the dependent variable, ignoring effects which operate indirectly through the effect which the independent variable has on other independent variables.

Usefulness

When the focus is on the prediction of X_0, rather than causal analysis, usefulness is clearly the measure of greatest interest. Usefulness actually has a closer relationship to a partial correlation coefficient than does β'_j; it can be shown that dividing the usefulness of X_j by $1 - \bar{R}^2$ gives the squared partial correlation between X_0 and X_j, holding all other variables constant. Since $1 - \bar{R}^2$ is constant for a given regression equation, it follows that the usefulnesses of the predictor variables in a regression equation are proportional to these squared partial correlations.

It follows directly from Formula 4 that β'_j equals the validity of the orthogonal component of X_j (i.e., the square root of the usefulness of X_j), divided by the standard deviation of the same orthogonal component (when all the original variables are expressed in standard-score form). Thus, if two variables are equally useful, the one with the larger β' weight has the orthogonal component with the smaller variance.

The hypothesis that a predictor variable has zero usefulness in the population is equivalent to the hypothesis that the variable

has a population beta weight of zero, so the significance tests of these two hypotheses are the same. The parametric test of this hypothesis is an F test, described in McNemar (1962, p. 284) and elsewhere. The F value given by this test equals $r_{0 \cdot j(p)}^2$ (which is the sample usefulness of the variable in question), multiplied by the fraction $(N - n - 1)/(1 - R^2)$. This fraction is, of course, constant for all the variables in a given regression equation. Hence the F statistic is equivalent to usefulness as a measure of the relative importance of the variables in a given regression equation. If a worker has access to a computer program which computes this F value for each predictor variable (if the test available is a t test, then t^2 equals F), then he can readily find each variable's usefulness by dividing F by the above fraction.

$\beta'_j \rho_{0j}$

It can be shown that $\bar{R}^2$ can be calculated from β' weights by the formula[6]

$$\bar{R}^2 = \beta'_1 \rho_{01} + \beta'_2 \rho_{02} + \cdots + \beta'_n \rho_{0n}$$

This formula has suggested to several writers (Chase, 1960; Hoffman, 1960; personal communications from several sources) that $\beta'_j \rho_{0j}$ must be a measure of the "importance" of X_j, since it totals to $\bar{R}^2$ when summed across all the variables in the regression equation, and all measures of importance have this property when predictor variables are uncorrelated. Ward (1962) raised a question concerning the value of the measure; in defense, Hoffman (1962) called it the unique measure of the "independent contribution" of X_j. Ward's position will be restated and elaborated, since the present position is in basic agreement with it.

Although it is the province of an author to assign a name like "measure of independent contribution" to any statistic he proposes, this particular name has accumulated a good deal of "surplus meaning" by virtue of the powerful properties which it has when predictor variables are uncorrelated, as in analysis-of-variance designs, where its meaning is highly specific. Partly as a review, the following is a list of meanings which "independent contribution" has when predictor variables are uncorrelated:

1. The squared validity of X_j.
2. The usefulness of X_j.
3. β'^2_j.
4. The amount the variance of the regression equation would drop if X_j were removed from the equation, expressed as a proportion of σ_0^2.
5. The amount the covariance between the regression equation and X_0 would drop if X_j were removed, expressed as a proportion of σ_0^2.
6. The increase in the variance of $X_{0(p)}$ when X_j is removed from the equation, expressed as a proportion of σ_0^2.
7. The average difference, expressed as a proportion of σ_0^2, between σ_0^2 and the variance of X_0 in subpopulations in which X_j is held constant.

This list attempts to include all of the major properties which most readers consciously or unconsciously associate with the term "independent contribution." It is thus of considerable interest to note that $\beta'_j \rho_{0j}$ has none of these properties. As a minor exception, $\beta'_j \rho_{0j}$ has Property 5 if the remaining variables in the regression equation are not reweighted after removal of X_j, but this is not a property of any particular interest.

Although all of the measures in the above list do sum to $\bar{R}^2$ when predictor variables are uncorrelated, this fact alone does not justify the use of a measure, simply on the grounds that it sums to $\bar{R}^2$ even when predictor variables are intercorrelated. It would be better to simply concede that the notion of "independent contribution to variance" has no meaning when predictor variables are intercorrelated. The meaninglessness of $\beta'_j \rho_{0j}$ as a measure of importance is further underscored by the fact that it can be zero, or even negative, in cases in which X_j contributes substantially to the prediction of X_0.

Englehart's Measure

Englehart assigned a "contribution to variance" not only to each predictor variable, but also to the joint effect of each pair of predictor

[6] See Theorem 9 of the document cited in Footnote 2.

variables. He based his system on the formula

$$\bar{R}^2 = \beta'_1{}^2 + \beta'_2{}^2 + \cdots + \beta'_n{}^2 + 2\beta'_1\beta'_2\rho_{12} + 2\beta'_1\beta'_3\rho_{13} + \cdots + 2\beta'_{n-1}\beta'_n\rho_{(n-1)n}$$

Each of the first n terms in this sum is labeled the "contribution to variance" of the corresponding predictor variable, while each of the last $[n(n-1)]/2$ terms is called the contribution of the "joint effect" of two variables. This analysis was accepted by McNemar (1962, p. 176).

The criticisms of this measure are similar to those of $\beta'_j\rho_{0j}$; the measure has none of the most important properties that a "contribution to variance" has when variables are uncorrelated. The concept of the "joint contribution to variance" of two predictor variables might connote to some readers a measure of the amount that $\bar{R}^2$ would drop if the two predictor variables were somehow made uncorrelated. This connotation would be incorrect; in fact, if β'_j, β'_k, and ρ_{jk} are all positive, so that the "joint contribution to variance" of X_j and X_k is positive in Englehart's system, then $\bar{R}^2$ would actually *increase* if ρ_{jk} were zero.

