# Issues in Multiple Regression[1]

Robert A. Gordon

ABSTRACT

Controlling for variables implies conceptual distinctness between the control and zero-order variables. However, there are different levels of distinctness, some more subtle than others. These levels are determined by the theoretical context of the research. Failure to specify the theoretical context creates ambiguity as to the level of distinctness, and leads to the partialling fallacy, in which one controls for variables that are not distinct in terms of appropriate theory. Although this can occur in using any control procedure, it is especially likely to occur in multiple regression, where high-order partial regression coefficients are routinely obtained in order to determine the relative importance of variables. Four major ways in which these regression coefficients can be seriously misleading are discussed. Although warnings concerning multicollinearity are to be found in statistics texts, they are insufficiently informative to prevent the mistakes described here. This is because the problem is essentially one of substantive interpretation rather than one of mathematical statistics per se.

The use of control variables is now a hallmark of sophisticated research. That this is so is due mainly to Kendall and Lazarsfeld's classic 1950 paper on partialling with categoric data, and to computers, which have removed restraints on calculating partials of almost any practical order, particularly for continuous data.[2] However, there is a fallacy in the uncritical use of partials that is easy to commit and which becomes more likely the higher the order of the partial. The purpose of this paper is to describe this fallacy and how it operates, especially in multiple regression, the case evidently most in need of clarification.

[1] Work on this paper was supported by research grant MH 10698-01, from the National Institute of Mental Health. The present version was part of a longer paper, "Issues in Multiple Regression and the Ecological Study of Delinquency" (Department of Social Relations, Johns Hopkins University, 1966), the remainder of which appeared in the *American Sociological Review* for December, 1967, under the title, "Issues in the Ecological Study of Delinquency." Readers interested in detailed substantive examples of points made in the present paper may wish to consult that paper. Concerning the material covered here, the author is grateful to the following Johns Hopkins University colleagues, with whom valuable discussions were held at one time or another: Leon J. Gleser and Joseph L. Gastwirth, Department of Statistics; Carl F. Christ, Department of Political Economy; and Arthur L. Stinchcombe, E. O. Schild, and James Fennessey, Department of Social Relations. This paper has also benefited from helpful comments made by the referees.

## THE PARTIALLING FALLACY

The introduction of a control variable into a relationship always implies a theoretical context, although in practice the context itself is often left unspecified. When experienced researchers fail to state the theoretical context explicitly, it is because they feel that it is sufficiently obvious. Often, they are right. Some researchers, however, have been misled by this silence, and they are unaware of how necessary it is to be conscious of the theoretical implications underlying any use of control variables. As though to emphasize this aspect of partialling, Kendall and Lazarsfeld referred to the control variable as the "test" variable. This clearly implied that a hypothesis was being tested, and thereby that a theory, however modest, was being invoked. Some researchers, however, engage

[2] Patricia L. Kendall and Paul F. Lazarsfeld, "Problems of Survey Analysis," *Continuities in Social Research; Studies in the Scope and Method of "The American Soldier,"* ed. Robert K. Merton and Paul F. Lazarsfeld (Glencoe, Ill.: Free Press, 1950), pp. 133–96.

in what is actually atheoretical partialling, as though the only hypothesis to be tested were the purely statistical one of whether the zero-order relationship could survive the application of any conceivable control.

The object, of course, is not simply to destroy an observed relationship but, rather, to see whether it can be destroyed by controlling for a variable that has been hypothesized to be potentially relevant and conceptually distinct within the theoretical context in which one has chosen to operate. Without a theory, however, there is no way of telling what is conceptually distinct and what is not. Consequently, variables are often introduced as controls that are not meaningfully different in terms of what would constitute an appropriate theory.[3] So closely do these variables approach being identical with one of the variables already in the zero-order relationship that controlling for them becomes tantamount to partialling that relationship out of itself.

An investigator studying the effect of socioeconomic status (SES) on delinquency, for example, could argue that median education *is* different from median rent and that it *is* reasonable to examine the relationship between either variable and delinquency free of the effects of the other variable. This is true as far as it goes, but it implies an extremely narrow and highly specialized theoretical focus. This smaller question should not be confused with a hypothesis concerning the relationship between SES and delinquency when one has two or more equally valid indicators of SES. Although one might wish to inspect for some reason the partial correlations between quantitative ability and verbal ability, on the one hand, and academic achievement, on the other, this would be a poor way to test whether ability in general is related to academic achievement. Clearly, partialling implies distinctness between the control and zero-order variables, but as each of these two examples shows, for a given set of data there can be different levels of distinctness. On one level education and rent are indistinguishable as indexes of SES; on another level they are two different variables with somewhat different properties and significance. If in working with these variables the theoretical context is left implicit, the investigator may find that he has committed himself to a theory—or to a level of distinctness—that he did not intend and that he would not support upon deeper consideration.

The fact that theory is quite undeveloped in his area does not excuse the researcher from deciding which of these major directions he wishes his readers to follow in understanding his results. Even an explicit postponement of the decision is preferable to an ambiguous presentation that could be construed either way. All too often, investigators are so unclear in their own minds why they are partialling that it is impossible to determine their intended level of distinctness. In this way they enjoy the methodological security of the microscopic level—in that one is always entitled to examine a partial if he wishes—while leaving their readers with impressions concerning the macroscopic level. Should attention be called to their indiscriminate partialling, they are apt to find themselves suddenly

[3] Disputes over variables left uncontrolled can also result from this absence of explicit theory. One party may define a variable globally, for example, "urbanism," so that it includes the variables typically correlated with living in a city, such as higher income and education, and therefore not control for the latter, whereas others may construe "urbanism" as a state of mind independent of income and education. Failure of the first party adequately to define urbanism in his study will precipitate attacks by the others on his omission of "obvious" controls. Actually, the entire dispute would be over a matter of definition entirely, and it should be conducted, if at all, on the semantic-esthetic-theoretical level and not on the methodological level. Fear of being so attacked serves as an incentive for controlling everything the investigator can lay hold of, whether appropriate or not, when the appropriate remedy would be for him to specify clearly the working theory that is guiding his research. These working theories can legitimately be quite modest—indeed, given the state of social science, they can hardly avoid being modest. We suspect that unwarranted embarrassment over their modesty keeps investigators from more often formulating working theories explicitly.

convinced that they had intended the narrower focus all along. Naturally, potential critics of such studies are reluctant to take a stand when the question of whether there is even an issue is itself so slippery. Consequently, sociology that is conceptually blurred accumulates, unchallenged, in the literature.

A somewhat more subtle version of the partialling fallacy is likely to be committed in multivariate studies that present all of the possible highest-order partials between each one of a large set of independent variables and the same dependent variable. Apparently, this practice also draws inspiration from Kendall and Lazarsfeld, although the procedures they advocated are actually quite different in logic. Kendall and Lazarsfeld's procedures *assume* knowledge concerning the presumed causal priority of the variables—they are not intended to provide that knowledge. Roughly, they address the question, "Is variable $A$ causally prior to $B$, or is it irrelevant?" and not the question, "Is variable $A$ causally prior to $B$, or is $B$ causally prior to $A$?" Yet it appears to be the latter question that is being posed when researchers calculate all possible highest-order partials to see which variable will emerge with the largest partial. Nothing in the Kendall and Lazarsfeld paper justifies using each independent variable in turn as a test variable for each other independent variable.

For one thing, the outcome of such a procedure is strongly influenced by small sampling or measurement errors when the independent variables are themselves highly correlated.[4] Moreover, the "intervening variable" and "spurious correlation" interpretations are not the only ones possible when covariation proves controllable by introducing a third variable. The choice between these two standard interpretations depends upon the causal ordering of the variables: whether the test variable is understood to be causally intermediate between the zero-order variables or whether it is antecedent to both of them. A third possibility, that it is to some degree causally *identical* with one of them, is completely overlooked in the classic discussions of partialling. Probably this is because no one expects researchers to employ such a variable as a control. That they might unwittingly do so fails to be anticipated. Yet, given that typical validities and reliabilities of social science measures are in the neighborhood of .70 and .85, respectively, for instruments constructed so as to *maximize* these values, it is not surprising that the factorial and causal equivalence of variables sometimes goes unrecognized when the correlations between them are generally lower than these.

When two variables are equivalent, they will both be equally valid to some degree, and controlling for one of them amounts to controlling for valid covariation. This makes as much sense as controlling for a parallel form of the same instrument. The presentation of all possible highest-order partials is a sure indication that the researcher has not thought through the theoretical connections among his variables. Once embarked upon, such a mechanical procedure is quite apt to lead him to control for variables whose covariation is largely valid.

Finally, there is no statistical rule for attributing controlled covariation to the influence of one rather than another of the independent variables, regardless of the disparity in size between their partial correlations. The question (of whether variable $A$ is prior to $B$ or $B$ is prior to $A$) is simply not answerable by this means.

An important property of the procedure of obtaining all possible highest-order partials is that the variables emerging with the largest partials will be those that are least redundantly represented. Conceivably, these could even be the variables that show the poorest zero-order associations with the common dependent variable. In such a case, an investigator is apt to conclude that his sophisticated statistical analysis has un-

[4] For an excellent discussion of this point, see H. M. Blalock, Jr., "Correlated Independent Variables: The Problem of Multicollinearity," *Social Forces*, XLII (December, 1963), 233–37.

covered the true importance of a variable that was otherwise obscured by the more superficial zero-order associations. Actually, nothing could be further from the truth. There is nothing more fundamental about a partial, as compared to a zero-order association, unless a good theory makes it so. The reasoning we are criticizing, for example, is reduced to the absurd when we realize that one could calculate all of the highest-order partials among the variables of a highly interrelated set and erroneously conclude, when the low partials fail to be significantly different from zero, that none of them was related to any other one.

Although the temptation to commit the partialling fallacy is greater in the case of continuous data—where it is more convenient to obtain partials of a high order—it should be emphasized that all control procedures are equally susceptible, including those for categoric data and for experiments.

## SOME UNAPPRECIATED ASPECTS OF MULTIPLE REGRESSION

When employing measures of association, investigators will calculate all possible highest-order partials for a given dependent variable only some of the time. Multiple regression, however, which is very popular in sociology, leads to highest-order partials automatically and invariably. For this reason, multiple regression is extremely susceptible to the partialling fallacy. Even worse, it is possible to commit this fallacy in a number of different ways in multiple regression, and because the partial regression coefficient is more difficult to understand intuitively than the partial correlation coefficient, these ways tend to be even more insidious. In the main body of this paper, we shall show in detail how this comes to be. But first, a few general remarks about the partial regression coefficient.

