3

Multiple Regression/ Correlation With Two or More Independent Variables

3.1 INTRODUCTION: REGRESSION AND CAUSAL MODELS

In Chapter 2 we examined the index of linear correlation between two variables, the Pearson product moment correlation r and the regression equation for estimating Y from X. Because of the simplicity of the two-variable problems, we did not need to go into detail regarding the interpretive use of these coefficients to draw substantive inferences. The inferences were limited to the unbiased estimation of their magnitudes in the population; the assertion, in the case of the regression coefficient, that one variable was, in part, related to or dependent on the other; and the demonstration of the significance of the departure of the coefficients from zero. When we move to the situation with more than one independent variable, however, the inferential possibilities increase more or less exponentially. Therefore, it always behooves the investigator to make the underlying theoretical rationale and goals of the analysis as explicit as possible. Fortunately, an apparatus for doing so has been developed in the form of the analysis of causal models. Because the authors advocate the employment of these models in virtually all investigations conducted for the purpose of understanding phenomena (as opposed to simple prediction), this chapter begins with an introduction to the use of causal models. A more complete presentation is found in Chapter 12.

3.1.1 What Is a Cause?

Conceptions of causality and definitions of cause and effect have differed among proponents of causal analysis, some offering no explicit definitions at all. Causal analysis as a working method apparently requires no more elaborate a conception of causality than that of common usage. In our framework, to say that X is a cause of Y carries with it four requirements:

1. X precedes Y in time (temporal precedence).
2. Some mechanism whereby this causal effect operates can be posited (causal mechanism).
3. A change in the value of X is accompanied by a change in the value of Y on the average (association or correlation).
4. The effects of X on Y can be isolated from the effects of other potential variables on Y (non-spuriousness or lack of confounders).

When X or Y is a quantitative variable (e.g., dollars, score points, minutes, millimeters, percentile ranks), the meaning of value is obvious. When X is a categorical scale (i.e., a collection of two or more qualitative states or groups), a change in value means a change from one state to another (e.g., from Protestant to Catholic or Protestant to non-Protestant, from depressed to not depressed, or from one diagnosis to another). When Y is a dichotomy (schizophrenia-nonschizophrenia), a change in value on the average means a change in proportion (e.g., from 10% schizophrenia for some low value of X to 25% schizophrenia for some higher value).

The third proposition should not be simplified to mean, If you change X, Y will change. This may, of course, be true, but it need not be. First, it may not be possible to manipulate X. For example, boys have a higher incidence of reading disability than girls; here sex (X) causes reading disability (Y), but it is meaningless to think in terms of changing girls into boys. Second, even when X can be manipulated, the way it is manipulated may determine whether and how Y changes, because the nature of the manipulation may defeat or alter the normal causal mechanism whereby X operates.

The models that we are employing have their roots in the path-analytic diagrams developed by the geneticist Sewell Wright (1921) for untangling genetic and nongenetic influences. These are often currently referred to as structural models or structural equation models. The purpose of the models is to make explicit exactly what the investigator has in mind about the variables and the meaning of their interrelationships. As such, they contribute to the clarity and internal consistency of the investigation. It should be recognized at the outset, however, that a causal model may never be established as proven by a given analysis; all that may be said is that the data are to some extent consistent with a given model or that they are not. Thus, the value of a given model is determined as much by the logic underlying its structure as by the empirical demonstrations of the fit of a given set of data to the model.[1]

3.1.2 Diagrammatic Representation of Causal Models

The basic rules for representing a causal model are quite simple.[2] Causal effects are represented by arrows going from the cause to the effect (the "dependent" variable). Usually, by convention, the causal flow is portrayed as going from left to right. In a simple model the independent variables are considered *exogenous* or predetermined variables. These variables are taken as given, and the model requires no explanation of the causal relationships among them. The relationships among these variables are represented by curved double-headed arrows connecting each pair.

To illustrate the use of a causal diagram, let us expand the academic example employed in Chapter 2. The investigator has collected the data on number of publications and time (expressed in number of years) since Ph.D. to determine the influence of productivity (as indexed by publications) and seniority (time since Ph.D.) on academic salaries. The resulting causal diagram is shown in Fig. 3.1.1. In this simple model we assert that academic salary is in part determined by time since Ph.D. and in part by publications. These latter two variables may be correlated with each other, but no causal explanation is offered for any relationship between them. However, salary is assumed not to cause changes in numbers of publications nor in time since Ph.D.

---

[1] The logical frame and historical development of causal models are discussed further in Section 12.1.

[2] This initial discussion is limited to elementary models and omits consideration of the effects of unmeasured causes and the assumptions underlying the models, for which see Chapter 12.

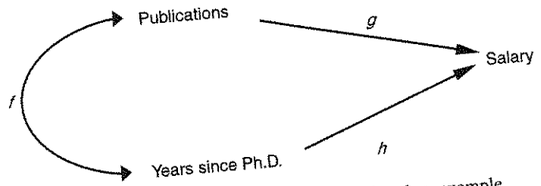


FIGURE 3.1.1 Causal model of academic salary example.

3.2 REGRESSION WITH TWO INDEPENDENT VARIABLES

To provide the estimates of effects required by our causal model we need a weight for each of our exogenous variables whose application will account for as much of the variance of our dependent variable as possible. Recalling that the regression equation, $Y = B_X + B_0$, was designed to be such an estimate for a single independent variable, we may anticipate that a similar procedure may produce the appropriate weights for two independent variables.

For example, suppose we have gathered the data in Table 3.2.1 to estimate the model for academic salaries presented in Fig. 3.1.1.[3] The correlation between salary (Y) and time since Ph.D. (X_1) is .710 and B_{Y1} is therefore .710($7889.77/4.577) = $1224 per year. The correlation between salary and number of publications (X_2) is .588, and its regression coefficient is therefore .588($7889.77/13.823) = $336 per publication (Table 3.2.1). If X_1 and X_2 were uncorrelated, we could simply use B_{Y1} and B_{Y2} together to estimate Y. However, as might be expected, we find a tendency for those faculty members who have had their degrees longer to have more publications than those who more recently completed their education ($r_{12} = .657$). Thus, X_1 and X_2 are to some extent redundant, and necessarily their respective estimates, $\hat{Y}_1$ and $\hat{Y}_2$ will also be redundant. What we need to estimate Y optimally from both X_1 and X_2 is an equation in which their redundancy (or more generally the relationship between X_1 and X_2) is taken into account. The regression coefficients in such an equation are called *partial regression coefficients* to indicate that they are optimal linear estimates of the dependent variable (Y) when used in combination with specified other independent variables.[4] Thus, $B_{Y1.2}$ is the partial regression coefficient for Y on X_1 when X_2 is also in the equation, and $B_{Y2.1}$ is the partial regression coefficient for Y on X_2 when X_1 is also in the equation. The full equation is

CH03EX01

$$\hat{Y} = B_{Y1.2}X_1 + B_{Y2.1}X_2 + B_{0Y.12} \tag{3.2.1}$$

The partial regression coefficients or B weights in this equation, as well as the regression constant B_0, are determined in such a way that the sum of the squared differences between (actual) Y and (estimated) $\hat{Y}$ is a minimum. Thus, the multiple regression equation is defined by the same ordinary least squares criterion as was the regression equation for a single independent variable. Because the equation *as a whole* satisfies this mathematical criterion, the term *partial regression coefficient* is used to make clear that it is the weight to be applied to an independent

[3] Again, the number of cases has been kept small to enable the reader to follow computations with ease. No advocacy of such small samples is intended (see sections on precision and power). We also present population estimates of variance and *sd*, rather than sample values, in conformity with computer statistical packages.

In this and the remaining chapters the dependent variable is identified as Y and the individual independent variables as X with a numerical subscript, that is X_1, X_2, etc. This makes it possible to represent independent variables by their subscripts only, for example B_{YX_3} becomes B_{Y3}.

[4] Hereafter we may refer to bivariate statistics such as correlations or regression coefficients as "zero-order" coefficients, in contrast to partial coefficients when other IVs are in the equation.