Inferring Relative Regression Weights from Relative Validities

This section briefly mentions a series of papers which deal with a problem, or set of problems, which is not clearly defined. Although these papers have not traditionally been considered to be closely related to regression theory, they are mentioned briefly here since a regression solution can be proposed for at least one of the problems with which they deal. All of these papers deal in some fashion with the relationship between the weight of a variable in a weighted average of several variables, and the "importance" of the variable to that composite. The weights of the variables in the composite may have been chosen by any subjective or objective method. This freedom in the method of assigning weights distinguishes these papers from those mentioned in the previous section, which assume that regression weights are used. For all of the measures of importance referred to in the present section, increasing the weight of a variable increases its importance.

Some of these papers (Creager & Valentine, 1962; Richardson, 1941) simply propose statistical measures of importance, calling the proposed measure the "effective weight" or "contribution to variance" of the variable. (In every case, the latter term is subject in large part to the same criticisms made of the term in the previous section.) Others (Edgerton & Kolbe, 1936; Horst, 1936; Wilks, 1938) go one step further, first adopting one particular measure of importance, and then showing how the variables should be weighted so that all of the variables are equally important, or, more generally, how the variables should be weighted so that the measures of importance of the different variables are proportional to some specified set of numbers.

These methods, then, are intended to be used for weighting variables in situations in which regression equations cannot be derived because the criterion variable is not easily observable, even though it exists in some meaningful sense. Although all of these papers give careful descriptions of the statistical properties of the resulting composite, none gives an example of a practical situation in which the composite can be shown to have optimum properties. Most of the papers state that the procedure for specifying the desired relative sizes of the measures of importance would vary across situations. This statement, though true, has been allowed to obscure the fact that not one of the papers gives even the slightest hint, even for one situation, how one should go about specifying these values. In other words, none of the papers makes a convincing case for the practical value of the particular measure of importance proposed.

In approaching the present problem, it would seem that the first question to be asked is what a layman or a psychologist is likely to mean when he tells a psychometrician that he wants several variables to be weighted so that they are, for example, "equally important." Most commonly (though certainly not always), he probably means that he estimates the variables to correlate equally with some specified but unobservable criterion variable. In this case, use can be made of

the property of multiple regression equations—obvious from an inspection of the normal equations of regression theory—that the relative weights of the predictor variables in a multiple regression equation can be determined from a knowledge of only the relative (rather than absolute) sizes of the correlations of the predictor variables with a criterion variable. That is, if the validities of all the predictor variables in a multiple regression equation are multiplied by the same constant and the beta weights are then recomputed using the new validities and the old matrix of predictor intercorrelations, the relative sizes of the weights will be unchanged; each weight will simply be multiplied by the same constant by which the validities were multiplied.

This fact enables the estimation of the optimum relative weights of several predictor variables even when the criterion variable is not directly observable, provided there is some estimate of the relative validities of the variables. For example, suppose several observers are estimated to be equally accurate raters of some trait on which subjects are to be ranked. Suppose the ratings by these observers are available, and the problem is to find the optimum relative weights for a weighted average of the raters. A problem of this type was described by Dunnette and Hoggatt (1957). A solution to this problem could proceed as follows. An arbitrarily chosen number can be entered into a multiple regression computer program as the common validity of the raters, along with the observed standard deviations and intercorrelations of the raters, and along with an arbitrarily chosen value for the standard deviation of the criterion variable. The weights computed by the regression program are then the weights used to form a composite variable. If the user has entered into the program accurate estimates of the relative validities of the different variables, then this composite is optimum in the obviously important sense that it correlates higher with the unobservable "criterion" variable than any other composite using different relative weights.

The technique thus makes explicit the measure of importance (i.e., simple validity) which one should use in specifying the relative importance of the different variables. The reader should be cautioned that the technique does not apply—at least in the simple form outlined above—to another common situation in which the validity of each predictor variable is estimated from the correlation of that variable with other predictor variables, rather than from external data as in the above example.

In using the technique, the arbitrarily chosen validities should not be set so high that they are inconsistent with the observed intercorrelations of the variables. For example, if two variables correlate zero with each other, it can be shown that they cannot both correlate .9 with the same criterion variable. Setting validities so high that such an inconsistency appears does not distort the relative weights computed by the program, but will usually produce one of two otherwise puzzling results (depending on the computer program used): The program will fail to run, or it will compute and print a value of $\bar{R}$ above unity.

Estimates of the Validity of Regression Equations

Estimating the Validity of the Population Regression Equation

Let the term "population regression equation" refer to the equation developed in the entire population using predictor variables X_1, X_2, $\cdots$, X_n to predict X_0; the validity of this equation is measured by the population multiple correlation $\bar{R}$. Likewise, let the term "sample regression equation" refer to an equation using the same variables which is developed in any random sample from that population; the validity of this equation in that same sample is measured by the sample multiple correlation R.

Normally, a sample multiple correlation is higher than the corresponding population multiple correlation. In the extreme instance in which a regression equation using one predictor variable is developed in a sample of only two individuals, the sample multiple correlation is unity in all but trivial cases,

no matter what the population correlation is. In general, the same result occurs whenever the number of predictor variables n is one less than the sample size N. If n is greater than or equal to N, then the solution is indeterminate; infinitely many sets of weights will yield sample multiple correlations of unity.

It is often useful to describe the validity of a regression equation in terms of its mean square error; this quantity is the mean of the squared differences between each person's true criterion score and the prediction of that score made by the regression equation.

