DOCUMENT RESUME

ED 166 203 TM 007 712

AUTHOR Jovick, Thomas D.
TITLE A Monte Carlo Investigation of Spuriously Inflated Regression Estimates.
PUB DATE Mar 78
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Correlation; *Critical Path Method; Goodness of Fit; Hypothesis Testing; Mathematical Models; *Multiple Regression Analysis; Predictor Variables; Sampling; Simulation; *Suppressor Variables
IDENTIFIERS *Monte Carlo Methods

ABSTRACT

A Monte Carlo simulation was used to ascertain the degree of inflation that can occur in regression estimates when samples contain randomly occurring instances of a pattern among correlations called cooperative suppression. Ten thousand samples of scores on three variables were randomly drawn from a population in which the correlations among the variables were prespecified such that cooperative suppression did not exist. Cooperative suppression occurred in nearly 48% of the samples but the incidence of regression coefficients that were grossly discrepant from the population parameters was rare. The implications for multiple linear regression and a method of causal investigation called path analysis are discussed. (Author/CTM)

A Monte Carlo Investigation of Spuriously Inflated Regression Estimates

Thomas D. Jovick

March 1978

Center for Educational Policy and Management
University of Oregon

Prepared for presentation at the Annual Meeting of the American Educational Research Association, March 27-31, 1978, Toronto, Canada.

Table of Contents

A MONTE CARLO INVESTIGATION OF SPURIOUSLY INFLATED REGRESSION ESTIMATES

Thomas D. Jovick

Introduction

The fields of educational research and evaluation have recently expressed interest in the method of path analysis as a means of determining causal relationships among variables in survey data. Like any other statistical procedure, it possesses properties which have not yet been fully ascertained. In using multiple linear regression as its basic analytical tool, path analysis relies heavily on sample regression estimates to reflect the validity of, and make causal interpretations about, a hypothesized model of causal relationships among variables. Little has been documented on the likelihood of obtaining regression estimates spuriously inflated beyond their population counterparts. Such inflations may occur when the pattern of positive and negative signs of the correlations among the variables produces a statistical phenomenon called cooperative suppression (Cohen and Cohen, 1975).

This study provides documentation bearing on spurious instances of cooperative suppression and the resulting inflation of regression coefficients in three-variable regression equations, and discusses implications of the results for path analysis. It attempts to provide such data by addressing four concerns:

1. How often will the pattern among correlations that is characteristic of cooperative suppression tend to occur in a sample just by chance? Most simply the answer involves a record of the proportion of time one expects to find samples in which r_{12} is negative and r_{Y1} and r_{Y2} are both positive.

In those instances where the pattern does occur:

2. How often will the negative correlation between independent variables be significant, allowing one to infer the presence of genuine cooperative suppression?
3. How much overestimation can be expected in the estimates when the pattern comprising cooperative suppression appears? This would tell how much the inflated estimates in general tend to deviate from their respective population values.
4. What size inflations generally occur by chance and how much of a problem are they? Do researchers have much reason to be concerned about those estimates being grossly inflated and therefore grossly misleading?

Although path analysis may employ equations involving several independent variables, this paper will limit itself to those with only two independent variables. The immediate tie to path models, then, is with those using a series of three-variable relationships.

As an attempt to inform these concerns, this study focused on the random occurrence of the cooperative suppression pattern among correlations in a two-independent-variable regression equation and the ensuing influence on the regression estimates. The analysis involved a two-stage process. First, a Monte Carlo simulation program generated 10,000 samples of 100 "scores" for each variable randomly drawn from a population in which the correlations among the three variables had been pre-specified such that cooperative suppression did not exist. For each sample the program calculated correlations and regression

estimates and recorded their signs and if they were significant for alpha = .05 (one-tailed). The second stage involved inspecting the data generated by the Monte Carlo program. This was done using programs in the Statistical Package for the Social Sciences (SPSS) to examine in more detail particular samples in which cooperative suppression effects occur.

Cooperative Suppression and Its Effects on Regression Estimates

Path analysis has been used quite extensively in cross-sectional sociological research as a means of testing theories. Although it cannot prove causality, it does purport to determine whether a pattern of intercorrelations among a set of variables can be meaningfully explained by a particular theoretical formulation about ordered relationships. The causal validity of the model itself rests primarily upon substantive empirical and conceptual considerations (Kerlinger and Pedhazur, 1973; Amick and Walberg, 1975; Namboodiri, et al, 1975).

The basic inferential tool crucial to the method is multiple linear regression which allows one to examine the magnitudes and directions of effects and their statistical significance while controlling for mutual influences among independent variables. The hypothesized causal model itself can be represented by a set of multiple linear regression equations.*

*Because variables in the behavioral sciences are often expressed in arbitrary scales, not much substantive information about a path analytic model is conveyed by non-standardized regression weights, which specify that a 1.0-point change in the independent variables causes b points change in the dependent variable. This is because the different scale ranges of the independent variables obscure the importance of different variables relative to one another when the nonstandardized b-weights are used.

For this reason the standardized regression weights, β's or betas, are used as path coefficients to represent the direct effect of independent on dependent variables. Each coefficient estimates the amount of change in standard deviation units of the dependent variable that is produced by a 1-standard deviation change in the respective independent variable (Amick and Walberg, 1975).

It is crucial to realize that although the R^2 and betas for any equation in the model are central to the analysis, their values rely on the correlations among the variables. The way in which correlations of different magnitudes and signs form different patterns of relationships will dictate the character of the regression estimates. A variety of patterns of correlations among dependent and independent variables can exist, each of which has implications for the magnitude of the regression estimates and their substantive interpretation.