We have indicated that, to the degree the variables of a set are highly interrelated, numerous, and conceptually similar, we approach being able to produce a very small partial correlation between any two of them by controlling for the rest. Similar circumstances affect the partial regression coefficient in nearly the same way. This comes about as follows.

As redundant independent variables are successively introduced into a regression problem, their common predictive value gets averaged, in a weighted manner, over all of their regression coefficients. As a result, all of their regression coefficients decline in absolute value. At the same time, the multiple correlation increases only a trivial amount with each new variable, reflecting the fact that little new information is being added; that the multiple correlation cannot decrease indicates that the common predictive value is conserved, although it does get spread out over more and more regression coefficients, each becoming smaller and smaller as new redundant variables are fed into the problem.

Continuing with our examination of the regression coefficient, we note that, if at any point a new variable is added that is uncorrelated with previous independent variables, then the regression coefficients of the previous variables will be unaffected. Of course, it would be possible to then add more variables that are redundant with respect to this new variable but not redundant with respect to the earlier set, so that the regression coefficient of the new variable is reduced, but not the regression coefficients of the earlier variables.

The argument developed above helps us to realize that among the independent variables there could occur two or more subsets of variables, the members of which were redundant (strongly correlated) with variables in their own subsets but relatively independent (weakly correlated) with respect to variables in other subsets. It becomes immediately apparent that, under these circumstances, the relative size of a variable's regression coefficient depends to a considerable extent upon the *number* of other variables in its subset. And if all variables were redundant to the same degree with others in their subset, unrelated to the same degree with variables in other subsets, and all were equally related to the depend-

ent variable, then differences in the size of the regression coefficients between the variables of one subset and those of another subset would depend entirely upon the relative numbers of variables in the two subsets.

These conditions are, of course, quite special. However, they serve to bring into sharp relief processes that operate as well in the analysis of real data. In conjunction with these conditions, there are three others that also affect the regression coefficient. The main sections of this paper are devoted to illustrating all of these effects.

The problems we are about to discuss are usually alluded to in statistics texts under the heading of "multicollinearity." However, discussions of multicollinearity in statistical texts tend to be tantalizingly brief. Typically, they emphasize only that as independent variables become closely related the standard errors of regression coefficients become extremely large, leading to estimates of the regression coefficients that are unstable and hence unlikely to reappear even approximately in another sample from the same population.[5] Often, mention is made also of the fact that perfect correlation between two or more variables makes it impossible to solve the normal equations uniquely (or to invert the correlation matrix) and that, to the extent lack of perfect correlation is purely the result of random error, the entire solution is simply a spurious reflection of that error. (With double-precision computer arithmetic, rounding or truncation errors are no longer the problem they once were.) Both of these undesirable outcomes require extremely high correlations between independent variables. The effects that concern us here, however, can be produced as well by lower degrees of intercorrelation between predictor variables, and their main influence happens not to be exerted through the standard error of the regression coefficient. Statistics texts focus upon conditions of extremely high correlation because it is at that point that the resulting problems become most nearly *statistical* ones. The issues discussed in this paper, however, are basically substantive in nature. Consequently, although continuous with the problem of multicollinearity as treated in statistical texts, they are not of statistical interest per se, and therefore they are not adequately treated in any statistical source known to us. Persons sophisticated in statistics are quick to recognize this continuity, and they typically respond with a shrug when it is brought to their attention. However, we have known occasions when sophisticated consultants have advised clients to adopt procedures that lead to the very errors in multiple regression that this paper warns against, simply because the consultants were insensitive to the substantive implications of the problems brought to them. Since it is difficult to communicate complex problems to consultants so that they are alerted to all relevant implications, it is important that consumers of statistical advice be made aware of common pitfalls. Consultants, too, might benefit from having the following issues made more salient.

To help distinguish between the different effects to be discussed, we shall hereafter use the term "redundancy" to refer to *high correlation* between two or more independent variables, regardless of the exact number of variables, and the term "repetitiveness" to refer to the *number* of redundant independent variables, regardless of the degree of redundancy. This will enable us to emphasize either aspect of the situation, as necessary, in attempting to demonstrate just how, for one thing, multicollinearity operates.

*The effect of differential repetitiveness.*—In Table 1 we have created three correlation matrixes designed to display some of the

[5] In sociology, Blalock has devoted more attention to this problem than anyone else (*ibid.*). Numerous other references to multicollinearity appear in his *Causal Inferences in Nonexperimental Research* (Chapel Hill: University of North Carolina Press, 1964). See especially pp. 48, 66–67, and 87 ff., for comments relevant to this paper's concerns. Other helpful warnings are to be found in Edward E. Cureton, "Validity," *Educational Measurement*, ed. E. F. Lindquist (Washington, D.C.: American Council on Education, 1951), pp. 690–93.

properties to which we refer. These matrixes include independent variables only. The first one, matrix A, illustrates the effect of unequal repetitiveness on the regression coefficient. All five of the independent variables constituting matrix A are equally good predictors, one at a time, of the dependent variable, in that all correlate .60 with that variable. All are redundant to the same degree, .80, with others in their subset, and are unrelated to the same degree, .20, with those of the other subset. However, one subset contains three variables, and the other contains only two variables. The effect of this inequality is to create a substantial difference between the standardized regression coefficients of the two subsets; this difference is produced entirely by the difference in density of sampling between the two domains of content implied by the two subsets.[6] Obviously, it cannot be attributed in any meaningful sense to the relative importance of the variables when we do not even know what these variables are and when they all have exactly the same correlation with the dependent variable.

In the last three columns of Table 1 we present the standard errors of the regression coefficients, based upon a hypothetical sample size of 100, and the associated $t$-tests of the hypothesis that each regression coefficient is equal to zero. These standard errors, based on a reasonable sample size, are not unduly large, and there is little difference between the standard errors of the two subsets. However, the $t$-tests for the subset of three fail to reach significance, whereas those for the subset of two are significant. This outcome, again, is due entirely

TABLE 1

THE EFFECTS OF DIFFERENTIAL REPETITIVENESS (BASED ON HYPOTHETICAL SAMPLES OF 100)

| Example | Correlations between Independent Variables | | | | | | | | $r_{yi}$ | $b_{yi}$ | $s_{b_{yi}}$ | $t$ | $p <$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Matrix A: | .. | .8 | .8 | .2 | .2 | | | | .6 | .19 | .11 | 1.71 | N.S. |
| Subsets of | .8 | .. | .8 | .2 | .2 | | | | .6 | .19 | .11 | 1.71 | N.S. |
| 3 and 2 | .8 | .8 | .. | .2 | .2 | | | | .6 | .19 | .11 | 1.71 | N.S. |
| | .2 | .2 | .2 | .. | .8 | | | | .6 | .27 | .10 | 2.70 | .01 |
| | .2 | .2 | .2 | .8 | .. | | | | .6 | .27 | .10 | 2.70 | .01 |
| Matrix B: | .. | .8 | .8 | .2 | .2 | .2 | | | .6 | .16 | .08 | 1.95 | N.S. |
| Subsets of | .8 | .. | .8 | .2 | .2 | .2 | | | .6 | .16 | .08 | 1.95 | N.S. |
| 3, 2, and 1 | .8 | .8 | .. | .2 | .2 | .2 | | | .6 | .16 | .08 | 1.95 | N.S. |
| | .2 | .2 | .2 | .. | .8 | .2 | | | .6 | .23 | .07 | 3.13 | .01 |
| | .2 | .2 | .2 | .8 | .. | .2 | | | .6 | .23 | .07 | 3.13 | .01 |
| | .2 | .2 | .2 | .2 | .2 | .. | | | .6 | .41 | .05 | 9.03 | .001 |
| Matrix C: | .. | .8 | .8 | .8 | .2 | .2 | .2 | .2 | .6 | .12 | .09 | 1.41 | N.S. |
| Subsets of | .8 | .. | .8 | .8 | .2 | .2 | .2 | .2 | .6 | .12 | .09 | 1.41 | N.S. |
| 4, 3, and 1 | .8 | .8 | .. | .8 | .2 | .2 | .2 | .2 | .6 | .12 | .09 | 1.41 | N.S. |
| | .8 | .8 | .8 | .. | .2 | .2 | .2 | .2 | .6 | .12 | .09 | 1.41 | N.S. |
| | .2 | .2 | .2 | .2 | .. | .8 | .8 | .2 | .6 | .16 | .08 | 1.97 | N.S. |
| | .2 | .2 | .2 | .2 | .8 | .. | .8 | .2 | .6 | .16 | .08 | 1.97 | N.S. |
| | .2 | .2 | .2 | .2 | .8 | .8 | .. | .2 | .6 | .16 | .08 | 1.97 | N.S. |
| | .2 | .2 | .2 | .2 | .2 | .2 | .2 | .. | .6 | .40 | .05 | 8.83 | .001 |

Note.—Multiple correlations, when $r_{yi} = .6$: for Matrix A, $R = .815$; Matrix B, $R = .906$; Matrix C, $R = .910$. The $t$-tests are based on more decimal places than are shown in the table.

[6] All of the regression coefficients discussed in this paper are standardized ones. Often, such $\beta$ weights are accompanied by an asterisk, but since there can be no confusion on this point, we omit the asterisk.

to the difference in density of sampling variables, and it comes about mainly through the difference in the magnitude of the regression coefficients rather than through the slight difference in their standard errors.

Matrix B illustrates the effect of adding a third subset containing only one variable. Because this one variable is not repetitive with any other, its regression coefficient is much larger than the rest, although it, too, is no more strongly correlated with the dependent variable than any other variable. The results of the *t*-tests follow accordingly. Since this new variable is not redundant with variables of the other two subsets, their regression coefficients are hardly affected by its presence. An incidental effect of the new variable is to raise the multiple correlation from .815, for matrix A, to .906, for matrix B; and the slight decrease in the size of the standard errors from their values for matrix A results from this improvement in the multiple correlation. It is true here that the standard error of the new variable is approximately only half the size of the standard errors of the other variables; however, the fact that the regression coefficients of the subset of three are not significant again, whereas those for the subset of two are significant still, is determined far more by the relative size of the regression coefficients than by the relative size of the standard errors. Both the effect on the standard error and on the regression coefficient favor the statistical significance of the less repetitive variables, but it is the regression coefficient, of course, upon which interpretations of relative importance are based, statistical considerations aside.