The expected value of the mean square error in a sample of size N in which a regression equation is developed is equal to $(N-n-1)/N$ times the mean square error in the population of the population regression equation. Therefore, the reciprocal of this fraction times the sample mean square error is an unbiased estimator of the population mean square error. Since the latter mean square error equals the population variance of the component of X_0 orthogonal to the predictors, it is denoted by ${\sigma_{0(p)}}^2$, as mentioned earlier. Similarly, the sample mean square error is

$$s_{0(p)}^2 = \frac{1}{N}\sum_i (x_{i0} - \hat{x}_{i0})^2$$

where $\hat{x}_{i0}$ is the predicted score of person i on X_0, as made by the sample regression equation. Thus, the formula for an unbiased estimator of ${\sigma_{0(p)}}^2$ is

$$\hat{\sigma}_{0(p)}^2 = \frac{N}{N-n-1} s_{0(p)}^2 \qquad [11]$$

An examination of the derivation of this estimator (Graybill, 1961, p. 111) shows it to be unbiased even if none of the usual assumptions of linearity, homoscedasticity, and normality holds, although without these assumptions little can be said about its efficiency.

A special case of Formula 11 is the case in which $n = 0$. In this case, the prediction of any individual's criterion score is the sample mean, so that the sample mean square error is the sample variance of the criterion variable. By the same reasoning, the population mean square error is the population variance of the criterion variable. Hence, the familiar formula which states that $N/(N-1)$ times the sample variance gives an unbiased estimate of the population variance is simply the special case of the previous formula in which $n = 0$.

The population mean square error is related to the multiple correlation by the formula

$$\bar{R} = \sqrt{1 - ({\sigma_{0(p)}}^2/{\sigma_0}^2)} \qquad [12]$$

which is merely the translation into present notation of a familiar formula taught in most undergraduate statistics courses. Wherry (1931) suggested that $\bar{R}$ could be estimated by substituting into this formula the estimates of ${\sigma_{0(p)}}^2$ and ${\sigma_0}^2$ described in the previous two paragraphs. He further pointed out that the ratio between these two estimates is a function of the sample multiple correlation. The resulting formula can be found in McNemar (1962, p. 184) and elsewhere.

Although the Wherry formula is based on unbiased estimators of ${\sigma_0}^2$ and ${\sigma_{0(p)}}^2$, in itself it is not an unbiased estimator of $\bar{R}$ in the strict statistical sense, contrary to McNemar (1962, p. 184) and others. However, this is no grounds for criticism of the formula, since it has long been known that any unbiased estimator of $\bar{R}$ has properties which make it clearly inferior to certain biased estimators. By definition, an estimator is unbiased only if the mean of an infinite number of estimates from independent random samples equals the parameter estimated, no matter what the value of that parameter is. When $\bar{R}$ is less than 1, no estimate of $\bar{R}$ based on finite samples can be perfect, that is, yield exactly correct estimates of $\bar{R}$ in every sample. Therefore, if $\bar{R} = 0$, an estimator which is unbiased must be one which can yield negative estimates in some samples. But this means that a statistic, in order to be an unbiased estimator of $\bar{R}$, must be able to assume values which the parameter estimated cannot assume, since $\bar{R}$ is greater than or equal to zero. Clearly, whenever an estimate of $\bar{R}$ is negative, the estimate can be improved by estimating $\bar{R}$ to be zero, since zero is always closer to the true $\bar{R}$

than is the negative estimate. But although this modification of the estimation procedure is obviously an improvement, it no longer yields an unbiased estimate of $\bar{R}$. Therefore, unbiased estimators of $\bar{R}$ are clearly not the best estimators. For similar reasons, unbiased estimators of $\bar{R}^2$ are of no practical interest. These points were made by Olkin and Pratt (1958), who nevertheless developed an unbiased estimator of $\bar{R}^2$.

Estimating the Validity of a Sample Regression Equation

Lord (1950) and Nicholson (1948) have pointed out that the Wherry formula has often been misinterpreted as an estimator of the true validity (i.e., the validity in the population) of a multiple regression equation developed in a sample. Actually, it overestimates this validity. The Wherry formula estimates, instead, the validity of the *population* regression equation, which was the equation developed in the entire population rather than in a sample. This equation, by definition, has a higher validity in the population than any other linear equation using the same predictor variables. In general, the weights in a sample regression equation will not be exactly equal to the weights in the population equation. The sample regression equation will then necessarily have a lower validity in the population than will the population regression equation. Thus, the Wherry formula generally overestimates the population validity of a sample regression equation, because it actually estimates a parameter (the validity of the population regression equation) which is higher than this validity.

The same distinction must be made if the validities of the two types of regression equation are measured in a second (cross-validation) random sample from the population. If any regression equation, based either on the entire population or on a random sample from that population, is applied to a cross-validation sample, then the expected mean square error of that equation in the cross-validation sample equals the mean square error of that equation in the population. Since the regression equation developed in a sample does not predict as well in the population as does the population regression equation, it would therefore also not be expected to predict as well in a random cross-validation sample from the population.

Hence, three different mean square errors must be distinguished. The smallest, and normally the easiest to observe, is the sample mean square error of the equation developed in that same sample. The second and next smallest is the population mean square error of the equation developed in the entire population; this equals the expected mean square error of that equation in any random sample. The third and largest is the mean square error of a sample regression equation in the population, which equals the expected mean square error of such an equation in a cross-validation sample. The Wherry formula is based on a formula which estimates the second of these three mean square errors from the first, while a prediction of cross-validity requires predicting the third from the first. In his original article, Wherry failed to distinguish between the second and third of these mean square errors, and the resulting confusion still appears in even the most recent standard sources, such as Guilford (1965, p. 401) and Guion (1965, pp. 163–164). Since the Wherry formula was thus often misused to attempt to predict the cross-validity of a sample regression equation, it was often observed to overestimate this quantity (i.e., to underestimate the mean square error).