Cohen and Cohen (1975) describe the pattern characteristic of cooperative suppression as one of the most attractive for researchers to find.* Its appeal is in the characteristic way the estimates get enhanced beyond what one would expect from the correlations between Y and each independent variable alone. For the two independent variable case, cooperative suppression comprises cases in which the independent variables correlate positively with Y but negatively with each other.

* Cohen and Cohen (1975) describe three "attractive" patterns whose appeal is in the characteristic way the regression estimates get enhanced in magnitude beyond what the correlations between Y and each independent variable would lead one to expect. The patterns come under the general label of suppression and are conveniently identified when each beta weight falls outside the range 0 to r_{Y1}.

Classical suppression occurs when $r_{Y2} = 0$, $r_{Y1} > 0$, and $r_{12} = 0$. Although X_2 is unrelated to Y, using it in the regression equation increases R_2 beyond its value had only X_1 been used. The absolute value of the betas are larger than the simple correlations with Y; in particular, although X_2 is uncorrelated with Y, its beta weight does not equal zero.

In net suppression, although all correlations are positive, X_2 suppresses a portion of the variance in X_1 that is irrelevant to (uncorrelated with) Y and thereby increases R^2 beyond what it would be if only X_1 were used in the equation. The beta weights will fall outside the range 0 to r_{Y1} with the addition that the beta for the suppressor variable will be opposite in sign of its r_{Y1}.

Examples of Cooperative Suppression

Cohen and Cohen (1975) describe an instance arising in personnel selection. A Director of Personnel, as an attempt to establish a means of selecting sales persons from among a pool of applicants, draws a sample of current salespersons and obtains ratings of their overall success in sales performance. Interview data suggest that social aggressiveness and habits and skills in record keeping each constitute a major determinant of sales successes. Measures of these two variables are administered to the sample. Results reveal that the correlation between social aggressiveness (X_1) and sales success (Y) is .29, between record keeping (X_2) and sales success (Y) is .24, and $r_{12} = -.30$ indicating that those high on social aggressiveness tend to be low on record keeping skills.

In this example, high social aggressiveness tends to go along with high sales success but also with low record keeping skills, which itself is incompatible with high sales success. When a person is high on social aggressiveness, he also tends to register low record keeping skills; that low standing contaminates or suppresses the true relationship (correlation) between social aggressiveness and sales success. In order to determine the unique relationships to the dependent variables one must control for the suppressing influence of the low standing on the other independent variable. When that is done, the relationship between X_1 and Y increases beyond the size of their zero order correlation.

The argument about cooperative suppression applies also to X_2. That is, the relationship between X_2 and Y is also suppressed by the negative

correlation between X_1 and X_2 and similarly becomes enhanced when controlling for X_1.

Randomly Occurring Cooperative Suppression: Hazards to Causal Inference

If cooperative suppression does not exist in the population, a chance still exists that it will be found in samples. Out of a very large number of independent samples drawn from the same population, a number will exhibit cooperative suppression by chance alone. A certain proportion of those instances will contain significant inverse correlations between the independent variables, and lead one to infer that cooperative suppression indeed does exist in the population and that the inflated regression estimates reflect the true relationships. The remainder will contain nonsignificant negative r_{12}'s and not suggest any such inference. In either case, the negative correlations will still occur and act to inflate the estimates even though the parameter for the correlations is zero.

The potential for hazardous inferences in such samples is obvious. The enhanced values of the estimates give the investigator the false impression his independent variables explain a good deal of the variance in the dependent variable and are important and major causal influences because of the large betas they exhibit.

The prospects for path analysis are disturbing for a variety of reasons. Those using the method traditionally appear indifferent to the magnitudes of the total proportions of variance explained by the independent variables in

each regression equation. However, what looks like large path coefficients may in total explain little of the variation in the dependent variable. Furthermore, the analysts tend to accept at face value the relative magnitudes of the path coefficients as a basis for assessing the causal importance or unimportance of direct connections between variables. More disturbing is the fact that some equations in the model may contain enhanced estimates whereas others do not. As a consequence, one part of the model may contain inflated estimates which appear large and substantial in contrast to uninflated estimates in another part of the model. The researcher's interpretations will then reveal "important" relationships among certain variables whose estimates occurred merely as a function of chance fluctuations in sign and magnitude in the correlation.

Cooperative Suppression: Theoretical and Mathematical Considerations

This section will demonstrate how the partialling process in multiple linear regression enhances the R^2 and beta weights to their proper magnitudes. In order to provide a more encompassing perspective on the problem, the discussion will deal with the instances for which the independent variables are uncorrelated, then those for which they are positively correlated, and finally to the focus of the study, those for which they are negatively correlated.

Effect of Cooperative Suppression on the Total Proportion of Variance Explained

When a dependent variable is regressed onto two uncorrelated independent variables, X_1 and X_2, the total proportion of Y variance explained is

simply the sum of the squared correlations of each X_1 with Y or

$$R^2 = r^2_{Y1} + r^2_{Y2} \quad (1)$$

Because X_1 and X_2 are uncorrelated, each squared correlation reflects the unique contribution of the variance in the particular X to the Y variance (Kerlinger and Pedhazur, 1973; Amick and Walberg, 1975).

