CHAPTER 11

DESCRIPTIVE STATISTICS FOR: MULTIVARIATE DISTRIBUTIONS

INTRODUCTION TO MULTIVARIATE ANALYSES

In this chapter four of the most common multivariate statistical methods for assessing the meaning of variables will be discussed: (1) the procedure known as elaboration is popular when data are presented in a contingency table format. Generally the level of measurement of such variables is nominal or, at best, ordinal; (2) partial correlation is a statistical procedure whereby the effect of a third variable on a bivariate relationship is mathematically removed; (3) multiple regression entails constructing a multiple regression equation in which the value of a dependent variable can be predicted from several independent variables; and (4) multiple correlation enables an analyst to determine how much of the variation in a single dependent variable is explained by a host of independent variables. The latter three techniques are ordinarily reserved for interval-ratio level measurement data.

The examination of bivariate relationships is ordinarily an intermediate phase between univariate and multivariate data analyses. The next step is to ferret out the substantive implications of the outcomes so that some causal inferences can be made. So salient is this implicit cause-effect framework that the terms "causal analysis" and "multivariate analysis" are often used synonymously. To impregnate the need for multivariate analysis let us briefly review what a bivariate relationship, like the Pearsonian r, tells us. Assuming the underlying scores are linearly related, r tells us if an association exists, the magnitude of the association, and the direction of the association. No matter how strong (or even perfect) the correlation is, without the systematic introduction of other variables (sometimes called control variables or test factors) into the analysis we cannot authoritatively decide if the original covariation is real or spurious. In other words, a causal connection between two variables is only one possible explanation among others.

To illustrate the importance and necessity of multivariate analysis suppose the following facts were reported to you: (1) In those regions of Europe with many storks the birth rate is high. The greater the number of "birds" the higher the number of births. Would it be sensible for me to argue that storks are responsible for babies? (2) The amount of property damage resulting from a fire is associated with the number of fire engines ending up at the fire. Could we conclude the fire engines cause the damage? (3) The death rate is much higher among hospitalized patients than among nonhospitalized people. Should we conclude that when sick it would be ill-advised to go to the hospital? The answer to all three queries would be an emphatic no. To say yes would be illogical and at the same time contradict common sense. In a statistical sense we would search for additional variables to help explain the original correlation since the bivariate associations, no matter how convincing, are not sufficient to assume a causal connection between them.[1] There are at least six possible explanations for the correlation between two variables. Let us briefly review these.

Possible Explanations for the Association between Two variables X and Y.[2]

1. The Causal Explanation. It may turnout that Y(the dependent variable) is a function of X(the independent variable), that is, that one variable (X) is the cause of the other (Y). Take, for example, Boyle's law

in thermodynamics: the pressure of a gas kept at a constant temperature varies inversely with the volume of the gas. In the social sciences one is hard pressed to discover similar invariate relationships, associations that are termed determinate. Even those explanations that are of the causal type are usually stochastic in nature, that is, the relationship between X and Y generally holds, or most of the time is witnessed, but rarely is the connection invariate. The relationship between formal education and annual income, social class and political party preference, age and conservatism tend to be stochastic. For simplistic sake the connection between X and Y which takes on this causal form may be diagrammed as:

$$X \longrightarrow Y$$

2. The Joint Result Explanation. Sometimes X and Y are related because both are associated with a third variable. Hence, X and Y are correlated because a third factor is the common cause. A classic illustration is the observed relationship between the number of fire engines (X) and the dollar damage (Y) of the fire. It is not that one causes the other, but that both are jointly affected by the same third variable (Z), namely, the anticipated or actual severity of the fire. In short, if the fire is severe many fire trucks are sent to the site and, at the same time, the potential dollar damage is quite great. Schematically we have the following situation:

$$\begin{tikzcd} & X \\ Z \arrow[ur] \arrow[dr] & \\ & Y \end{tikzcd}$$

3. The Intervening Effects Explanation. Superficially (and statistically) this explanation resembles the former one. It differs from the joint explanation in terms of the temporal placement of the third variable. Rather than the third variable being antecedent to both X and Y, the third variable (Z) comes

between (i.e., *intervenes*) the original two. For example, political behavior (voting) is correlated with social class. Specifically, the higher one's social class the greater one's tendency to vote. A third variable which has proven to alter the original association is political interest. In other words, if political interest is held constant (controlled in the language of statisticians), then the original correlation is dramatically reduced. Substantively this means that there is little difference among members of different social strata when comparisons are made between/among people with the same interest in politics. In time sequencing, political interest comes after social class and before voting behavior making it an intervening variable. In schematic form:

$$X \longrightarrow Z \longrightarrow Y$$

4. *The Interacting Effects Explanation*. With the interacting effects explanation the third variable may antecede or intervene X and Y but unlike the former reasoning, Z has *differential effects* on the X-Y nexus. In other words, only under certain conditions of Z does X have a particular effect. For example, it has been demonstrated that there is a positive correlation between socio-economic status and sexual permissiveness. However, when church attendance is controlled, (i.e., divided into high and low church attenders) the relationship is small or non-existent for the frequent church attenders (Z_2) and positive for the infrequent church attenders (Z_1). In short, attitudes toward sexual permissiveness interact (are differentially influenced) with church frequenting. Diagrammatically:

$$\begin{tikzcd}
 & Z_1 \arrow[dr] & \\
X \arrow[ur] \arrow[dr] & & Y \\
 & Z_2 &
\end{tikzcd}$$

5. *The Chance or Sampling Fluctation Explanation*. It is always possible that the correlation between variables is due to the idiosyncrasies of the sampling process, even when probability samples are drawn. This is a pervasive explanation for the bivariate association but can be minimized with a knowledge of sampling theory.

6. *The Related Observations Explanation*. An axiom underlying statisstical theory is that observations must be independent of each other. If, for example, one studied various features of the income tax system in many different states, the correlations between the states would be undoubtedly high, if not perfect, not because the inter-state comparisons were causally connected but because they were part of a common system, namely, the same federal tax system.[3] In brief, observations must not be related to each other in this fashion but be genuinely independent.

The *purpose* of multivariate analysis is to clarify and elaborate the meaning of bivariate associations. Literally, "multivariate" implies the analysis of many ("multi") variables ("variates"). In practice it refers to a minimum of three variables with which the researcher works. Multivariate analysis is used to determine which of the six possible explanations listed above is plausible. To decide which of the six explanations is most "correct" we use multivariate analysis for the first four possibilities, a knowledge of sampling theory for deciding if the outcomes are due to the nuances of sampling, and assure ourselves that the observations are independent to avoid the sixth possibility.[4] The major *functions of multivariate analysis* are: 1) *control*, 2) *interpretation*, and 3) *prediction*.

The first function, control, is statistical in character. In quasi-experimental designs (e.g., sample surveys) where experimental and control groups are not feasible, statistical control is substituted for experimental control. The second function, interpretation , is achieved by studying the time order of the X, Y, Z variables in order to decide which is antecedent, intervening, and consequent. The third function, prediction, is used when we wish to explain (the variation in) the dependent variable from a host of theoretically salient explanatory (independent) variables.[5] The manner in which each each of these functions is statistically carried out will be discussed.

Control

In laboratory experimental designs control is achieved by physically allocating subjects to two or more groups. Ordinarily each subject is randomly placed in either the control or experimental group. In brief, control is built right into the research design of the investigation. In quasi-experimental designs, it is highly impractical and often impossible to exert this same kind of manipulation. To accomplish its equivalency, control is instituted after the data have been collected during the data analysis stage of the study. In this latter instance control is statistical vis-a-vis physical. Two different traditions of quantitative control exist in the statistical literature, 1) subgroup comparison (sometimes called sub-classification) and 2) partial correlation.[6]

Crosstabulation

Subgroup comparison is typically accomplished through the crosstabulation of variables. Crosstabulation techniques may be conceptually equated to matching procedures in experimental methodologies. Under the latter circumstances variance control is achieved by comparing groups that are presumably the same: Through such matching the groups under examination are made more or less equivalent prior to the introduction of the experimental stimulus (independent variable). In quasi-experimental procedures this "matching" takes place at the analytical phase and entails dividing the subjects into homogeneous subgroups according to the categories of the control variable. Generally, only control variables correlated with the independent and dependent variables are selected as controls and the original bivariate association is reexamined within each of the control variable's subdivisions.

Introduction to Elaboration

This section is devoted to a perspective for multivariate analysis that is particularly appropriate for survey data analysis. The technique is referred to as "the Columbia school", "the Lazarsfeld method","the elaboration model" or "the interpretation method". This varied nomenclature derives from the fact that the goal is to elaborate the empirical relationships among variables in order to interpret that relationships in the manner developed by the late Paul Lazarsfeld (1900-1976) at Columbia University. The purpose of the elaboration model, generally a non-mathematical procedure, is to comprehend the association between two variables, a bivariate relationship, through the simultaneous introduction of theoretically relevant variables. It was developed primarily through the medium of contingency tables, but the logic is both applicable and useful with other statistical techniques.[7]

Steps Involved in Elaboration.