Lord and Nicholson, working independently, found that an unbiased estimator of the population mean square error of a regression equation developed in a sample of size N is

$$\frac{N+n+1}{N-n-1} s_{0(p)}^2 \qquad [13]$$

In their derivations, Lord and Nicholson assumed that the conditional distributions of X_0 are normal, have a common variance, and are linearly related to the predictor variables. They further assumed that the scores observed on the predictor variables are fixed by the investigator, rather than sampled randomly from a population (the usual case in psychology). If we replace the assumption of fixed scores by the assumption of random

sampling, and replace the other assumptions by the assumption that scores on all variables form a multivariate normal distribution, then the estimator comparable to Formula 13 is

$$\frac{N-2}{N-n-2}\cdot\frac{N+1}{N-n-1}\,s_{o(o)}{}^2 \qquad [14]$$

This estimator gives still larger estimates of the mean square error than does Formula 13, except in the trivial case in which $n = 0$, in which instance the two estimates are identical. Formula 14 is an algebraic rearrangement of a formula given by Stein (1960, p. 427). An independent derivation by the present author, which is longer but which requires less mathematical competence to follow, is given under Theorem 18 of the document cited in Footnote 2.

Unfortunately, there is as yet no practical means of establishing a confidence interval around estimates computed from either Formula 13 or 14; empirical work suggests that their standard errors might be quite large.

An extremely important property of Formulas 13 and 14 is that the estimated true validity of a sample multiple regression equation is very low (and the mean square error very high) when the number of predictor variables is large in relation to the number of people in the sample on which the equation was derived. This is often observed in practice. For example, Guttman (1941, p. 360) constructed a regression equation with 84 predictor variables using 136 subjects, and observed a correlation with the criterion variable of .73 in the initial sample and .04 in a cross-validation sample. Thus, it is often better to use fewer predictor variables, or to use a different prediction method altogether, than to use a regression equation with an extremely large number of variables. For example, using the same data, Guttman observed a cross-validity of .20 using a regression equation with 21 variables, and a cross-validity of .31 for a simple item-analytic technique.

Relation between the Mean Square Error and the Correlation Coefficient

The points to be made in this section are relevant to measures other than multiple regression equations, even though they will be phrased in regression terms.

Formula 12 shows the well-known means of translating the mean square error of a population regression equation into the coefficient of correlation between the regression equation and the criterion variable. However, when the mean square error of a *sample* regression equation is computed in a cross-validation sample or in the population as a whole, neither Formula 12 nor any other formula provides an exact translation of that mean square error into a coefficient of correlation between the regression equation and the criterion variable. This is because two regression equations can have different mean square errors yet correlate equally with the criterion variable. Suppose that two sample regression equations are based on the same predictor variables, and that the weights within the first equation have exactly the same relative sizes as the weights within the second. Then the two equations will correlate equally with the criterion variable. But if the actual sizes of the weights, or the size of the additive constant a, differ between the two equations, then the two equations will have different mean square errors in the population or in a cross-validation sample.

Just as two regression equations with different mean square errors may correlate equally with the criterion variable, parallel reasoning shows that two equations with the same mean square error may have different correlations with the criterion variable. Thus a cross-validity mean square error computed in a second sample, or estimated by Formula 13 or 14, cannot be translated exactly into a correlation coefficient by formulas analogous to Formula 12. If N is large, however, the familiar formula for translating a mean square error into a correlation coefficient should give a good approximation.

When a regression equation or other measure is used to select the m individuals estimated to be highest on the criterion variable and m is fixed by the situation, then the value of the measure depends only upon the relative, not absolute, scores of the individuals on that measure. Since the correlation coefficient likewise depends only on relative scores, it

follows that, in this situation, the correlation between the measure and the criterion variable gives a more realistic statement of the value of the measure than does the mean square error. On the other hand, when m is not predetermined, whether a person is selected depends upon his absolute score on the measure rather than upon the relationship of that score to the scores of other individuals. In this situation, the value of a measure is a function of the actual difference between a person's estimated and true criterion scores; therefore, the mean square error is a more appropriate index of the value of the measure than is the correlation coefficient. For this reason, Formulas 13 and 14 are most useful in situations in which the number of people to be selected by a measure is flexible rather than predetermined. (None of this is meant to imply, however, that either the multiple correlation coefficient or the mean square error is *proportional* to the value of a test battery, as value is measured in decision theory terms.)

Statistical Criteria for Selecting Predictor Variables

Formulas 13 and 14 show the desirability of selecting a small number of predictor variables for use in a regression equation when v, the total number of available variables, is large. This section discusses several methods which have been proposed for doing this. All methods discussed below involve complex computational manipulations, such as inverting the $v \times v$ correlation matrix of predictor variables. Although modern computers use approximate methods to perform these calculations, the calculations are so complex that either the computational time or the rounding errors increase rapidly as v increases. As a very rough rule, when v is larger than approximately 50–100, item-analytic methods discussed by Darlington and Bishop (1966) are preferable to any of the methods discussed below, both because they are simpler computationally and because tests constructed by those methods have been demonstrated to perform better on cross-validation when the number of people in the test-construction sample is small. Roughly speaking, the methods discussed below are of most interest when $10 < v < 50$, but may be of value when $5 < v < 100$.

If an investigator wishes to predict a criterion variable by a regression or least squares technique, and he has available v possible predictor variables, he can use in the equation all v variables (so that n, the number of predictor variables used in the regression equation, is equal to v). Or, he can discard all v predictor variables (so that $n = 0$) and use the sample mean of the criterion variable as the prediction of each person's criterion score. Or, he can use a regression equation with less than v predictors (so that $v > n > 0$). Since he can choose independently whether to include or exclude each of the v variables, he is faced with 2^v possible alternative sets of predictor variables, plus sets formed by lumping together several variables and then entering them in a regression equation as one variable.

