DOCUMENT RESUME

ED 144 897 TM 018 179

AUTHOR Thompson, Bruce
TITLE Interpreting Regression Results: Beta Weights and Structure coefficients are Both Important.
PUB DATE 4 Apr 92
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of covariance; •Analysis of variance; Heuristics; Mathematical Models; •Regression (Statistics); •Research Methodology; •Test Interpretation
IDENTIFIERS •Beta Weights; Linear Models; •Structure Coefficients

ABSTRACT

Various realizations have led to less frequent use of the "OVA" methods (analysis of variance--ANOVA--among others) and to more frequent use of general linear model approaches such as regression. However, too few researchers understand all the various coefficients produced in regression. This paper explains these coefficients and their practical use in formulating interpretations of regression results. A small heuristic data set of 20 subjects is used to make the discussion more concrete and accessible. It is argued that sensible interpretation of regression results usually must invoke an examination of both beta weights and structure coefficients. Six tables and two figures illustrate the discussion. Three appendices provide details of the calculations, and there is a 20-item list of references. (Author/SLD)

Betastru.wpl 4/9/92

Interpreting Regression Results:

beta weights and Structure Coefficients are Both Important

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

Paper presented at the annual meeting (session #20.19) of the American Educational Research Association, San Francisco, April 21, 1992.

Abstract

Various realizations have led to less frequent use of OVA methods, and to more frequent use of general linear model approaches such as regression. However, too few researchers understand all the various coefficients produced in regression. The paper explains these coefficients and their practical use in formulating interpretations of regression results. A small heuristic data set is employed to make the discussion more concrete and accessible. It is argued that sensible interpretation of regression results usually must invoke an examination of both beta weights and structure coefficients.

One reason why researchers may be prone to categorizing continuous variables (i.e., converting intervallic scaled variables down to nominal scale) is that some researchers unconsciously and erroneously associate ANOVA (Fisher, 1925) with the power of experimental designs. Researchers often value the ability of experiments to provide information about causality; they know that ANOVA can be useful when independent variables are nominally scaled and dependent variables are intervallic scaled; they then begin to *unconsciously* identify the analysis of ANOVA with design of an experiment.

It is one thing to presume an ANOVA analysis when an experimental design is performed. It something quite different to assume an experimental design was implemented (and that causal inferences can be made) just because an ANOVA analysis is performed. These sorts of illogic, in which design and analysis are confused with each other, are all the more pernicious, because they tend to arise unconsciously and thus are not readily perceived by the researcher (Cohen, 1968).

Humphreys (1978, p. 873) notes that:

> The basic fact is that a measure of individual differences is not an independent variable, and it does not become one by categorizing the scores and treating the categories as if they defined a variable under experimental control in a factorial designed analysis of variance.

Similarly, Humphreys and Fleishman (1974, p. 468) note that categorizing variables in a non-experimental design using an ANOVA analysis "not infrequently produces in both the investigator and his audience the illusion that he has experimental control over the independent variable. Nothing could before wrong."

These sorts of confusion are especially disturbing when the researcher has some independent or predictor variables that are intervallic scaled, and decides to convert them to nominal scale, just to be able to perform some ANOVA analysis. As Cliff (1987, p.
130) notes, the practice of discarding variance on intervallic scaled predictor variables to perform OVA analyses creates problems in almost all cases:

> Such divisions are not infallible; think of the persons near the borders. Some who should be highs are actually classified as lows, and vice versa. In addition, the "barely highs" are classified the same as the "very highs," even though they are different. Therefore, reducing a reliable variable to a dichotomy makes the variable more unreliable, not less.

Nor do enough researchers realize that the practice of discarding variance on an intervallic scaled predictor variables to perform OVA analyses "makes the variable more unreliable, not less" (Cliff, 1987, p. 130), which in turn lessens statistical power against Type II error. Perdhazur (1982, pp. 452-453) makes the point, and explicitly presents the ultimate consequences of bad practice in this vein:

> categorization of attribute variables is all too frequently resorted to in the social sciences••• It is possible that some of the conflicting evidence in the research literature of a given area may be attributed to the practice of categorization of continuous variables… Categorization leads to a loss of information, and consequently to a less sensitive analysis.

It is the IQ dichotomy or trichotomy in the computer, and not the Intervallic scaled IQ data with an SEM of 3 sitting and collecting dust on the shelf, which will be reflected in the ANOVA printout.

These various realizations have led to less frequent use of OVA methods, and to more frequent use of general linear model approaches such as regression (Edgington, 1974; Elmore & Woehlke, 1988; Goodwin & Goodwin, 1985; Willson, 1982) and canonical correlation analysis (Thompson, 1991). However, too few researchers understand all the linkages and uses of the various coefficients (e.g., , part and partial , and bet weights, and structure coefficients) produced in regression.

The present paper has two purposes: (a) to e plain the various coefficients produced in a regression analysis, and (b) to discuss the relative merits of interpreting beta weights as against structure coefficients. Table 1 presents the hypothetical data for 20 subjects that will be employed to make this discussion more concrete. The analysis was performed with the SPSS commands presented in Appendix A; thus the interested reader can readily reproduce or further explore these results.

INSERT TABLE 1 ABOUT HERE.

All three cases employ V1 as the dependent variable. Four different types of cases of regression analyses are presented: use of (a) a single predictor variable (V2); (b) perfectly uncorrelated predictor variables (V2, V3, and V4); (c) correlated predictor variables (V5, V6, and V7) with no suppressor effects; and (d) correlated predictor variables (V5, V6, and VS) with suppressor effects present.