Table elaboration entails a systematic procedure involving three steps:[8]

(1) Two variables, generally an independent (causal) and dependent (effect), are cross-classified. While the number of categories, levels, or

conditions in each may vary, for heuristic purposes we will confine all variables to dichotomies (i.e., two sub-divisions in each). This creates what is called a 2 x 2 table (two categories of the independent variable and two categories of the dependent variable). The analyst assesses the relationship between X and Y by percentaging the table appropriately and/ or computing an appropriate measure of association.[9]

(2) A third variable (Z), called a test factor or control variable, is introduced and the original relationship is decomposed into two partial tables, one for each level, category, or condition of the control variable. This procedure is called stratifying the 2 x 2 table and creates a 2 x 2 x 2 table when dichotomized again. The selection of the third variable is based upon the researcher's theoretical framework and is a logical operation.

(3) The analyst evaluates the effect of Z. Two types of comparisons can be made: (1) Compare the relationships in the partial tables with the original. (2) Compare the relationships in one partial table with the relationship in the other. In evaluating the effect of Z several different patterns can emerge.

Patterns of Elaboration.[10]

Depending on the strength and direction of the relationships revealed in the zero order, conditional, and marginal tables several statistical patterns (called cases here) may occur.[11] In the cases that follow Q will be used as the association coefficient although other correlation coefficients could be used. Yule's Q was chosen for its computational simplicity as well as appropriateness for the subsequent variables' levels of measurement.[12] The notation X, Y, Z refers to the independent, dependent, and control variable, respectively.

Case 1: $Q_{YX} = Q_{YX.Z_1} = Q_{YX.Z_2}$

If this pattern of relationships emerged we would conclude that X and Y are independent of or not associated with or affected by Z. If this configuration were to occur after many different control variables or test factors were introduced we would conclude the X was causally linked to Y. On such occasions the X-Y relationships is upheld whether we observe it with Z (or Z's) varying or held constant. Hence the magnitude of the association in the partial tables ($Q_{YX.Z_1}$ and $Q_{YX.Z_2}$) is equal to the magnitude in the zero-order one (Q_{YX}).

As Figure 11.1 indicates, if the partial associations are the same as the original one, a condition termed replication is manifest, regardless of whether the test factor is antecedent (comes before X) or intervening (comes between X and Y). If the original association was upheld with the introduction of various control variables we would ultimately conclude the initial correlation was genuine and not spurious.

(Figure 11.1 here)

Case 2: $Q_{YX} \neq 0$

$Q_{YX.Z_1} = Q_{YX.Z_2} = 0$

When this pattern is manifest we conclude that X and Y are either joint results of Z or that Z is an intervening variable with regard to X and Y. Without further evidence we cannot tell which of these two possibilities is the more plausible one. The reason is that the two possible explanations differ according to how X and Z are related. In other words, nothing is known about the causal ordering or causal linkages between the variables from the statistical results themselves. To make some sense out of this pattern a theoretical frame-

FIGURE 11.1

THE ELABORATION PARADIGM[13]

Partial Relationships Compared with Original	Test Factor's Relationship to X and Y			
	Antecedent	Intervening	Type Analysis	Notation
Same	replication			$Q_{YX}=Q_{YX \bullet Z_1}=Q_{YX \bullet Z_2}$
Less than or None (0)	explanation	interpretation	M	$Q_{YX}\neq 0$; $Q_{YX \bullet Z_1}=Q_{YX \bullet Z_2}$
Split (different)	specification	prediction	P	$Q_{YX}\neq 0$; $Q_{YX \bullet Z_1}\overset{-}{>}Q_{YX}$ $Q_{YX \bullet Z_2}=0$

work or model is invoked. The time ordering of X and Z provide a clue. Referring to Figure 11.1, when the partial associations compared to the zero-order one are less than or equal to zero two patterns of elaboration can be identified: 1) explanation and 2) interpretation. Which of these two operates is a function of the test factor's time placement (i.e., whether it is antecedent or intervening).

1) explanation. This is the term used to describe a spurious relationship between X and Y since, when X is controlled, the original association is "explained away" or "washed out". Two conditions are required for this: 1) the test factor must be antecedent to both X and Y, and 2) the partial relationships must be zero or substantially less than was found in the original. 2) interpretation. The statistical results for both interpretation and explanation are identical. Our theoretical reasoning allows us to differentiate one from the other. Does Z come before both X and Y? If it does "explanation" is the appropriate term. If, on the other hand, Z comes "in between" or intervenes we have "interpreted" the mechanism through which the relationship occurs or the variable which mediates the X-Y nexus.

Case 3: $Q_{YX} \neq 0$

$$Q_{YX.Z_1} \geq Q_{YX}$$

$$Q_{YX.Z_2} = 0$$

This situation is referred to as interaction or specification. It reveals a direct relationship between X and Y only when Z has a certain value. It not only matters whether Z is held constant or not, but also matters at what level Z is held constant. In short, at one level of Z (Z_1 in example) we observe a relationship between X and Y but at another level (Z_2 in example) there is none. This would be revealed if the Q value in one partial table was approximately 0 but in another was significantly greater than 0.

As Figure 11.1 denotes, regardless of whether or not the test factor is antecedent or intervening, if the partial relations are split (i.e., one partial the same or greater, the other less than or zero when compared with the origiaal) a pattern known as interaction has occurred. Two forms of interaction are revealed: (1) specification when the test factor is antecedent to X and Y, and (2) prediction when the test factor intervenes between X and Y. The term "specification" is self-descriptive. The researcher has specified those circumstances (under that level of Z) under which the relationship holds or does not. The label "prediction" denotes the interaction between X and Y when Z is intervening.

Tabular Examples For Each of The Elaboration Patterns.

Example 1

Suppose we take a random sample of 156 U.S. communities to determine whether race (X) and delinquency rates (Y) are correlated.[14] Assume we provide "good" operational definitions of our variables and ultimately categorize race into two subcategories--white and black--and delinquency rate into "high" and "low". Using the principle of the joint occurrence of attributes as our underlying rationale (Chapter 10)we simultaneously classify the units of analysis into Table 11.1. This is a 2 x 2 frequency and percentage contingency table. Our task is to determine if, how much, and the direction of the association.

(Table 11.1 Here)

As a first step in untangling the bivariate relationship, percentages, appropriately computed, should be entered and compared. Since race is the independent variable and delinquency the dependent variable the column marginal totals, 69

TABLE 11.1 Zero-order or Original Table

CROSS-TABULATION OF DELINQUENCY RATES BY RACE

		Race (X)			
		Black		White	
		n	%	n	%
Delinquency Rates (Y)	High	45	65	30	34
	Low	24	35	57	66
		69	100%	87	100%

E = 31%
Q = .56

TABLE 11.2 First Order Partial or Conditional Tables

CROSS-TABULATION OF DELINQUENCY RATES BY RACE AND SES

		a. High SES (Z_1)				b. Low SES (Z_2)			
		Race (X)				Race (X)			
		Black		White		Black		White	
		n	%	n	%	n	%	n	%
Delinquency Rates (Y)	High	3	14	9	14	42	87.5	21	87.5
	Low	18	86	54	86	6	12.5	3	12.5
		21	100%	63	100%	48	100%	24	100%
		E = 0% Q = 0				E = 0% Q = 0			

and 87, will be the percentage base and the comparison will be made across categories of the independent variable.As can be seen in the table, epsilon (ε), the percentage difference value, is 31. Because it is greater than zero we know an association exists but we do not know how strong it is. To assess the correlation one of the nominal measures of association would be desirable (e.g., Ø, V, Q). Suppose we use Q, substituting the cell frequencies here into the formula in endnote #12. The association in the whole table is computed:

$$Q = \frac{(45)(57) - (30)\ (24)}{(45)\ (57) + (30)\ (24)} = \frac{2565 - 720}{2565 + 720} = \frac{1845}{3285} = .56$$

The next stage is to decide if the association is "real" or perhaps explained away by another variable (recall there are six different ways to account for a relationship between X and Y). Theory should serve as a guide at this juncture. Social scientists are well aware that socio-economic conditions affect many different pathological conditions (like delinquency) and know that both historical and contemporary socio-economic conditions of whites and blacks are far from identical. In line with this rationale socio-economic status is introduced as a control variable with two categories, high and low.[15] Since the control variable is dichotomized we will examine the original association under two different homogeneous circumstances; (1) delinquency rates for whites and blacks from high SES's will be compared (Table 11.2a) and (2) delinquency rates for whites and blacks from low SES's will be compared (Table 11.2b)These two new tables are called *partial tables*;"partial" since they represent only part of the whole) or *conditional tables* ("conditional" since they display separate conditions of the third variable) because when the same cell frequencies in each are added together they will produce the identical frequency that appeared in the original(sometimes called *zero order*) table. For example, summing n_{11}(3) in the high SES table and

n_{11}(42) in the low SES table for blacks yields a total of 45 which is cell frequency n_{11} in the origiaal table.