In general, it is impractical to compute all of the 2^v or more possible regression equations and then estimate the validity of each equation, so it is necessary to follow some simpler procedure in choosing a final regression equation. The remainder of this section discusses several such procedures.

Selecting Variables to Minimize Sampling Errors of Beta Weights

If the regression of X_0 on the predictor variables is linear and if the conditional distributions of X_0 have a common standard deviation, then the sampling distribution of each sample beta weight b_j has mean β_j (hence b_j is an unbiased estimator of β_j) and a conditional standard deviation

$$\sigma_{0(\mathrm{p})} / \sqrt{N}\, s_{j(\mathrm{p})} \qquad [15]^7$$

[7] Although this expression as a whole is a population value, it contains the sample value $s_{j(\mathrm{p})}$. This usage will be new to many psychologists, although it is standard practice in some branches of statistical theory. Briefly, the expression as a whole is the standard deviation of the sampling distribution of b_j *in those samples which have a given value of* $s_{j(\mathrm{p})}$, rather than in all samples. This point is further clarified in the discussion at the beginning of Part II of the document cited in Footnote 2. Some of the points made briefly in the remainder of the present paragraph are also expanded there.

An estimate of Expression 15, using the estimate of $\sigma_{0(p)}^2$ given by Formula 11, is computed by most standard multiple regression computer programs. If we can also assume normality of the conditional distributions of X_0 (or if N is large enough so that the central limit theorem applies), then dividing $b_j - \beta_j$ by this estimate of Expression 15 yields a statistic with a t distribution with $N - n - 1$ df. When β_j is set equal to zero, this test is equivalent to the F test mentioned earlier. Although many texts fail to mention it, Bartlett (1933, esp. pp. 277–278) has shown that use of both Expression 15 and the t test is appropriate when the sample values on the predictor variables are determined by the random sampling procedure common in psychology, as well as in the case (more commonly discussed by statisticians) in which those values are fixed by the investigator.

Because the quantities N and $\sigma_{0(p)}$ in Expression 15 are the same for all the variables in any one regression equation, the standard errors of the several b_js in the equation are inversely proportional to the values of $s_{j(p)}$, which are the observed standard deviations of the orthogonal components of the various predictor variables. The quantity $s_{j(p)}$ is generally large if X_j has low correlations with the other predictor variables and small if X_j is highly correlated with the other predictors; therefore, the weights of the variables which have the lowest correlations with other predictor variables are generally the weights which are least subject to sampling errors.

Cureton (1951a, pp. 12–15; 1951b, p. 691) referred to the number of variables which must be removed from a set of predictor variables in order to leave the remaining variables reasonably uncorrelated with each other as the number of "approximate linear restraints" in the set. He recommended that variables be removed or combined so as to eliminate the approximate linear restraints, in order to maximize $s_{j(p)}$ for each j and thus minimize the sampling errors of beta weights in the regression equation. His recommendation was apparently accepted by Guilford (1954, p. 404). It will be shown, however, that Formulas 13 and 14 raise considerable doubt as to the value of this strategy.

Consider a situation with three predictor variables, X_1, X_2, and X_3. Suppose the initial-sample validity of the regression equation using X_1 and X_2 equals the initial-sample validity of the equation using X_1 and X_3, but suppose that r_{12}^2 is lower than r_{13}^2. In this situation, an investigator following Cureton's recommendation would prefer using the former of the two equations, since the estimated standard errors of the beta weights are lower in that equation, even though the initial-sample validities of the two equations are the same.

On the other hand, when Formula 13 or 14 is used to estimate the true validity of a regression equation, intercorrelations of the predictor variables are ignored except insofar as they affect the initial-sample validity. Hence, in the above example, the predicted cross-validities of the two regression equations would be the same, despite the differences in the estimated sampling errors of the beta weights caused by the difference between r_{12}^2 and r_{13}^2.

We thus have the paradoxical situation that the sizes of the errors in estimates of beta weights do not enter into the estimation of the true validity of a regression equation.

The solution to this paradox lies in the nature of the correlation between the two sample beta weights within the same equation (i.e., the correlation we would observe if we drew infinitely many equal-sized independent random samples from the population, computed two regression weights b_j and b_k in each sample, and then correlated b_j with b_k across the samples). In a regression equation with n variables, the correlation between any two weights b_j and b_k equals -1 times the partial correlation between X_j and X_k, partialing out the other $n - 2$ predictor variables.[8] This correlation is

[8] This statement assumes linearity and homoscedasticity but not normality. The statement is exactly correct only if values on the predictor variables are fixed by the experimenter; if instead they are sampled randomly (the most common case in psychology), then we use, instead of all samples, the subset of samples with given values of

defined as the correlation between the components of X_j and X_k orthogonal to the other $n - 2$ predictor variables. In the present example, in which $n = 2$, there are no other predictor variables; hence, the correlation between b_1 and b_2 is $-r_{12}$. Thus in this example, if r_{12} is positive, the correlation between the two sample beta weights is negative. Since sample beta weights are unbiased estimates of the corresponding population weights, this means that if r_{12} is positive, an overestimation of one beta weight will tend to be found in the same sample with an underestimation of the other beta weight. Further, the higher the value of r_{12}, the more probable it is that this relationship exists.