Matrixes A and B, with their subsets of unequal numbers of equally good predictors, demonstrate that if a particular construct, such as socioeconomic status, is represented by many variables, and another construct by one or a few variables, the predictive value of SES would be spread thinly over several regression coefficients, while the predictive value of the other construct would be concentrated in only one or two coefficients, thus giving the impression that the SES variables were less strongly related to the dependent variable than the variables representing the second construct.[7] Under circumstances such as these, it could even happen that the regression coefficients of the construct having the *weaker* relationship with the dependent variable would attain statistical significance when the remainder did not, simply as a result of its being less repetitively represented. For example, in matrix C of Table 1, the correlation between the eighth variable (a subset of one) and the dependent variable could drop as low as .40, and it would still yield a significant regression coefficient of .187, which would be higher than those associated with variables accounting for more than twice as much of the dependent variable's variance. The regression coefficients of each of the first four variables would equal, in this case, .133, and of each of the next three variables, .174.

Ironically, the more important domain, in terms of total predictive value, is apt to be the one that is repetitively oversampled, both because its effects are likely to be more pervasive and because researchers will be inclined to devote more attention to it. The outcome of reducing the correlation between just the eighth variable and the dependent variable in matrix C to .40 indicates, equally

[7] An example of this effect is discussed in Gordon, "Issues in the Ecological Study of Delinquency," *op. cit.* (n. 1 above). It is drawn from a study by Bernard Lander in which *four* SES variables were included with *two* supposed anomie variables (see his *Towards an Understanding of Juvenile Delinquency* [New York: Columbia University Press, 1954]). Within the SES subset, the average absolute correlation was .77; within the anomie subset, it was .76. Redundancy within sets was thus nearly identical. The intrasubset correlations in our examples are intended to approximate these values. Although the absolute correlations between Lander's two subsets averaged .53, rather than the .20 for matrix A, this stronger relation between his subsets, if substituted into matrix A, would reduce the difference between the two sets of regression coefficients in our example by only 34 per cent. This tendency in Lander's matrix toward reduction in contrast, moreover, is offset by the greater repetitiveness within his larger subset, which contained four variables instead of the three in the larger subset of matrix A.

ironically, that odds-and-ends type variables of less consequence—included perhaps more because they happened to be at hand than because of their theoretical relevance—could then appear to be the more important.

Matrix C, it will be noted, was produced simply by adding another similar variable to each of the two major subsets of matrix B. It is intended to illustrate what is wrong with an attempt to test a hypothesis concerning the relative importance of two types of variable (represented by the two major subsets in matrixes A and B) by adding to the regression two new variables deemed to be representative of each type.[8] We see from the regression coefficients of matrixes B and C that, if these new variables were truly typical of their type, the outcome could not possibly be otherwise. If one pie is to be divided among a larger number and another pie among a smaller number, no matter how often we add one to the number for each pie, it will never alter the fact that the portions from the first pie will be smaller than those from the second. Similarly, the regression coefficients of the larger subset in matrix C continue to be smaller than those of the smaller subset, so that the status quo derived from matrix B remains essentially unaffected.

*The effect of heterogeneity among correlations with the dependent variable.*—In the examples of matrixes A, B, and C, all of the predictors were equally correlated with the dependent variable. Obviously, even if these correlations were equal in the population, they would not all be exactly equal in a sample, and there could even be real differences among their population values that were nevertheless so small as to lead us to regard them subjectively as practically equal. Whether or not we would regard them as equal, differences among these correlations do affect the regression coefficients, especially when there are redundant subsets among the predictors. However, the magnitude of the effect depends heavily upon whether the differences appear between variables in *different* subsets or between variables in the *same* subset. This is illustrated in Table 2, which shows what happens when the correlations with the dependent variable for matrix A, all of which were .60, are systematically varied. Column 1 simply repeats, for purposes of comparison, the rele-

[8] For examples, see David J. Bordua, "Juvenile Delinquency and 'Anomie': An Attempt at Replication," *Social Problems*, VI, No. 3 (1958–59), 230–238; Roland J. Chilton, "Continuity in Delinquency Area Research: A Comparison of Studies for Baltimore, Detroit, and Indianapolis," *American Sociological Review*, XXIX, No. 1 (1964), 71–83; and the discussion of these papers in Gordon, *op. cit.*

TABLE 2

THE EFFECTS OF FOUR DIFFERENT SETS OF CORRELATIONS WITH THE DEPENDENT VARIABLE WHEN THE INDEPENDENT VARIABLES ARE INTERCORRELATED AS IN MATRIX A

| VARIABLE | SET OF CORRELATIONS WITH DEPENDENT VARIABLE, $r_{yi}$ | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| 1 | .60 | .55 | .60 | .60 |
| 2 | .60 | .55 | .60 | .60 |
| 3 | .60 | .55 | .60 | .60 |
| 4 | .60 | .60 | .55 | .60 |
| 5 | .60 | .60 | .55 | .55 |
| | Corresponding Regression Coefficients, $b_{yi}$ | | | |
| 1 | .19 | .17 | .19 | .19 |
| 2 | .19 | .17 | .19 | .19 |
| 3 | .19 | .17 | .19 | .19 |
| 4 | .27* | .28* | .24* | .38* |
| 5 | .27* | .28* | .24* | .13 |
| | Standard Errors of Regression Coefficients, $s_{b_{yi}}$ | | | |
| 1 | .11 | .12 | .12 | .11 |
| 2 | .11 | .12 | .12 | .11 |
| 3 | .11 | .12 | .12 | .11 |
| 4 | .10 | .11 | .11 | .10 |
| 5 | .10 | .11 | .11 | .10 |
| | Corresponding Multiple Correlation Coefficients, $R_{y.12345}$ | | | |
| | .815 | .781 | .783 | .803 |

* Significant at .05 level or better. All tests have 94 degrees of freedom, based on hypothetical samples of 100.

vant data for the original matrix A problem. It will be recalled that matrix A consisted of a subset of three and a subset of two equally redundant predictors. Columns 2 and 3, respectively, show the effect of lowering the correlations with the dependent variable to .55, first for just the entire subset of three, and then for just the entire subset of two. In either case, the effect on the regression coefficients is small. However, when *just one* of the correlations involving a variable from the subset of two is lowered to .55, so that a difference in their relation with the dependent variable now appears between variables within the *same* redundant subset, the effect on the regression coefficients is pronounced (see column 4). In this situation, the regression coefficient of the variable with the higher correlation (here, the fourth variable) acquires most of the predictive value of its subset. In our example, its regression coefficient so closely approaches the values held by those for the subset of one in matrixes B and C that it might be said to behave as though it actually were a subset of one. Once again we point out that the effect on the standard errors of the regression coefficients, in any of the situations depicted in Table 2, is so small that it often fails to appear within two decimal places.

Our attitude toward the greater prominence given the fourth variable over its subset partner, the fifth variable, by this slight difference in their correlations with the dependent variable, could well be that this is quite proper. After all, if a dependent variable correlates slightly better with one of two variables that are themselves highly correlated with each other, it could mean that it is fundamentally more like that variable than like the second. Even so, the question could be raised as to whether a mode of analysis that transforms an 11:12 relationship (in the correlations with the dependent variable) or a 5:6 relationship (in terms of variance accounted for) into a 1:3 relationship (in the regression coefficients) provides the most helpful picture of the data. In any event, our acceptance of this outcome would be sharply revised if the outcome were based on observed values that did not reflect the true parameter values—for example, if the true correlations with the dependent variable were equal, but the observed correlations were not, or if the observed correlations reversed the direction of the true difference between the absolute correlations of two predictors with the dependent variable.[9]

It is difficult to say how likely these kinds of sampling fluctuations are to occur in practice. However, we might note that, for an observed correlation of .60 and a sample size of 100, the .95 confidence interval ranges from .46 to .71. In the present context, it suffices to point out that their effects on regression coefficients, should they occur, can be surprisingly strong.

Even when differences like those in column 4 of Table 2 reflect the true values of correlations, however, there is no reason to be complacent concerning their effect vis-à-vis the regression coefficients of variables in other subsets. In our example, the fourth variable is not correlated any more strongly with the dependent variable than the three variables constituting the larger subset, yet its regression coefficient is twice as large as any of theirs. In columns 2 and 3 of Table 2, we showed that, even when correlations with the dependent variable differed between entire subsets, the effect was not so great as this. This helps us to realize that if the correlation of the fourth variable with the dependent variable were in fact somewhat larger than the corresponding correlations of the subset of three, we would be tempted to regard its much larger regression coefficient as an appropriate reflection of this stronger relation with the dependent variable. In actuality, the greater part of the magnitude of the difference between regression coefficients would be due to differential relations with the dependent variable *within the subset of two* and would have nothing whatsoever to do with variables in the subset of three. Just how much of the difference

[9] It should be kept in mind that we are concerned with the absolute values of the correlations with the dependent variables, and not their algebraic values.

between regression coefficients would be due to differences within subsets and how much to differences between subsets would depend on the particular magnitudes involved, of course. Our point is simply that a very substantial part of this difference may be due to causes that are quite irrelevant to giving an accurate picture of the relative importance of the variables.

Of course, these differences in correlation with the dependent variable could appear among the variables either of the more repetitive (larger) subset or of the less repetitive (smaller) subset. In either case, they would tend to concentrate the predictive value of the entire subset in the regression coefficients of a smaller number of variables. Should the highest correlation with the dependent variable be attained by only one predictor, then the subset's predictive value would tend to concentrate in just one regression coefficient. Depending on whether such differences appeared in connection with the smaller or the larger subset, they would either add to or detract from the effect of differential repetitiveness on the variables involved.

*The effect of unequal redundancy between subsets.*—Obviously, the degree of redundancy (or level of internal correlation) can differ between subsets that are nevertheless clearly recognizable by their high internal correlations and which are identical with respect to repetitiveness (the number of predictors involved). This constitutes the simplest situation of the four that we shall discuss.