Normally one finds correlations greater than zero between the independent variables, in which case formula (1) no longer applies. When all correlations are positive, r^2_{Y1} and r^2_{Y2} no longer reflect unique contributions of X_1 and X_2 to the Y variance. Rather, part of the proportion of Y variance explained by one also involves part of the proportion explained by the other. By not partialling out this redundant variance, one risks inferring that each is explaining a greater proportion of Y than it really is.

Essentially, the partialling process is one of extracting from one of the independent variables all information in it contributed by the other independent variable; then, the proportion explained by one independent variable plus the proportion explained by the other after the first has been partialled from it combine to give the total proportion of Y variance explained. Inspection of the formula for the total proportion of variance demonstrates how this happens.

When a correlation exists between the independent variables formula (2) or (3), which are equivalent to each other, must be used (Kerlinger and Pedhazur, 1973), although each reduces to formula (1) when $r_{12} = 0$. Formula (2) is

$$R^2 = r^2_{Y1} + r^2_{Y(2.1)} \quad (2)$$

where r is the squared semi-partial correlation between X_2 and Y controlling for X_1. It gives the proportion of variance added by X_2 explaining Y after taking into account that amount contributed by X_1. Alternately, R^2 is also given by the following formula:

$$R^2 = r^2_{Y2} + r^2_{Y(1.2)} \qquad (3)$$

where r is the squared semi-partial correlation of X_1 with Y controlling for X_2. The formulas for the semi-partials themselves are the key to how and where this shared variance gets extracted.

$$r_{Y(2.1)} = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1 - r^2_{12}}} \qquad (4)$$

$$r_{Y(1.2)} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r^2_{12}}} \qquad (5)$$

In formula (4) for example, when r_{12} is positive, $r_{Y1}r_{12}$ is also positive and gets subtracted from r_{Y2} thus taking into account the sharing of variance in Y due to the correlation between the independent variables (Kerlinger and Pedhazur, 1973).

With cooperative suppression, the presence of a negative correlation lends a peculiar twist to this notion of extracting variance shared between independent variables. When r_{12} is negative, r^2_{Y1} and r^2_{Y2} again no longer reflect unique proportions of Y variance explained by X_1 and X_2 respectively. The negative correlation, however, indicates that the independent variables

are mutually suppressing some of the variance in Y each explains by itself. Rather than extracting shared variance, the partialling process adds this "hidden" part of the unique variance back into each independent variable's zero-order relationship with the dependent variable.

Referring to formula (4) again for illustrative purposes, we see that the influence of the negative correlation takes place in the term $r_{Y1}r_{12}$ of the numerator. When r_{12} is negative, $r_{Y1}r_{12}$ is also negative but its absolute value gets added to r_{Y2}, in effect increasing the correlation between X_2 and Y after controlling for X_1.

Table 1 presents some fictitious data to illustrate the phenomenon characteristic of cooperative suppression. The left half presents the estimates when no correlation exists between the independent variables and serves as a comparison for what happens to them as the correlation becomes increasingly negative.

The table shows that, in all instances, the R^2 increases as a function of taking into account the inverse relationship in the independent variables. As the correlation becomes more negative, obviously a greater portion of the unique relationship between independent and dependent variables is suppressed; when that inverse relationship is taken into account, the R^2 increases more and more above what would be expected if no correlation existed between the independent variables. For example, when r_{Y1} and r_{Y2} equal .2, one doesn't expect the total proportion of variance to be greater than .08. Yet, as a function of a negative r_{12}, the R^2 increases beyond this by a minute amount to .088 when $r_{12} = -.1$ and to a more substantial size of .2 when $r_{12} = -.6$.

TABLE 1 (Continued)

r_{12} = Zero						r_{12} Negative				
r_{Y1}	r_{Y2}	R^2	$\beta_{Y1.2}$	$\beta_{Y2.1}$	Standard Error of Betas	r_{12}	R^2	$\beta_{Y1.2}$	$\beta_{Y2.1}$	Standard Error of Betas
.400	.400	.320	.400	.400	.080	-.100	.360	.440	.440	.080
						-.150	.380	.470	.470	.080
						-.200	.400	.500	.500	.080
						-.250	.430	.530	.530	.080
						-.300	.460	.570	.570	.080
						-.400	.530	.670	.670	.080
.150	.250	.090	.150	.250	.100	-.100	.093	.180	.270	.100
						-.150	.098	.190	.280	.100
						-.200	.104	.210	.290	.100
						-.250	.111	.230	.310	.100
						-.300	.118	.250	.320	.100
						-.400	.136	.300	.370	.100
.200	.400	.200	.200	.400	.090	-.100	.220	.240	.420	.090
						-.150	.230	.270	.440	.090
						-.200	.240	.290	.460	.090
						-.250	.260	.320	.480	.090
						-.300	.270	.350	.510	.090
						-.400	.310	.430	.570	.090

TABLE 1

FICTITIOUS DATA ILLUSTRATING INFLATIONARY CHARACTERISTICS IN COOPERATIVE SUPPRESSION: INFLATION IN REGRESSION ESTIMATES DUE TO NEGATIVE r_{12} COMPARED TO $r_{12} = 0$ HOLDING r_{Y1} AND r_{Y2} CONSTANT

r_{12} = Zero						r_{12} Negative				
r_{Y1}	r_{Y2}	R^2	$\beta_{Y1.2}$	$\beta_{Y2.1}$	Standard Error of Betas	r_{12}	R^2	$\beta_{Y1.2}$	$\beta_{Y2.1}$	Standard Error of Betas
.200	.200	.080	.200	.200	.097	-.100	.088	.220	.220	.097
						-.150	.094	.240	.240	.098
						-.200	.100	.250	.250	.100
						-.250	.110	.270	.270	.100
						-.300	.110	.290	.290	.100
						-.400	.130	.330	.330	.100
						-.600	.200	.500	.500	.120
.300	.300	.180	.300	.300	.090	-.100	.200	.330	.330	.090
						-.150	.210	.350	.350	.090
						-.200	.230	.380	.380	.090
						-.250	.240	.400	.400	.090
						-.300	.260	.430	.430	.090
						-.400	.300	.500	.500	.090
						-.600	.450	.750	.750	.090