(Table 11.2 Here)

To determine the appropriate explanation for the original association the same kinds of procedures are applied to the partial tables as were applied to the original table. First, percentage the table appropriately and compare in the opposite direction that the percentages total 100. Second, compute a measure of association to determine the correlation in the entire table. Performing these operations we discover that epsilon equals zero in the "high SES" partial table (Table 11. and zero in the "low SES' partial table (Table 11.2b). Whenever ε equals zero we are informed that no association exists. As a check we go ahead and compute Q. Substituting the table data into the Q formula:

$$Q_H = \frac{(3)(54) - (9)(18)}{(3)(54) + (9)(18)} = \frac{162-162}{162+162} = \frac{0}{324} = 0$$

$$Q_L = \frac{(42)(3) - (21)(6)}{(42)(3) + (21)(6)} = \frac{126-126}{126+126} = \frac{0}{252} = 0$$

How do we interpret this combination of outcomes? Before answering this query directly let's again consider the range of alternatives that could occur when comparing a zero-order table with partial tables. There are three possible configurations that could emerge: 1) the partial associations may be identical or nearly so to the zero order association, 2) the partial associations may vanish, and 3) the partial associations may be different from one another. Suppose alternative number 1 had occured (i.e., $Q = Q_{Z_1} = Q_{Z_2} = .56$). Since the Q values are identical it appears that the control variable has no influence upon the X-Y nexus, a condition called replication. If the introduction of many different control variables reveals the same outcome the conclusion would be that X and Y are causally related. A researcher can never be absolutely sure this is true because the number of control variables is theoretically infinite. However, for the present, if several control variables yielded values identical (or nearly so) to the original bivariate association one could not logically say they had an impact upon the initial relationship.

Suppose alternative number 2 occured which, in fact, did. The original correlation of .56 completely disappeared in both partial tables. What are we to make of this pattern? In answering this query one must decide on the time order of the variables under examination. Technically a distinction is made between antecedent (that variable which comes first), intervening (that variable that comes in-between the two variables), and consequent (that variable whose variation you wish to examine, most often this is the dependent variable) variables. The consequent or dependent variable is delinquency rate (temporally, logically, and theoretically this comes after the other two). Next we must decide, which comes first: race or socio-economic status? Race, of course, is fixed at the time of conception and does not change during one's lifetime while one's socio-economic status can be altered. Hence, race is antecedent and SES intervening. Because the control variable completely accounts for the initial relationship between race and delinquency (no difference exists between X and Y within each category of Z) we take refuge in the intervening effect explanation. Conceptually, race has its influence through the medium of social class. The original relationship cannot be labled spurious because race is still temporally prior to SES. What is critical in explaining the association in situation number two is the temporal sequencing of X, Y, and Z.

TABLE 11.3 Marginal Table

CROSS-TABULATION OF SOCIAL CLASS BY RACE

		Race (X)			
		Black		White	
		n	%	n	%
SES (Z)	High	21	30	63	72
	Low	48	70	24	28
		69	100%	87	100%

E = 42%
Q = -.71

TABLE 11.4 Marginal Table

CROSS-TABULATION OF DELINQUENCY RATES BY SES

		SES (Z)			
		High		Low	
		n	%	n	%
Delinquency Rates (Y)	High	12	14	63	87.5
	Low	72	86	9	12.5
		84	100%	72	100%

E = 73.5%
Q = -.95

To illustrate a *spurious* connection between X and Y suppose we commenced the analysis with SES and delinquency rates cross tabulated. Table 11.4 presents this arrangement. The question is: Does SES affect delinquency? Are the two variables associated? The same principles of contingency table analysis are applied. The percentage difference here is 73.5 and Yule's Q = -.95. SES and delinquency are highly correlated. Is the correlation real? We introduce race as a control (Table 11.2) and compare the association in the original table, Q = -.95, with the association in the two partial tables.[16] The Q's in the partial tables are both 0. What does this configuration mean? The answer depends upon the time sequencing of the variables. Because race is antecedent to SES and delinquency and because the associations disappear in the partial tables we say the original relationship is spurious. It is spurious because the control variable which antecedes the original two completely causes the original association to vanish. Contrast this situation with our illustration of the relationship between race and delinquency with SES controlled. Even though the correlation disappeared the *time ordering* of the X, Y, Z variables is different leading us to a different substantive conclusion. The spurious outcome is called *explanation* whereas the intervening outcome is called *interpretation.*

Suppose the third alternative occurred, that is, the partial tables' values were different from each other. Had this been witnessed the *interaction effects* explanation would have been invoked. This means that the control variable and the antecedent variable interact in such a fashion that under differential conditions of the control variable different outcomes occur. If Z is antecedent to X and this pattern occurred the technical explanation would be called *specification*; if Z is consequent to X the pattern is referred to as *prediction*.

Results and Interpretation for Example 1.

The bivariate association betwene X (race) and Y (delinquency rates) is $\neq 0$ since $Q = .56$ and $\varepsilon = 31$. The association between X and Y in both partial tables is $= 0$ since $Q = 0$ and $\varepsilon = 0$. These results highlight that elaboration case #2 is operating. There is no direct correlation between X and Y but a relationship appears in the zero-order table because both X and Y are associated with Z. The correlation between X (race) and Z (SES) produced a $Q = -.71$ (Table 11.3) while the correlation between Y (delinquency rates) and Z(SES) produced a $Q = -.95$ (Table 11.4). Hence it matters a great deal whether we observe the relationship between X and Y with Z varying (as in the zero-order table) or with Z held constant (as in the conditional tables). In particular, the relationship between X and Y will only appear when Z is allowed to vary. Since race antecedes--comes before--SES it would not be proper to argue that SES causes race. Instead, race affects SES which in turn affects delinquency rates.

(Tables 11.3 & 11.4)

In our three variable examples, even though SES "washed out" the original association (refer to partial tables 11.2a & 11.2b) race is still temporally prior to SES. Hence, SES helps us "interpret" the original association but does not make it spurious.

The Lazarsfeld Accounting Formula.

Paul Lazarsfeld has advanced an equation for helping us summarize and grasp the meaning of many multivariate problems.[17] It reads as follows:

$$\Delta YX = \Delta YX.Z1 + \Delta YX.Z2 + \left[\frac{N}{(N1)\ (N2)}\right] \Delta XZ \Delta YZ$$

This formula tells us that the original association between X and Y (ΔYX) can be accounted for by the conditional or partial associations ($\Delta YX.Z1$ and $\Delta YX.Z2$) plus the marginal associations (ΔXZ and ΔYZ) and a ratio between the total N and the N's in the partial tables $(N/(N_1)$ (N_2).

The formula is presented in terms of deltas (Δ's) which represents the difference between observed and expected cell frequencies (Chapter 10). In a 2 x 2 table all deltas are identical except for their sign. Although the formula is not primarily for computation--it is provided since it facilitates understanding multivariate problems--we will demonstrate its application. In all cases, the substituted delta value will be the one produced in cell n_{11} of the respective tables. Hence,

$$11.82 = 0 + 0 + \left[\frac{156}{(84)(72)}\right] (16.15)(28.38)$$

$$11.82 = 11.82$$

The accounting formula tells us that X and Y are not related to each other except through the fact that each is related to Z. In this situation $\Delta YX.Z1$ and $\Delta YX.Z2$ equal zero and vanish from the equation. However, each variable is related to Z and because of this the marginal tables display a departure from independence (as can be seen by the Δ values of 16.15 and 28.38). Such a situation allows us to conclude that Z and Y are not causally related or that their association is spurious. It might also be said, however, that X and Y are related only to the extent that they both measure the same thing, namely, Z.

Example 2.

Suppose we examine the relationship between the number of fire engines (X) and the amount of fire damage (Y).[18] Computing an association coefficient (Table 11.5) indicates a positive correlation, Q = .54. Substantively, the more trucks that respond the greater the damage. One might hastily assume the trucks themselves caused the damage. However, an antecedent test factor (Z), the size of

TABLE 11.5 Zero Order or Original Table

CROSS-TABULATION OF AMOUNT OF DAMAGE BY NUMBER OF FIRE ENGINES

		Number of Fire Engines (X)			
		One		Two or More	
		n	%	n	%
Amount of Damage (Y)	Under $10,000	1050	70	287	41
	$10,000 or More	450	30	413	59
		1500	100%	700	100%

E = 29
Q = .54

TABLE 11.6 First Order Partial or Conditional Tables

CROSS-TABULATION OF AMOUNT OF DAMAGE BY NUMBER OF FIRE ENGINES AND NUMBER OF ALARMS

		a. One-Alarm Fire (Z_1)				b. Two or More-Alarm Fire (Z_2)			
		Number of Fire Engines (X)				Number of Fire Engines (X)			
		One		Two or More		One		Two or More	
		n	%	n	%	n	%	n	%
Amount of Damage (Y)	Under $10,000	950	95	190	95	100	20	100	20
	$10,000 or More	50	5	10	5	400	80	400	80
		1000	100%	200	100%	500	100%	500	100%
		E = 0 Q = 0				E = 0 Q = 0			

TABLE 11.7 Marginal Table

CROSS-TABULATION OF NUMBER OF ALARMS BY NUMBER OF FIRE ENGINES

		Number of Fire Engines (X)			
		One		Two or More	
		n	%	n	%
Number of Alarms (Z)	One	1000	67	200	29
	Two or More	500	33	500	71
		1500	100%	700	100%

E = 38
Q = .67

TABLE 11.8 Marginal Table

CROSS-TABULATION OF AMOUNT OF DAMAGE BY NUMBER OF ALARMS

		Number of Alarms (Z)			
		One		Two or More	
		n	%	n	%
Amount of Damage (Y)	Under $10,000	1140	95	200	20
	$10,000 or more	60	5	800	80
		1200	100%	1000	100%

E = 75
Q = .97

the fire measured in terms of the number of alarms, explains away the original association since both partials (Tables 11.6a and 11.6b) have Q's = 0. The marginal tables, X with Z (Table 11.7) suggest that the more severe the fire the more fire engines that respond (Q = .67) and the larger the fire (Z) the greater the damage (Y), Q = .97 (Table 11.8).