If all predictors correlate equally with the dependent variable—thus eliminating this source of disturbance—then unequal redundancy will produce larger regression coefficients, having smaller standard errors, for the variables in the less redundant subset. Although the resulting effects on both the regression coefficient and its standard error tend to be moderate, they are not insignificant. For example, in five out of six hypothetical problems that we constructed, based upon sample sizes of 100, these effects combine to prevent the regression coefficients of the more redundant subset from attaining significance. In all cases, those of the less redundant subset were significant. Table 3 presents the results from these six problems, which were varied systematically as to difference in redundancy (two levels) and strength of correlation between the subsets (three levels). It can be seen that the

TABLE 3

THE EFFECTS OF TWO LEVELS OF UNEQUAL REDUNDANCY AT THREE LEVELS OF BETWEEN-SUBSET CORRELATION IN A FOUR-VARIABLE MATRIX (ALL CORRELATIONS WITH THE DEPENDENT VARIABLE HAVE BEEN SET EQUAL TO .60)

| CORRELATIONS WITHIN EACH TWO-VARIABLE SUBSET | CORRELATIONS BETWEEN SUBSETS, $r_{13}=r_{14}=r_{23}=r_{24}$ | | |
|---|---|---|---|
| | .2 | .5 | .6 |
| | Regression Coefficients, $b_{y1}=b_{y2}$; $b_{y3}=b_{y4}$ | | |
| Subset I: $r_{12}=.7$ | .288 | .234 | .222 |
| Subset II: $r_{34}=.8$ | .264 | .204 | .180 |
| Difference | .024 | .030 | .042 |
| Subset I: $r_{12}=.7$ | .288 | .246 | .228 |
| Subset II: $r_{34}=.9$ | .252 | .192 | .156 |
| Difference | .036 | .054 | .072 |
| | Standard Errors of Regression Coefficients, $s_{b_{y1}}=s_{b_{y2}}$; $s_{b_{y3}}=s_{b_{y4}}$ | | |
| Subset I: $r_{12}=.7$ | .08 | .10 | .11 |
| Subset II: $r_{34}=.8$ | .10 | .12 | .13 |
| Subset I: $r_{12}=.7$ | .09 | .10 | .11 |
| Subset II: $r_{34}=.9$ | .14 | .16 | .18 |
| | $t$-Tests of Regression Coefficients | | |
| Subset I: $r_{12}=.7$ | 3.4* | 2.3* | 2.0* |
| Subset II: $r_{34}=.8$ | 2.7* | 1.7 | 1.4 |
| Subset I: $r_{12}=.7$ | 3.4* | 2.4* | 2.0* |
| Subset II: $r_{34}=.9$ | 1.8 | 1.2 | 0.9 |

* Significant at .05 level or better. All tests have 95 degrees of freedom, based on hypothetical samples of 100.

effects grow stronger (*a*) the greater the difference in redundancy between the subsets, and (*b*) the stronger the correlation between subsets.

It is somewhat of a challenge to devise a plausible illustration of this particular effect that would also be intuitively regarded as leading to an obviously erroneous conclusion.[10] Imagine, however, a battery of predictors of success in some performance that requires both verbal and mathematical ability. Assume that the number of verbal measures equals the number of mathematical measures, that the correlations between the two subsets of measures are uniform, and that all correlate equally well with the dependent variable.

In this situation, it could occur that all of the verbal measures correlate more highly with each other than do the mathematical measures with each other, as a result of the more specific nature of the latter. For example, a course in analytical geometry could be more different from one in probability theory than a course in history from one in English literature. (In factor-analytic terms, a greater part of the valid variance of the mathematical measures would be represented in specific factors and a smaller part in a common factor than in the case of the verbal measures. To eliminate the possibility of a strong general factor that could absorb even the specifics, it might be necessary to stipulate that the population be relatively homogeneous in intelligence.)

At first glance, it seems appropriate that the greater predictive value of the mathematical measures would be reflected in regression coefficients larger than those of the verbal measures. However, as the result of crowding into the analysis multiple measures of a domain that is more constricted than the mathematical domain—a quite reasonable step from the standpoint of enhancing reliability—it could happen that none of the regression coefficients belonging to the verbal measures would reach statistical significance. If one were unaware of the more fundamental dimensions underlying the predictors, of how they were organized, of the difference in redundancy, and of the effect of these upon a regression analysis, one might easily conclude that verbal ability was less important than mathematical ability, overlooking the fact that if only one verbal measure had been employed it surely would have been significant. Thus, differences in redundancy can bring about exactly the same result caused by differences in repetitiveness. The main difference is that the illusion of fairness is greater for the pure form of the present effect, in that the number of variables is the same for each domain (or subset). Both effects, of course, can be simultaneously operative.

*The effect of other possible variations in correlations.*—It is only when correlations are uniform within each separate subset, and uniform between them as well, so that subset boundaries are clearly defined, that we can conveniently speak of redundancy and of differences in redundancy between sub-

[10] An approximate example can be found in the data of Lander, Bordua, and Chilton (all cited above), where we can identify equal-sized subsets of unequal within-set redundancy by considering variables in pairs. These pairs are made up on the basis of each variable's highest correlation (in absolute value) with any other variable. Just as for sociometric choices, these relations may or may not be reciprocal. Thus, in their three studies their four SES variables form two stable and reciprocal subsets, but their two anomie variables do not. One SES subset is comprised of education and rent ($r$ = .89, .78, and .89, for each study in order), the other of substandard housing and overcrowding ($r$ = .86, .83, and .93). In Lander's study only, the two anomie variables represent each other's most highly correlated variable, and then, in contrast to the values of .89 and .86 for the SES pairs, their correlation is only −.76. Consequently, on a pairwise basis, the anomie subset for Lander's study is less redundant than the SES subsets. To an even greater degree, this disparity appears in the Bordua and Chilton studies too, so that the highest correlation of any anomie variable is always less than the highest correlation of any SES variable.

It should be emphasized that this illustration from real data focuses only on the correlations *within* these subsets of size two as a source of redundancy, as though the between-set correlations were all equal. In fact, in the data from the three studies, they are not all equal. As will be shown in the next section of this paper, this complicates the situation considerably.

sets. Even slight departures from uniformity within any one of these groups of correlations will affect the regression coefficients in ways that no measure of redundancy can anticipate.

Three plausible measures of redundancy are: (1) the average absolute correlation of each predictor with the rest, (2) the average of the squares of correlations, and (3) the squared multiple correlation (SMC) of each predictor with all of the remainder. (Obviously, if the correlations *between* subsets are all equal, only the off-diagonal correlations *within* subsets need be considered for the averages.) The two averages have the advantage of being unaffected by the number of variables when the correlations are uniform, thus enabling us to define redundancy independently of repetitiveness. Moreover, they are easy to calculate. Because the SMC incorporates the contributions of redundancy—in the pure sense—and of repetitiveness too, it summarizes accurately the total overlap of each predictor with the others. However, this quality makes it unsuitable for drawing a heuristic distinction between redundancy and repetitiveness.

Even when the correlations with the dependent variable are all identical, the more sophisticated SMC measure does not fully determine the relative strengths of the regression coefficients. Those familiar with the computations will recognize that this is because the SMC depends upon only the value of a main diagonal element of the inverse of the matrix of correlations between predictors (see equation [6], below). It is with the variations among the predictor correlations underlying this further indeterminacy that this section is concerned. Since subsets with precise boundaries rarely occur in real data, the manner of characterizing this variation is somewhat arbitrary. This section, therefore, could quite reasonably have had many other headings, such as "variations within subsets" and "variations in the correlations between subsets."

It is best to proceed by introducing explicitly the inverse of the correlation matrix. Many persons performing multiple regression will be aware that there exist several methods for computing the regression coefficients, all of them tedious. They are probably also aware, nowadays, that, except for certain special cases, any rectangular matrix can be inverted and that this also is a tedious procedure. Few of them, however, will ever have seen an inverse matrix—fewer still will have actually inverted a matrix by hand computation. This is unfortunate because matrix inversion is one of the ways to solve multiple regression problems and, as it turns out, the elements of the inverse matrix can be expressed in an intuitively meaningful way that enables one to observe what is going on in a multiple regression better than at any other stage of the calculations.

For three independent variables, the regression coefficients can be obtained from the following matrix multiplication:

$$\begin{vmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{vmatrix} \begin{vmatrix} r_{y1} \\ r_{y2} \\ r_{y3} \end{vmatrix} = \begin{vmatrix} b_{y1.23} \\ b_{y2.13} \\ b_{y3.12} \end{vmatrix}, \quad (1)$$

where the $r_{yi}$ are the correlations with the dependent variable, and the $C$'s are elements of the inverse matrix.[11] The matrix multiplication simply expresses the following three operations in ordinary algebra:

$$\begin{aligned} C_{11}r_{y1} + C_{12}r_{y2} + C_{13}r_{y3} &= b_{y1.23}, \\ C_{21}r_{y1} + C_{22}r_{y2} + C_{23}r_{y3} &= b_{y2.13}, \quad (2) \\ C_{31}r_{y1} + C_{32}r_{y2} + C_{33}r_{y3} &= b_{y3.12}. \end{aligned}$$

In view of our simplifying assumption that all of the correlations with the dependent variable are equal or, in other words, that $r_{y1} = r_{y2} = r_{y3} = r_{yc}$, each of the above three operations could be expressed in the

[11] Except for omitting asterisks for the standardized regression coefficients, our notation follows that employed in Helen M. Walker and Joseph Lev, *Statistical Inference* (New York: Henry Holt & Co., 1953), pp. 332–34. This reference also provides a fuller development of the matrix algebra of multiple regression, as well as a parallel presentation in ordinary algebra.

following form, here illustrated only for the first:

$$r_{yc}(C_{11} + C_{12} + C_{13}) = b_{y1.23}. \quad (3)$$

Inasmuch as the correlations with the dependent variable are not now of interest, it is clear that the sum of the elements in each row wholly determines the regression coefficient of the variable associated with that row of the inverse (and also of the correlation) matrix. Consequently, the impact of just the independent variables on the regression coefficient could be expressed in the form of such a sum for each row. At some later point, these sums could then be scaled down to true regression coefficients simply by multiplying each of them by $r_{yc}$.

As it stands, the inverse matrix is not very helpful. Its elements have no immediately apparent interpretation, and the familiar meaning attached to the original correlation coefficients will have become hopelessly obscured in the course of the involved calculations necessary to obtain the inverse. True, the relative magnitudes of the elements and their locations in the matrix will suggest points at which things are happening of relatively greater or lesser importance—but the nature of these happenings will be unknown.

However, each element of the inverse can be expressed in terms of more familiar quantities. Let us take first the main diagonal elements, $C_{ii}$, for our three-variable example. If $s_{i.jk}$ is used to represent the square root of the residual variance left when any independent variable is predicted by all of the remaining independent variables or, in other words, if

$$s_{i.jk} = \sqrt{1 - R^2_{i.jk}}, \quad (4)$$

where $R_{i.jk}$ represents the multiple correlation, then $C_{ii}$ is the reciprocal of this residual variance, or

$$C_{ii} = \frac{1}{(s_{i.jk})(s_{i.jk})}. \quad (5)$$

It might be pointed out, incidentally, that one can obtain all of the SMC's for a group of variables from the inverse matrix, since for any variable $i$,

$$\text{SMC} = 1 - \frac{1}{C_{ii}}. \quad (6)$$

For true correlation matrixes, that is, matrixes that observe the requirements for consistency among correlation coefficients, $C_{ii}$ will always be positive.[12] Its range is from 1.0 to infinity.