The phenomenon apparently occurs whether or not r_{Y1} and r_{Y2} are equal. For example, when r_{Y1} equals .2 and r_{Y2} equals .4, one typically expects to find an R^2 not larger than .2; but as r_{12} becomes increasingly negative (-.1 to -.4), the R^2 deviates increasingly above .2 from .22 to .31. Ghiselli (1964, p. 311) noted the same general trend as part of a discussion about prediction studies.

Effect of Cooperative Suppression on the Beta Weights (Path Coefficients)

Table 1 suggests that, like R^2, the beta weights also increase as a function of partialling out the irrelevant portion of variance in the independent variables. The increase appears to be more dramatic than it is for R^2. For example, in the absence of suppression effects when r_{Y1} and r_{Y2} equal .3, one would normally expect each beta to be no larger than .3; but, as r_{12} becomes increasingly negative from -.1 to -.6, each beta deviates above expectation from .33 to .75. Similarly, when r_{Y1} = .2 and r_{Y2} = .4, one normally expects $\beta_{Y1.2}$ to be no larger than .2 and $\beta_{Y2.1}$ to be no larger than .4; yet, as r_{12} increases in negativity from -.1 to -.4, $\beta_{Y1.2}$ increases from .24 to .43 and $\beta_{Y2.1}$ increases from .42 to .57. The formulas for the beta weights provide some insight as to why this happens.

For the two independent variable case,

$$\beta_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} \quad (6)$$

and

$$\beta_{Y2.1} = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r_{12}^2} \qquad (7)$$

Note their similarity with the formulas for the semi-partial correlation coefficients; particularly, the numerators are identical to those for the corresponding semi-partial.

When r_{12} equals zero, $\beta_{Y1.2}$ reduces to r_{Y1} and $\beta_{Y2.1}$ reduces to r_{12}. Normally, with r_{12} positive, the betas will be less than their respective correlations because the amount of variance shared between the independent variables gets partialled out of each. This is algebraically manifested in the subtraction of $r_{Y2}r_{12}$ from r_{Y1} in (6) and $r_{Y1}r_{12}$ from r_{Y2} in (7). In cooperative suppression the absolute value of the quantities $r_{Y1}r_{12}$ and $r_{Y2}r_{12}$ get added in the numerator, in effect enhancing the magnitude of the betas beyond r_{Y1} and r_{Y2}, respectively. Thereby, the true magnitudes of the relationships between independent and dependent variables are brought to the surface.

Procedure

The initial step was to generate, for three variables (X_1, X_2, and Y) a sample of scores which were randomly selected from a population in which the three correlations, r_{12}, r_{Y1} and r_{Y2}, are of a pre-specified magnitude. Kaiser and Dickman (1962) present a method for randomly generating a sample correlation matrix from a given population matrix.

Invoking their procedure for 10,000 samples of size n = 100* and using the same population matrix of correlations, a Monte Carlo computer program was developed to generate three sampling distributions, one for each correlation. The population parameters for the correlations**, their respective beta weights and the total proportion of variance (R^2) are:

	parameter value
r_{12}	.0000
r_{Y1}	.2000
r_{Y2}	.3500
β_{Y1}	.2000
β_{Y2}	.3500
R^2	.1625

For each sample, regression estimates were calculated and whether or not each value was positive and statistically significant was checked. The magnitude and direction of the differences $\beta_{Y1} - r_{Y1}$ and $\beta_{Y2} - r_{Y2}$ were also recorded to demonstrate the discrepancies between the betas and their respective correlations in each sample. All this information was stored on tape for later access.

---

*This study limited the sample size to 100 to keep it in the realm of sample sizes accessible to educational studies yet sill in the area normally used in path analytic studies. In view of the difficulties in obtaining large samples in most educational research, an N of 100 seemed an appropriate size.

**For population values for r_{Y1}, r_{Y2} and r_{12}, see Appendix A.

The Monte Carlo method therefore provided the basic data from which to select out and analyze the incidence of cooperative suppression effects. Canned programs in the Statistical Package for the Social Sciences (SPSS) were then used to examine this data in more depth.

Results

Characteristics of the Overall Distribution

Across all samples, the averages for each correlation and regression estimate nearly equalled the population parameters and the standard deviations were quite small. Kurtosis and skewness deviated little from zero indicating close approximations in form to the normal distribution. These data are presented in Table 2.