Four Regression Situations and Their Effects on Regression Results

1. Using a Single Predictor Variable (V2)

The simplest regression case involves the use of only a single predictor variable. For example, one might wish to predict height of adults using information about the subjects' heights at two years of age. There are two possible reasons why one might wish to employ egression in this case, or in other cases as well.

Fir t, one might have data on both the predictor and dependent variables for an acceptably large (e.g., 2,000 adults now aged 21) and representative sample of subjects. One might wish to employ their data to derive a system of weighting scores on the predictor variable such that an optimal prediction of the dependent variable is produced. Then the system of weighting the predictor variable might be generalized for use with different persons whom we believe are similar to those from whom we derived our original weighting system, but for whom we do not have or cannot acquire scores on the dependent variable (e.g., children who are now aged 2, for whom the height at age 21 cannot yet be determined with certainty). This application of regression focuses on *prediction.* We are interested in obtaining accurate prediction, but do not care very much as to why the prediction works.

Second, a certain theory might predict that a certain variable should predict a certain dependent variable with a given degree of accuracy. If we have data on both variables for an acceptably large sample that we believe to be representative of some group about which we wish to generalize, then we can employ regression to test our theory. This application of regression focuses on *explanation.* Here we wish to be able to make good predictions, even for persons for whom we already have data on even the dependent variable, but our primary emphasis is on *understanding why* our *prediction works in the way that it works.*

A Venn diagram of data involving height at age 2 and height at age 21 for a large sample of people might look something like the Case A Venn diagram in Figure 1. The overlap of the circles suggests that the predictor variable and the criterion variable overlap considerably, as reflected in the r^2 statistic that evaluates this overlap. Such a result suggests that scores on the predictor variable would do a reasonably good job of predicting scores on the dependent variable.

INSERT FIGURE 1 ABOUT HERE.

The Venn diagram is a representation of the data from a *group* or *aggregate* perspective. It also possible to conceptualize the di ta at an *individual* level, case by case. The individual case perspective requires that the weighting system used in the regression analysis must be made explicit. Conventional regression analysis employs two types of weights: an additive constant ("a") applied to every case and a multiplicative constant ("b") applied to the predictor variable for each case. Thus, the weighting system takes the form of a regression equation:

$$Y \leftarrow\!\!\text{----}\ \hat{Y} = a + b\ (X)$$

For example, it is known that the following system of weights works reasonably well to predict height at age 21 from height at age 2:

$$Y \leftarrow\!\!\text{----}\ \hat{Y} = 0 + 2.0\ (X)$$

Thus, an individual that is 27" tall at age 2 is predicted to have a height of 54" (0 + 2.0 x 27 = 0 + 54 = 54) at age 21.

The regression problem can also be conceptualized using a scattergram plot. The line of best fit to the data points is a graphical representation of the regression equation, i.e., the regression line actually is the regression equation (and vice versa). The "a" weight: is the point on the vertical Y axis that the regression line crosses the Y axis when X is O; this is called the intercept. The "b" weight is the slope (i.e., change in rise change in run) of the regression line, e.g., the line changes in "b" units of Y. for every changes of 1 unit of X (or 2 times "b" units of Y for every 2 units of change in X, etc.).

An alternative form of the prediction equation involves first converting both variables into Z score form (i.e., scores transformed to have a mean of 0 and an of 1.0 via the algorithm

$Z = ((X - \bar{X})/SD_x)$. When all the variables are in Z score form, the "a" weight is still present, but it is always zero. Therefore, the regression equation simplifies to the form:

$$Z_y \longleftarrow \hat{Y} = + \beta (Z_x)$$

Note that the multiplicative weight for this case is always distinguished from the multiplicative weight for the non-standardized scores by referring to the weights for Z scores as β weights (as against "b" weights). It happens that for a two variable regression problem the β weight to predict Z_y with Z_x is the bivariate correlation coefficient between the two variables (of course, so is the β weight to predict Z_x with Z_y, since $Xy_x = Xxy$).

"b" and β weights can readily be transformed back and forth with the equation:

$$\text{"b"} = \beta\,(SDy/SDx) \text{ or } \beta = \text{"b"}\,(SOx/SOy)$$

As the formulas imply, "b" and β will be equal when (a) either is zero or (b) the two variables• standard deviations are equal. Of course, the formulas also imply that "b" and β always have the same signs, since the SDs can't be negative, so they can't influence the signs of the weights.

When two variables are uncorrelated, Xxy = "b" - β. In this case the predictor has no linear predictive value. Since the regression line always yields the optimal prediction from the predictive data in hand, the "a" weight in such a case will always be $\bar{Y}$, and each person's $\hat{Y}_i$ = "a" = Y. Upon reflection, this seams perfectly sensible. If IQ scores and shoe sizes are perfectly

uncorrelated for adults, and you are told the shoe sizes of adults and are asked to predict the IQ score of each person, your best prediction is simply to estimate that each and every person's IQ is 100.

Table 2 presents the bivariate correlation matrix associated with the Table 1 heuristic data. Given these results, the prediction equation would be:

$$Z_Y \longleftarrow \hat{Y}_i = +.0878\ (Z_X)$$

It also happens that regression lines (and all other regression functions) always pass through the means of all variables. Since the means of both V1 and V2 for the Table 1 data are 50, the point where the regression line passes through the Y axis is 50.0, and thus "a" equals 50. Furthermore, since for these data both SDy and SD_X are equal, for these data the "b" multiplicative weight also equals β equals +.0878. These dynamics are illustrated in the Figure plot of the data and the regression line that best fits the data. Note that the regression line is relatively flat, since the correlation coefficient (and "b" and β, for these data) is nearly zero.

INSERT TABLE 2 AND FIGURE 2 ABOUT HERE.