An off-diagonal element, $C_{ij}$, is defined as follows:

$$C_{ij} = -\frac{r_{ij.k}}{(s_{i.jk})(s_{j.ik})} \quad (i \neq j). \quad (7)$$

Thus, the numerator contains a familiar partial correlation coefficient, of an order that is always two less than the order of the matrix. Since we are now ignoring correlations with the dependent variable, the minus sign in formula (7) indicates that, when the partial correlation in the numerator is positive, this quantity will be subtracted from the main diagonal in forming the row sum, thus reducing our hypothetical regression coefficient. Of course, in real applications, the magnitudes and signs of the correlations with the dependent variable would have to be taken into account, in accordance with formula (2). It might be pointed out that the 1.0 in the numerator of the main diagonal entries, $C_{ii}$, is consistent with the definition of the $C_{ij}$, in that the partial cor-

[12] On the consistency relation, see *ibid.*, pp. 344–45. This consistency requirement governs the construction of the examples used in this paper, which were created simply by writing down correlation coefficients and then inverting the resulting matrixes to obtain the usual multiple regression statistics. It was not necessary to generate the raw data implied by the correlations, although much trial and error was involved in arriving at suitable illustrations. Some care is required, however, in order to avoid examples that violate the consistency rule. The appearance of negative elements in the main diagonal of the inverse is one indication that consistency has been violated. A negative element there implies an SMC greater than 1.0 and, hence, "negative" residual variance. One should also avoid creating $r_{yi}$ that are so large that a multiple correlation greater than 1.0 is implied for the example problem.

relation of a variable with itself is always 1.0.[13]

Inversion of the appropriate correlation matrix, incidentally, provides a convenient computer method for obtaining all of the highest-order partial correlations between variables of a set simultaneously, according to the formula

$$r_{ij.kl\ldots} \text{ etc.} = \frac{-C_{ij}}{\sqrt{C_{ii}C_{jj}}}. \qquad (8)$$

We can now examine the elements of the inverse matrix, together with their more detailed versions as expressed in formulas (5) and (7), in order to see what is going on within certain problems in which small differences in correlations between independent variables produce disproportionately large differences in regression coefficients.

Table 4 presents three matrixes. The correlations appearing in the first of these, matrix L, resemble closely the correlations between the four variables in an example of real data.[14] This matrix, as did the original, contains two clearly delineated two-variable subsets. The only alteration is that we have substituted a correlation of .71—approximately their mean—for each of the four heterogeneous correlations between the two two-variable subsets in the original. By then introducing only a very simple change into the between-set correlations, we can retain the relevance of real data, without their complexity, while observing the effect of variation among these correlations. This change appears in matrix M, where the correlations of the first variable with the third and fourth have been raised from .71 to .73. In each of the examples based on the three matrixes in Table 4, the hypothetical correlations with the dependent variable have been set equal to .50.

In matrixes L and M the internal correlations of the less redundant subset are .86 and of the more redundant, .89. In L, the larger regression coefficients naturally accompany the variables of the less redundant subset. This is the effect of unequal redundancy that was discussed earlier. Before going on to the more complicated matters of this section, let us note how this comes about, first by examining the ordinary inverse, and then the detailed inverse.

In the ordinary inverse, we note that the between-subset $C_{ij}$ are all equal, at $-.474$. Therefore, only the within-subset $C_{ij}$ need be considered. Of these, we look first at the main diagonal elements. $C_{11}$ and $C_{22}$, of the more redundant subset, are both larger than $C_{33}$ and $C_{44}$. This would produce larger regression coefficients for the *more* redundant subset, an effect opposite to that which actually occurs. The remaining within-subset elements, $C_{12}$, $C_{21}$, $C_{34}$, and $C_{43}$, which carry negative signs, determine the final outcome. This is because $C_{12}$ and $C_{21}$ are both absolutely larger than $C_{34}$ and $C_{43}$. In the pure case of the effect, this difference is always large enough to more than offset the difference between the two sets of main diagonal elements. Consequently, the more redundant subset will always have the smaller regression coefficients. The detailed inverse makes it clear why this is so.

The difference between the two sets of main diagonal elements reflects the greater overlap (higher SMC's) of the more redundant variables. In the detailed inverse,

[13] A derivation of equations (5) and (7) for the case with two independent variables, but for unstandardized coefficients, is given in A. Hald, *Statistical Theory with Engineering Applications* (New York: John Wiley & Sons, 1952), pp. 640–42. Because Hald uses only two variables, it is not readily apparent that many of the terms are actually partials when more than two variables are involved. The conversion to standardized data, however, is quite simple. Equation (6) is derived in K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering* (New York: John Wiley & Sons, 1965), p. 450. From this, equation (5) is readily obtained. Equation (8) is stated without proof in Robert G. D. Steel and James H. Torrie, *Principles and Procedures of Statistics* (New York: McGraw-Hill Book Co., 1960), p. 301. See also Cyril H. Goulden, *Methods of Statistical Analysis* (New York: John Wiley & Sons, 1952), chap. viii. From this equation, both equations (5) and (7) can be obtained with the help of (6). We regret that, although these relationships are fairly well known, we cannot cite a proof for the general case of equation (7).

[14] The reference is to the four SES variables in Lander, *op. cit.*

TABLE 4

MATRIXES L, M, AND N; EXAMPLES OF THE TIPPING EFFECT

($r_{yi} = .50$; $N = 100$)

| Matrix | Correlations and Inverse of Correlations between Independent Variables | | | | $b_{yi}$ | $s_b$ | $t$ | $p <$ |
|---|---|---|---|---|---|---|---|---|
| L: Correlations... | ... | .89 | .71 | .71 | .146 | .195 | .75 | N.S. |
| | .89 | ... | .71 | .71 | .146 | .195 | .75 | N.S. |
| | .71 | .71 | ... | .86 | .156 | .176 | .89 | N.S. |
| | .71 | .71 | .86 | ... | .156 | .176 | .89 | N.S. |
| L: Inverse..... | 5.166 | −3.925 | − .474 | − .474 | ..... | ..... | ..... | ..... |
| | −3.925 | 5.166 | − .474 | − .474 | ..... | ..... | ..... | ..... |
| | − .474 | − .474 | 4.202 | −2.941 | ..... | ..... | ..... | ..... |
| | − .474 | − .474 | −2.941 | 4.202 | ..... | ..... | ..... | ..... |
| L: Detailed inverse*..... | $+\frac{1}{(.4400)^2}$ | $-\frac{.760}{(.4400)(.4400)}$ | $-\frac{.102}{(.4400)(.4879)}$ | $-\frac{.102}{(.4400)(.4879)}$ | | | | |
| | $-\frac{.760}{.1936}$ | $+\frac{1}{(.4400)^2}$ | $-\frac{.102}{(.4400)(.4879)}$ | $-\frac{.102}{(.4400)(.4879)}$ | | | | |
| | $-\frac{.102}{.2147}$ | $-\frac{.102}{.2147}$ | $+\frac{1}{(.4879)^2}$ | $-\frac{.700}{(.4879)(.4879)}$ | | | | |
| | $-\frac{.102}{.2147}$ | $-\frac{.102}{.2147}$ | $-\frac{.700}{.2380}$ | $+\frac{1}{(.4879)^2}$ | | | | |
| M: Correlations.. | ... | .89 | .73 | .73 | .116 | .199 | .58 | N.S. |
| | .89 | ... | .71 | .71 | .176 | .192 | .91 | N.S. |
| | .73 | .71 | ... | .86 | .156 | .176 | .89 | N.S. |
| | .73 | .71 | .86 | ... | .156 | .176 | .89 | N.S. |
| M: Inverse..... | 5.397 | −3.920 | − .622 | − .622 | ..... | ..... | ..... | ..... |
| | −3.920 | 5.032 | − .382 | − .382 | ..... | ..... | ..... | ..... |
| | − .622 | − .382 | 4.230 | −2.913 | ..... | ..... | ..... | ..... |
| | − .622 | − .382 | −2.913 | 4.230 | ..... | ..... | ..... | ..... |
| M: Detailed inverse*..... | $+\frac{1}{(.4305)^2}$ | $-\frac{.752}{(.4305)(.4458)}$ | $-\frac{.130}{(.4305)(.4862)}$ | $-\frac{.130}{(.4305)(.4862)}$ | | | | |
| | $-\frac{.752}{.1919}$ | $+\frac{1}{(.4458)^2}$ | $-\frac{.083}{(.4458)(.4862)}$ | $-\frac{.083}{(.4458)(.4862)}$ | | | | |
| | $-\frac{.130}{.2093}$ | $-\frac{.083}{.2167}$ | $+\frac{1}{(.4862)^2}$ | $-\frac{.689}{(.4862)(.4862)}$ | | | | |
| | $-\frac{.130}{.2093}$ | $-\frac{.083}{.2167}$ | $-\frac{.689}{.2364}$ | $+\frac{1}{(.4862)^2}$ | | | | |

* Denominators below the main diagonal show products of numbers in denominators above the diagonal.