TABLE 2

MEANS, STANDARD DEVIATIONS, KURTOSIS, SKEWNESS, RANGES, MINIMA AND MAXIMA FOR CORRELATIONS AND REGRESSION ESTIMATES ACROSS ALL SAMPLES (N = 10,000)

	r_{12}	r_{y1}	r_{Y2}	β_{Y1}	β_{Y2}	R^2
Mean	.002	.20	.347	.20	.344	.176
Standard Deviation	.099	.097	.087	.091	.086	.066
Kurtosis	-.045	-.103	-.011	-.092	.033	.052
Skewness	-.017	-.111	-.193	-.099	-.172	.357
Range	.759	.748	.629	.728	.626	.453
Minimum	-.383	-.161	.014	-.172	.009	.004
Maximum	.376	.587	.643	.556	.635	.457

Table 3 presents the distributions in terms of the percentage of cases falling into ± 1, 2, and 3 standard deviation intervals around the parameter. The midpoint for each segmented line has been set equal to the appropriate population parameter and standard deviation units marked off on each side. The parameter values appear in parentheses adjacent to the name of each estimate. Consult Table 2 for the appropriate value of the standard deviation for the correlations and regression estimates. Entries are the percentage of samples falling in the specified interval, e.g., 33.8% of the sample r_{12}'s were within 1 standard deviation below the parameter and 13.4% were between -1 and -2 standard deviation.

When correlations are sampled from a population in which the relationship is other than zero, the sampling distribution tends to be skewed, as suggested by Table 3. As a further check on the performance of the Monte Carlo program, distributional properties for r_{Y1} and r_{Y2} were compared with those for the normal distribution with mean equal zero and standard deviation equal 1, that is N(0,1). In order to express such distributions in terms of a normal distribution, first it was necessary to employ the Fisher r to Z transformation of the correlations. The transformation is:

$$Z = 1/2 \log\left(\frac{1 + r_{xy}}{1 - r_{xy}}\right)$$

The resulting sampling distribtuion will be approximately normal with a standard deviation of $\sqrt{\frac{1}{N-3}}$.

The distributions of the transformed r_{Y1} and r_{Y2} values for the 10,000 samples closely approximated the normal distribution.

TABLE 3

PERCENTAGE OF ALL SAMPLES FALLING INTO ± 1, 2 AND 3 STANDARD DEVIATION INTERVALS AROUND THE PARAMETER FOR CORRELATIONS AND REGRESSION ESTIMATES (N = 10,000)

	-3 SD	-2 SD	-1 SD	Parameter Value	+1 SD	+2 SD	+3 SD
r_{12} (0.0)	2.1	13.4	33.8	34.2	14.3	2.0	
r_{Y1} (.20)	2.4	13.5	33.3	34.5	14.3	1.8	
r_{Y2} (.35)	2.6	13.8	32.9	35.4	13.5	1.4	
β_{Y1} (.20)	2.3	13.6	33.3	34.3	14.3	2.0	
β_{Y2} (.35)	2.6	14	33	35	13.3	1.6	
R^2 (.1625)	.3	10.2	33.5	34.8	16.8	3.8	

Incidence of Cooperative Suppression

Out of the 10,000 samples, the pattern characteristic of cooperative suppression occurred in 48% or 4,763 samples. Of these 91.2% (4,343) had a nonsignificant r_{12}, which is 43.4% of all 10,000; thus, about 43 out of 100 cases had the pattern characteristic of cooperative suppression but would not allow one to infer it exists in the population. The remaining 8.8% (420) or 4.2% of the entire 10,000 cases had a significant r_{12} for a one-tailed test with alpha = .05; that is, about 4 out of 100 samples did allow one to infer the presence of cooperative suppression, even though it did not actually exist in the population.

Discrepancies Between Betas and Correlations

When the pattern of cooperative suppression has been found, regardless of whether or not r_{12} is significant, one expects the beta weights to be larger than their respective correlations, and this proved to be true for the present data. The information given below illustrates the general sizes of the discrepancies, even though none actually existed in the population.

In each sample generated during the Monte Carlo routine, the correlations, r_{Y1} and r_{Y2}, were subtracted from their corresponding beta weights, β_{Y1} and β_{Y2}. Table 4 presents the means, standard deviations, ranges, minima and maxima for the resulting differences in estimates for patterns of cooperative suppression only. The mean differences depict the amount of inflation that tends to occur as a function of the randomly occurring negative correlations between the independent variables. In general, the mean differences across all patterns of cooperative suppression are slight.

The range of differences in regression estimates suggests the possibility that an occasional case may yield marked discrepancies, but their standard deviations suggest that the occurrence of maximum differences attained in this study is rare (5 to 7 standard deviations above the mean difference).

TABLE 4

ALL SAMPLES WITH PATTERN OF COOPERATIVE SUPPRESSION: MEANS, STANDARD DEVIATIONS, RANGES, MINIMA AND MAXIMA FOR DIFFERENCES BETWEEN CORRELATIONS AND RESPECTIVE BETAS (N = 4,763)

	$\beta_{Y1} - r_{Y1}$	$\beta_{Y2} - r_{Y2}$
Mean	.027	.016
Standard Deviation	.022	.015
Range	.146	.120
Minimum	.000	.000
Maximum	.146	.120

Table 5 presents the means, standard deviations, ranges, minima and maxima for the differences in estimates for patterns of cooperative suppression with a significant r_{12}. Although their discrepancies were larger than those for cases with a nonsignificant r_{12}, these instances comprised only about 9% of all cooperative suppression patterns. In light of the small standard deviations, this information makes it clear that the odds were low for obtaining a sample with a cooperative suppression pattern in which the r_{12} was significant and in which the estimates deviated markedly from the parameter.