(Tables 11.5, 11.6, 11.7, and 11.8 Here)

Since the size of the fire is temporally prior to X we say that the original association is causally spurious, that is, we've "explained" the X-Y association.

Example 3.

Our third example begins with a slight association between anomia and residency, Table 11.9, Q = .05.[19] We decide to introduce race (Z) as a test factor since there is reason to believe that race is associated with both residency (X) and anomia (Y). The original table is decomposed into two partial tables, one for each condition of race, Tables 11.10a and 11.10b.

When this is done we observe that among whites, urban dwellers are more likely to experience anomia than rural dwellers, Q = .36. The same is not true among blacks. Here, rural blacks are slightly more likely to experience anomia, Q = -.06. When the pattern of size and/or direction differ in the conditional tables we say we have specified the relationship.

(Tables 11.9, 11.10, 11.11, and 11.12 Here)

Summary

The above exposition is a brief introduction to multivariate analysis using cross tabulation as a surrogate control mechanism because with certain kinds of research designs (i.e, sample survey) the physical allocation of subjects is either impractical or impossible. The patterns of elaboration just discussed are "ideal typical" in character. In actual empirical research the results are generally not so convincing since the relationship between two variables is often

TABLE 11.9

CROSS-TABULATION OF ANOMIA BY RESIDENCY

		Residency (X)			
		Urban		Rural	
		n	%	n	%
Anomia (Y)	High	257	41	116	39
	Low	369	59	184	61
		626	100%	300	100%

E = 2
Q = .05

TABLE 11.10

CROSS-TABULATION OF ANOMIA BY RESIDENCY BY RACE

		a. Whites (Z_1)				b. Blacks (Z_2)			
		Urban		Rural		Urban		Rural	
		n	%	n	%	n	%	n	%
Anomia (Y)	High	70	37	19	21	187	43	97	46
	Low	120	63	70	79	249	57	114	54
		190	100%	89	100%	436	100%	211	100%
		E = 16 Q = .36				E = 3 Q = -.06			

TABLE 11.11

CROSS-TABULATION OF RACE BY RESIDENCY

		Residency (X)			
		Urban		Rural	
		n	%	n	%
Race (Z)	White	190	30	89	30
	Black	436	70	211	70
		626	100%	300	100%

E = 0
Q = .02

TABLE 11.12

CROSS-TABULATION OF ANOMIA BY RACE

		Race (Z)			
		White		Black	
		n	%	n	%
Anomia (Y)	High	89	32	284	44
	Low	190	68	363	56
		279	100%	647	100%

E = 12
Q = -.25

accounted for by several variables, rarely is a single one so profound in its affect. Furthermore, while the focus has been upon first order partial tables it is possible to have second, third, and n^{th} order partial tables in which case two or more variables are simultaneously controlled. In any event, the analyst would control for all relevant variables and the selection of these factors is a logical and theoretical consideration, the only statistical guideline being that the control variable be related to both the independent and dependent variables.

Elaboration is a multivariate technique for ferreting out the meaning of the relationship between two variables by systematically introducing variables thought to be in part, or in whole, responsible for the initial association.

The meaning of the outcomes is explained in terms of the elaboration paradigm presented in Figure 11.1.

For those who wish to pursue the subtleties and nuances that have purposively been neglected your attention is called to Morris Rosenberg's classic book The Logic Of Survey Analysis[20].

PARTIAL CORRELATION

The second tradition of statistical control is commonly referred to as partial correlation, a statistical technique that mathematically adjusts the original bivariate covariation so that the influence of the control variable(s) is/are removed. Partial correlation provides a single summary index of association to describe the relationship between two variables while adjusting for the effects of one or more additional variables. The additional variables are called control variables.

Conceptually, partial correlation is somewhat analogous to cross-tabulation with test factors. However, the nature of control, as we will see, is different. The same generic query is asked here as in elaboration, namely, what influence does variable Z have upon X and Y? Whereas multivariate cross-tabulation procedures can be employed with observations at any measurement level, partial correlation is generally reserved for interval/ratio level data. Moreover, cross-tabulation is limited by the (often) severe reduction in cases since each partial table will have only part of the total number of cases. When the variables contain several categories this attrition of cases can sometimes become statistically problematic. With partial correlation, this liability does not occur.

Like elaboration, partial correlation can be used in a variety of ways to understand and clarify the relationships between three or more variables. For example, partial correlation techniques facilitate disclosing spurious relationships, locating intervening variables, and are useful in helping the researcher make certain kinds of causal inferences.

The Logic of Partial Correlation. The formula for the most basic partial correlation coefficient, technically a first order partial correlation, reads:

$$r_{YX.Z} = \frac{r_{YX} - (r_{YZ})(r_{XZ})}{\sqrt{(1-r_{YZ}^2)(1-r_{XZ}^2)}}$$

In statistical notation, Y = dependent variable, X = independent variable, and Z = control variable. The formula tells us that we want to know the correlation between Y and X with Z controlled or held constant.[21]

Let us decompose the partial correlation formula to better see what each term represents.[22] The numerator involves:

$$r_{YX} - (r_{YZ})(r_{XZ})$$

Note that the product $(r_{YZ})(r_{XZ})$, resembles a combined measure of the effects of Z on both X and Y, is subtracted from the original XY correlation. While not exactly true, the product term is a kind of average r^2, indicating the average proportion of variation in X and Y accounted for by Z. This implies that we only consider the covariation of X and Y for the portion of the respective variances that remains after Z has operated on both X and Y.

The denominator of the formula is a geometric mean representing the average value of the coefficients of nondetermnination:

$$\sqrt{(1 - r_{YZ}^2)(1 - r_{XZ}^2)}$$

The coefficient of nondetermination is the proportion of unexplained variation, (e.g., $1-r_{XZ}^2$ is the proportion of variation in X not explained by Z while $1-r_{YZ}^2$ is the proportion of variation in Y not explained by Z). This means that the partial correlation between X and Y controlling for Z is the correlation between the residuals of the regressions of X on Z and Y on Z.

A graphic illustration should clarify these statements. Figures 11.2a and 11.2b represent the regression of X on Z and Y on Z, respectively. The vertical lines from the points to the regression line represent residuals or variation unexplained by Z. Suppose we plot new points for X and Y taking the distande of points from the regression lines as new scores for X and Y. These "new scores" are residuals and are plotted in Figures 11.3a and 11.3b. By constructing a new scattergram for "residual Y" on "residual X" scores we would produce Figure 11.4. Because the points in the scattergram represent the same cases, we can compute a correlation coefficient for these residual scores. This correlation coefficient is the partial correlation coefficient, $r_{YX.Z}$.

It is unnecessary to complete the process of finding the residual scores and then computing r for them since the computing formula does exactly that. The purpose of illustrating this process is to give you a better appreciation of what it means to remove the influence of variable Z. Reflecting on this process should convince you that Z is not held constant in the same manner as it was in contingency table elaboration. The test factor Z is allowed to vary and is taken

FIGURE 11.2

REGRESSIONS OF X ON Z AND Y ON Z[23]

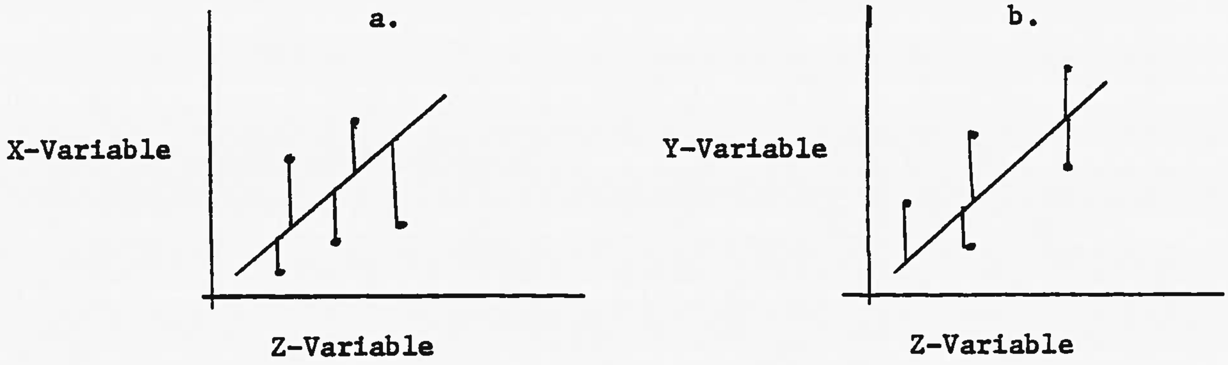


FIGURE 11.3

RESIDUAL X AND Y SCORES[24]