Table 3 presents related concepts from the perspective of the individual scores of the 20 subjects. Since we select the regression equation to yield the best possible prediction of Y for the group as a whole, on the average, then it is no surprise that the mean "e" score is always zero. This is part of an operational definition of a "best fit" position for the regression line.

INSERT TABLE 3 ABOUT HERE.

Since $\hat{Y}$. scores are derived by weighting (with "a" and "b" or with *13* weights) and then summing the weighted values of the "observed" variables, $\hat{Y}$ scores are "synthetic" or "latent" variables. A set of "e'' scores are defined as the Y. scores minus the $\hat{Y}$ scores; "e" scores are also synthetic variables. Thus, a regression analysis always involves k observed variables plus two additional synthetic variables. Indeed, the whole analysis focuses on the synthetic variables.

The sum of squares of the Y scores (.147) (i.e., the explained variance in Y) plus the sum of squares of the "e" scores (18.857) (i.e., the unexplained variance in Y) exactly (within rounding error) equals the sum of squares total (19.000). We can even look at the "e" scores to find the person who most deviates from the regression line (person #16). In Figure 2 the "e" scores are the distance, *always in vertical units of Y (since Y is* what *we* care *about, we focus of the entire analysis on Y units),* of a given score from the regression line. And the sum of squares explained divided by the sum of squares *of* Y tells us the proportion of Y that we can explain with the predictors, i.e., the B^2.

Table 4 makes these and some other important points. As might be expected, since their areas in the Venn diagram by definition never overlap at all, the correlation of the "e" scores and the $\hat{Y}$ scores is always zero. By the same token, the multiple correlation of Y with the predictors as a set (e.g., R1.234) always exactly equals the bivariate between Y and $\hat{Y}$, since $\hat{Y}$ is all the useful part of any and all the predictors with all the useless parts of the predictors deleted.

INSERT TABLE 4 ABOUT HERE.

2. Using Perfectly Uncorrelated Predictor Variables (V2. V3. and V4)

Regression analysis is also relatively straightforward in the case of multiple predictors that are perfectly uncorrelated. This sounds like an improbable occurrence, but in practice happens quite frequently, as when we employ certain kinds of scores from factor analysis (Thompson, 1983) or when we use planned contrasts in a balanced ANOVA model (Thompson, 1985, 1990).

In a sense, the use of a single predictor is a special case of having multiple predictor variables that are uncorrelated with each other, and many of the same dynamics occur. For example, when there is a single predictor, or when multiple predictor variables are perfectly uncorrelated with each, the o f each predictor with the dependent variable is that predictor's individual weight. This is illustrated in the Table 5 results involving the prediction of V1 with perfectly uncorrelated predictors V2, V3, and V4.

INSERT TABLE 5 ABOUT HERE.

Table 5 also presents the structure coefficient (r_5) for each predictor variable. A structure coefficient (Thompson & Borrello, 1985) is the correlation of a predictor with Y, and is very useful in giving us a better understanding of what the synthetic variable, derived by weighting the observed variables, actually is. As Thompson and Borrello (1985) emphasize, a predictor can have a $\mathscr{B}$ weight of zero, but can actually be an exceptional powerful predictor variable. One must always look at both and structure coefficients when evaluating the importance of a predictor.

Table 6 makes clear that something else intriguing happens when the predictors are perfectly uncorrelated, i.e., the sum of the r^2 · s for the predictors (each representing how much of the dependent variable a predictor can explain) will equal the R^2 involving all the predictors, since in this case the predictors do not overlap at all with each other. This is illustrated in Figure 1. Thus, .0077 plus .1440 plus .0471 equals the R^2 of 19.86%.

INSERT TABLE 6 ABOUT HERE.

3. <u>Using Correlated Predictor Variables (V5, V6. and V7) with No Suppressor Effects</u>

Things get appreciably more complicated when the predictors overlap with each other. The $\mathscr{B}$ weights for given predictors no longer equal the r's for the same predictors, as reflected in Table 5. As reflected in Table 6, the r's no longer sum to R^2, i.e., the sum, .5094 does not equal the R^2 of 49.575%. And notice how in Table 5 variable V7 has a near-zero weight (+.082372) and an r_5 of +.6238.

4. <u>Using Correlated Predictor Variables (V5. V6. and V8) with Suppressor Effects Present</u>

However, appreciably more complicated dynamics occur when suppressor effects are present in the data. As defined by Pedhazur (1982, p. 104), "A suppressor variable is a variable that has a zero, or close to zero, correlation with the criterion but is correlated with one or more than one of the predictor variables." Variable VB in variable set V5, V6, and V8 as predictors of Vl involve something of this dynamic, as reflected in the Table 2 correlation coefficients. Notice in Table 6 that the sum of the ^{2} values is .3468, but the B^2 value for these data *is 54.677%, which is larger* than the sum of the ^{2} values!

Suppressor effects are quite difficult to explain in an intuitive manner. Horst (1966) gives an example that is relatively accessible. He describes the prediction of pilot training success during World War II using mechanical, numerical and spatial abilities, each measured with paper and pencil tests. The verbal scores had very low correlations with the dependent variable, but had larger correlations with the other two predictor, since they were all measured with paper and pencil tests, i.e., measurement artifacts inflate correlations among traits measures with similar methods. As Horst (1966, p. 355) noted, "Some verbal ability was necessary in order to understand the instructions and the items used to measure the other three abilities."

Including verbal ability scores in the regression equation in this example actually serves to remove the contaminating influence of the predictor from the other predictors, which effectively increases the R^2 value from what it would be if only mechanical and spatial abilities were used as predictors. The verbal ability variable has negative weights in the equation. As Horst (1966, p. 355) notes, "To include the verbal score with a negative weight served to suppress or subtract irrelevant ability, and to discount the scores of those who did well on the test simply because of their verbal ability rather than because of abilities required for success in pilot training."