TABLE 5

SAMPLES WITH PATTERN OF COOPERATIVE SUPPRESSION AND SIGNIFICANT r_{12}: MEANS, STANDARD DEVIATIONS, RANGES, MINIMA AND MAXIMA FOR DIFFERENCES BETWEEN CORRELATIONS AND RESPECTIVE BETAS (N = 420)

	$\beta_{Y1}-r_{Y1}$	$\beta_{Y2}-r_{Y2}$
Mean	.072	.044
Standard Deviation	.022	.018
Range	.124	.113
Minimum	.022	.007
Maximum	.146	.120

Magnitudes of the Regression Estimates

In conjunction with the expectation that the betas would be larger than the correlations, the regression estimates in patterns of cooperative suppression were expected to overestimate their parameters. That is, because of the enhancing effect of the negative r_{12}, the mean for the β_{Y1}'s, β_{Y2}'s and the R^2's were expected to be larger than their population values. This was found not to be true in the data of this study. The sets of information below elucidate different aspects of the data concerning the sizes of the sample estimates.

Table 6 presents the means, standard deviations, ranges, minima and maxima for the regression estimates for the 4,763 patterns of cooperative suppression. The means for all three estimates appear to be quite accurate approximations to the parameters.

TABLE 6

ALL SAMPLES WITH PATTERN OF COOPERATIVE SUPPRESSION: MEANS, STANDARD DEVIATIONS, RANGES, MINIMA AND MAXIMA FOR REGRESSION ESTIMATES (N = 4,763)

	β_{Y1}	β_{Y2}	R^2
Mean	.207	.348	.167
Standard Deviation	.084	.087	.065
Range	.490	.588	.412
Minimum	.007	.047	.015
Maximum	.497	.635	.427

Although the minimum and maximum values attained for all correlations and estimates deviate markedly from the population parameters, the incidence of such extreme values was rare. Over all 4,763 instances of cooperative suppression, about 70% of the sample β_{Y1}'s, β_{Y2}'s, and R^2's fell within $\pm$ 1 standard deviation of their parameters and about 95-98% fell within $\pm$ 2 standard deviations of their parameters.

When cast in light of all 10,000 samples, these percentages become reduced by about half. Generally, about 32-34% are within $\pm$ 1 standard deviation of their parameters and 45-47% are within $\pm$ 2 standard deviations of their parameters.

Table 7 presents the means, standard deviations, ranges, minima and maxima for the 420 cases of cooperative suppression with significant r_{12}'s,

The distributional data indicates that about 64-67% of these 420 estimates were within ± 1 standard deviation of their population parameters and 87-97% were within ± 2 standard deviations. This suggests that the estimates for what one would infer to be genuine cooperative suppression situations still tend to fall close to their parameters even though they are subject to inflationary effects due to statistically significant negative correlations between the independent variables.

TABLE 7

SAMPLES WITH PATTERN OF COOPERATIVE SUPPRESSION AND SIGNIFICANT r_{12}: MEANS, STANDARD DEVIATIONS, RANGES, MINIMA AND MAXIMA FOR REGRESSION ESTIMATES (N = 420)

	β_{Y1}	β_{Y2}	R^2
Mean	.214	.351	.152
Standard Deviation	.078	.091	.064
Range	.396	.466	.341
Minimum	.040	.110	.024
Maximum	.436	.576	.365

Out of all 10,000 samples, about 3% had a significant pattern of cooperative suppression with estimates within ± 1 standard deviation of their population parameters and about 4% with estimates within ± 2 standard deviations.

Magnitudes of the Correlation

Briefly, then, the inflated betas in cooperative suppression patterns were larger than their correlation counterparts but still tended to be rela-

tively good estimates of their own parameters. This could mean only that the sample correlations, r_{Y1} and r_{Y2}, in these patterns of cooperative suppression tended to underestimate their respective parameters. In the discussion below, different aspects of the data concerning sizes of the correlations show this to be true.

Table 8 presents the means, standard deviations, ranges, minima and maxima for the correlations for the 4,763 patterns of cooperative suppression. The mean for r_{12} was about 1 standard deviation below the population parameter, but that was expected since all selected cases had no positive r_{12}'s. The means for r_{Y1} and r_{Y2} were somewhat lower than their parameters.

TABLE 8

ALL SAMPLES WITH PATTERN OF COOPERATIVE SUPPRESSION: MEANS, STANDARD DEVIATIONS, RANGES, MINIMA AND MAXIMA FOR CORRELATIONS (N = 4,763)

	r_{12}	r_{Y1}	r_{Y2}
Mean	-.077	.181	.331
Standard Deviation	.058	.086	.088
Range	.340	.475	.604
Minimum	-.340	.000	.028
Maximum	.000	.475	.632

Although the minimum and maximum values attained for all correlations deviated markedly from the parameters, the frequency of such extreme values was low. Over the 4,763 cases, about 70% of the negative r_{12}'s were 1 standard deviation below zero, 95% were 2 standard deviations below and 99.8%

were 3 standard deviations below. About 67-70% of the sample r_{Y1}'s and r_{Y2}'s fell within $\pm$ 1 standard deviation of their parameters and about 95-99% fell within $\pm$ 2 standard deviations.

When cast in light of the 10,000 samples, those percentages become reduced by about half. For r_{Y1} and r_{Y2} about 32-34% were within $\pm$ 1 standard deviation of their parameters and 45-57% within $\pm$ 2 standard deviations. More of the r_{Y1}'s and r_{Y2}'s that randomly occurred together with the negative r_{12}'s were below than above their parameters. Because of this slight discrepancy in distributional properties for correlations and betas and the consequent average underestimation by the correlations, the amount of inflation due to the negative r_{12} was, on the average, sufficient to make the averages of the sample betas accurate population approximations.