TABLE 3.2.1
Seniority, Publication, and Salary Data on 15 Faculty Members

	Time since Ph.D. (X_1)	No. of Publications (X_2)	Salary (Y)	
	3	18	$51,876	
	6	3	54,511	
	3	2	53,425	$r_{Y1} = .710$ ($r^2_{Y1} = .505$)
	8	17	61,863	$r_{Y2} = .588$ ($r^2_{Y2} = .346$)
	9	11	52,926	$r_{12} = .657$
	6	6	47,034	
	16	38	66,432	$\hat{Y}_1 = \$1{,}224X_1 + \$43{,}659$
	10	48	61,100	
	2	9	41,934	$\hat{Y}_2 = \$336X_2 + \$46{,}357$
	5	22	47,454	
	5	30	49,832	
	6	21	47,047	
	7	10	39,115	
	11	27	59,677	
	18	37	61,458	
M	7.67	19.93	$53,046	
sd	4.58	13.82	$8,166	

variable (IV) when one or more specified other IVs are also in the equation. Thus $B_{Y1.2}$ indicates the weight to be given X_1 when X_2 is also in the equation, $B_{Y2.13}$ is the X_2 weight when X_1 and X_3 are in the equation, $B_{Y4.123}$ is the X_4 weight when X_1, X_2, and X_3 are also used in the equation for Y, and so on. The weights for the IVs taken together with B_0 constitute the necessary constants for the linear regression equation.

When the regression equation is applied to the IV values for any given observation, the result will be an estimated value of the dependent variable ($\hat{Y}$). For any given set of data on which such an equation is determined, the resulting set of $\hat{Y}$ values will be as close to the observed Y values as possible, given a single weight for each IV. "As close as possible" is defined by the least squares principle.

For our example of estimating salary (Y) from time since Ph.D. (X_1) and number of publications (X_2), the full regression equation is

$$\hat{Y}_{12} = \$983X_1 + \$122X_2 + \$43{,}082, \tag{3.2.2}$$

where $983 is the partial regression coefficient $B_{Y1.2}$ for X_1 and $122 is the partial regression coefficient $B_{Y2.1}$ for X_2. The redundancy of information about Y carried by these two variables is reflected in the fact that the partial regression coefficients ($983 and $122) are each smaller in magnitude than their separate zero-order Bs ($1,224 and $336). We may interpret $B_{Y2.1} = \$122$ directly by stating that, for any given time since Ph.D. (X_1), on the average each additional publication is associated with an increase in salary of only $122 rather than the $336 that was found when time since Ph.D. were ignored. The $B_{Y1.2} = \$983$ may be similarly interpreted as indicating that, for faculty members with a given number of publications (X_2), on the average each additional year since Ph.D. is associated with an increase in salary of $983 rather than the $1224 that was found when number of publications was ignored. From a purely statistical point of view, these changes are a consequence of the redundancy of the two causal variables

(i.e., the tendency for faculty who have had their Ph.D.s longer to have more publications ($r_{12} = .657$); the partialing process controls for this tendency.[5] Viewed through the lens of causal analysis we see (particularly in the case of number of publications) how seriously we can be misled about the causal impact of a variable when we fail to include in our model other important causes. This, then, is an instance in which we have failed to consider the need to *isolate* the effects of a presumably causal variable from other correlated potential causes (Bollen, 1989).

Thus far, we have simply asserted that the regression equation for two or more IVs takes the same form as did the single IV case without demonstrating how the coefficients are obtained. As in the case of presenting correlation and regression with one IV, we initially standardize the variables to eliminate the effects of noncomparable raw (original) units. The regression equation for standardized variables[6] is

$$\hat{z}_Y = \beta_{Y1.2} z_1 + \beta_{Y2.1} z_2 \tag{3.2.3}$$

Just as r_{YX} is the standardized regression coefficient for estimating z_Y from z_X, $\beta_{Y1.2}$ and $\beta_{Y2.1}$ are the standardized partial regression coefficients for estimating z_Y from z_1 and z_2 with minimum squared error.

The equations for $\beta_{Y1.2}$ and $\beta_{Y2.1}$ can be proved via differential calculus to be

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

$$\beta_{Y2.1} = \frac{r_{Y2} - r_{Y1} r_{12}}{1 - r_{12}^2}. \tag{3.2.4}$$

A separation of the elements of this formula may aid understanding: r_{Y1} and r_{Y2} are "validity" coefficients, that is, the zero-order (simple) correlations of the IVs with the dependent variable. r_{12}^2 represents the variance in each IV shared with the other IV and reflects their redundancy. Thus, $\beta_{Y1.2}$ and $\beta_{Y2.1}$ are partial coefficients because each has been adjusted to allow for the correlation between X_1 and X_2.

To return to our academic example, the correlations between the variables are $r_{Y1} = .710$, $r_{Y2} = .588$, and $r_{12} = .657$. We determine by Eq. (3.2.4) that

$$\beta_{Y1.2} = \frac{.710 - (.588)(.657)}{1 - .657^2} = .570,$$

$$\beta_{Y2.1} = \frac{.588 - (.710)(.657)}{1 - .657^2} = .213,$$

and that the full regression equation for the standardized variables is therefore

$$\hat{z}_Y = .570 z_1 + .213 z_2.$$

[5]The terms *holding constant* or *controlling for*, *partialing the effects of*, or *residualizing* some other variables(s) indicate a mathematical procedure, of course, rather than an experimental one. Such terms are statisticians' shorthand for describing the average effect of a particular variable for any given values of the other variables.

[6]We employ the greek symbol β for the standardized coefficient in order to be consistent with the literature and with the earlier edition. It should not be confused with the other use of this symbol to indicate Type II errors of inference.

Once $\beta_{Y1.2}$ and $\beta_{Y2.1}$ have been determined, conversion to the original units is readily accomplished by

$$B_{Y1.2} = \beta_{Y1.2} \frac{sd_Y}{sd_1}$$

$$B_{Y2.1} = \beta_{Y2.1} \frac{sd_Y}{sd_2}. \tag{3.2.5}$$

Substituting the values for our running example (Table 3.2.1), we find

$$B_{Y1.2} = .570 \left(\frac{\$7622}{4.42} \right) = \$983$$

$$B_{Y2.1} = .213 \left(\frac{\$7622}{13.35} \right) = \$122.$$

Because we are again using the original units, we need a constant B_0 that serves to adjust for differences in means. This is calculated in the same way as with a single IV:

$$\begin{aligned} B_0 &= M_Y - B_{Y1.2} M_1 - B_{Y2.1} M_2 \\ &= \$53{,}046 - \$983(7.67) - \$122(19.93) \\ &= \$43{,}082. \end{aligned} \tag{3.2.6}$$

The full (raw score) regression equation for estimating academic salary is therefore

$$\hat{Y}_{12} = \$983 X_1 + \$122 X_2 + \$43{,}082,$$

and the resulting values are provided in the third column of Table 3.3.1 later in this chapter.

The partial regression coefficients, $B_{Y1.2} = \$983$ and $B_{Y2.1} = \$122$, are the empirical estimates, respectively, of h and g, the causal effects of our independent variables accompanying the arrows in the causal diagram (Fig. 3.1.1).

3.3 MEASURES OF ASSOCIATION WITH TWO INDEPENDENT VARIABLES

Just as there are partial regression coefficients for multiple regression equations (equations for predicting Y from more than one IV), so are there partial and multiple correlation coefficients that answer the same questions answered by the simple product moment correlation coefficient in the single IV case. These questions include the following:

1. How well does this group of IVs together estimate Y?
2. How much does any single variable add to the estimation of Y already accomplished by other variables?
3. When all other variables are held constant statistically, how much of Y does a given variable account for?

3.3.1 Multiple *R* and R^2

Just as r is the measure of association between two variables, so the multiple R is the measure of association between a dependent variable and an optimally weighted combination of two or

more IVs. Similarly, r^2 is the proportion of each variable's variance shared with the other, and R^2 is the proportion of the dependent variable's variance (sd_Y^2) shared with the optimally weighted IVs. Unlike r, however, R takes on only values between 0 and 1, with the former indicating no relationship with the IVs and the latter indicating a perfect relationship. (The reason that Rs are always positive becomes clear shortly.) The formula for the multiple correlation coefficient for two IVs as a function of the original rs is

$$R_{Y.12} = \sqrt{\frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2}}. \tag{3.3.1}$$

A similarity between the structure of this formula and the formula for β coefficients may lead the reader to suspect that R may be written as a function of these coefficients. This is indeed the case; an alternative formula is

$$R_{Y.12} = \sqrt{\beta_{Y1.2}r_{Y1} + \beta_{Y2.1}r_{Y2}}. \tag{3.3.2}$$

For the example illustrated in Table 3.1.1 the multiple correlation is thus, by Eq. (3.3.1),

$$R_{Y.12} = \sqrt{\frac{.5047 + .3455 - 2(.710)(.588)(.657)}{1 - .4313}},$$
$$= \sqrt{.5300} = .728$$

or by Eq. (3.3.2),

$$R_{Y.12} = \sqrt{.570(.710) + .213(.588)},$$
$$= \sqrt{.5300} = .728.$$

(We again remind the reader who checks the previous arithmetic and finds it "wrong" of our warning in Section 1.2.2 about rounding errors.)