This last example makes a very important point: The latent or synthetic variables analyzed in all Parametric methods are always more than the sum of their constituent parts. If we only look at observed variables, such as by only examining a series of bivariate r's, we can easily under or overestimate the actual effects that are embedded within our data. We must use analytic methods that honor the complexities of the reality that we purportedly wish to study--a reality in which variables can interact in all sorts of complex and counterintuitive ways.

beta versus Structure Coefficients

Debate over the relative merit of emphasizing beta weights as against structure coefficients during interpretation has been fairly heated (Harris, 1989, 1992). The position taken here is that the thoughtful researcher should always interpret either (a) both the beta weights and the structure coefficients (b) both the beta weights and the bivariate correlations of the predictors with Y.

It has been noted by Pedhazur (1982, p. 691) that structure coefficients "are simply zero-order correlations of independent variables with the dependent variable divided by a constant, namely, the multiple correlation coefficient. Hence, the zero-order correlations provide the same information." Thus, the structure r's and the predictor-dependent variable r's will lead to identical interpretations, because they are merely expressed in a different metric. Because $r_3 = r_{x \text{ with YHAT}} / R$, structure r's and predictor dependent variable r•s will always have the same sign, since R cannot be negative, and will equal each other only when R=0.0 or R=1.0.

Although the interpretation of predictor-dependent variable correlations will lead to the same conclusions as interpretations of ·s, some researchers have a stylistic preference for structure coefficients. As Thompson and Borrello (1985, p. 208) argue,

> it must be noted that interpretation of only the bivariate correlations seems counterintuitive. It appears inconsistent to first declare interest in an omnibus system of variables and then to consult values that consider the variables taken only two at a time.

The squared predictor-dependent variable correlation coefficients inform the researcher regarding the proportion of Y variance explained by the predictor. Squared structure coefficients inform the researcher regarding the proportion of $\hat{Y}$ (i.e., only the explained portion of Y) variance explained by the predictors.

Some researchers object to interpreting structure coefficients, because they are not affected by the collinearity (i.e., the correlations) among predictor variables. Beta weights, on the other hand, are affected by correlations among the predictors, and therefore may change if these correlations change or if the variables in a study are added or deleted in replications. These are not instrinsic weaknesses.

Since science is about the business of generalizing relationships across subjects, across variables and measures of variables, and across time, in some respects it is desirable that structure coefficients are not impacted by collinearity. On the other hand, when the variables in a study are fixed for the researcher's purposes, then one is less troubled by the impacts of collinearity among a widely accepted and fixed se of predictors. Thus, the utility of statistics varies somewhat from problem to problem or situation to situation.

Other researchers are troubled by the fact that structure r's are inherently bivariate. One response is that all conventional parametric methods are correlational, i.e., are special cases of canonical correlation analysis (Knapp, 1978), and that even a multivariate method such as canonical can be conceptualized as a bivariate statistic (Thompson, 1991). Indeed, R itself is a bivariate statistic, albeit one involving a synthetic variable, since R is the Pearson between Y and $\hat{Y}$. It should also be noted that r_s is really not completely bivariate, in that it is a correlation involving $\hat{Y}$, and $\hat{Y}$ is a synthetic or latent variable involving all the predictors variables.

Interpreting only beta weights is not sufficient, except in the one variable case, since then X = beta and Xs = 1.0 (unless B=0.0). Together, the beta weights and the structure coefficients tell the researcher which case applies as regards the data. Three possibilities exist, as reflected in the Figure 1 diagrams.

Case #1. When the betas of multiple predictors each equal the predictors' respective r's with Y (and each $r_5 = r_{y\ \text{with}\ X}/R$ = beta/R), then the researcher knows that the predictors are uncorrelated. In this case interpreting betas, structure coefficients, or predictor-dependent variable correlations will all lead to the same conclusions regarding the importance of predictor variables.

Case #2. When all predictors have nonzero betas and nonzero structure coefficients (or r's with Y), then predictor variables overlap with each other, i.e., are multicollinear. The R^2 will be less than the sum of the r^2's.

Case #3. When a predictor has, at the extreme, a zero structure Coefficient (and a zero correlation with Y), but a nonzero beta weight, then suppressor effects are present. Only by consulting more than one set of results will one really understand the data.

References

Cliff, N. (1987). Analyzing multivariate data. san Diego: Harcourt Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Edgington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. American Psychologist, , 25-26.

Elmore, P.B., & Woehlke, P.L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 11(9), 19-20.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, England: Oliver and Boyd.

Goodwin, L. O., & Goodwin, W. L. (1985). statistical techniques in AEBJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 1!(2), 5-11.

Harris, R.J. (1989). A canonical cautionary. Multivariate Behavioral Research, , 17-39.

Harris, R.J. (1992, April). structure coefficients versus scoring coefficients as bases for interpreting emergent variables in multiple regression and related technique§. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Horst, (1966). Psychological measurement and prediction. Belmont, CA: Wadsworth.

Humphreys, L.G. (1978). Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. Journal of Educational Psychology, 2Q, 873-876.

Humphreys, L.G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. Journal of Educational Psychology, , 464-472.

Knapp, T. R. (1978). canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, , 410-416.

Pedhazur, E. J. (1982). Multiple regression in behavioral research; Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston.

Thompson, B. (1983, January). The calculation of factor scores: An alternative. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. [Order document #04501 from National Auxiliary Publication Service, P.O. Box 3513, Grand Central Station, NY, NY 10163]

Thompson, B. (1985). Alternate methods for analyzing data from experiments. Journal of Experimental Education, , 50-55.