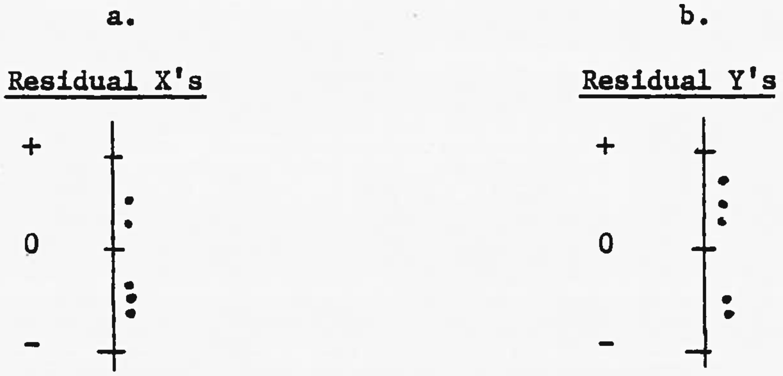
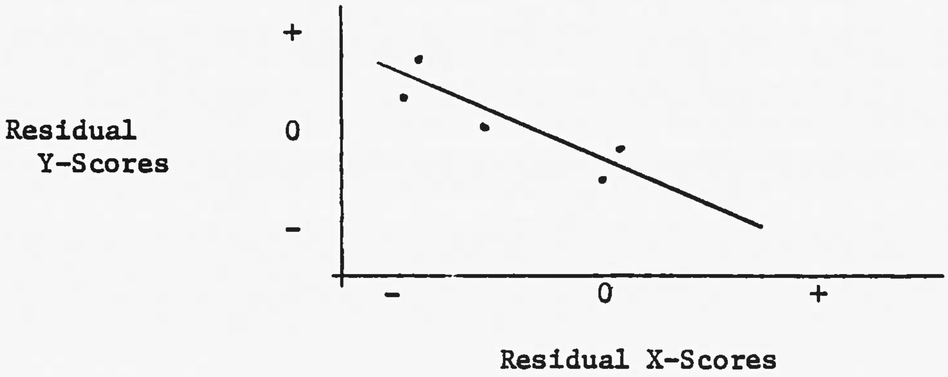


FIGURE 11.4

SCATTERGRAM FOR RESIDUAL Y ON RESIDUAL X[25]



into account by assessing the correlation between X and Y for only that component of their correlation that remains after Z has had its affect. While the technique for computing the partial r is different, the same result is obtained as that for table elaboration (i.e., Z is held constant).

Example 1. Table 11.13 presents hypothetical data (although these data are modeled after an actual empirical investigation by Ritterband and Silberstein) on mean achievement level (X, the independent variable), number of disorders (Y, the dependent variable), and percent blacks in the student body (Z, the control variable).[26] Since the partial correlation formula assumes the Pearson product moment r's have already been calculated, the first step is to obtain zero-order coefficients for r_{YX}, r_{YZ}, and r_{XZ}. While we will not demonstrate the actual computations, a simple computing formula for r will be presented so the reader can verify the values presented here. The formula reads:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\left[N\Sigma X^2 - (\Sigma X)^2\right] \left[N\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

The right hand side of the Table 11.13 presents the three zero-order correlation coefficients. A zero-order correlation coefficient is one containing no control variables. A brief explanatory note is in order. The correlation, r, between mean achievement level (X) and number of disorders (Y) is -.36. This moderate negative correlation suggests that disorders are more frequent in those schools with low achievement levels and the converse. The correlation between percentage black (Z) and the number of disorders (Y) is +.54. Apparently, the larger the black population the greater the number of disorders, and the converse. The correlation between the percentage black (Z) and mean achievement level (X) is -.63 and indicates that where the black population is large, the mean achievement level is low, and vice versa.

TABLE 11.13

MEAN ACHIEVEMENT LEVEL (X), NUMBER OF DISORDERS (Y), AND PERCENTAGE BLACK (Z) IN TEN HYPOTHETICAL SCHOOLS[27]

	School	Y	X	Z	Zero-order correlations:
$\Sigma Y=35$	A	5	65	72	$r_{YX}= -.36$
$\Sigma Y^2=173$	B	2	72	55	$r_{XZ}= -.63$
$\overline{Y}=3.5$	C	8	90	60	
$s_Y=2.25$	D	4	76	92	$r_{YZ}= +.54$
$\Sigma X=912$	E	1	97	38	Standardized & understandardized coefficients:
$\Sigma X^2=85{,}946$	E	5	105	59	b*YX.Z= -.03
$\overline{X}=91.2$	G	5	84	93	b*YZ.X= .52
$s_X=16.65$	H	3	93	12	$b_{YX.Z}= .004$
$\Sigma Z=539$	I	2	121	24	
$\Sigma Z^2=35{,}723$	J	0	109	34	$b_{YZ.X}= .045$
$\overline{Z}=53.9$					
$s_Z=25.83$					

Given this configuration of correlation coefficients, we suspect that the original association between school achievement level (X) and frequency of disorders (Y) is spurious. Why? Because both mean achievement and percentage black are correlated with the number of disorders in the student body, r= -.36 and r= +.54, respectively. To statistically determine the validity of our suspicion it is necessary to compute the correlation between X and Y with the effects of Z removed. This is done through the partial correlation formula presented earlier. Substituting the appropriate values into the formula, we have:

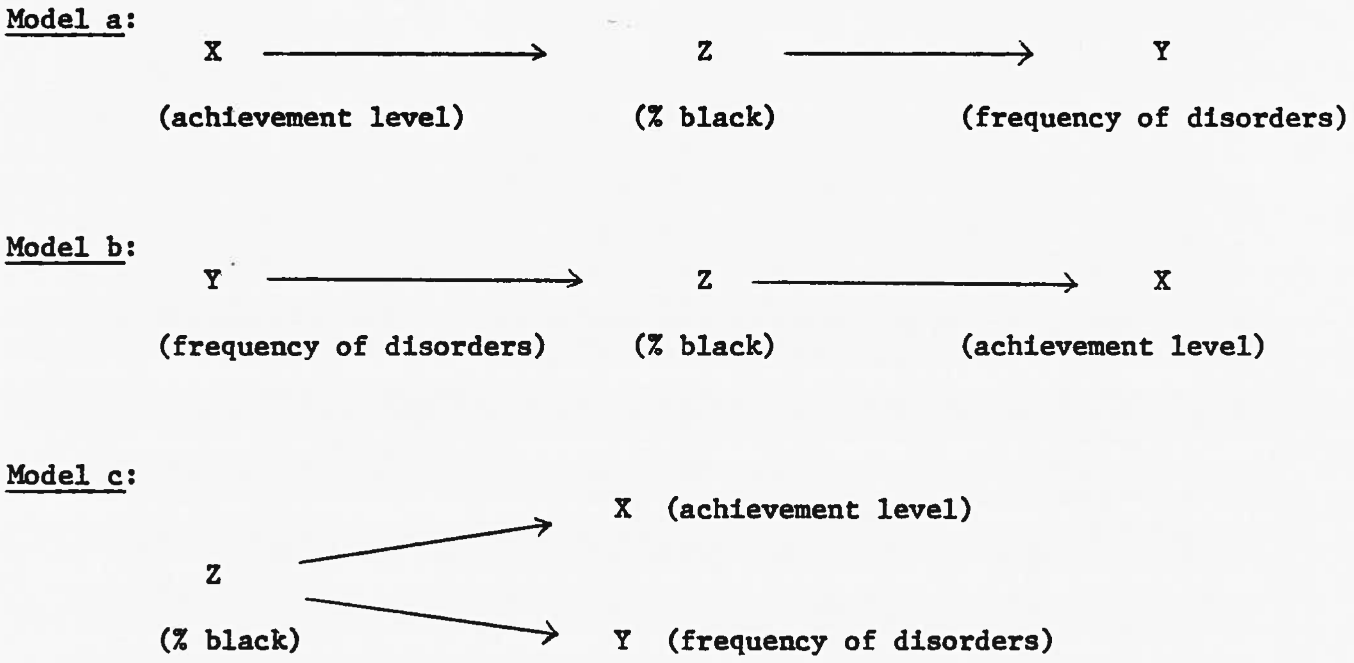
$$r_{YX.Z}= \frac{-.36 - (.54)(-.63)}{\sqrt{\left[(1-(.54)^2\right]\left[(1-(-.63)^2\right]}} = -.03$$

The partial correlation reveals a dramatic reduction in the magnitude of the original association. When Z was permitted to vary r= -.36; when Z was held constant, it diminished to virtually 0 (r = -.03).

<u>What are we to make of this</u>? Before concluding that the bivariate association is spurious (which is implied in the statistical result) we need to make certain assumptions about the causal structure linking the three variables. As with the elaboration paradigm, the time ordering of the variables becomes important. Let us consider via diagrams (Figure 11.5) three possible causal structures.

FIGURE 11.5

THREE POSSIBLE CAUSAL STRUCTURES LINKING X, Y, AND Z[28]



Consider model a first.Note that this model would assume achievement level affects the percent black. This reasoning would be plausible if students were permitted to choose the high schools they attended and if black students were disproportionately prone to select schools with low academic achievement. Such conditions as these are unrealistic and this allows us to dismiss the first alternative.

Model b implies that the frequency of disorders affects the percent black. This is contrary to what we know since the black population anteceded the disorders. Hence, this causal structure is also judged untenable.

Model c assumes that the percentage black preceded and is related to both the achievement level of the school as well as the frequency of disorders. This model is both theoretically and statistically compatible with our results. Theoretically, the percentage black in each school affects the average achievement in the respective schools and the percentage black also affects the frequency of disorders. Statistically, the original correlation (between X and Y) virtually disappears when percent black is controlled. Consequently, it is appropriate to conclude that the original association is spurious.