We saw in Chapter 2 that the absolute value of the correlation between two variables $|r_{YX}|$ is equal to the correlation between Y and $\hat{Y}_X$. The multiple correlation is actually definable by this property. Thus, with two IVs,

$$R_{Y.12} = r_{Y\hat{Y}_{12}}, \tag{3.3.3}$$

and taking the example values in Table 3.3.1 we see that indeed $r_{Y\hat{Y}_{12}} = .728 = R_{Y.12}$. That $r_{Y\hat{Y}_{12}}$ and hence $R_{Y.12}$ cannot be negative can be seen from the fact that by the least squares criterion $\hat{Y}$ is as close as possible to Y.

The reader will again recall that r_{XY}^2 is the proportion of variance of Y shared with X. In exact parallel, $R_{Y.12}^2$ is the proportion of sd_Y^2 shared with the optimally weighted composite of X_1 and X_2. These optimal weights are, of course, those provided by the regression equation used to estimate Y. Thus,

$$R_{Y.12}^2 = \frac{sd_{Y.12}^2}{sd_Y^2} \tag{3.3.4}$$
$$= \frac{5549^2}{7622^2} = .5300;$$

TABLE 3.3.1
Actual, Estimated, and Residual Salaries

	1 Y	2 $\hat{Y}_1$	3 $\hat{Y}_{12}$	4 $Y - \hat{Y}_{12}$	5 $\hat{X}_{2.1}$	6 $X_2 - \hat{X}_{2.1}$	7 $Y - \hat{Y}_1$
	$51,876	$47,332	$48,223	$3,653	10.68	7.32	$4,544
	54,511	51,005	49,345	5,166	16.63	−13.63	3,506
	53,425	47,332	46,275	7,150	10.68	−8.68	6,093
	61,863	53,454	53,016	8,847	20.59	−3.59	8,409
	52,926	54,678	53,268	−342	22.58	−11.58	−1,752
	47,034	51,005	49,710	−2,676	16.63	−10.63	−3,971
	66,432	63,249	63,437	2,995	36.46	1.54	3,183
	61,100	55,903	58,757	2,343	24.56	23.44	5,197
	41,934	46,107	46,144	−4,210	8.70	.30	−4,173
	47,454	49,781	50,676	−3,222	14.64	7.36	−2,327
	49,832	49,781	51,651	−1,819	14.64	15.36	51
	47,047	51,005	51,537	−4,490	16.63	4.37	−3,958
	39,115	52,229	51,180	−12,065	18.61	−8.61	−13,114
	59,677	57,127	57,183	2,494	26.54	.46	2,550
	61,458	65,698	65,281	−3,823	40.42	−3.42	−4,240
M	$53,046	$53,046	$53,046	$0	19.93	0	$0
sd	$7,622	$5,415	$5,552	$5,227	8.77	10.07	$5,365

Correlations

	$\hat{Y}_1$	$\hat{Y}_{12}$	$X_2 - \hat{X}_{2.1}$
Y	$.710 = r_1$	$.728 = R_{Y.12}$	$.161 = sr_2$
$Y - \hat{Y}_1$	$0 = r_{(Y.1)1}$	.051	$.228 = pr_2$

that is, some 53% of the variance in salary (Y) is linearly accounted for by number of years since doctorate (X_1) and number of publications (X_2) in this sample.

Again in parallel with simple correlation and regression the variance of the residual, $Y - \hat{Y}_{12}$, is that portion of sd_Y^2 not linearly associated with X_1 and X_2. Therefore (and necessarily),

$$r_{\hat{Y}(Y-\hat{Y})} = 0, \tag{3.3.5}$$

and since such variances are additive,

$$sd_Y^2 = sd_{\hat{Y}_{12}}^2 + sd_{Y-\hat{Y}_{12}}^2. \tag{3.3.6}$$

It should also be apparent at this point that a multiple R can never be smaller than the absolute value of the largest correlation of Y with the IVs and must be almost invariably larger. The optimal estimation of $\hat{Y}_{12}$ under circumstances in which X_2 adds nothing to X_1's estimation of $\hat{Y}$ would involve a 0 weight for X_2 and thus $R_{Y.12}$ would equal $|r_{Y1}|$, the absolute value of r_{Y1}. Any slight departure of X_2 values from this rare circumstance necessarily leads to some (perhaps trivial) increase in $R_{Y.12}$ over $|r_{Y1}|$.

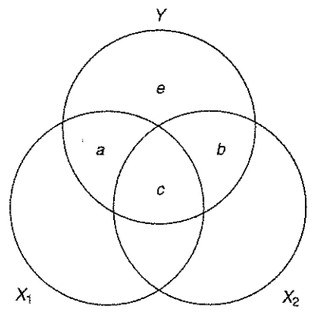
As with bivariate correlation the square root of the proportion of Y variance *not* associated with the IVs is called the *coefficient of* (multiple) *alienation*. This value is $\sqrt{1 - R^2} = \sqrt{1 - .5300} = .686$ for these data.

3.3.2 Semipartial Correlation Coefficients and Increments to R^2

One of the important problems that arises in MRC is that of defining the contribution of each IV in the multiple correlation. We shall see that the solution to this problem is not so straightforward as in the case of a single IV, the choice of coefficient depending on the substantive reasoning underlying the exact formulation of the research questions. One answer is provided by the semipartial correlation coefficient sr and its square, sr^2. To understand the meaning of these coefficients, it is useful to consider the "ballantine." Recall that in the diagrammatic representation of Fig. 2.6.1 the variance of each variable was represented by a circle of unit area. The overlapping area of two circles represents their relationship as r^2. With Y and two IVs represented in this way, the total area of Y covered by the X_1 and X_2 circles represents the proportion of Y's variance accounted for by the two IVs, $R^2_{Y.12}$.

Figure 3.3.1 shows that this area is equal to the sum of areas designated a, b, and c. The areas a and b represent those portions of Y overlapped uniquely by IVs X_1 and X_2, respectively, whereas area c represents their simultaneous overlap with Y. The "unique" areas, expressed as proportions of Y variance, are squared semipartial correlation coefficients, and each equals the increase in the squared multiple correlation that occurs when the variable is added to the other IV.[7] Thus

(3.3.7)
$$a = sr_1^2 = R^2_{Y.12} - r^2_{Y2},$$
$$b = sr_2^2 = R^2_{Y.12} - r^2_{Y1}.$$



$$r^2_{Y1} = a + c$$
$$r^2_{Y2} = b + c$$
$$R^2_{Y.12} = a + b + c$$

FIGURE 3.3.1 The ballantine for X_1 and X_2 with Y.

[7] Throughout the remainder of the book, whenever possible without ambiguity, partial coefficients are subscripted by the relevant independent variable only, it being understood that Y is the dependent variable and that all other IVs have been partialed. In this expression (i) indicates that X_i is not included in the variables X_1 to X_k that are being partialed. Thus, $sr_i = r_{Y(i.)12\ldots(i)\ldots k)}$, the correlation between Y and X_i from which all other IVs in the set under consideration have been partialed. Similarly, R without subscript refers to $R_{Y.12\ldots k}$.