Table 9 presents the correlation means, standard deviations, ranges, minima and maxima for the 420 cases of cooperative suppression with significant r_{12}'2 when using a one-tailed test and alpha = .05. The means for the correlations are even farther below their parameters than they are for all instances with the cooperative suppression pattern or those with the pattern and non-significant r_{12}'s. This distributional data shows that 60-63% of the r_{Y1}'s and r_{Y2}'s were within $\pm$ 1 standard deviation of their population parameters and 91-98% within $\pm$ 2 standard deviations. As before, these percentages suggest the correlations tended to fall closely to their parameters, but they are somewhat misleading by themselves because a disproportionate percentage of of cases fall at the lower end of the distribution. Although this observation was noted above, it is more marked for these 420 instances.

TABLE 9

SAMPLES WITH PATTERN OF COOPERATIVE SUPPRESSION AND SIGNIFICANT r_{12}: MEANS, STANDARD DEVIATIONS, RANGES, MINIMA AND MAXIMA FOR CORRELATIONS (N = 420)

	r_{12}	r_{Y1}	r_{Y2}
Mean	-.205	.142	.305
Standard Deviation	.035	.080	.092
Range	.174	.393	.464
Minimum	-.340	.003	.070
Maximum	-.166	.396	.534

Nearly 3% of all 10,000 cases had a significant pattern of cooperative suppression with correlations within $\pm$ 1 standard deviation of their parameters and about 4% within $\pm$ 2 standard deviations.

Summary

From a population in which cooperative suppression was absent and r_{Y1} = .2, r_{Y2} =.35 and r_{12} = 0, 10,000 random samples of size 100 were generated for a 2 independent-1 dependent variable system and correlations and regression estimates calculated for each sample. Nearly 48% of the samples yielded the pattern characteristic of cooperative suppresion, i.e. r_{Y1} and r_{Y2} positive in sign r_{12} negative. Only about 9% of these, however, were found to have a statistically significant negative correlation between independent variables for alpha = .05 (one-tailed test). This comprised about 4% of all 10,000 samples and reflects the overall incidence in which one would incorrectly

infer that cooperative suppression exists in this particular population. The remaining 91% of the samples with patterns of cooperative suppression, or 43.3% of all 10,000, would not have allowed such an inference.

Although the betas were larger than their counterpart correlations, the observation of β_{Y1}'s and β_{Y2}'s grossly discrepant from the r_{Y1}'s and r_{Y2}'s respectively was rare. Furthermore, the betas did not, on the average, overestimate their parameters. On the contrary, the means for β_{Y1}, β_{Y2} and R^2 quite accurately approximated their parameters, and the majority of sample values for each tended to cluster close to their parameters.

The means for r_{Y1} and r_{Y2} were lower than their parameters and their distributions were somewhat weighted on the low end. Apparently, a disproportionate number of r_{Y1}'s and r_{Y2}' falling below their parameters occurred in samples with negative r_{12}'s.

The averages for the correlations were not highly discrepant from their parameters nor were the distributions highly skewed. The largest discrepancies and unbalance in the distributions occurred in those samples of cooperative suppression with significant r_{12}'s.

In general, then, there occurred a slight discrepancy in the distributional properties for correlations and betas, the correlations being positively skewed, and an average underestimation of the parameters by the sample correlations. The magnitudes of inflation attributable to the randomly occurring negative r_{12}'s, though slight, were on the average sufficient to make the means for the sample betas accurate approximations of their parameters.

Implications for Path Analysis

The original concern was over the possibility of a path analyst empirically finding a pattern of cooperative suppression among some variables in the model although it did not exist among them in the population. The related causal effects which the analyst would infer to be large would in fact have occurred as the result of inflationary effects attributable to the presence of a randomly occurring negative correlation between the independent variables. Unknowingly, the researcher would conclude that these relationships in the model reflected causal influences much larger than actually existed. The results of this study do have some implications about such ppssibilities but tentatively must be limited to causal models with two independent variable equations and a population in which variables are moderately related and cooperative suppression is nonexistent.

The possibility of finding the pattern is good but the need for concern over grossly inflated estimates appears minimal. If the pattern occurred among a set or sets of variables in the model, the betas would not tend to be grossly exaggerated. This apparently is true regardless of the magnitude of the negative r_{12}. Although large estimates did appear as a function of the randomly occurring negative r_{12}, the possibility of obtaining estimates of such magnitudes was low. Indeed, the majority of values for the regression estimates tended to congregate near their parameter values. Therefore, the data suggest that the path analyst need not be concerned over the inflationary effects on the betas and total proportion of variance explained should a pattern of cooperative suppression unexpectedly appear in the sample data.

Implications for Future Research

To determine the generalizability of the results, two courses of action should be taken. One is to perform a variety of analyses of this same type using a sample size of 100 and two independent variables but with different positive population correlations between independent and dependent variables. Beyond that, studies could also vary the magnitude of a negative population correlation between independent variables. One would then be investigating the incidence of suppression and magnitudes of estimates that ensue when cooperative suppression does exist in the population.

The second is to repeat all those analyses with differing correlations for different sample sizes. In particular, smaller sample sizes would be more meaningful to educational research since researchers are often constrained by the necessity to use units of analysis of which they can obtain only a small number, such as classrooms and schools.

BIBLIOGRAPHY

Amick, D.L. and Walberg, H.J. Introductory Multivariate Analysis for Educational, Psychological, and Social Research. Berkeley, Calif.: McCutchan Publishing Corporation, 1975.