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. Educational and Psychological Measurement, , 917-928.

Thompson, B. (1990, April). Planned versus unplanned and orthogonal versus non-orthogonal contrasts: The neo-classical perspective. Paper presented at the annual meeting of the American

Educational Research Association, Boston. (ERIC Document Reproduction Service No. ED 318 753)

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, (2), 80-95.

Thompson, B., & Borrello, G.M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, , 203-209.

Willson, v. L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9, 5-10.

Table 1
Heuristic Data for 3 cases

ID	Vl	V2	V3	V4	VS	V6	V7	V8
1	49.553	48.473	51.610	49.338	49.162	49.718	49.488	50.240
2	50.094	48.812	50.537	51.545	50.576	49.640	49.925	51.286
3	50.799	49.152	49.732	49.890	50.386	49.662	49.889	50.641
4	50.778	49.491	49.195	48.786	49.646	50.297	51.399	51.116
5	50.296	49.830	48.927	49.338	50.579	49.924	49.732	49.904
6	51.420	50.170	48.927	50.662	50.598	50.704	50.303	50.223
7	49.582	50.509	49.195	51.214	48.595	49.350	48.549	49.095
8	50.345	50.848	49.732	50.110	49.087	51.979	49.566	48.004
9	49.988	51.188	50.537	48.455	50.386	48.923	49.148	51.652
10	50.860	51.527	51.610	50.662	50.806	50.068	49.481	49.781
11	49.753	50.170	48.927	50.662	49.768	51.384	49.325	48.400
12	50.491	50.509	49.195	51.214	51.681	49.026	50.357	49.841
13	48.415	50.848	49.732	50.110	48.873	49.657	50.294	49.378
14	49.474	51.188	50.537	48.455	51.746	48.945	51.679	50.997
15	49.506	51.527	51.610	50.662	49.755	50.467	50.510	50.224
16	47.166	48.473	51.610	49.338	48.393	49.058	47.365	49.210
17	50.480	48.812	50.537	51.545	50.857	48.217	50.556	51.488
18	51.158	49.152	49.732	49.890	50.760	50.537	50.344	49.275
19	49.067	49.491	49.195	48.786	49.834	50.541	51.022	50.030
20	50.778	49.830	48.927	49.338	48.512	51.904	51.070	49.216
Mean	50.000	50.000	50.000	50.000	50.000	50.000	50.000	50.000
SD	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2
Bivariate correlation Matrix

	Vl	V2	V3	V4	VS	V6	V7
V2	.0878	1.0000					
V3	-.3795	.0000	1.0000				
V4	.2170	.0000	.0000	1.0000			
V5	.4819	.1757	-.0053	.1247	1.0000		
V6	.2903	.1426	-.3929	-.0795	-.3758	1.0000	
V7	.4392	.1525	-.3123	-.1864	.4213	•1671	1.0000
V8	.1740	-.1400	.2691	-.1437	.5089	-.6302	.3542

Table 3
Observed and Synthetic Variable Scores
Predicting V1 with V2

Case	V1	-MV1	dev	devsq	V2	YHAT	-MYHAT	dev	devsq	e	e2
1	49.553	50.0	-0.449	0.201	48.473	49.866	50.0	-0.134	0.018	-0.313	0.098
2	50.094	50.0	0.093	0.009	48.812	49.896	50.0	-0.104	0.011	0.198	0.039
3	50.799	50.0	0.797	0.636	49.152	49.926	50.0	-0.074	0.006	0.873	0.763
4	50.778	50.0	0.776	0.603	49.491	49.955	50.0	-0.045	0.002	0.823	0.677
5	50.296	50.0	0.294	0.087	49.83	49.985	50.0	-0.015	0.000	0.311	0.097
6	51.420	50.0	1.419	2.012	50.17	50.015	50.0	0.015	0.000	1.405	1.974
7	49.582	50.0	-0.419	0.176	50.509	50.045	50.0	0.045	0.002	-0.463	0.214
8	50.345	50.0	0.343	0.118	50.848	50.075	50.0	0.074	0.006	0.270	0.073
9	49.988	50.0	-0.014	0.000	51.188	50.104	50.0	0.104	0.011	-0.116	0.014
10	50.860	50.0	0.858	0.737	51.527	50.134	50.0	0.134	0.018	0.726	0.527
11	49.753	50.0	-0.248	0.062	50.17	50.015	50.0	0.015	0.000	-0.262	0.069
12	50.491	50.0	0.489	0.240	50.509	50.045	50.0	0.045	0.002	0.446	0.199
13	48.415	50.0	-1.587	2.517	50.848	50.075	50.0	0.074	0.006	-1.660	2.754
14	49.474	50.0	-0.528	0.278	51.188	50.104	50.0	0.104	0.011	-0.630	0.398
15	49.506	50.0	-0.495	0.246	51.527	50.134	50.0	0.134	0.018	-0.628	0.395
16	47.166	50.0	-2.836	8.040	48.473	49.866	50.0	-0.134	0.018	-2.700	7.290
17	50.480	50.0	0.478	0.229	48.812	49.896	50.0	-0.104	0.011	0.584	0.341
18	51.158	50.0	1.157	1.337	49.152	49.926	50.0	-0.074	0.006	1.232	1.519
19	49.067	50.0	-0.935	0.873	49.491	49.955	50.0	-0.045	0.002	-0.888	0.789
20	50.778	50.0	0.776	0.603	49.83	49.985	50.0	-0.015	0.000	0.793	0.629
Total	1000.00			19.00	1000.00	1000.00			0.147	0.000	18.857
Mean	50.00				50.00	50.00				0.000	

Table 4
Correlation coefficients Among Two Observed
and Two Synthetic Variables

	V1	YHAT	E	V2
V1	1.0000	.0878	.9961**	.0878
YHAT	.0878	1.0000	.0000	1.0000**
E	.9961**	.0000	1.0000	.0000
V2	.0878	1.0000**	.0000	1.0000

Note. $R_{Y.X} = r_{Y.\hat{Y}}$.