This fact becomes important when considered in conjunction with the effect of different combinations of errors in the two beta weights on the validity of the regression equation. When two predictor variables are positively correlated, then, if an error is made in estimating one beta weight, the adverse effect of this error on validity can be lessened by an error in the opposite direction in estimating the other beta weight. The higher the correlation between the two variables, the greater is the ability of errors in opposite directions to compensate for each other in the prediction of criterion scores, since the predictors are increasingly "substitutable" for each other.[9] In the extreme case in which two variables of equal variance correlate perfectly, any two pairs of beta weights with the same sum are completely equivalent to each other. For example, in this extreme instance, weights of .8 and $-.2$, of .3 and .3, and of $-.1$ and .7 are all equivalent since each pair sums to .6.[10]

Hence, when r_{12} is positive, the sampling errors of the two beta weights are larger than if $r_{12} = 0$, but the errors tend to be in opposite directions and such errors tend to compensate for each other in a manner lacking when $r_{12} = 0$.

Errors in beta weights also tend to compensate for each other when r_{12} is negative. Again, in the extreme case in which $r_{12} = -1$, the compensation is perfect. Thus, it is precisely in those situations in which errors in the estimates of beta weights tend to be largest that the adverse effect of these errors on validity is minimized by the pattern in which the errors tend to occur. This is true for regression techniques with any number of variables.[11] Therefore an investigator seeking to choose which of several regression equations has the highest true validity need not concern himself directly with sampling errors of beta weights. He should simply choose the equation for which Formula 13 or 14 predicts the lowest cross-validation mean square error, based on the number of individuals in the initial sample, the number of predictor variables in an equation, and the initial-sample validity (expressed in terms of the initial-sample mean square error).

The laws described above also explain the paradoxical but common finding that when predictor variables are highly correlated, regression equations developed in two different random samples from the same population often have widely different weights, yet both equations predict about equally well in both samples. In a typical example of this effect in operation, a psychologist using the Graduate Record Exam Verbal Aptitude Test and the Miller Analogies Test to predict a criterion of success in graduate school found that the regression equation developed in one half of his sample gave a high positive weight to the GRE and a near-zero weight to the MAT, while the equation developed in the other half of his sample did exactly the opposite, giving the MAT a high positive weight and the GRE a near-zero weight. Yet each equation worked almost as well in the

the predictor variables. See Theorem 16 of the document cited in Footnote 2.

[9] The last two statements follow directly from a formula proven by Guttman (1941, p. 305). It is given without proof as Theorem 17 of the document cited in Footnote 2.

[10] The truth of this statement is unaffected by the fact that standard methods of deriving multiple regression weights break down when two predictor variables are perfectly correlated.

[11] These few paragraphs are an attempt to reconcile Formulas 13 and 14 with facts which at first seem to contradict them, using the simplest case of two predictor variables. A more rigorous development of the reasoning presented would simply amount to proofs of those formulas.

other half of the sample as in the half in which it was originally derived.

Removing Variables with Small Beta Weights

It was shown above that the predictor variables with the smallest β' weights in a population regression equation are not necessarily those whose removal would cause the smallest drop in the population validity of that equation. The same relationship clearly holds between initial-sample beta weights and initial-sample validity.

Formulas 13 and 14 show that when two regression equations have the same number of predictor variables, the one which has the higher initial-sample validity has the higher estimated true validity. Since removing variables with the smallest beta weights is not the most efficient way to achieve the highest possible initial-sample validity after the removal of a given number of predictor variables, it follows that this is also not the best way to maximize true validity.

Stepwise Regression

The foregoing discussion has made it clear that the only statistics relevant to selecting predictor variables from a larger number of variables are the initial-sample validity, N, and n for each of the possible regression equations formed from different combinations of the variables. The technique of stepwise regression, for which computer programs are widely available, has the desirable property that it uses only these statistics.

This technique selects variables for a regression equation one at a time. Selecting first the most valid predictor variable, it then selects that variable which when combined with the first is the most useful—that is, the one which adds the most to the multiple correlation and which thus yields the best two-predictor equation among those equations which contain the first variable selected. The extent to which the multiple correlation would be increased by a variable is determined by computing the validity of the orthogonal component or some mathematically equivalent statistic for the predictor variable being considered. The technique then selects by the same criterion the variable which combines with the first two variables to produce the best three-predictor equation. Subsequent variables are selected in a similar manner. Variables can also be removed if they are found to be no longer useful.

The process can be stopped when the initial-sample validity of the equation approaches that computed using all available variables, or when adding the most useful remaining variable produces no statistically significant increase in the multiple correlation by the significance test mentioned earlier. Significance tests are not normally appropriate for this purpose, however, since addition of a variable to a regression equation does not normally require a definite rejection of the hypothesis that fewer variables would suffice. Perhaps the best strategy is to use Formula 13 or 14 to evaluate each of the regression equations calculated by a stepwise regression computer program, and then to select the one equation which appears best by this criterion. Of course, Formula 13 or 14 will then underestimate the mean square error of the equation so selected for the same reason that the correlation between a test and a criterion variable can be expected to shrink if the test was selected from a large number of tests on the basis of this correlation.

Factor Analysis of Predictor Variables

Another technique for reducing the number of predictor variables is to factor analyze the set of all available predictors and then use some of the resulting factors in a regression equation in place of the original variables. This section discusses the conditions under which this procedure or variations of it are likely to improve the prediction of the criterion variable.

If the number of factors extracted equals the original number of predictor variables, then it can be shown that the multiple regression equation constructed to predict the criterion variable from the factors is equivalent to the comparable equation constructed from the original variables. The two equations will make identical predictions for any individual since the weight given to each original variable in the equation based on factors exactly equals the weight given that same variable

in the regression equation based on the original variables.[12] Therefore, any improvement resulting from the use of factors as predictors can occur only when the number of factors used is smaller than the number of original predictor variables.

Because of this, often only the few factors which account for the most variance are used (cf. Horst, 1941, pp. 437–444). However, from a purely mathematical standpoint, it could conceivably happen that the factor which accounts for the least variance in the predictor variables could correlate perfectly with the criterion variable, and all other factors could correlate zero with the criterion. Therefore, if factor analysis is a possibility, it is important to consider, from the nature of the variables being factored, whether this or a similar result is likely to occur.