Anderson, J.G. and Evans, F.B. Causal models in educational research: Recursive models. American Educational Research Journal, 1974, 11 (1), 29-39.

Blalock, H.M. Jr. Causal Models in the Social Sciences. Chicago: Aldine, 1971.

Cohen, J. and Cohen, P. Applied Multiple Regression/Correlation Analysis in the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1975.

Conger, A.J. A revised definition for suppressor variables: A guide to their identification and interpretation. Educational and Psychological Measurement, 1974, 34, 35-46.

Conger, A.J. and Jackson, D.N. Suppressor variables, prediction, and the interpretation of psychological relationships. Educational and Psychological Measurement, 1972, 32, 579-599.

Duncan, O.D. Path analysis: Sociological examples. American Journal of Sociology, 1966, 72, 1-16.

Duncan, O.D. Contingencies in constructing causal models. In E.F. Borgatta (Ed.), Sociological Methodology: 1969. San Francisco: Jossey-Bass, 1969, 74-112.

Duncan, O.D. Introduction to Structural Equation Models. New York: Academic Press, 1975.

Ghiselli, E.E. Theory of Psychological Measurement. New York: McGraw-Hill Book Company, 1964.

Goldberger, A.S. and Duncan, O.D. Structural Equation Models in the Social Sciences. New York: Seminar Press, 1973.

Hays, W.L. Statistics. New York: Holt, Rinehart and Winston, 1963.

Heise, D.R. Problems in path analysis and causal inference. In E.F. Borgatta (Ed.), Sociological Methodology: 1969. San Francisco, Calif.: Jossey-Bass, 1970, Chapter 2.

Hoffman, P.J. Generating variables with arbitrary properties. Psychometrika, 1959, 24 (3), 265-267.

Kaiser, H. and Dickman, K. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika, 1962, 27 (2), 179-182.

Kerlinger, P.R. and Pedhazur, E.J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, 1973.

Land, K.C. Principles of path analysis. In E.F. Borgatta (Ed.), Sociological Methodology: 1969. San Francisco: Jossey-Bass, 1969.

Lohnes, P.R. and Colley, W.W. Introduction to Statistical Procedures with Computer Exercises. New York: John Wiley and Sons, Inc., 1968.

Madaus, G.F., Woods, E.M. and Nuttall, R.L. A causal model analysis of Bloom's Taxonomy. American Educational Research Journal, 1973, 10 (4), 253-262.

Namboodiri, N.K., Carter, L.F. and Blalock, H.M. Jr. Applied Multivariate Analysis and Experimental Designs. New York: McGraw-Hill Book Company, 1975.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. and Brent, D.H. Statistical Package for the Social Sciences. New York: McGraw-Hill Book Company, 1975.

Werts, C.E. and Linn, R.L. Path analysis: Psychological examples. Psychological Bulletin, 1970, 74, 193-212.

Wherry, R.J. Sr., Naylor, J.C., Wherry, R.J. Jr. and Fallis, R.F. Generating multiple samples of multivariate data with arbitrary population parameters. Psychometrika, 1965, 30 (3).

Wright, S. The method of path coefficients. Annals of Mathematical Statistics, 1934, 5, 161-215.

APPENDIX A

Population Values for r_{Y1}, r_{Y2} and r_{12}

It is interesting and pertinent to path analysis to look at a population in which some degree of relationship between each independent variable and the dependent variable actually does exist. The use of path analysis presumably deals with a plausibly accurate model in which the relationships have been formulated on the basis of other empirical results in conjunction with theoretical substance. Chances are that some form of the hypothesized relationships do indeed exist in the population.

Other aspects of this study favored the use of population correlations for r_{Y1} and r_{Y2} greater than zero. For N = 100 and alpha = .05 (one-tailed), the critical size for the correlation is .165; most r's generated under a population correlation which equal zero would be considerably smaller than this. The small sizes of a majority of the sample r_{Y1}'s and r_{Y2}'s would have been substantively uninteresting and produced regression estimates whose magnitudes would probably be ignored in a lot of path analytic studies. Therefore, making the population parameters for r_{Y1} and r_{Y2} greater than zero was judged appropriate.

From a purely technical standpoint, sampling from a population in which these two correlations are positive and the correlation between the independent variables is zero increases the chances of randomly obtaining patterns of cooperative suppression. Because the study intended to examine the incidence of cooperative suppression in samples when in fact it did not exist in the

population, the population parameter for r_{12} was not made negative. Two factors primarily determined the choice of parameter values for r_{Y1} and r_{Y2}. In order to obtain a majority of correlations in the neighborhood of those typically found in studies, it was decided not to make the degree of relationship in the population too large. To make the study more interesting, the parameters for r_{Y1} and r_{Y2} were made different from each other. With the above considerations in mind, this investigation used arbitrarily selected population parameters as follows: $r_{12} = 0$, $r_{Y1} = .20$, $r_{Y2} = .35$.

A Monte Carlo Investigation of Spuriously Inflated Regression Estimates

THOMAS D. JOVICK, Center for Educational Policy and Management

This study used a Monte Carlo simulation to ascertain the degree of inflation that can occur in regression estimates when samples contain randomly occurring instances of a pattern among correlations called cooperative suppression. Ten thousand samples of "scores on three variables were randomly drawn from a population in which the correlations among the variables were prespecified such that cooperative suppression did not exist. Cooperative suppression occurred in nearly 48% of the samples but the incidence of regression coefficients grossly discrepant from the population parameters was rare. Discussion centers around the implications for multiple linear regression and a method of causal investigation called path analysis.