A formula for sr for the two IV case may be given as a function of zero-order rs as

$$sr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r^2_{12}}}$$

(3.3.8) and

$$sr_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1 - r^2_{12}}}.$$

For our running example (Table 3.2.1), these values are

$$sr_1 = \frac{.710 - .588(.657)}{\sqrt{1 - .657^2}} = .430,$$
$$sr_1^2 = .1850$$

or, by Eq. (3.3.7)

$$sr_1^2 = .5305 - .3455 = .1850.$$

For X_2,

$$sr_2 = \frac{.588 - .710(.657)}{\sqrt{1 - .657^2}} = .161$$
$$sr_2^2 = .0258,$$

or, by Eq. (3.3.7),

$$sr_2^2 = .5305 - .5047 = .0258.$$

The semipartial correlation sr_1 is the correlation between all of Y and X_1 from which X_2 has been partialed. It is a *semi*partial correlation because the effects of X_2 have been removed from X_1 but not from Y. Recalling that in this system "removing the effect" is equivalent to subtracting from X_1 the X_1 values estimated from X_2, that is, to be working with $X_1 - \hat{X}_{1.2}$, we see that another way to write this relationship is

(3.3.9)
$$sr_1 = r_{Y(X_1 - \hat{X}_{1.2})}.$$

Another notational form of sr_1 used is $r_{Y(1.2)}$, the 1.2 being a shorthand way of expressing "X_1 from which X_2 has been partialed," or $X_1 - \hat{X}_{1.2}$. It is a convenience to use this dot notation to identify which is being partialed from what, particularly in subscripts, and it is employed whenever necessary to avoid ambiguity. Thus $i \cdot j$ means *i from which j is partialed.* Note also that in the literature the term *part* correlation is sometimes used to denote semipartial correlation.

In Table 3.3.1 we present the $X_2 - \hat{X}_{2.1}$ (residual) values for each case in the example in which salary was estimated from publications and time since Ph.D. The correlation between these residual values and Y is seen to equal .4301, which is sr_1; and $.4301^2 = .1850 = sr_1^2$, as before.

To return to the ballantine (Fig. 3.3.1) we see that for our example, area $a = .1850$, $b = .0258$, and $a + b + c = R^2_{Y.12} = .5305$. It is tempting to calculate c (by $c = R^2_{Y.12} - sr_1^2 - sr_2^2$) and interpret it as the proportion of Y variance estimated jointly or redundantly by X_1 and X_2. However, any such interpretation runs into a serious catch—there is nothing in the mathematics that prevents c from being a negative value, and a negative proportion of

variance hardly makes sense. Because c is not necessarily positive, we forgo interpreting it as a proportion of variance. A discussion of the circumstances in which c is negative is found in Section 3.4. On the other hand, a and b can never be negative and are appropriately considered proportions of variance; each represents the increase in the proportion of Y variance accounted for by the addition of the corresponding variable to the equation estimating Y.

3.3.3 Partial Correlation Coefficients

Another kind of solution to the problem of describing each IV's participation in determining R is given by the *partial* correlation coefficient pr_1, and its square, pr_1^2. The squared partial correlation may be understood best as that proportion of sd_Y^2 *not* associated with X_2 that *is* associated with X_1. Returning to the ballantine (Fig. 3.3.1), we see that

$$pr_1^2 = \frac{a}{a+e} = \frac{R_{Y.12}^2 - r_{Y2}^2}{1 - r_{Y2}^2}$$

(3.3.10)

$$pr_2^2 = \frac{b}{b+e} = \frac{R_{Y.12}^2 - r_{Y1}^2}{1 - r_{Y1}^2}.$$

The a area or numerator for pr_1^2 is the squared semipartial correlation coefficient sr_1^2; however, the base includes not all the variance of Y as in sr_1^2 but only that portion of Y variance that is not associated with X_2, that is, $1 - r_{Y2}^2$. Thus, this squared partial r answers the question, How much of the Y variance that is not estimated by the other IVs in the equation is estimated by this variable? Interchanging X_1 and X_2 (and areas a and b), we similarly interpret pr_2^2. In our faculty salary example, we see that by Eqs. (3.3.10)

$$pr_1^2 = \frac{.5305 - .3455}{1 - .3455} = \frac{.1850}{.6545} = .2826$$

$$pr_2^2 = \frac{.5305 - .5046}{1 - .4312} = \frac{.0259}{.5688} = .0455$$

Obviously, because the denominator cannot be greater than 1, partial correlations will be larger than semipartial correlations, except in the limiting case when other IVs are correlated 0 with Y, in which case $sr = pr$.

pr may be found more directly as a function of zero-order correlations by

$$pr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r_{Y2}^2}\sqrt{1 - r_{12}^2}}$$

(3.3.11)

$$pr_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1 - r_{Y1}^2}\sqrt{1 - r_{12}^2}}.$$

For our example

$$pr_1 = \frac{.710 - .588(.657)}{\sqrt{1 - .3455}\sqrt{1 - .4312}} = .5316$$

and $pr_1^2 = .5316^2 = .2826$, as before;

$$pr_2 = \frac{.588 - .710(.657)}{\sqrt{1 - .5047}\sqrt{1 - .4312}} = .2133$$

and $pr_2^2 = .2133^2 = .0455$, again as before.

In Table 3.3.1 we demonstrate that pr_2 is literally the correlation between X_2 from which X_1 has been partialed (i.e., $X_2 - \hat{X}_{2.1}$) and Y from which X_1 has also been partialed (i.e., $Y - \hat{Y}_1$). Column 6 presents the partialed X_2 values, the residuals from $\hat{X}_{2.1}$. Column 7 presents the residuals from Y_1 (given in column 2). The simple correlation between the residuals in columns 6 and 7 is .2133 $= pr_2$ (the computation is left to the reader, as an exercise). We thus see that the partial correlation for X_2 is literally the correlation between Y and X_2, each similarly residualized from X_1. A frequently employed form of notation to express the partial r is $r_{Y2.1}$, which conveys that X_1 is being partialed from both Y and X_2 (i.e., $r_{(Y.1)(2.1)}$), in contrast to the semipartial r, which is represented as $r_{Y(2.1)}$.

Before leaving Table 3.3.1, the other correlations at the bottom are worth noting. The r of Y with $\hat{Y}_1$ of .710 is identically r_{Y1} and necessarily so, since $\hat{Y}_1$ is a linear transformation of X_1 and therefore must correlate exactly as X_1 does. Similarly, the r of Y with $\hat{Y}_{12}$ of .728 is identically $R_{Y.12}$ and necessarily so, by definition in Eq. (3.3.3). Also, $Y - \hat{Y}_1$ (that is, $Y \cdot X_1$) correlates zero with $\hat{Y}_1$, because when a variable (here X_1) is partialed from another (here Y), the residual will correlate zero with any linear transformation of the partialed variables. Here, $\hat{Y}_1$ is a linear transformation of X_1 (i.e., $\hat{Y}_1 = B_1X_1 + B_0$).

Summarizing the results for the running example, we found $sr_1^2 = .1850$, $pr_1^2 = .2826$ and $sr_2^2 = .0258$, $pr_2^2 = .0522$. Whichever base we use, it is clear that number of publications (X_2) has virtually no *unique* relationship to salary, that is, no relationship beyond what can be accounted for by time since doctorate (X_1). On the other hand, time since doctorate (X_1) is uniquely related to salary (sr_1) and to salary holding publications constant (pr_1) to a quite substantial degree. The reader is reminded that this example is fictitious, and any resemblance to real academic departments, living or dead, is mostly coincidental.

3.4 PATTERNS OF ASSOCIATION BETWEEN *Y* AND TWO INDEPENDENT VARIABLES

A solid grasp of the implications of all possible relationships among one dependent variable and two independent variables is fundamental to understanding and interpreting the various multiple and partial coefficients encountered in MRC. This section is devoted to an exposition of each of these patterns and its distinctive substantive interpretation in actual research.

3.4.1 Direct and Indirect Effects

As we have stated, the regression coefficients $B_{Y1.2}$ and $B_{Y2.1}$ estimate the causal effects of X_1 and X_2 on Y in the causal model given in Fig. 3.4.1, Model A. These coefficients, labeled f and g in the diagram, are actually estimates of the *direct effects* of X_1 and X_2, respectively. Direct effects are exactly what the name implies—causal effects that are not mediated by any other variables in the model. All causes, of course, are mediated by some intervening mechanisms. If such an intervening variable is included, we have Model B shown in Fig. 3.4.1. In this diagram X_1 is shown as having a causal effect on X_2. Both variables have direct effects on Y. However, X_1 also has an *indirect* effect on Y via X_2. Note that the difference between Models A and B is not in the mathematics of the regression coefficients but in the understanding of the substantive causal process.