TABLE 4—*Continued*

| Matrix | Correlations and Inverse of Correlations between Independent Variables | | | | | $b_{yi}$ | $s_{b_{yi}}$ | $t$ | $p <$ |
|---|---|---|---|---|---|---|---|---|---|
| N. Correlations:. . | .... | .80 | 0.00 | 0.00 | 0.00 | .386 | .080 | 4.80 | .001 |
| | .80 | .... | 0.00 | 0.00 | .10 | .142 | .081 | 1.76 | N.S. |
| | 0.00 | 0.00 | .... | ..80 | 0.00 | .278 | .079 | 3.49 | .001 |
| | 0.00 | 0.00 | .80 | .... | 0.00 | .278 | .079 | 3.49 | .001 |
| | 0.00 | .10 | 0.00 | 0.00 | .... | .486 | .048 | 10.02 | .001 |
| N: Inverse..... | 2.829 | −2.286 | 0.000 | 0.000 | .229 | ..... | ..... | ..... | ..... |
| | −2.286 | 2.857 | 0.000 | 0.000 | − .286 | ..... | ..... | ..... | ..... |
| | 0.000 | 0.000 | 2.778 | −2.222 | 0.000 | ..... | ..... | ..... | ..... |
| | 0.000 | 0.000 | −2.222 | 2.778 | 0.000 | ..... | ..... | ..... | ..... |
| | .229 | − .286 | 0.000 | 0.000 | 1.029 | ..... | ..... | ..... | ..... |

N: Detailed inverse*.....

$$+\frac{1}{(.5946)^2}-\frac{.804}{(.5946)(.5916)}-\frac{0.000}{(.5946)(.6000)}-\frac{0.000}{(.5946)(.6000)}-\frac{-.134}{(.5946)(.9858)}$$

$$-\frac{.804}{.3518}+\frac{1}{(.5916)^2}-\frac{0.000}{(.5916)(.6000)}-\frac{0.000}{(.5916)(.6000)}-\frac{.167}{(.5916)(.9858)}$$

$$-\frac{0.000}{.3568}-\frac{0.000}{.3550}+\frac{1}{(.6000)^2}-\frac{.800}{(.6000)(.6000)}-\frac{0.000}{(.6000)(.9858)}$$

$$-\frac{0.000}{.3568}-\frac{0.000}{.3550}-\frac{.800}{.3600}+\frac{1}{(.6000)^2}-\frac{0.000}{(.6000)(.9858)}$$

$$-\frac{-.134}{.5862}-\frac{.167}{.5832}-\frac{0.000}{.5915}-\frac{0.000}{.5915}+\frac{1}{(.9858)^2}$$

we see that, since these elements always have 1.0 in the numerator, the difference is produced by the smaller residual standard deviations in the denominators of the more redundant pair, according to equations (4) and (5).

The remaining within-subset elements, $C_{12}$, $C_{21}$, $C_{34}$, and $C_{43}$, have the same denominators as their main diagonal elements. This leaves the outcome to be determined entirely by the numerators of these remaining elements, the numerators of the main diagonal elements being fixed. These critical numerators consist of the partial correlations between members of the same subset. Since all other things (the between-subset zero-order correlations) are equal, the more redundant subset, with the higher zero-order correlations, has the higher partial correlation. Thus, in this example, the zero-order correlations of .89 and .86 give rise to the second-order partials of .76 and .70, respectively. This partial correlation determines the quantity to be subtracted from the main diagonal elements and thus accounts for the fact that the more redundant subset always has the smaller regression coefficients.

However, the principal demonstration of the present section has to do with the changes from .71 to .73 in some of the correlations in going from matrix L to matrix M. These changes drastically alter the relative sizes of the first two regression coefficients. For matrix M, $b_{y1.234}$ is only two-thirds as large as $b_{y2.134}$, which is also larger than the regression coefficients of the two less redundant variables, $b_{y3.124}$ and $b_{y4.123}$. Once again, the difference between the standard errors of $b_{y1.234}$ and $b_{y2.134}$ is negligible. Although none of the $t$-tests reaches significance in this example, it is evident, with such a substantial difference between the values of $t$ for the first and second variables,

that the results of the test could easily straddle the threshold of significance for a similar problem with less overlap between all of the variables. The addition of a fifth variable, relatively unrelated to the rest, could also lower the standard errors enough to accomplish the same result.

The fact that $b_{y2.134}$ becomes *larger* than $b_{y1.234}$ can be traced to the difference between the partial correlations in the numerators of $C_{13}$, $C_{14}$, $C_{23}$, and $C_{24}$. A potential for reversing the direction of this difference exists as a result of the difference between the residual standard deviations in the denominators of $C_{11}$ and $C_{22}$, the main diagonal elements. This difference between the denominators—caused by the difference between $R^2_{1.234}$ and $R^2_{2.134}$—favors $b_{y1.234}$ being larger than $b_{y2.134}$ instead of vice versa. The $r_{yi}$ control the relative weighting of these two influences, according to equation (2); and if they were not all equal in the present case, the result could be quite different. In our example, the influence of the partial correlations depends on $r_{y3}$ and $r_{y4}$, while that of the SMC's depends upon $r_{y1}$ and $r_{y2}$.

Because the changes to .73 in matrix M are symmetrical with respect to the third and fourth variables, they hardly affect $b_{y3.124}$ and $b_{y4.123}$ at all. It is only the predictive value of the first two variables that is disturbed, and even this remains roughly constant in total ($b_{y1.234}$ plus $b_{y2.134}$). Only its distribution between variables one and two is altered, as though the balance of predictive value were tipped sharply in favor of variable two by the slight change introduced into the correlations.

The third example in this series, matrix N, is intended to illustrate several things. First, it shows how a correlation that is relatively unimportant in appearance (.10 in this case), perhaps far distant in the matrix from the subset it is affecting and hence easy to ignore, can also bring this sharp tipping effect about. In this example, the values of $t$ do straddle the conventional significance point, and the regression coefficient of the second variable is only 37 per cent as large as that of the first variable. Furthermore, the first variable is made to look more important than the third and fourth variables, although its zero-order correlations are all identical to theirs. This greatly enhanced importance of the first variable comes about indirectly through a minor correlation of *another* variable (the second) with which the first just happens to be paired.

Although the regression model assumes the correlations between independent variables to be fixed, this viewpoint is of little comfort if the decision to adopt one rather than another of several quite different interpretations of the data depends heavily upon correlations that are statistically unreliable, substantively inconsequential, or both. For the disturbing correlation of .10 in the present example, and an assumed sample size of 100, the .95 confidence interval ranges from −.10 to .28. In the case of the real data upon which this example is based, the four correlations between the two subsets all had different values, ranging from .68 to .76. With a real sample size of 155, none of the *relevant* comparisons between those correlations is statistically significant, yet their differences would influence strongly the outcome of a regression analysis in which they were the correlations for the independent variables. In a case like this, where all four of the between-subset correlations have different values, to the extent they fail to tip one subset because the $r_{yi}$ of the other subset are small relative to the $r_{yi}$ of the first, they can more easily succeed in tipping the other subset, for which the relevant $r_{yi}$ will then be large.

It should not be assumed, because of the large number of zero correlations that it contains, that the example of matrix N is in any way peculiar. Actually, the zeroes are conservative in their influence. If $r_{13}$, $r_{14}$, $r_{23}$, and $r_{24}$ (with their symmetrical counterparts) were all changed from 0.0 to .60, for example, $b_{y1.2345}$ would equal .553 and $b_{y2.1345}$ would equal .034, simply as the result of the .10 correlation. The second regression coefficient would then be only 6 per cent,

instead of 37 per cent, as large as the first. Clearly, the more highly all of the variables are correlated, the more accentuated the tipping effect.

Of course, the more conspicuous the disturbing correlation, the stronger the tipping effect also—and the more justifiable the interpretations affected by it. Therefore, this aspect of the problem requires no further demonstration, since the difference between .10 and zero is sufficiently compelling as an example of a trivial difference in correlations. But the within-subset correlations, $r_{12}$, $r_{21}$, $r_{34}$, and $r_{43}$, could be other than .80, which is a rather high correlation. What happens to the tipping effect when various other possible values are substituted for the .80 correlations in matrix N is explored, therefore, in Table 5. This table shows that the more highly correlated the variables in question, the more susceptible they are to being tipped. If in all these cases we regard the difference between a correlation of zero and a correlation of .10 as the cause, then it is clear that the magnitude of the effect can range rather widely; throughout most of this range, however, we would regard the effect as disproportionate to the cause.

The reason for this increasing susceptibility lies with the $C_{ij}$ that occupy positions in the inverse matrix corresponding to the positions of the correlations between the subset members themselves—in the present example, $C_{12}$ and $C_{21}$. Other things remaining fixed, as the correlation between the first two variables increases, so will their partial correlation. Since this partial constitutes the numerator of $C_{12}$ and $C_{21}$, this means that the numerators of $C_{12}$ and $C_{21}$ will approach the value of 1.0 in the numerators of $C_{11}$ and $C_{22}$. And since, by definition, variables one and two are in the same subset, they will tend to have similar SMC's, and hence the denominators of all four of these elements will tend to be close in value. Taken together, these factors cause $C_{12}$ and $C_{21}$ to approach $C_{11}$ and $C_{22}$ in absolute value. Subtraction of the former pair from the latter pair thus tends more and more—as the within-set correlations increase—to cancel entirely the contribution of the main diagonal elements to the row sum, leaving this sum to be determined more and more fully by the partial correlations in the numerators of the remaining $C_{ij}$ of these rows.

These remaining $C_{ij}$—in the present example only $C_{15}$ and $C_{25}$, since $C_{13} = C_{14} =$

TABLE 5

ROW SUMS OF INVERSE OF MATRIX N (FOR $b_{y2.1345}$ AND $b_{y1.2345}$) AS A FUNCTION OF THE CORRELATIONS WITHIN SUBSETS, $r_{12}$, $r_{21}$, $r_{34}$, AND $r_{43}$*

| Correlations within Subsets, $r_{12}=r_{21}=r_{34}=r_{43}$ | Sum for $b_{y2.1345}$ | Sum for $b_{y1.2345}$ | Ratio, $\frac{b_{y2.1345}}{b_{y1.2345}}$ | Sum for $b_{y3.1245}$ and for $b_{y4.1235}$† |
|---|---|---|---|---|
| .10 | .816 | .918 | .89 | .909 |
| .20 | .737 | .852 | .86 | .834 |
| .40 | .603 | .759 | .79 | .714 |
| .60 | .476 | .714 | .67 | .625 |
| .80 | .285 | .772 | .37 | .556 |
| .90 | 0.000 | 1.000 | .00 | .526 |
| .99 | −9.091 | 10.000 | −.91 | .502 |

* Row sums are given because of the difficulty of finding a value of $r_{yi}$ that would serve for all of the examples and yet not violate the restrictions on consistency between correlations. The ratios between these sums, of course, are the same as the ratios between actual regression coefficients, given that all of the $r_{yi}$ within each separate problem are equal.

† The sums for these two coefficients are identical, of course.

$C_{23} = C_{24} = 0.0$—also react to the above changes so as to enhance the tipping effect. First of all, we shall look at their denominators.

As the correlation between the first two variables increases, naturally so do the SMC's of both, leading to smaller residual standard deviations in the denominators of all of the elements in their rows and columns. This gives greater weight to the partials in the numerators of the remaining $C_{ij}$, making these $C_{ij}$ larger in size. Any difference between the partials of the first row and the partials of the second row is then reflected in a larger difference between $C_{1j}$ and $C_{2j}$. This effect is especially telling whenever $C_{1j}$ and $C_{2j}$ are opposite in sign, as are $C_{15}$ and $C_{25}$ in the example of matrix N.