$r_3 = X.\hat{y}$.

$r_{e.\hat{Y}}$ always= 0.

Table 5
Regression Results for Predicting V1
with V1, V2 and V3, or V5, V6 and V7, or V5, V6 and V8

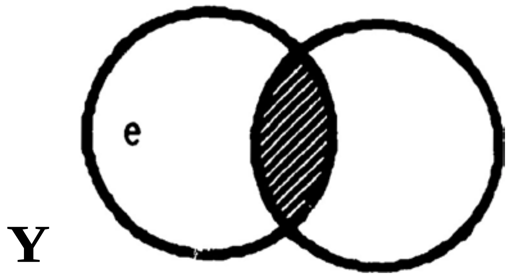
Set	beta	r	partial	structure
V2	0.08786	0.0878	0.0977	*0.1970*
V3	-0.379456	-0.3795	-0.3903	*-0.8511*
V4	0.216955	*0.2170*	0.2356	*0.4866*
V5	0.641788	*0.4819*	0.5791	*0.6844*
V6	0.517727	*0.2903*	0.5287	*0.4123*
V7	0.082372	*0.4392*	0.0865	*0.6238*
V5	0.584123	*0.4819*	0.5971	*0.6517*
V6	0.716874	*0.2903*	0.6359	*0.3926*
V8	0.328547	*0.1740*	0.3310	*0.2354*

regcomp.wk1

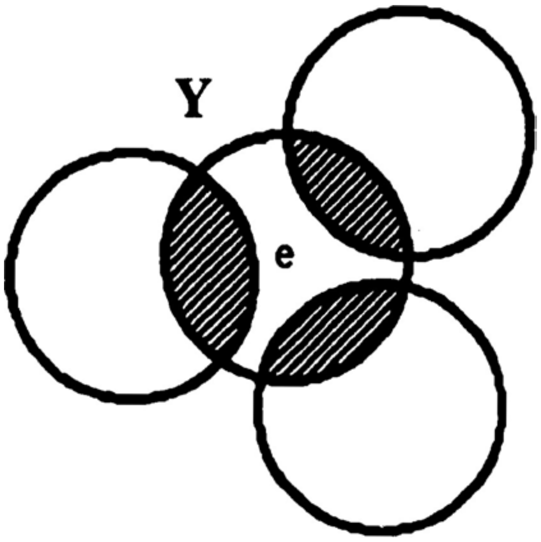
Table 6
Results Associated with Table 1 Data
and the Prediction of V1 with Variable Sets of Size k=3

Predictor/ Sum	$r_{\text{Y with P}}$	$r^2_{\text{Y with P}}$
V2	0.0878	0.0077
V3	-0.3795	0.1440
V4	0.2170	0.0471
Sum		0.1988
VS	0.4819	0.2322
V6	0.2903	0.0843
V7	0.4392	0.1929
sum		0.5094
V5	0.4819	0.2322
V6	0.2903	0.0843
V8	0.1740	0.0303
sum		0.3468

Figure 1

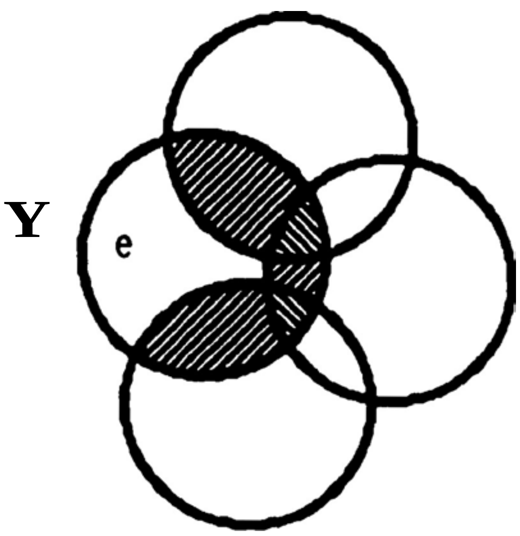


Case #1: One Predictor



Case *2:

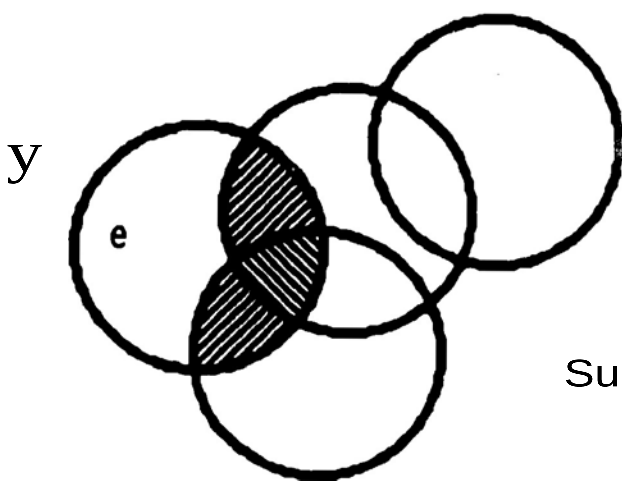
Multiple Uncorrelated Predictors



Case •3:

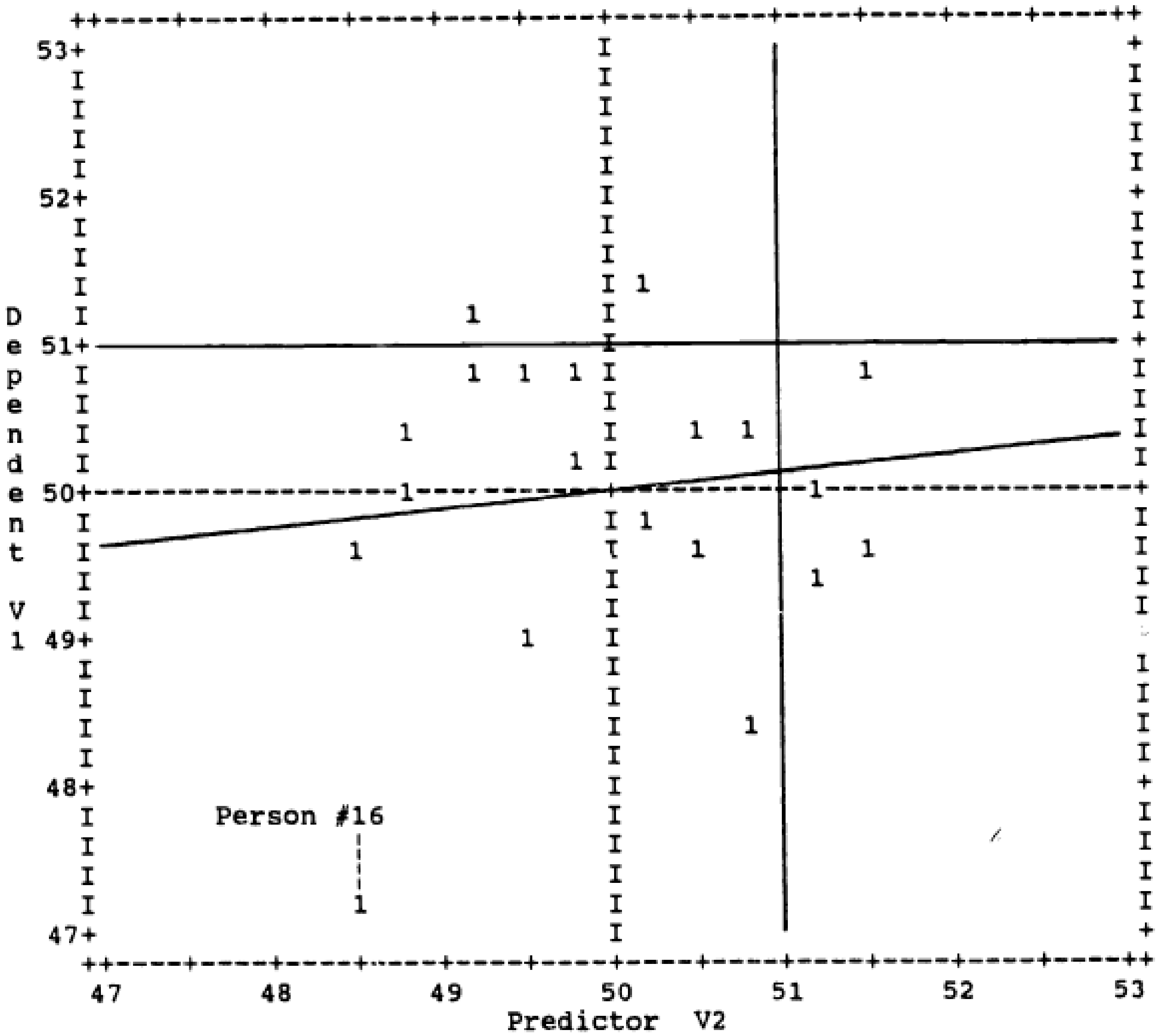
Multiple Correlated Predictors

Suppressor Variable



Case #4:
Suppressor Effects

Figure 2
V1 Correlated With V2



APPENDIX A
SPSS Program to Analyze Table 1 Data

```
TITLE 'CHECK OUTPUT FROM GENNEW.FOR' DATA
LIST FILE•ABC /1
  ID V1 TO VB (F4.0,8F8.3)
LIST VARIABLES n ALL/CASES=500/FORMAT=NUMBERED
SUBTITLE '1. UNCORRELATED PREDICTORS'
REGRESSION VARIABLES=V1 TO V8/DESCRIPTIVE=ALL/DEPENDENT=V1/
  ENTER V2/ENTER V3/ENTER V4
compute yhat=45.607930+(.087844*V2) compute
e=v1-yhat
print formats yhat e (f10.5) list
variables=id v1 yhat e v2
correlations variables=v1 yhat e V2/statistics=all
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/
  ENTER V2/ENTER V4/ENTER VJ
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/
  ENTER V3/ENTER V4/ENTER V2
compute yhat1=53.733930+(.087844*V2)-(.379495*V3)+(.216975*V4)
compute e1=V1-yhat1
correlations variables=V1 TO V4 yhat1 e1/STATISTICS=ALL
PLOT /TITLE 'V1 Correlated With V2'
  /HORIZONTAL='Predictor  V2' REFERENCE (50) MIN(47) MAX(SS)
  /VERTICAL-'Dependent  V1' REFERENCE (50) MIN(47) MAX(SS)
  /PLOT=V1 WITH V2
PARTIAL CORR VARIABLES=V1 WITH V2 BY V3, V4 (2)
PARTIAL CORR VARIABLES=V1 WITH VJ BY V2, V4 (2)
PARTIAL CORR VARIABLES=V1 WITH V4 BY V2, V3 (2)
SUBTITLE '2. PREDICTORS POSITIVELY CORRELATED'
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/
  ENTER V5/ENTER V6/ENTER V7
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/
  ENTER V5/ENTER V7/ENTER V6
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/ ENTER
  V6/ENTER V7/ENTER VS
compute yhat1=-12.097163+(.641816*V5)+(.517747*V6)+(.082382*V7)
compute e1=V1-yhat1
correlations variables=V1 V5 TO v7 yhat1 e1/STATISTICS=ALL
PARTIAL CORR VARIABLES=V1 WITH VS BY V6, V7 (2)
PARTIAL CORR VARIABLES=l V1 WITH V6 BY VS, V7 (2)
PARTIAL CORR VARIABLES=V1 WI H V7 BY VS, V6 (2)
SUBTITLE '3. SUPPRESSOR VARIABLE EFFECTS'
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/ ENTER
  V5/ENTER V6/ENTER VB
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/
  ENTER V5/ENTER V8/ENTER V6
REGRESSION VARIABLES=V1 TO VB/DESCRIPTIVE=ALL/DEPENDENT=V1/ ENTER
  V6/ENTER VB/ENTER VS
compute yhat1=-31.480230+(.584149*V5)+(.716902*V6)+(.328556*VB)
compute e1=V1-yhat1
correlations variables=V1 V5 V6 V8 yhat1 el/STATISTICS=ALL
PARTIAL CORR VARIABLES=Vl WITH V5 BY V6, V8 (2)
PARTIAL CORR VARIABLES=Vl WITH V6 BY VS, V8 (2)
PARTIAL CORR VARIABLES=Vl WITH V8 BY V5, V6 (2)
```