The advantage of Model B, if it is valid, is that in addition to determining the *direct* effects of X_1 and X_2 on Y, one may estimate the *indirect* effects of X_1 on Y as well as the effect of X_1 on X_2. This latter (h) in Model B is, of course, estimated by the regression coefficient of X_2 on X_1, namely B_{21}. The direct effects, f and g, are the same in both Models A and B and

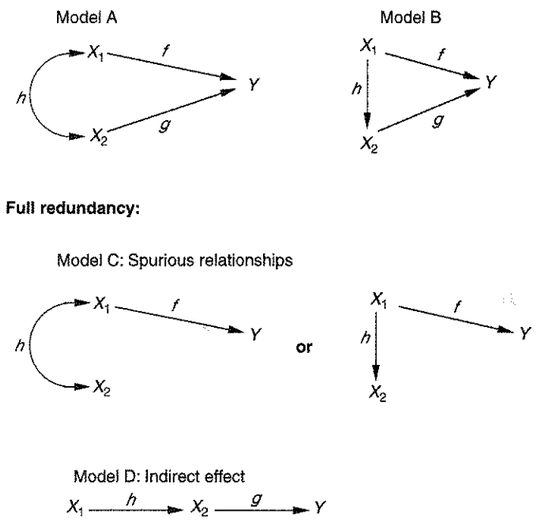


FIGURE 3.4.1 Representation of relationships between Y and two IVs.

are estimated by the sample regression coefficients for X_1 and X_2 from the equation for Y. The relationship between two exogenous variables, h in Model A, is conventionally represented by the correlation between the variables. The magnitude of the indirect effect of X_1 on Y in Model B may also be estimated by a method described in Chapter 11.

3.4.2 Partial Redundancy

We have included Models A and B under the rubric *partial redundancy* because this is by far the most common pattern of relationship in nonexperimental research in the behavioral sciences. It occurs whenever $r_{Y1} > r_{Y2}r_{12}$ and $r_{Y2} > r_{Y1}r_{12}$ [see Eqs. (3.2.4), (3.3.8), and (3.3.11)], once the variables have been oriented so as to produce positive correlations with Y. The sr_i and β_i for each IV will be smaller than its r_{Yi} (and will have the same sign) and thus reflect the fact of redundancy. Each IV is at least partly carrying information about Y that is also being supplied by the other. This is the same model shown by the ballantine in Fig. 3.3.1. We consider another situation in which r_{Y2} is negative in the next section.

Examples of Model A two-variable redundancy come easily to mind. It occurs when one relates school achievement (Y) to parental income (X_1) and education (X_2), or delinquency (Y) to IQ (X_1) and school achievement (X_2), or psychiatric prognosis (Y) to rated symptom severity (X_1) and functional impairment (X_2), or—but the reader can supply many examples of his or her own. Indeed, redundancy among explanatory variables is the plague of our efforts to understand the causal structure that underlies observations in the behavioral and social sciences.

Model B two-variable redundancy is also a very common phenomenon. Some substantive examples are given in Fig. 3.4.2. Here we see that age is expected to produce differences in

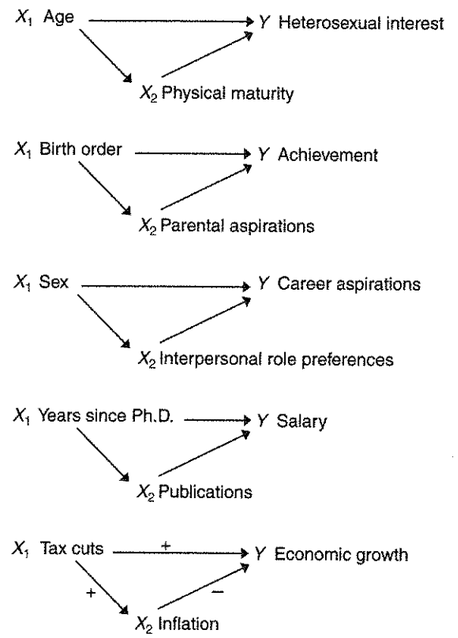


FIGURE 3.4.2 Examples of causal Model B.

physical maturity in a sample of school children and that each is expected to cause differences in heterosexual interest. The presence of an arrow from age to heterosexual interest implies that physical maturity is not the only reason why heterosexual interest increases with age. Birth order of offspring is expected to produce differences in parental aspirations and both are causally related to achievement. We might expect sex differences in interpersonal role preferences and that both of these variables will produce differences in career aspirations. Also, for our running example, we expect the passage of time since Ph.D. to produce increases in the number of publications and increases in both of these variables to produce increases in salary.

3.4.3 Suppression in Regression Models

In each of the causal circumstances we have discussed, we expect the *direct* effects of the variables to be smaller than the zero-order (unpartialed) effects. In addition, we anticipate an *indirect* effect of our X_1 variables to take place via the X_2 variables. Although partial redundancy is the most commonly observed pattern for causal Models A and B, it is not the only possible model. *Suppression* is present when either r_{Y1} or r_{Y2} is less than the product of the other with r_{12}, or when r_{12} is negative (assuming, as throughout, positive r_{Y1} and r_{Y2}). In this case the partialed coefficients of X_1 and X_2 will be larger in value than the zero-order coefficients and one of the partialed (direct effect) coefficients may become negative.

The term *suppression* can be understood to indicate that the relationship between the independent or causal variables is hiding or suppressing their real relationships with Y, which would be larger or possibly of opposite sign were they not correlated. In the classic psychometric literature on personnel selection, the term suppression was used to describe a variable (such as verbal ability) X_2 that, although not correlated with the criterion Y (e.g., job performance), is correlated with the available measure of the predictor X_1 (e.g., a paper and pencil test of job skills) and thus adds irrelevant variance to it and reduces its relationship with Y. The inclusion of the suppressor in the regression equation removes (suppresses) the unwanted variance in X_1, in effect, and enhances the relationship between X_1 and Y by means of $B_{Y1.2}$. This topic is discussed again in Chapter 12.

For a substantive example, suppose a researcher is interested in the roles of social assertiveness and record-keeping skills in producing success as a salesperson. Measures of these two characteristics are devised and administered to a sample of employees. The correlation between the measure of social assertiveness (X_1) and sales success (Y) is found to be $+.403$, the correlation between record keeping (X_2) and $Y = +.127$ and $r_{12} = -.305$, indicating an overall tendency for those high on social assertiveness to be relatively low on record keeping, although each is a desirable trait for sales success. Because $-.305 < (.403)(.127)$ we know that the situation is one of suppression and we may expect the direct effects (the regression and associated standardized coefficients) to be larger than the zero-order effects. Indeed, the reader may confirm that the β coefficients are .487 for social assertiveness and .275 for record keeping, both larger than their respective correlations with Y, .403 and .127. The coefficients may be considered to reflect appropriately the causal effects, the zero-order effects being misleadingly small because of the negative relationship between the variables.

A Model B example of suppression may be found in the (overly simple) economic model shown in Fig. 3.4.2, in which tax cuts are expected to produce increases in economic growth but also inflation. Because inflation is expected to have negative effects on economic growth, one can only hope that the direct positive effects of the tax cuts on economic growth will exceed the indirect negative effect attributable to the effect on inflation.

Suppression is a plausible model for many homeostatic mechanisms, both biological and social, in which force and counterforce tend to occur together and have counteractive effects. The fact that suppression is rarely identified in simple models may be due to the difficulty in finding appropriate time points for measuring X_1, X_2, and Y. Suppression effects of modest magnitude are more common in complex models. Material suppression effects are likely to be found in analyses of aggregate data, when the variables are sums or averages of many observations and R^2s are likely to approach 1 because of the small error variance that results in these conditions. Tzelgov and Henik (1991) provide an extensive discussion of conditions under which suppression occurs.

3.4.4 Spurious Effects and Entirely Indirect Effects

Model C in Fig. 3.4.1 describes the special case in which $r_{Y2} = r_{Y1}r_{12}$. This model is of considerable interest because it means that the information with regard to Y carried by X_2 is *completely* redundant with that carried by X_1. This occurs whenever the B, sr, and pr coefficients for X_2 are approximately zero. This occurs when their numerators are approximately zero (i.e., when $r_{Y2} \approx r_{12}r_{Y1}$). For the causal model the appropriate conclusion is that X_2 is not a cause of Y at all but merely associated (correlated) with Y because of its association with X_1. In some fields such as epidemiology, X_1 is referred to as a *confounder* of the relationship between X_2 and Y. (But note the appropriate considerations before drawing such a conclusion from sample results, as discussed in Section 3.7.) A great many analyses are carried out precisely to

determine this issue—whether some variable has a demonstrable effect on Y when correlated variables are held constant or, alternatively, whether the variable's relationship to Y is (or may be) spurious. Thus, for example, a number of investigations have been carried out to determine whether there is a family size (X_2) influence on intelligence (Y) independent of parental social class (X_1), whether maternal nutrition (X_2) has an effect on infant behavior (Y) independent of maternal substance use (X_1), whether the status of women (X_2) in various countries has an effect on fertility rate (Y) independent of economic development (X_1), or indeed whether any of the X_2 effects on Y shown in Fig. 3.4.2 are nil. Generally, the question to be answered is the "nothing but" challenge. Is the relationship between Y and X_2 nothing but a manifestation of the causal effects of X_1?