Granting that the partial correlations in the numerators of the remaining $C_{ij}$ will seldom be of the first order, the way in which they are affected can best be suggested by examining the most influential first-order partial for $r_{15}$, namely, $r_{15.2}$, and the numerator of the usual formula for calculating this partial, $r_{15} - r_{12}r_{25} = 0.0 - (.80)(.10)$. Clearly, as the within-subset correlation $r_{12}$ increases, this entire expression will become increasingly negative, leading to an increasingly *positive* $C_{15}$ (see eq. [7]). $C_{25}$, on the other hand, will remain negative, because the numerator of $r_{25.1}$, its own corresponding partial, which is $r_{25} - r_{12}r_{15} = .10 - (.80)(0.0)$, remains positive. (The denominators of the pairs of partials we are examining are virtually identical, and so they can be ignored.)

Although tipping is apt to be especially strong when these $C_{ij}$ elements become opposite in sign (note the effect of $C_{15}$ and $C_{25}$ on their row sums in matrix N), the effect does not require this opposition. In the case of the previous example, matrix M, the tipping effect did not depend upon any sign changes. For that matrix, relevant illustrations would be for the numerators of, say, elements $C_{13}$ and $C_{23}$. ($C_{14}$ and $C_{24}$ behave identically to these two.) For the appropriate first-order partials, $r_{13.2}$ and $r_{23.1}$, the respective numerators would appear as follows: $r_{13} - r_{12}r_{23} = .73 - (.89)(.71) = .098$; and $r_{23} - r_{12}r_{13} = .71 - (.89)(.73) = .060$. Although .098 and .060 are both small and positive, and the difference between them is also extremely small, it is their *ratio* that counts. This ratio, of .060 to .098, is .612—almost exactly the same as the ratio of .614 of $C_{23}$ to $C_{13}$ (that is, of −.382 to −.622) in this matrix. (Although our point is that the numerators of these first-order partials almost entirely determine the numerators of $C_{13}$ and $C_{23}$, and that the latter numerators are chiefly responsible for the difference between $C_{13}$ and $C_{23}$, that .612 is as close as it is to .614 is partly due to coincidence. Generally, although close, these values would not be so nearly identical.) The other first-order partials for these $C_{ij}$, $r_{13.4}$ and $r_{23.4}$, are much less influential. The ratio between their numerators, for example, $r_{13} - r_{14}r_{34} = .73 - (.73)(.86) = .102$ and $r_{23} - r_{24}r_{34} = .71 - (.71)(.86) = .099$, is .973. This is so close to 1.0 that it would tend to produce almost identical values for $C_{13}$ and $C_{23}$. These examples indicate how susceptible some analyses are to being strongly influenced by even the most minute changes in the detailed inverse matrix.

The interested reader will also note the additional small assist given to $C_{13}$ and $C_{14}$ of matrix M from having .4305 as a factor in their denominators instead of the .4458 of $C_{23}$ and $C_{24}$. This stems from the higher SMC of variable one, of course.

Returning to matrix N and Table 5, we note that eventually $b_{y2.1345}$ passes through the zero point and becomes negative. When the within-subset correlations are in the range .90–1.0, the absolute values of the two regression coefficients become extremely large. Their algebraic sum, however, remains practically constant. This illuminates the behavior of multiple regression coefficients in curvilinear regression and another mistake that is sometimes made.

Whenever a quadratic (curvilinear) component is introduced as a new independent variable into a correlation matrix, its absolute correlation with the linear component

of the same variable will be very high—usually between .96 and .99. According to the effect of repetitiveness, the two components might be expected to divide the predictive value of either one alone approximately equally between them, with the addition of some small gain from the improvement in fit. This would yield regression coefficients for each component of the variable approximately half the size of the linear component's alone, suggesting that there might be a problem in testing for significance with the usual *t*-test.

However, because of the susceptibility of two such highly correlated predictors to both the tipping effect and the effect of unequal correlation with the dependent variable (which occur despite the near perfect correlation), this neat division of the predictive value into two equal parts is unlikely to come about. Instead, one component will have either slightly lower correlations than the other with the remaining independent variables, or a slightly greater correlation with the dependent variable, or both, and the predictive value of the pair will tip markedly in its favor. Because of the pair's extremely high correlation, and the fact that these two effects are so potent, one of the two regression coefficients is usually passed right through the zero point to assume a high negative value, whereas the other assumes a high positive value. Although the algebraic sum of the two exceeds that of the one only to the extent there is an improvement in fit, both coefficients become much larger in absolute value than the regression coefficient of the linear component alone, leading to the superficial impression that with allowance for curvilinearity the true importance of the predictor has been uncovered.

In this situation, the appropriate statistical test compares the error sum of squares from the curvilinear analysis with the error sum of squares from the noncurvilinear analysis to see whether the former is significantly smaller than the latter. In effect, the predictor's two regression coefficients in the curvilinear analysis are tested at one stroke, rather than individually, as with the usual *t*-test, which would be wrong to apply in this case.[15]

Let us return to matrix N one last time in order to say a word about the effect of negative and zero correlations. It is well known that most social science correlation matrixes are entirely positive, or can be made so by reflecting the appropriate variables. However, exceptions can occur. When they do, the effect of a negative correlation on the multiple regression can be much stronger than its absolute magnitude would lead one to expect. Some zero correlations, seemingly innocent in appearance, have the same result. Take, for example, the zero correlation, $r_{15}$, in matrix N. It occurs between two variables, one and five, that are both correlated in the same direction with the same other variable, namely, variable two. The fact that their joint positive correlations with two are not reflected in a positive correlation between them indicates that they both contain variance that is nega-

[15] Lander appears to have tested these two components in his curvilinear analysis with the *t*-test, as though they were independent regression coefficients, because he reports different levels of significance for the two. The correct test would yield only one level of significance, and it would apply simultaneously to both components. The correct test, incidentally, continues to reject the hypothesis of non-curvilinearity for Lander's data—tentative exploration with the wrong test suggests that it lacks power, so that when it is significant the correct test will be significant too. (We are indebted to Leon J. Gleser for looking into this question of power.) Lander made much of the curvilinear relationship between delinquency and percentage nonwhite. It seemed to him to indicate that delinquency was at a maximum in census tracts that were more heterogeneous, because anomie was greater there. Although our findings concerning the fact of curvilinearity concur with his, this does not imply any indorsement of his interpretation of the shape of the curve or of its cause. Because our delinquency rate data are only close approximations of his (see Gordon, *op. cit.*), and it is difficult in any case to follow his description of what he did, no attempt was made to check this part of his analysis in more detail. For a good discussion of tests of regression coefficients, see Jerome C. R. Li, *Statistical Inference II; the Multiple Regression and Its Ramifications* (Ann Arbor, Mich.: Edwards Bros., Inc., 1964), pp. 185–86.

tively correlated between them, so that the total correlation averages out to zero. Since one and five both correlate with two in the same direction, the only way they can contain mutually negatively correlated variance, without altering the sign of either $r_{12}$ or $r_{25}$, is if it is in a direction orthogonal to that component of two's variance expressed in its correlations with both of them. Since both one and five contain variance that is orthogonal to two, the presence of this variance reduces their correlations with two because it constitutes a part of their total variance that two cannot possibly account for. Consequently, when either is partialled out of the relationship the other has with two, the partials become *larger* than the observed zero-order correlations, $r_{12}$ and $r_{25}$. And when two is partialled out of $r_{15}$, the partial becomes negative, thus giving rise to a positive $C_{15}$. Had $r_{15}$ been negative instead of zero, all of these statements would still apply, only more strongly.

The best way to visualize these relations is to imagine all three variables located with respect to two orthogonal factors. Variable two would lie collinear with one factor, and variables one and five would lie 90 degrees apart from each other (if uncorrelated; if negatively correlated, between 90 and 180 degrees apart), one loading positively and the other negatively on the second factor. Variable two would lie between them, forming acute angles with both. Variables one and five thus cancel the second factor's variance in each other, in effect rotating each toward the first factor and strengthening the correlation of each with two. In this type of situation, one and five play the role of suppressor variables with respect to each other, suppressing the contaminating variance of the second factor.[16]

Because correlations like $r_{15}$, especially when negative, cause other correlations in their row and column to be in effect higher than they appear to be (that is, the relevant first-order partials are higher than the zero-orders), they are thus capable of creating situations of much higher redundancy than the unwary investigator might realize. Therefore, it is necessary to be aware of their possible presence and of what they can do, should one wish to inspect a matrix to see to what extent it might be subject to the effects that have been described in this paper. These suppressor relations could, for example, drastically increase susceptibility to tipping, or cause tipping in a direction opposite to that which one might ordinarily expect on the basis of the zero-order correlations.

In the examples of matrixes L, M, and N, it so happens that the tipping effect works to diminish the apparent importance of the variable with the higher SMC. Consequently, even if the magnitude of the effect seems excessive, at least its direction appears to conform to one's expectations concerning the outcome of partialling, namely, that the most overlapping variable will usually be the most adversely affected. In order to dispel any impression that this is necessarily always the case, we present matrix Z, in Table 6. After inspecting this matrix, it may come as a surprise to some that the first variable possesses the highest SMC. Not only is this true, but it is also the variable that benefits most from the strong tipping effect induced by having set $r_{14}$ equal to zero instead of .10.

## GENERAL COMMENTS ON MULTIPLE REGRESSION

The examples in the preceding section show that small variations among the correlations of a highly related set can create large variations among their regression coefficients. It is hard to imagine any substantive importance that could be attached to such small differences between correlations, yet data analysts are quite apt to attach substantive importance to the larger differences between regression coefficients that they produce. Particularly likely to be misleading are those comparisons between regression coefficients that pit a variable

[16] For a good discussion of suppressor variables and multiple regression, see J. P. Guilford, *Fundamental Statistics in Psychology and Education* (New York: McGraw-Hill Book Co., 1965), pp. 403–8.

winning the sharp internal competition within one such subset against a losing variable from some other similarly competitive subset. When this happens, the zero-order difference between the two can be either exaggerated or minimized. Whichever way it goes, the outcome will be determined by considerations that have little to do with the relation between the two variables. Instead, it will reflect mainly each of their relations to separate sets of other variables—often, quite trivial aspects of those relations.

The generally accepted view that partial regression coefficients express the relative importance of variables has contributed to this uncritical way of looking at them by its not having taken sufficient account of the nature of that importance. Oftentimes, what is actually being compared, if anything, is the local importance of a variable in its own domain with the local importance of another variable belonging to some other domain. It is quite doubtful that sociologists are always seeking such a domain-bound conception of importance whenever they employ this method. Furthermore, as the examples show, even comparisons between variables within the same domain or subset can be extremely sensitive to minor disturbances. Adding to the confusion is the fact that the method itself does not distinguish between comparisons that are sensitive to local contexts and those that reflect more uniformly the total context of all of the independent variables.