Example 2. This illustration reveals the manner in which partial correlation can aid in identifying explanatory variables.

To illustrate, consider the following correlation matrix --a table containing the correlation values among variables-- showing Pearson r's between combinations of eunomia, education, and income (Table 11.14).

TABLE 11.14

CORRELATION MATRIX FOR EUNOMIA , EDUCATION, AND INCOME

	(Y)	(X)	(Z)		
	Eunomia	Education	Income	$\bar{X}$	s
(Y) Euncmia	1.00	.40	.25	13.3444	2.4197
(X) Education		1.00	.47	10.0629	3.7367
(Z) Income			1.00	5718.2	314.31

Eunomia, a psychological state of well being, correlates with education and income to the tune of .40 and .25, respectively. We want to determine the r between eunomia and education with the effect of income removed and the r between eunomia and income with education statistically controlled. Substituting the simple r's from the correlation matrix into the formula we have:

$$r_{YX.Z} = \frac{.40 -(.25)(.47)}{\sqrt{\left[1 - (.25)^2\right]\left[1 - (.47)^2\right]}} = .33$$

Before interpreting this partial correlation, let us compute the r between eunomia and income with the effect of education removed. The same computing formula may be used with the appropriate substitutions for the new control variable. Hence,

$$r_{YZ.X} = \frac{.25 - (.40)(.47)}{\sqrt{\left[1 - (40)^2\right]\left[1 - (.47)^2\right]}} = .08$$

The r between eunomia and education is .40. When the influence of income is removed from the bivariate association the r between eunomia and education is reduced to .33. Resorting to the PRE interpretation, we may say that education accounts for 16 % ($r^2 = .40^2 = .16$) of the variation in eunomia but when the effects of income are removed producing an $r_{YX.Z} = .33$, we say that 11% ($r^2_{YX.Z} = (.33)^2 = .1089$) of the variation in eunomia is explained by education after removing the influence of income. Thus, some of the bivariate association between education and eunomia is due to income but the relationship is not greatly altered when income is systematically controlled.

The r between eunomia and income is .25. When the influence of education is partialled out from the original correlation the r between eunomia and

and income is .08. Whereas income originally accounted for about 6% ($.25^2$=.0625) of the variation in eunomia, when the influence of education is removed only about ½% ($.08^2$ = .0064) of the variation in eunomia is explained by income. Thus, income as an explanatory variable does not have a great deal of independent influence on Y. On the other hand, since the r between education and eunomia is upheld when income is controlled, we begin assessing its direct causal influence on the dependent variable.

Guidelines for Interpreting Partial Correlation Coefficients.[29] Certain statistical conditions can be identified which will make a partial r zero or prohibit it from being zero. Three such "rules of thumb" will be advanced:

1) When the sign of r_{YX} (the original zero-order correlation) is not consistent with the sign of the $(r_{XZ})(r_{YZ})$ product, $r_{YX.Z}$ cannot be zero. Suppose r_{YX} = .45 and r_{XZ} = .30; r_{YZ} = -.25. The product of $(r_{XZ})(r_{YZ})$ = (.30) (-.25) = -.075. The partial r = .57 (.45 - (.30)(-.25)/ $\sqrt{[1-(.30^2)][(1-(-.25^2)]}$. Of course, if one, but not both of the product terms is negative, the product must also be negative (e.g., $(r_{XZ})(r_{YZ})$ = (.30)(-.25) = -.075). If r_{YX} is positive,the numerator (when r's display different signs, on the right) becomes the sum of two positive quantities and hence cannot be zero. Similarly, if both r_{XZ} and r_{YX} are negative or if both are positive, the product $(r_{XZ})(r_{YX})$ must be positive. If r_{YX} is negative and when r_{XZ} and r_{YZ} have different signs, the numerator on the right becomes the sum of two negative numbers and, again, cannot be zero. This discussion suggests that scrutinizing the signs of the zero-order correlations may suffice to determine some circumstances when the partial r cannot be zero.

2) If $(r_{YZ})(r_{XZ}) = r_{YX}$, then $r_{YX.Z}$=0. If the product term equals the same value as the original bivariate association, then the partial correlation

will be zero since the left-hand side of the equation is subtracted from the right hand side. For example, if $r_{YX} = .60$ whereas $r_{XZ} = .20$ and $r_{YZ} = .30$, then the numerator of the formula must equal zero (i.e., $.60 - (.20)(.30) = .60 - .60 = 0$). In practice, then, when r_{YX} and $(r_{YZ})(r_{XZ})$ are about equal, the partial association itself will be approximately zero. Of course, the denominator must also be taken into account. Since the denominator--the geometric mean of two coefficients of nondetermination--is virtually always less than unity(1.00), the numerator alone yields an underestimate of the first order partial (except when r_{YZ} and $r_{XZ} =0$). Also,the numerator alone underestimates the first-order partial correlation coefficient more when the correlations of X and Y with the control variable Z are strong than when they are weak.

3) The partial correlation $r_{YX.Z} \geq r_{YX}$ if $r_{XZ} =0$ or if $r_{YZ} = 0$. This is so since the numerator is just r_{YX}. If $r_{XZ} = 0$, $r_{YZ} =.50$, and $r_{YX} =.70$ the numerator of the partial r would be: $.70 - (0)(.50) = .70 - 0 - .70$. Furthermore, if one, but not both, of the correlations of the two variables (X and Y) is zero, one of the terms in the denominator is 1.0 and the other will be less than 1.0. If r_{YX} is divided by a quantity less than 1.0, the quotient must exceed r_{YX}. For example,

$$\frac{.70-(0)(.50)}{\sqrt{(1-0^2)(1-0^2)}} = \frac{.90}{\sqrt{(1)(1)}} = \frac{.90}{1} = .90$$

The implications of this configuration of correlation coefficients is that control variables uncorrelated or only weakly correlated with either or both the original variables cannot produce a first-order partial significantly different from the initial bivariate association.

Higher-Order Partial Correlation Coefficients.

The partial correlation coefficients computed above involve a single control variable and are termed *first-order partials*. The order of the partial correlation is determined by the number of control variables. If more than a single control variable is introduced, the order of the control is determined by the *number* of control variables. For example, with two variables it is called a *second order partial correlation*, with three control variables it is called a *third order partial correlation*, etc.

A formula for a second order partial correlation coefficient looks like this:

$$rYX.ZW = \frac{{}^{r}YX.Z - ({}^{r}YW.Z)\ ({}^{r}XW.Z)}{\sqrt{(1-{}^{r}YW.Z^{2})(1-{}^{r}XW.Z^{2})}}$$

Notice that this "higher-order" partial is based on the same principles as the first-order partial. The generic formula can apply to partials of any order. However, partial correlation coefficients beyond the third order are rarely used for two major reasons: (1) it is unlikely that four *non*-interacting variables can be isolated, and (2) higher order partial correlations tend to be cumulatively affected by measurement error.

Summary

An important difference between partial correlation techniques and those of cross-tabulation is that in the former, unlike the latter, a *single* statistical index reflecting the degree of correlation between two variables controlling for a third (in a first order partial correlation) is produced. In cross-tabulation, there are as many summary indices as there are categories of the control variable. One disadvantage of the partial correlation approach is when the partial associations vary from one level of the control variable to the next. This condition is known as *statistical interaction*. In these circumstances, the partial r averages out the different partial correlations and may be detrimental to the substantive meaning of the data. Like the elaboration paradigm presented in the discussion of cross-tabulation the same interpretation, depending upon the partial correlations in comparison to the original and the time order of the control variable in relation to the independent variable, is applicable.

MULTIPLE REGRESSION AND MULTIPLE CORRELATION

In most social science research it is insufficient and overly simplistic to predict the values of a dependent variable with just one independent variable. For a more complete and realistic model most situations require using several independent variables. For example, demographers have noted that population growth in a particular geo-political region is explained by four variables: 1) birth rate, 2) death rate, 3) immigration rate, and 4) emigration rate. Similarly, a sensible explanation of college grade point average (GPA) would include several indices as predictors (e.g., high school GPA, college entrance exam scores, IQ scores, etc.).

The simultaneous influence of multiple explanatory variables on a single dependent variable can be assessed by deriving a multiple regression equation. Such a model facilitates describing the linear relationship between the dependent and independent variables. The multiple regression equation is an extension of the simple bivariate model (Ch. 10) and reads:

$$Y' = a + b_1X + b_2Z \ . \ . \ . \ b_kX_k$$

Where: b_1X = controlling for Z . . . X_k (holding their values fixed), Y is linearly related to X, with slope b_1

b_2Z = controlling for X . . .X_k, Y is linearly related to Z, with slope b_2

a = Y intercept (i.e., the value of Y when X . . . X_k = 0)

Y' = predicted value of Y given known values of the independent variables

b_1 and b_2 are called partial regression coefficients or the regression line for each independent variable, controlling for the other variable(s).