When the variables being factored contain substantial error variance, it is well known that this error tends to be concentrated in the factors which account for the least variance, with a resultant increase in the reliability and therefore the validity of the other factors. In such a situation, the strategy of using in a regression equation only the several factors which account for the most variance would have much to recommend it.

The situation is different when highly reliable variables, such as age, sex, or census data, are used. In such situations, it could happen that factors which account for very little variance in the predictor variables are highly useful in predicting the criterion. For example, if two of the original variables are highly correlated, such as subject's age and age of the subject's next younger sibling (assuming he has one), a factor consisting of the difference between these two scores would "account for" very little variance in the original two variables. Yet this difference might well have had an important effect on the subject's childhood and therefore might correlate more highly with an external criterion than, say, a factor consisting of the sum of the two ages, which accounts for far more variance in the original predictor variables. In such situations, two alternative strategies are especially worthy of consideration: stepwise regression (discussed above), and stepwise regression using all of the factors rather than the original variables. When factors are uncorrelated, it follows from Formula 3 that the latter procedure simply involves selecting the factors most highly correlated with the criterion variable. Both these strategies have the desirable property that only the usefulness of each predictor variable is considered in the selection of variables, and this property is not shared by any strategy which considers only the factors which account for the most variance in the original set of predictor variables. An empirical comparison of the two stepwise methods has been made by Burket (1964).

[12] See the discussion under Theorem 11 of the document cited in Footnoote 2.

Summary

Basic Formulas

The beta weight and usefulness of a predictor variable in a multiple regression equation are expressed simply in terms of the properties of the component of the variable orthogonal to the other predictor variables.

Suppressor Variables

A variable receives a negative weight in a regression equation if the ratio between its correlation with the error in the rest of the equation, and its correlation with the criterion variable, exceeds a certain amount.

The relations possible among sets of variables are so complex that when a variable with a positive correlation with the criterion variable receives a negative weight in a regression equation, it is generally very difficult or impossible to determine, from the content of the variables, whether the negative weight is "unreasonable."

Measures of the "Importance" of a Predictor Variable

When the predictor variables in a multiple regression equation are intercorrelated, the "contribution to variance" of a predictor variable cannot be interpreted in the same way that it can be interpreted when predictor variables are uncorrelated. In the latter case, the phrase has essentially the same meaning it has in analysis-of-variance designs.

If the usefulness of a predictor variable is defined as the amount that the squared multiple correlation would drop if the variable were removed, then rank ordering the predictor variables in a regression equation gives different orders depending on whether the ranking is by validity, by usefulness, or by the absolute value of the beta weight. This is true even if all variables have the same standard deviation.

From the sizes of the weights in a multiple regression equation predicting a specified dependent variable from several independent variables, it is sometimes possible to measure the size of the "effect" which each of the independent variables has on the dependent variable.

Two measures of "importance," which sum to $\bar{R}^2$ when summed across all variables in a regression equation, have little practical value.

Inferring Relative Regression Weights from Relative Validities

The relative sizes of the weights in a regression equation can be computed from the relative validities (correlations with the criterion variable) of the predictor variables, even if the actual validities are unknown. This provides an exact solution to a common practical problem.

Estimates of the Validity of Regression Equations

The Wherry formula estimates the validity of the multiple regression equation developed in a population from the validity of an equation developed in a sample.

Statistics which give strictly unbiased estimates of a population multiple correlation coefficient are of no practical interest.

The Wherry formula has been used widely but incorrectly to estimate the cross-validity of a regression equation developed in a sample. Two alternative formulas are the correct formulas for this situation.

These alternative formulas produce extremely low estimates of cross-validity when the number of predictor variables is large in relation to the number of cases used in developing the regression equation. This agrees with the results of empirical cross-validation studies. Therefore, cross-validity is sometimes enhanced by using fewer predictor variables or by using a different prediction method altogether.

Estimates of the true validity of a sample regression equation can be expressed either as correlation coefficients or as mean square errors. Unfortunately, estimates in one form cannot always be readily converted to the other form, despite the well-known formula relating the two in other situations. The correlation coefficient is more useful in "fixed quota" situations, and the mean square error is more useful in "flexible quota" situations.

Statistical Criteria for Selecting Predictor Variables

The method of "approximate linear restraints" is not the most effective method of selecting predictor variables, because of the highly paradoxical relationship between the validity of a regression equation and the sampling errors of beta weights in the equation.

The same analysis explains the fact that regression equations developed in two different random samples from the same population often have surprisingly different beta weights, yet in any one sample the two equations make very similar predictions and thus have very similar validities.

The method of removing from a regression equation variables with low beta weights is not the most effective method, because such variables are not necessarily those whose removal produces the smallest drop in the multiple correlation.

Stepwise regression and extensions thereof are defended.

Under certain conditions, factor analysis can be used to develop a few factors which contain most of the valid variance in a set of predictor variables; under other conditions this procedure is not recommended.

REFERENCES

ANDERSON, T. W. *Introduction to multivariate statistical analysis.* New York: Wiley, 1958.

BARTLETT, M. S. On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh,* 1933, **53**, 260–283.

BEATON, A. E. The use of special matrix operators in statistical calculus. (Research Bulletin No. 64-51) Princeton, N. J.: Educational Testing Service, 1964.

BLALOCK, H. M., JR. *Causal inferences in nonexperimental research.* Chapel Hill: University of North Carolina Press, 1964.

BURKET, G. R. A study of reduced rank models for multiple prediction. *Psychometric Monographs,* 1964, No. 12.

BURKS, B. S. On the inadequacy of the partial and multiple correlation technique. *Journal of Educational Psychology,* 1926, **17,** 532–540, 625–630.