For some time now, sociologists have been aware that the values of regression coefficients are not immutable and that they can be greatly affected by changes in the selection of independent variables to be included in an analysis. They have continued, however, to regard regression coefficients as being meaningful within the context of the particular problem in which they appear. Our attempts to describe the inner workings of regression showed four ways in which this assumption could be seriously in error. Even for a given set of variables, there is a sense in which the comparisons being made can be grossly unfair and misleading.

The question naturally arises as to whether there are any conditions under which the effects that we have described do not matter. It is certainly clear that as the level of correlation and the number of variables increase, conditions become more critical. Especially as the number of variables increases, the rationale for the presence of any particular variable is apt to grow more tenuous. Unlike factor analysis, multiple regression is not an all-purpose method for data reduction.[17] If posed in terms of levels of correlation between independent variables, we suspect that the answer to the question is that when the correlations are so low that they do not matter, then partialling itself will not matter and a zero-order analysis would serve as well.

In a more important form, the same question is raised again by the many examples of

TABLE 6

MATRIX Z

($r_{yi}$ = .50; $N$ = 100)

| Variable | Correlations between Independent Variables | | | | SMC | $b_{yi}$ | $s_{b_{yi}}$ | $t$ | $p <$ |
|---|---|---|---|---|---|---|---|---|---|
| 1..... | ... | .80 | .80 | .00 | .719 | .34 | .14 | 2.5 | .02 |
| 2..... | .80 | ... | .80 | .10 | .714 | .10 | .13 | 0.7 | N.S. |
| 3..... | .80 | .80 | ... | .10 | .714 | .10 | .13 | 0.7 | N.S. |
| 4..... | .00 | .10 | .10 | ... | .038 | .48 | .07 | 6.6 | .001 |

[17] Analyses such as Chilton's, for example, which employs eighteen independent variables and uses highly correlated census tract data, are almost certainly of no practical value (see Chilton, *op. cit.*, p. 80).

regression having been successfully applied in other fields, such as economics. Apparently, the manner in which these fields employ regression must differ, in some fundamental way, from its less sophisticated use in sociology. Much of this success is accounted for by studies that employ regression for predictive purposes only. Although prediction often entails making comparisons between regression coefficients, the criticisms set forth here do not apply in their case. This is because those comparisons are intended to achieve pragmatic rather than theoretical objectives. Their aim is to eliminate superfluous variables rather than to test theoretical hypotheses. However, many other studies plainly do seek a better understanding of the relations between variables. The following examples of this may help us to understand why they do not run afoul of misleading effects.

Cows, acres, and men were employed as independent variables in a study of dairy farm income. The regression coefficients showed them to be important in the order listed. Nonetheless, it is absolutely clear that *no matter what* the rank order of cows in this problem, and *no matter how small* its regression coefficient turned out to be, no one would claim that cows are irrelevant to dairy farm income. One would as soon conceive of a hog farm without hogs. Although men turned out to be the factor of production that was least important in this problem, no one would claim either that men are not in fact essential.[18]

Another study examined various body measurements, together with age, sex, and race, in order to arrive at size standards for children's clothing. Height and girth at hips were found to be most important, with age contributing no net effect. It is part of the charm of this example that the more tautological elements emerged, and rightfully so, at the expense of the fundamentally more causative variable. Even so, no one would argue, on the basis of this finding, that a child's body size is unrelated to his age.[19]

The relatively strong theory that surrounds the zero-order relationships in this pair of examples contributed to their success in two ways. First, it prevented the investigators from mistakenly dismissing some variables as being of no theoretical importance, had they been tempted to do so.[20] Second, it meant that the zero-order relationships were so thoroughly understood that the investigators really did intend to move beyond them into the more intimate analysis of partialling. The decision to operate at this microscopic level means that one is interested in differences between variables no matter how highly correlated and similar they may be.[21]

These observations lead us to conclude that successful work with regression coefficients is characteristically pitched at the finer of the levels of distinctness distinguished earlier in this paper. Consequently, the advantages of strong theory and of understanding, if indeed these are separable, would always be present. This conclusion is consistent with the views of Ezekiel and Fox, who stress over and over the necessity for "careful logical analysis, and the need both for good theoretical knowledge of the field in which the problem lies and for thorough technological knowledge of the elements involved in the particular problem."[22] Furthermore, workers in other fields

[18] Mordecai Ezekiel and Karl A. Fox, *Methods of Correlation and Regression Analysis* (New York: John Wiley & Sons, 1959), p. 181.

[19] *Ibid.*, p. 454. See especially chap. xxv, which contains many other examples.

[20] Lander, for example, dismissed the importance of SES in relation to delinquency, not realizing that the reason his SES variables failed to produce significant regression coefficients was that they were so repetitively and redundantly represented (see n. 3 above). Had the theoretical connection between SES and delinquency been as strong, say, as that between age and size of child, he would have been forced to think twice before writing off the relevance of SES.

[21] Had Lander deliberately aimed at such a fine-grained analysis, it would not have occurred to him to dismiss anything.

[22] *Op. cit.*, p. 458. See also p. 432, where they counsel years of experience with the type of data to

seem to be more aware than sociologists are of the pitfalls in regression analysis. This is reflected in efforts in those fields to develop techniques that go beyond simple examination of the regression coefficients. Thus, with the recent and important exception of Blalock's work, there has been nothing in sociology like Wright's path coefficients in biometrics or Frisch's exhaustive bunch maps in econometrics.[23] Merely processing one's data through a stepwise regression routine guarantees nothing in the way of protection. It may in fact induce a false sense of security. Although a stepwise analysis does examine an important subset of all of the relations between variables, and any additional information is apt to be helpful, of itself it is no substitute for understanding of the depth recommended by Ezekiel and Fox. The advantage of the more sophisticated techniques appears to lie in the production of just this kind of understanding, rather than in some mechanical but ingenious circumvention of the problems connected with ordinary regression coefficients. Through this increased understanding, the investigator is led naturally toward working at the finer level of distinctness, even if this had not been his original intention.[24]

Lest there be some misunderstanding, let us make absolutely clear that we have not been condemning the method of multiple regression in general. There remain many situations in sociology for which regression is an excellent tool of analysis.[25] We do condemn, however, those applications of regression coefficients that seek to determine

---

be investigated. Other, special uses of regression, to which our criticism is not meant to apply, are cited in chap. xxiv. Norman Draper and Harry Smith are equally emphatic about the need for understanding. See chap. viii, which is excellent, in their *Applied Regression Analysis* (New York: John Wiley & Sons, 1966). Their chap. vi is also the best discussion we know of the various stepwise procedures.

[23] Since this paper was first drafted, Otis Dudley Duncan has revived interest in path coefficients. See his "Path Analysis: Sociological Examples," *American Journal of Sociology*, LXXII, No. 1 (1966), 1–16. In addition to Duncan's paper, bibliographic references to Sewall Wright's work may be found in Blalock, *Causal Inferences*, p. 193; and in John W. Tukey, "Causation, Regression, and Path Analysis," *Statistics and Mathematics in Biology*, ed. Oscar Kempthorne and Others (Ames: Iowa State College Press, 1954), chap. iii. On bunch maps, see Ragnar Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems* (Oslo: University Institute of Economics, 1934). For a recent application of bunch maps, see Richard Stone, *The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, 1920–1938* (Cambridge: Cambridge University Press, 1954), Vol. I, esp. chap. xix. Besides Stone, other discussions of Frisch's work and of alternative approaches may be found in Tjalling Koopmans, *Linear Regression Analysis of Economic Time Series* (Netherlands Economic Institute No. 20 [Haarlem: De Erven F. Bohn N. V., 1937], Part II; and Harold T. Davis, *The Analysis of Economic Time Series* (Bloomington, Ind.: Principia Press, 1941), pp. 195–97.

[24] True, had Lander, in the example we have criticized, been able to employ even stepwise analysis, he might have noted that after "owner occupancy" and "substandard housing," all five of the remaining variables together added less than 1 per cent to the explained variance. This would have protected him from undertaking the analysis that he did—but it would also have changed the complexion of the problem and told him nothing concerning the five variables that were excluded. Furthermore, because a stepwise analysis will continue to accept variables as long as they contribute a worthwhile increment to explained variance, it offers no protection against the effects described in this paper whenever the correlations between independent variables are low. Yet, those effects can still occur, albeit in a vastly attenuated form. Even in the example of Lander's data, it would have accepted two variables strongly saturated with SES. Both of them would have to split the predictive value of SES between them, in potential contrast to any additional worthwhile predictor representing some orthogonal domain all by itself.

[25] Many good illustrations of the flexibility of this method, and of applications of the general linear hypothesis, are to be found in Li, *op. cit.*, and in Robert A. Bottenberg and Joe H. Ward, Jr., *Applied Multiple Linear Regression* (Lackland Air Force Base, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, 1963). The latter is especially valuable from the standpoint of instructing students and may be obtained from the U.S. Dept. of Commerce, Clearinghouse for Federal, Scientific and Technical Information, Springfield, Virginia 22151, by referring to the number AD 413 128 and enclosing a check for $2.75 made out to "National Bureau of Standards, CFSTI."

the relative importance of variables in the manner of the examples we have cited. Goldberger's comment on this practice is one of the best we have run across, and yet, we feel, it is worded not strongly enough: "The whole point of *multiple* regression . . . is to *try* to isolate the effects of the individual regressors, by 'controlling' on the others. Still, when orthogonality is absent the concept of the contribution of an individual regressor remains inherently ambiguous."[26] This warning, furthermore, applies not just to regression analysis but to all the known control procedures, including those for categoric data and for experiments.

At best, the foregoing remarks will prove helpful in avoiding technical errors. There is little that can be said, unfortunately, concerning the avoidance of theoretical errors. Even though investigators conscientiously consider what level of distinctiveness would be appropriate for an analysis, the possibility of committing all of the above fallacies will probably remain as an outcome of bad theorizing. Things regarded as similar may not be similar, and things regarded as different may not be different. Understanding means correct understanding.

JOHNS HOPKINS UNIVERSITY

[26] Arthur S. Goldberger, *Econometric Theory* (New York: John Wiley & Sons, 1964), p. 201. We are grateful to Clinton S. Herrick for bringing this passage to our attention.