Prediction. Our earlier discussion of simple linear regression will serve as a prelude to the more complex procedure known as *multiple (linear) regression*. Regression implies prediction once again while multiple denotes that more than one predictor variable is employed in the data analysis. This regression equation is referred to as a *linear, additive* model. It is additive because the effect of X (b_1X) is added to the effect of Z (b_2Z). The property of additivity is most salient (and will be demonstrated later in this section) since it means that X and Z do *not* interact in their effects upon Y.[30] The task is to predict the dependent variable's values from a knowledge of several (from two to K) independent variables' values. In both social research and the practical affairs of everyday life more often than not, more than one variable is accountable for the variation in another variable (Y). Through the construction of a multiple regression equation the simultaneous effects of several independent variables on a dependent variable can be assessed. This equation, which describes the amount of linear relationship between the causal and effect variables, can be more elaborately written:

$$Y' = a_{Y.XZ} + b_{YX.Z}X + b_{YZ.X}Z \ . \ . \ .$$

In this equation Y' = the dependent variable, X and Z = independent variables, $a_{Y.XZ}$ = Y-intercept, and $b_{YZ.X}$ = partial regression coefficients (slopes) of the regression line for each independent variable, controlling for the other variables. Like its simple regression counterpart, the mathematical constants a and b are estimated so that the average square error in prediction is minimized using the least squares criterion.

The Mechanics of Computing the Multiple Regression Equation. Two distinct stages are ordinarily entailed in computing the regression coefficients. First, from the raw data themselves we calculate all possible pairs of correlation coefficients among the variables.[31] This would be achieved by utilizing any of the various computing formulae for the Pearsonian r. In the present, we will commence discussion with the r's having already been determined. Second, we use the correlation coefficients to obtain the various values needed for the multiple regression equation. The raw data and related statistics appearing in Table 11.13 will be used.

For illustrative purposes we will take a hypothetical three variable case involving mean achievement level (X), number of disorders (Y), and percentage black (Z) in ten high schools. Our task is to predict the number of disorders (the dependent variable) from a knowledge of mean achievement level and percentage black (the independent variables). To construct the multiple regression equation, we need computing formulae to determine the standardized regression constants (denoted by b*). To calculate $b*_{YX.Z}$ we may use the following formula which uses the correlation coefficient among variables:

$$b*_{YX.Z} = \frac{r_{YX} - (r_{YZ})(r_{XZ})}{1 - (r_{XZ})^2}$$

Conceptually, $b*_{YX.Z}$ is a standardized regression value indicating how much of the variation in Y is accounted for by X when the contribution of Z is removed. In this sense b* coefficients are like partial correlation coefficients in that the contribution of other variables is held constant or controlled.[32] Generically, the first variable to the right of b* is the dependent variable (Y), the second variable (X) is the independent variable, and the variable used as a control (Z) is to the right of the dot. Hence, the general formula depicts a standardized regression formula for X and Y with Z held constant.

To compute $b*_{XY.Z}$ we substitute as follows:

$$b*_{YX.Z} = \frac{-.36 - (.54)(-.63)}{1 - (.-63)^2} = \frac{-.0198}{.6031} = -.03$$

The other regression coefficient, $b*_{YZ.X}$, is computed as follows:

$$b*_{YZ.X} = \frac{r_{YZ} - (r_{YX})(r_{XZ})}{1 - (r_{XZ})^2}$$

$$b*YZ.X = \frac{.54 - (-.36)(-.63)}{1 - (-.63)^2} = .\frac{3132}{6031} = .52$$

As a double check on our computations, a set of *normal equations (not* related to the normal curve) predicting the correlation between each variable and the dependent variable, are solved for the known b*'s. For a three variable problem, the normal equations are:

$$b^*_{YX.Z} + (r_{XZ})(b^*_{YZ.X}) = r$$

$$-.03 + (-.63)(.52) = -.36$$

$$(r_{XZ})(b^*_{YX.Z}) + b^*_{YZ.X} = r_{YZ}$$

$$(-.63)(-.03) + (.52) = .54$$

The normal equations confirm the accuracy of our original computations. Note that the respective r's in Table 11.13 are identical to these.

The standardized regression equation using percentage black and mean achievement level as predictors of the number of disorders would be expressed:[33]

$Y'_Z = -.03X_Z + .52Z_Z$ (the subscript Z indicates the values are expressed in standard deviation units)

With a standardized regression equation two things are accomplished: 1) the relative importance of the independent variables' explanatory power can be determined (e.g., in predicting Y, Z is a more important variable than X). Additionally, we can estimate the best prediction equation for the variables under scrutiny.

Interpreting the b coefficients*. The b* coefficients enable an analyst to compare the relative importance of one variable with the contribution of other variables in the regression equation. Since $b^*_{YZ.X}$ is larger than $b^*_{YX.Z}$ we know that a given change in the former regression value would produce a larger change in the dependent variable (in z score units). In fact, it is over seventeen times (.52 ÷ .03 = 17.33) as potent as the latter variable. For the present case we can say that for every increase of one standard deviation unit in the percentage black, the number of disorders increases by .52 standard deviation units; and with an increase of one standard deviation in mean achievement level, the number of disorders decreases by .03 standard deviation units. When the relative importance of the two predictor variables are considered, it is evident that percentage black contributes more ($b^* = .52$) to explaining student disorders than does mean achievement level ($b^* = -.03$).

Unstandardized Regression Coefficients. The disadvantage of standardized regression equations and coefficients is that we cannot make predictions on Y in terms of the original score units. Fortunately, it is relatively easy to convert standardized values into unstandardized values in the following manner:

$$b_{YX.Z} = s_Y/s_X \quad (b^*_{YX.Z}) = 2.25/16.65\ (-.03) = -.004$$

$$b_{YZ.X} = s_Y/s_Z \quad (b^*_{YZ.X}) = 2.25/25.83\ (.52) = .045$$

$$\begin{aligned} a_{Y.XZ} &= \bar{Y} - b_{YX.Z}\bar{X} - b_{YZ.X}\bar{Z} \\ &= 3.5 - (-.004)(91.2) - (.045)(53.9) \\ &= 3.5 - (-.3648) - (2.4255) \\ &= 1.44 \end{aligned}$$

Having computed both the unstandardized and the standardized regression coefficients (sometimes called beta weights) we may check our calculations by computing the standardized beta weights using the unstandardized regression coefficients.

$$b^*_{YX.Z} = b_{YX.Z}\ \frac{(s_X)}{(s_Y)} = -.004\left(\frac{16.65}{2.25}\right) = -.03$$

$$b^*_{YZ.X} = b_{YZ.X}\ \frac{(s_Z)}{(s_Y)} = .045\left(\frac{25.83}{2.25}\right) = .52$$

Consequently, the double check confirms the originally computed standardized beta values.

The unstandardized multiple regression equation reads:

$$Y' = 1.44 + (-.004)\ X + .045\ Z$$

Substantively, this multiple regression equation means that schools with an increase of one unit in the mean achievement level will witness a drop of .004 in the number of disorders. However, with an increase of one unit in the percentage of blacks the number of disorders will increase .045. The Y - intercept value, 1.44, indicates how many student disruptions to expect if there were no blacks and no mean achievement level. Practically speaking it would be virtually impossible to have a school system with no achievement. Nevertheless, mathematically, this is the manner in which the intercept value is interpreted.

Multiple Correlation. Multiple correlation, symbolized by a capital R, is the correlation between the dependent variable and all independent variables used in the analysis and is symbolized as $R_{Y.XZ}$. The variable to the left of the dot is the predicted (or dependent) variable. Like the simple coefficient of determination (r^2), R^2 (the coefficient of multiple determination) tells us how much variation in the dependent variable is explained by all independent variables in the multiple regression equation. Similarly, $1-R^2$, the coefficient of multiple non-determination, indicates how much of the variation in Y is due to other variables not included in the multiple regression formulation. Once the standardized regression coefficients have been determined the multiple correlation coefficient can easily be calculated from the simple r's and the standardized regression betas via the following formula:

$$R^2_{Y.XZ} = (b^*_{YX.Z})(r_{YX}) + (b^*_{YZ.X})(r_{YZ}) = (-.03)(-.36) + (.52)(.54) = .29$$

The coefficient of multiple determination for our data is computed to be:

$$R^2_{Y.XZ} = .29$$

To interpret this value, we may say that 29%(.29 x 100= 29%) of the variation in Y is accounted for by X and Z.

R^2 can also be directly calculated from the zero-order correlations themselves using this formula:

$$R^2_{Y.XZ} = \frac{r^2_{YX} + r^2_{YZ} - 2\, r_{YX}\, r_{YZ}\, r_{XZ}}{1 - r^2XZ}$$

$$\frac{(-.36)^2 + (-.54)^2 - 2(-.36)(.54)(-.63)}{1 - (-.63)^2} =$$

$$\frac{.1296 + .2916 - .2449}{.6031} = \frac{.1763}{.6031} = .29$$

This figure corroborates the earlier one using simple r's and standardized beta weights. R^2 reaches its maximum value when the independent variables are not intercorrelated. When *multi-collinearity* exists, that is, a high degree of association between the independent variables exists, the multiple R will not be much larger than that of the largest zero-order correlation. Ideally, although rarely the case in practice, the explanatory variables in the regression equation should be independent of each other ($r = 0$) if maximum predictive power is to be achieved. If the correlation between the independent variables is zero, the formula reduces to $R^2_{Y.XZ} = r^2_{YX} + r^2_{YZ}$ (and this notion can be extended to any number of independent variables in regression analysis).