Complete redundancy, however, does not always imply a spurious relationship. In Fig. 3.4.1, Model D we see a situation in which the partial coefficients for X_1 approach zero, indicating correctly that there is no *direct* effect of X_1 on Y. There is, however, an *indirect* effect that, according to the model, takes place entirely via X_2; that is, the effect of X_1 is mediated by X_2.

Many investigations are designed to answer questions about intervening mechanisms—for example, is the higher female (X_1) prevalence of depression (Y) entirely attributable to lower female income/opportunity structure (X_2)? Are ethnic (X_1) differences in achievement (Y) entirely due to economic deprivation (X_2)? Is the demonstrable effect of poor parent marital relationship (X_1) on delinquency (Y) entirely attributable to poor parent-child relationships (X_2)? In these cases the relationships between X_1 and Y cannot be said to be spurious but are nevertheless likely to have different theoretical implications and policy import when they are entirely redundant than when they are not.

As in the case of the comparison of Models A and B, the difference between Models C and D lie not in the coefficients but in one's understanding of the causal processes that gave rise to the coefficients. Again, one can only demonstrate consistency of sample data with a model rather than prove the model's correctness.

3.5 MULTIPLE REGRESSION/CORRELATION WITH k INDEPENDENT VARIABLES

3.5.1 Introduction: Components of the Prediction Equation

When more than two IVs are related to Y, the computation and interpretations of multiple and partial coefficients proceed by direct extension of the two-IV case. The goal is again to produce a regression equation for the k IVs of the (raw score) form

$$\hat{Y} = B_{Y1.23\ldots k}X_1 + B_{Y2.13\ldots k}X_2 + B_{Y3.12\ldots k}X_3 + \ldots + B_{Yk.123\ldots k-1}X_k + B_{Y0.123\ldots k}, \tag{3.5.1}$$

or, expressed in simpler subscript notation,

$$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_k + B_0,$$

or, as in the simple two variable equation, expressed in terms of the original Y plus the errors of prediction e:

$$Y = B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_kX_k + B_0 + e.$$

When this equation is applied to the data, it yields a set of $\hat{Y}$ values (one for each of the n cases) for which the sum of the $(Y - \hat{Y})^2$ values over all n cases will (again) be a minimum. Obtaining these raw-score partial regression weights, the B_i, involves solving a set of k simultaneous equations in k unknowns, a task best left to a computer program, although

Appendix 2 provides the method of hand calculation for MRC.) The purpose of the present section is to lay down a foundation for understanding the various types of coefficients produced by MRC for the general case of k independent variables, and their relationship to various MRC strategies appropriate to the investigator's research goals.

3.5.2 Partial Regression Coefficients

By direct extension of the one- and two-IV cases, the raw score partial regression coefficient B_i $(= B_{Yi.123...(i)...k})$ is the constant weight by which each value of the variable X_i is to be multiplied in the multiple regression equation that includes all k IVs. Thus, B_i is the average or expected change in Y for each unit increase in X_i when the value of each of the $k-1$ other IVs is held constant. β_i is the partial regression coefficient when all variables have been standardized. Such standardized coefficients are of interpretive interest when the analysis concerns test scores or indices whose scaling is arbitrary, or when the magnitudes of effects of variables in different units are to be compared.


CH03EX02

For example, let us return to the study in which we seek to account for differences in salary in a university department by means of characteristics of the faculty members. The two IVs used thus far were the number of years since each faculty member had received a doctoral degree (X_1) and the number of publications (X_2). We now wish to consider two additional independent variables, the gender of the professor and the number of citations of his or her work in the scientific literature. These data are presented in Table 3.5.1, where sex (X_3) is coded (scored) 1 for female and 0 for male, and the sample size has been increased to 62 as a more reasonable size for analysis. The correlation matrix shows that sex is negatively correlated with salary ($r_{Y3} = -.210$), women having lower salaries on the average than men. The number of citations (X_4) is positively associated with salary ($r_{Y4} = .550$), as well as with the other IVs. Sex correlates very little with the other IVs, except for a tendency in these data for women to be more recent Ph.D.s than men ($r_{13} = -.201$).

The (raw-score) multiple regression equation for estimating academic salary from these four IVs may be obtained from computer output or by the matrix inversion method of Appendix 2 (where this problem is used illustratively). It is $\hat{Y} = \$857X_1$ (time) $+ \$92.8$ (publications) $- \$918X_3$ (female) $+ \$202X_4$ (citations) $+ \$39{,}587$. These partial B_i coefficients indicate that for any given values of the other IVs, an increase of one in the number of citations is associated with a salary increase of \$202 $(= B_4)$; an increase of one unit in X_3, and hence the average difference in salary (holding constant the other IVs) is $-\$918$ (favoring men); and the effects of an additional year since degree (X_1) and an increase of one publication (X_2) are \$857 and \$93, respectively. Note also that $B_0 = \$39{,}587$ is the estimated salary of a hypothetical male professor fresh from his doctorate with no publications or citations, that is, all $X_i = 0$.

In this problem, the salary estimated by the four IVs for the first faculty member (Table 3.5.1) is

$$\begin{aligned}\hat{Y} &= \$857(3) + \$92.8(18) - \$918(1) + \$202(50) + \$39{,}587 \\ &= \$2{,}571 + \$1{,}670 - \$918 + \$10{,}100 + \$39{,}587 \\ &= \$53{,}007.^{8}\end{aligned}$$

The remaining estimated values are given in the last column of Table 3.5.1.

[8] Within rounding error.

TABLE 3.5.1
Illustrative Data With Four Independent Variables

I.D.	Time since Ph.D. (X_1)	No. of publications (X_2)	Sex (X_3)	No. of citations (X_4)	Salary (Y)	Estimated Salary
01	3	18	1	50	\$51,876	\$53,007
02	6	3	1	26	54,511	49,340
03	3	2	1	50	53,425	51,523
04	8	17	0	34	61,863	54,886
05	9	11	1	41	52,926	55,682
06	6	6	0	37	47,034	52,757
07	16	38	0	48	66,432	66,517
08	10	48	0	56	61,100	63,917
09	2	9	0	19	41,934	45,973
10	5	22	0	29	47,454	51,769
11	5	30	1	28	49,832	51,391
12	6	21	0	31	47,047	52,937
13	7	10	1	25	39,115	50,644
14	11	27	0	40	59,677	59,596
15	18	37	0	61	61,458	70,763
16	6	8	0	32	54,528	51,933
17	9	13	1	36	60,327	54,858
18	7	6	0	69	56,600	60,076
19	7	12	1	47	52,542	55,272
20	3	29	1	29	50,455	49,786
21	7	29	1	35	51,647	54,426
22	5	7	0	35	62,895	51,589
23	7	6	0	18	53,740	49,778
24	13	69	0	90	75,822	75,302
25	5	11	0	60	56,596	57,008
26	8	9	1	30	55,682	52,418
27	8	20	1	27	62,091	52,833
28	7	41	1	35	42,162	55,539
29	2	3	1	14	52,646	43,489
30	13	27	0	56	74,199	64,541
31	5	14	0	50	50,729	55,267
32	3	23	0	25	70,011	49,340
33	1	1	0	35	37,939	47,605
34	3	7	0	1	39,652	43,010
35	9	19	0	69	68,987	62,996
36	3	11	0	69	55,579	57,112
37	9	31	0	27	54,671	55,628
38	3	9	0	50	57,704	53,090
39	4	12	1	32	44,045	49,672
40	10	32	0	33	51,122	57,789
41	1	26	0	45	47,082	51,943
42	11	12	0	54	60,009	61,032
43	5	9	0	47	58,632	54,198
44	1	6	0	29	38,340	46,857
45	21	39	0	69	71,219	75,135
46	7	16	1	47	53,712	55,643
47	5	12	1	43	54,782	52,751
48	16	50	0	55	83,503	69,043