Appendix B
Calculation of a Partial correlation coefficient

```
r12.3                                                                    regcom2.wk1
(r12    -(r13      x  r23))/((1-  r13**2)**.5     x(1 - r23**2)**.5)
(0.087836-{-0.37945 x  0))/((1-  -0.37945**2)**.5 X(1 -   0**2)**.5)
{0.087836-(-0.37945 X   0))/((1-     0.143986)**.5 X(1 -      0)**.5)
(0.087836-       0        )/((    0.856013    )**.5 X(     1    )**.5)
(0.087836                 )/(        0.925209       X           1)
 0.087836                 /                     0.925209
 0.094936

r14.3
(r14    -(r13      x r34))/((1-   r13**')**.5     x(1 - r34**2)**.5)
(0.216955-(-0.37945 X   0))/((1--0.37945**2)**.5 X(1 -   0**2)**.5)
(0.216955-(-0.37945 X   0))/((1-    0.143986)**.5 x(1 -      0)**.5)
(0.2169 5-       0        )/((    0.856013   )**.5 X(     1    )**.5)
(0.21695!5                )/(        0.925209       x           1)
 0.216955                 /                     0.925209
 0.234492

r24.3
(r24    -(r23       x r34))/((1-  r23**2)**.5      x(1 - r34**2)**.5)
(       0-(       0 X   0))/((1-    0**2)**.5     X(1 -   0**2)**.5)
(       0-(       0 X   0))/((1-       0)**.5     X(1 -      0)**.5)
(       O-        0        )/((   1       )* /5    X(     1    )**.5)
(       0                  )/(       1             X           1)
        0                  /                   1
        0


(r12.3  -(r14.3   x  r24.3))/((1  -  14.3**2)**.5  x(1  -r24.3**2)**.5)
(0.094936-(0.234492   x   0))/((1-   0.234492**2)**.5x{1-   0**2)**.5)
(0.094936-(0.234492  x  0))/((1-    0.054986)**.5x(1-        0)**.5)
(0.094936-        o        )/((   0.945013     )**.5x(     1     )**.5)
(0.094936                  )/(       0.972117        X           1)
 0.094936                  /                    0.972117
```

Note. This partial correlation coefficient was derived using algorithms 5.2 and 5.3 from Pedhazur (1982, pp. 102 and 106, respectively). "**2" means raise to the second exponential power, i.e., square. "**.5" means raise to the .5 exponential power, i.e., take the square root.

Appendix C
Calculation of a Semi-Partial (or Part) Correlation Coefficient

:r 1(2.34): = SQRT $r^2$1(2.34) = $R^2$1.234 - $R^2$1.34
SQRT 0.00771 - 0.19877- 0.19106
0.08781
:r 1(3.24): = SQRT $r^2$1(2•34) = $R^2$1.234 - $R^2$1.24
SQRT 0.14399 - 0.19877- 0.05478
0.37946
:r 1(4.23): = SQRT $r^2$1(2•34) = $R^2$1.234 - $R^2$1.23
SQRT 0.04707 = 0.19877- 0.15170
0.21696

:r 1(5.67): = SQRT $r^2$1(5•67) = $R^2$1.567 - $R^2$1.67
SQRT 0.00474 = 0.49575- 0.49101
0.06885
:r 1(6.57): = SQRT $r^2$1(6•57) = $R^2$1.567 - $R^2$1.57
SQRT 0.19567 = 0.49575- 0.30008
0.44235
:r 1(7.56): = SQRT $r^2$1(7.56) = $R^2$1.567 - $R^2$1.56
SQRT 0.25443 = 0.49575- 0.24132
0.50441

:r 1(5.68): = SQRT $r^2$1(5.68) = $R^2$1.568 - $R^2$1.68
SQRT 0.05576 = 0.54677- 0.49101
0.23614
:r 1(6.58): = SQRT $r^2$1(6•58) = $R^2$1.568 - $R^2$1.58
SQRT 0.30769 = 0.54677- 0.23908
0.55470
:r 1(8.56): = SQRT $r^2$1(8•56) - $R^2$1.568 - $R^2$1.56
SQRT 0.25110 = 0.54677- 0.29567
0.50110

Note. These absolute values of part correlations were derived using algorithm 5.19 from Pedhazur (1982, p. 119).