R^2 can also be computed *from a combination of zero-order and first-order partial correlations* as in the formula:[34]

$$R^2_{Y.XZ} = r^2_{YX} + (1-r^2_{YX})\, r^2_{YZ.X} = (.36)^2 + \left[(1-(-.36)^2\right](.43^2) = .29$$

The R and R^2 Values. As with the simple r we may resort to the PRE interpretation which indicates the percentage of variation in the dependent variable (e.g., number of disorders) explained by all the independent variables (e.g., mean achievement level and percentage blacks) in the regression equation. To use this interpretation R must be squared (just like r had to be squared). R^2 is called the *coefficient of multiple determination* and its $1 - R^2$ counterpart the *coefficient of multiple non-determination*. The relationship between R and R^2 is, of course, $R^2 = \sqrt{R}$.

R² as a Proportional Reduction in Error Measure.[35] The coefficient of multiple determination (and its counterpart the coefficient of multiple nondetermination) is an analogue of the coefficient of determination (and its counterpart the coefficient on nondetermination) produced via simple correlational analysis. The generic PRE formula reads:

$$\frac{E_1 - E_2}{E_1}$$

Rule 1 (for E_1). The dependent variable is predicted in the absence of knowledge of other independent variables (e.g., X, Z, . . . k). The best predictor for each case would be the mean of Y, $\bar{Y}$.

Rule 2 (for E_2). To include data for the independent variables in the analysis the multiple regression equation: $Y' = a_{Y.XZ} + b_{YX.Z}X + b_{YZ.X}Z$

By substituting the appropriate X and Z values a prediction, Y', for each case can be obtained.

Prediction Errors. For rule 1, a prediction error is $Y - \bar{Y}$. The total prediction error is summarized by the sum of squared prediction errors. For the present data, this would be $(5 - 3.5)^2 + (2 - 3.5)^2 + . . . (0 - 3.5)^2$. Or, equivalently, $\Sigma Y^2 - (\Sigma Y)^2/N$. The total sum of squares, the prediction errors for rule 1, equals 50.5. For rule 2, a prediction error is Y - Y'. Table 11.15 displays the entire process, resulting in the sum of squared errors (SSE), the residuals squared and summed, equal to 35.7239.

Definition of Measure. The proportional reduction in error obtained by using the linear multiple regression equation $Y' = a + b_1X + b_2Z$ instead of $\bar{Y}$ as a predictor is:

$$R^2 = \frac{TSS - SSE}{TSS}$$

or

$$\frac{50.5 - 35.7239}{50.5} = .29$$

Not only does this discussion highlight the PRE logic of the multiple correlation coefficient, it also provides another computing formula for R^2.

When R^2 is computed from a sample, there is a tendency toward inflation of its value. Hence, to correct for this known bias, statisticians recommend a correction factor ($R^2{}_C$) be applied which tends to reduce the original R^2 value. The correction factor is:[36]

$$R^2{}_C = 1 - \frac{N - 1}{N-K-1} (1 - R^2)$$

Substituting the present data into the formula, we have the following substitutions: N = sample size, 10; k = the number of independent variables, 2; R^2, $(.29)^2$ = the original uncorrected coefficient of multiple determination; therefore,

$$R^2{}_C = 1 - \frac{10 - 1}{10-2-1} (1 - .29) = .09$$

The shrinkage in the corrected R^2 value is due to the fact that both N and K are small. This reduction is less substantial when N and K are large.

TABLE 11.15

ILLUSTRATING COMPUTATIONS OF SUM OF SQUARED ERRORS (RESIDUALS), SSE

School	Observed Y Value	Predicted Y Value Using multiple regression equation: $Y' = 1.44 + (-.004)X + .045Z$	Residual $Y - Y'$	Residual Squared $(Y - Y')^2$
A	5	4.42	0.580	0.3364
B	2	3.627	-1.627	2.6471
C	8	3.78	4.220	17.8084
D	4	5.276	-1.276	1.6281
E	1	2.762	-1.762	3.1046
F	5	3.675	1.325	1.7556
G	5	5.289	-0.289	0.0835
H	3	1.608	1.392	1.9377
I	2	2.036	-0.036	0.0013
J	0	2.534	-2.534	6.4212
			$\Sigma \approx 0$	$\Sigma =$ 35.7239

Summary

Multiple regression analysis is a statistical procedure whereby the simultaneous influence of multiple explanatory variables on a single dependent variable can be assessed. It is an extension of simple bivariate regression and is added to the effect of another (or others) in order to predict a dependent variable's value. Both standardized and unstandardized procedures and interpretations for the multiple regression equation were discussed.

Multiple correlation, the association between a dependent variable and all independent variables used in the analysis, was considered. Multiple correlation coefficients can be computed from simple r's and standardized regression coefficients, simple correlation coefficients alone, and simple r's and partial correlation coefficients. Each of these procedures was demonstrated. The interpretation of a multiple R is a PRE one. The proportional reduction in error interpretation of R was clearly demonstrated. Two vital concepts-- (1) coefficient of multiple determination (analogous to the coefficient of determination for r), and (2) coefficient of multiple non-determination (analogous to the coefficient of nondetermination)--were considered.

Important Concepts Discussed in This Chapter

Multivariate Distributions

Elaboration

Partial Correlation

Multiple Regression

Multiple Correlation

Causal Explanation

Joint Result Explanation

Intervening Effects Explanation

Interacting Effects Explanation

Chance or Sampling Fluctuation Explanation

Related Observations Explanation

Control

Interpretation

Prediction

Subgroup Comparison

Crosstabulation

Important Concepts Discussed in This Chapter (cont.)

Patterns of Elaboration

- Replication
- Explanation
- Interpretation
- Interaction (specification)
- Prediction

Partial Table

Conditional Table

Zero-Order Table

Partial Correlation Coefficients

Multiple Regression Equation

Linear, Additive Model

Standardized regression coefficients

Unstandardized regression coefficients

Coefficient of Multiple Determination

Coefficient of Multiple Non-determination

R^2 as a proportional reduction in error measure

Chapter 11 Endnotes

[1] Methodologists maintain that to establish causality, four conditions must be met: 1) there must be an association between the variables (e.g., cigarette smoking cannot be the cause of lung cancer if the incidence of lung cancer is essentially the same when smokers and non-smokers are compared); 2) the presumed causal variable must precede in time the presumed effect variable (e.g., if people developed lung cancer before ever having smoked it would be ridiculous to argue the smoking was the responsible agent); 3) the original association must not be spurious, that is, "explained away" or "vanish" when examined in the context of additional variables; and 4) there should be a theoretical rationale explicitly linking the variables together. Each condition is necessary in the sense that it must be present, but all must be simultaneously present to claim a cause-effect nexus. When all four are present we have the sufficient conditions for positing a cause/effect relationship between variables.

[11] The terminology for the various tables is as follows: a zero-order table is the original table in which X and Y have been crosstabulated; the partial tables are those which display the original X/Y relationship within categories of the test factors; the marginal tables are zero order tables in which the control variable is cross-tabulated with X (one table) and Z (another table).

[12] Suppose we label a 2 x 2 contingency table as follows:

		X	
		X_1	X_2
Y	Y_1	n_{11}	n_{12}
	Y_2	n_{21}	n_{22}

n_{11} refers to the number of observations at the intersection of row 1 and column 1; n_{12} refers to the number of observations at the intersection of row 1 and column 2; in general nrc stands for the number of cases at the intersection of a particular row and a particular column. Using these designations the formula for Q reads:

$$Q = (AD - BC) / (AD + BC) \quad \text{OR} \quad (n_{11})\,(n_{22}) / (n_{12})\,(n_{21})$$

[32] The similarities and differences between standardized beta weights (b*'s) and partial correlation coefficients ($r_{YX.Z's}$) are noteworthy. Both reflect the effect of an independent variable on a dependent variable when the effects of other independent variables are mathematically adjusted. The numerator of the standardized beta weights (sometimes called path coefficients) and the partial correlation coefficients is identical. b*, on the one hand, indicates the amount of change in the dependent variable (in standard score form) associated with a unit change in the independent variable when other independent variables are mathematically controlled. Hence, b* is an asymmetric measure. $r_{YX.Z}$, on the other hand, is a symmetric measure which expresses the relationship between two variables with the effect of another (or others) statistically removed. In short, the partial correlation coefficient provides a measure of the accuracy of prediction while beta coefficients provide an indicator of the relative importance of a variable in prediction. (Loether, H.J. & McTavish, D.G. *Descriptive and Inferential Statistics: An Introduction*, 317.

[33] We have expressed X, Y, and Z not as raw scores, but as standard scores (Z scores.) Thus, the multiple regression equation can be denoted as: $Y'_Z = b^*_{YX.Z}X_Z + Z_Z$ etc. When the regression equation is expressed in standard score form there is no need for an intercept constant, a, since $\bar{Y}_Z = 0$ (the mean of z-scores equals zero). Furthermore, the notation Y_Z actually refers to the predicted z-score value of the dependent variable Y, and X_Z and Z_Z refer to variables X and Z in standard score terms, respectively.

[34] The partial correlation coefficients are computed as follows:

$$r_{YX.Z} = \frac{-.36 - (-.63)(.54)}{\sqrt{[1-(-.63)^2][1-(.54)^2]}} = \frac{-.0198}{\uparrow \; .6536} = -.03$$

$$r_{YZ.X} = \frac{.54 -(-.36)(-.63)}{\sqrt{[1-(-.36)^2]\,[1-(-.63)^2]}} = \frac{.3132}{\uparrow \; .7245} = .43$$