TABLE 3.5.1 (continued)

I.D.	Time since Ph.D. (X_1)	No. of publications (X_2)	Sex (X_3)	No. of citations (X_4)	Salary (Y)	Estimated Salary
49	5	18	0	33	47,212	52,206
50	4	16	1	28	52,840	49,236
51	5	5	0	42	53,650	52,817
52	11	20	0	24	50,931	55,716
53	16	50	1	31	66,784	63,279
54	3	6	1	27	49,751	47,249
55	4	19	1	83	74,343	60,620
56	4	11	1	49	57,710	53,012
57	5	13	0	14	52,676	47,905
58	6	3	1	36	41,195	51,359
59	4	8	1	34	45,662	49,705
60	8	11	1	70	47,606	60,681
61	3	25	1	27	44,301	49,011
62	4	4	1	28	58,582	48,123
M	6.79	18.18	.56	40.23	$54,816	$54,816
sd	4.28	14.00	.50	17.17	$9,706	$6,840

Correlation Matrix

	Time since Ph.D. (X_1)	No. of publications (X_2)	Sex (X_3)	No. of citations (X_4)	Salary (Y)
Time since Ph.D.	1.000	.651	−.210	.373	.608
No. of publications	.651	1.000	−.159	.333	.506
Sex	−.210	−.159	1.000	−.149	−.201
No. of citations	.373	.333	−.149	1.000	.550
Salary	.608	.506	−.201	.550	1.000

Standardized Partial Regression Coefficients

The regression equation may be written in terms of standardized variables and β coefficients as

$$\hat{z}_Y = .378z_1 + .134z_2 - .047z_3 + .357z_4.$$

The β values may always be found from B values by inverting Eq. (3.2.5):

$$\beta_i = B_i \frac{sd_i}{sd_Y}. \tag{3.5.2}$$

for example, $\beta_4 = .202(17.17/9706) = .357$. As always, with standardized Y and IVs the intercept β_0 is necessarily zero, and thus may be omitted.

3.5.3 R, R^2 and Shrunken R^2

Multiple R and R^2

Application of the regression equation to the IVs yields a set of estimated $\hat{Y}$ values. The simple product moment correlation of Y with $\hat{Y}$ equals the multiple correlation; in this example,

$r_{Y\hat{Y}} = R = .709$. As with the one- or two-IV case, R^2 is the proportion of Y variance accounted for and $R^2 = sd^2_{\hat{Y}}/sd^2_Y = (6885)^2/(9706)^2 = .5032$.

R^2 may also be written as a function of the original correlations with Y and the β coefficients by extension of Eq. (3.3.2):

$$R^2_{Y.12\ldots k} = \sum \beta_i r_{Yi}, \tag{3.5.3}$$

where the summation is over the k IVs. Thus in the current example,

$$R^2_{Y.1234} = .378(.608) + .134(.506) - .047(.201) + .357(.550) = .5032,$$

as before.

Lest the reader think that this represents a way of apportioning the Y variance accounted for among the IVs (that is, that X_i's proportion is its $\beta_i r_{Yi}$), it is important to recall that β_i and r_{Yi} may be of opposite sign (under conditions of suppression). Thus, the suppressor variable on this interpretation would appear to account for a negative proportion of the Y variance, clearly a conceptual impossibility. The fact that $\beta_i r_{Yi}$ is not necessarily positive is sufficient to preclude the use of Eq. (3.5.3) as a variance partitioning procedure.

R^2 may also be obtained as a function of the β coefficients and the associations between the IVs as

$$R^2 = \sum(\beta_i^2) + 2\sum(\beta_i \beta_j r_{ij}) \tag{3.5.4}$$

where the first summation is over the k IVs, and the second over the $k(k-1)/2$ distinct pairs of IVs. In the current problem,

$$\begin{aligned} R^2_{Y.1234} = {} & .378^2 + .134^2 + .047^2 + .357^2 + 2[(.378)(.134)(.651) \\ & + (.378)(.047)(.210) + (.378)(.357)(.373) + (.134)(.047)(.159) \\ & + (.134)(.357)(.333) + (.047)(.357)(.150)] = .5032 \end{aligned}$$

This formula appears to partition R^2 into portions accounted for by each variable uniquely and portions accounted for jointly by pairs of variables, and some authors so treat it. However, we again note that any of the $k(k-1)/2$ terms $\beta_i \beta_j r_i$ may be negative. Therefore, neither Eq. (3.5.4) nor Eq. (3.5.3) can serve as variance partitioning schemes. This equation does, however, make clear what happens when all correlations between pairs of IVs equal 0. The triple-product terms will all contain $r_{ij} = 0$ and hence drop out, and $R^2 = \Sigma\beta_i^2 = \Sigma r^2_{Yi}$, as was seen for the two-IV case (Section 3.4.2).

Shrunken or Adjusted R^2: Estimating the Population ρ^2

The R^2 that we obtain from a given sample is not an unbiased estimate of the population squared multiple correlation, ρ^2. To gain an intuitive understanding of part of the reason for this, imagine the case in which one or more of the IVs account for no Y variance in the population, that is, $r^2_{Yi} = 0$ in the population for one or more X_i. Because of random sampling fluctuations we would expect that only very rarely would its r^2 with Y in a sample be exactly zero; it will virtually always have some positive value. (Note that although r can be negative, neither r^2 nor R^2 can be.). Thus, in most samples it would make some (possibly trivial) contribution to R^2. The smaller the sample size, the larger these positive variations from zero will be, on the average, and thus the greater the inflation of the sample R^2. Similarly, the more IVs we have, the more opportunity for the sample R^2 to be larger than the true population ρ^2. It is often desirable to have an estimate of the population ρ^2 and we naturally prefer one that is more

accurate than the positively biased sample R^2. Such a realistic estimate of the population ρ^2 (for the fixed model) is given by

$$\tilde{R}^2_Y = 1 - (1 - R^2_Y)\frac{n-1}{n-k-1} \tag{3.5.5}$$

This estimate is necessarily (and appropriately) smaller than the sample R^2 and is thus often referred to as the "shrunken" R^2. The magnitude of the "shrinkage" will be larger for small values of R^2 than for larger values, other things being equal. Shrinkage will also be larger as the ratio of the number of IVs to the number of subjects increases. As an example, consider the shrinkage in R^2 when $n = 200$ and cases where $k = 5$, 10, and 20 IVs, thus yielding k/n ratios of 1/40, 1/20, and 1/10, respectively. When $R^2 = .20$, the shrunken values will equal, respectively, .1794, .1577, and .1106, the last being a shrinkage of almost one-half. When $R^2 = .40$, the comparable values are, respectively, .3845, .3683, and .3330, smaller shrinkage either as differences from or proportions of R^2. For large ratios of k/n and small R^2, these shrunken values may be negative; for example, for $R^2 = .10$, $k = 11$, $n = 100$, Eq. (3.6.3) gives $-.0125$. In such cases, by convention, the shrunken R^2 is reported as zero.

It should be clear from this discussion that whenever a subset of IVs has been selected post hoc from a larger set of potential variables on the basis of their relationships with Y, not only R^2, but even the shrunken R^2 computed by taking as k the number of IVs selected, will be too large. This is true whether the computer capitalizes on chance by performing a stepwise regression, or the experimenter does so by selecting IVs with relatively larger r_{Yi}s. A more realistic estimate of shrinkage is obtained by substituting for k in Eq. (3.6.3) the *total* number of IVs from which the selection was made.