CHASE, C. I. Computation of variance accounted for in multiple correlation. *Journal of Experimental Education,* 1960, **28,** 265–266.

CREAGER, J. A., & VALENTINE, L. D., JR. Regression analysis of linear composite variance. *Psychometrika,* 1962, **27,** 31–38.

CURETON, E. E. Approximate linear restraints and best predictor weights. *Educational and Psychological Measurement,* 1951, **11,** 12–15. (a)

CURETON, E. E. Validity. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, D. C.: American Council on Education, 1951. (b)

DARLINGTON, R. B., & BISHOP, C. H. Increasing test validity by considering interitem correlations. *Journal of Applied Psychology,* 1966, **50,** 322–330.

DARLINGTON, R. B., & PAULUS, D. H. On the use of interaction terms in multiple regression equations. Paper presented at the meeting of the Educational Research Association of New York State, Albany, November, 1966. (Available from the author)

DUBOIS, P. H. *Multivariate correlational analysis.* New York: Harper, 1957.

DUNLAP, J. W., & CURETON, E. E. On the analysis of causation. *Journal of Educational Psychology,* 1930, **21,** 657–679.

DUNNETTE, M. D., & HOGGATT, A. C. Deriving a composite score from several measures of the same attribute. *Educational and Psychological Measurement,* 1957, **17,** 423–434.

EDGERTON, H. A., & KOLBE, L. E. The method of minimum variation for the combination of criteria. *Psychometrika,* 1936, **1,** 183–188.

ELASHOFF, R. M., & AFIFI, A. Missing values in multivariate statistics—I. Review of the literature. *Journal of the American Statistical Association,* 1966, **61,** 595–604.

ENGLEHART, M. D. The technique of path coefficients. *Psychometrika,* 1936, **1,** 287–293.

FEDERER, W. T., & ZELEN, M. Analysis of multifactor classification with unequal numbers of observations. *Biometrics,* 1966, **22,** 525–552.

GRAYBILL, F. A. *An introduction to linear statistical models.* Vol. 1. New York: McGraw-Hill, 1961.

GUILFORD, J. P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.

GUILFORD, J. P. *Fundamental statistics in psychology and education.* (4th ed.) New York: McGraw-Hill, 1965.

GUION, R. M. *Personnel testing.* New York: McGraw-Hill, 1965.

GULLIKSEN, H. *Theory of mental tests.* New York: Wiley, 1950.

GUTTMAN, L. Mathematical and tabulation techniques. Supplementary Study B. In P. Horst, *Prediction of personal adjustment.* (Bulletin No. 48) New York: Social Science Research Council, 1941.

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

HOFFMAN, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin,* 1960, **57,** 116–131.

HOFFMAN, P. J. Assessment of the independent contributions of predictors. *Psychological Bulletin,* 1962, **59,** 77–80.

HORST, P. Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika,* 1936, **1,** 53–60.

HORST, P. *Prediction of personal adjustment.* (Bulletin No. 48) New York: Social Science Research Council, 1941.

KENDALL, M. G. *A course in multivariate analysis.* London: Griffin, 1957.

LORD, F. M. Efficiency of prediction when a progression equation from one sample is used in a new sample. (Research Bulletin No. 50-40) Princeton, N. J.: Educational Testing Service, 1950. (Discussed by H. E. Brogden, Statistical theory and research design. *Annual Review of Psychology,* 1954, 5, 381.)

MCNEMAR, Q. *Psychological statistics.* (3rd ed.) New York: Wiley, 1962.

MONROE, W. S., & STUIT, D. B. Correlation analysis as a means of studying contributions of causes. *Journal of Experimental Education,* 1935, 3, 155–165.

NICHOLSON, G. E., JR. The application of a regression equation to a new sample. Unpublished doctoral dissertation, University of North Carolina, 1948. (Condensed in G. E. Nicholson, Jr., Prediction in future samples. In I. Olkin et al. (Eds.), *Contributions to probability and statistics.* Stanford: Stanford University Press, 1960.)

OLKIN, I., & PRATT, J. W. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics,* 1958, **29,** 201–211.

RICHARDSON, M. W. Supplementary Study D. In P. Horst, *Prediction of personal adjustment.* (Bulletin No. 48) New York: Social Science Research Council, 1941.

ROZEBOOM, W. W. Linear correlations between sets of variables. *Psychometrika,* 1965, **30,** 57–71.

SIMON, H. A. *Models of man: Social and rational. Mathematical essays on rational human behavior in a social setting.* New York: Wiley, 1957.

STEIN, C. Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics.* Stanford: Stanford University Press, 1960.

TURNER, M., & STEVENS, D. The regression analysis of causal paths. *Biometrics*, 1959, **15**, 236–258.

TURNER, M., MONROE, R. J., & LUCAS, H. L., JR. Generalized asymptotic regression and non-linear path analysis. *Biometrics*, 1961, **17**, 120–143.

WARD, J. H., JR. Comments on "The paramorphic representation of clinical judgment." *Psychological Bulletin*, 1962, **59**, 74–76.

WHERRY, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, 1931, **2**, 440–457.

WILKS, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 1938, **3**, 23–40.

WILLIAMS, E. J. *Regression analysis*. New York: Wiley, 1959.

WRIGHT, S. Correlation and causation. *Journal of Agricultural Research*, 1921, **20**, 557–585.

WRIGHT, S. The interpretation of multivariate systems. In O. Kempthorne et al. (Eds.), *Statistics and mathematics in biology*. Ames: Iowa State College Press, 1954.

WRIGHT, S. Path coefficients and path regressions: Alternative or complimentary concepts? *Biometrics*, 1960, **16**, 189–202. (a)

WRIGHT, S. The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics*, 1960, **16**, 423–445. (b)

(Received February 14, 1967)