3.5.4 *sr* and sr^2

The semipartial correlation coefficient sr and its square sr^2 in the general case of k IVs may be interpreted by direct extension of the two IV case. Thus sr_i^2 equals that proportion of the Y variance accounted for by X_i beyond that accounted for by the other $k - 1$ IVs, and

$$sr_i^2 = R^2_{Y.12\ldots i\ldots k} - R^2_{Y.12\ldots(i)\ldots k}, \tag{3.5.6}$$

(the parenthetical i signifying its omission from the second R^2), or the increase in the squared multiple correlations when X_i is included over the R^2 that includes the other $k - 1$ IVs, but excludes X_i. This may be thought of as the unique contribution of X_i to R^2 in the context of the remaining $k - 1$ IVs. As in the two-IV case, the semipartial r equals the correlation between that portion of X_i that is uncorrelated with the remaining IVs and Y:

$$\begin{aligned} sr_i &= r_{Y(i.12\ldots(i)\ldots k)} \\ &= r_{Y(X-\hat{X}_i.12\ldots(i)\ldots k)}. \end{aligned} \tag{3.5.7}$$

As might be expected, sr_i may also be written as a function of the multiple correlation of the other IVs with X_i,

$$sr_i = \beta_i \sqrt{1 - R^2_{i.12\ldots(i)\ldots k}}. \tag{3.5.8}$$

Neither sr_i nor sr_i^2 is provided as default output by most MRC computer programs; however, the term $1 - R^2_{i.12\ldots(i)\ldots k}$ is often provided. This term, called the variable's *tolerance*, alerts the data analyst to the level of redundancy of this variable with other predictors.[9] Occasionally sr_i^2 values are provided, possibly labeled as the "unique" contribution to R^2. When pr_i is available, sr_i^2 is readily determined by

$$sr_i^2 = \frac{pr_i^2}{1 - pr_i^2}(1 - R^2_{Y.123\ldots k}). \tag{3.5.9}$$

3.5.5 *pr* and pr^2

The partial correlation coefficient pr_i, we recall from the two-IV case, is the correlation between that portion of Y that is independent of the remaining variables, $Y - \hat{Y}_{12\ldots(i)\ldots k}$, and that portion of X_i that is independent of the (same) remaining variables, $X_i - \hat{X}_{i.12\ldots(i)\ldots k}$, that is,

$$\begin{aligned} pr_i &= r_{Yi.12\ldots(i)\ldots k} \\ &= r_{(Y-\hat{Y}_{12\ldots(i)\ldots k})(X_i-\hat{X}_{i.12\ldots(i)\ldots k})}. \end{aligned} \tag{3.5.10}$$

pr^2 is thus interpretable as the proportion of that part of the Y variance that is independent of the remaining IVs (i.e., of $1 - R^2_{Y.12\ldots(i)\ldots k}$) accounted for *uniquely* by X_i:

$$pr_i^2 = \frac{sr_i^2}{1 - R^2_{Y.12\ldots(i)\ldots k}} \tag{3.5.11}$$

It can be seen that pr_i^2 will virtually always be larger than and can never be smaller than sr_i^2, because sr_i^2 is the unique contribution of X_i expressed as a proportion of the *total* Y variance whereas pr_i^2 expresses the same unique contribution of X_i as a proportion of that *part* of the Y variance not accounted for by the other IVs.

3.5.6 Example of Interpretation of Partial Coefficients

Table 3.5.2 presents the semipartial and partial correlations and their squares for the salary example. We see that publications (X_2) accounts for 26% (r^2_{Y2}) of the salary variance, it accounts uniquely for only 1% of the salary variance ($sr_2^2 = .01$), and only 2% of the salary variance not accounted for by the other three variables ($pr_2^2 = .02$). Notice that in this example the partial coefficients of the four IVs are ordered differently from the zero-order correlations. Although time since Ph.D. taken by itself accounts for .37 (r^2_{Y1}) of the variance in salary, it uniquely

TABLE 3.5.2
Correlations of Predictors With Y

Predictor	r_{Yi}	r^2_{Yi}	sr_i	sr_i^2	pr_i	pr_i^2
X_1, Time since Ph.D.	.608	.370	.278	.077	.367	.135
X_2, No. of publications	.506	.256	.101	.010	.142	.020
X_3, Sex	−.201	.040	−.046	.002	−.065	.004
X_4, No. of citations	.550	.302	.328	.107	.422	.178

[9] Chapter 10 deals with this and other indices of IV intercorrelation in more detail.

accounts for only 8% of this variance, whereas citations, which alone accounts for 30% of the salary variance, accounts uniquely for 11%. The reason for this is the much greater redundancy of time since Ph.D. with other predictors (46%) as compared with citations (16%); (see next section).

3.6 STATISTICAL INFERENCE WITH *k* INDEPENDENT VARIABLES

3.6.1 Standard Errors and Confidence Intervals for *B* and β

In Section 2.8.2 of Chapter 2 we showed how to determine standard errors and confidence intervals for r and B in the two-variable case, provided that certain distributional assumptions are made. Similarly, one may determine standard errors for partial regression coefficients; that is, one may estimate the sampling variability of partial coefficients from one random sample to another, using the data from the single sample at hand.

The equation for estimating the standard error of B is particularly enlightening because it shows very clearly what conditions lead to large expected sampling variation in the size of B and hence in the accuracy one can attribute to any given sample B value. A convenient form of the equation for the standard error of B for any X_i is

$$SE_{B_i} = \frac{sd_Y}{sd_i}\sqrt{\frac{1}{1-R_i^2}}\sqrt{\frac{1-R_Y^2}{n-k-1}} \tag{3.6.1}$$

where R_Y^2 is literally $R_{Y.12\ldots k}^2$, and R_i^2 is literally $R_{i.12\ldots(i)\ldots k}^2$. The ratio of the sds, as always, simply adjusts for the scaling of the units in which X_i and Y are measured. Aside from this, we see from the third term that the size of the SE_B will decrease as the error variance proportion $(1-R_Y^2)$ decreases and its $df\,(=n-k-1)$ increase. (On reflection, this should be obvious.) Note that this term will be constant for all variables in a given regression equation. The second term reveals an especially important characteristic of SE_B, namely, that it increases as a function of the squared multiple correlation of the remaining IVs with X_i, R_i^2. Here we encounter a manifestation of the general problem of multicollinearity, that is, of substantial correlation among IVs. Under conditions of multicollinearity there will be relatively large values for at least some of the SE_Bs, so that any given sample may yield relatively poor estimates of some of the population regression coefficients, that is, of those whose R_i^2s are large. (See Chapter 10 for further discussion of this issue.)

In order to show the relationship given in Eq. (3.6.1) more clearly it is useful to work with variables in standard score form. B_i expressed as a function of standard scores is β_i. The standard error of β_i drops the first term from (3.6.1) because it equals unity, so that

$$SE_{\beta_i} = \sqrt{\frac{1-R_Y^2}{n-k-1}}\sqrt{\frac{1}{1-R_i^2}} \tag{3.6.2}$$

To illustrate the effects of differences in the relationships of a given X_i, with the remaining IVs, we return to our running example presented in Tables 3.5.1 and 3.5.2. In this example, number of publications and number of citations had very similar zero-order correlations with salary, .506 and .550, respectively. Their correlations with other IVs, especially time since Ph.D. differed substantially, however, with publications correlating .651 and number of citations correlating .373 with time. The squared multiple correlation with other IVs is .4330

for number of publications and .1581 for number of citations. Substituting these values into Eq. (3.6.2) we find

$$SE_{\beta_{Publications}} = \sqrt{\frac{1-.5032}{57}}\sqrt{\frac{1}{1-.4330}} = .0934(1.3280) = .124,$$

$$SE_{\beta_{Citations}} = \sqrt{\frac{1-.5032}{57}}\sqrt{\frac{1}{1-.1581}} = .0934(1.0899) = .102.$$

Thus we can see that the redundancy with other variables has not only reduced the β for publications (to .134 from $r_{Y.Publications} = .506$) as compared to citations (to .357 from $r_{Y.Citations} = .550$), it also has made it a less reliable estimate of the population value. In contrast, the β for sex, although smaller in size than that for citations, .047 versus .357, has a slightly smaller SE, .096 versus .102. Sex shared 5% of its variance with the other IVs, whereas citations shared 16% of its variance with the other IVs.

Converting from these back to the SE_B we find

$$SE_{BPublications} = 9706/14.0(.124) = 85.9$$
$$SE_{BCitations} = 9706/17.17(.102) = 57.5$$

In Section 2.8.2 of Chapter 2, we showed how to compute and interpret confidence intervals in simple bivariate correlation and regression. For MRC, we proceed in the same way, using our faculty salary example. For the regression coefficients, the B_i, we found the standard errors for publications and citations to be, respectively, 85.9 and 57.5. The *margin of error* (*me*) for B_i is $t_c(SE_{Bi})$, where t_c is the multiplier for a given confidence interval for the error *df*. Most frequently 95% confidence intervals are reported in the literature. However, 80% *CI* may provide a more realistic feeling for the likely population value in some cases.

See the regression equation in Section 3.5.2 for the B values (93 and 202) in what follows. Using the approximate critical value of t for $\alpha = .20$, $t_c = 1.3$ as the multiplier, the 80% *me* for publications $= 1.3(85.9) = 112$, so the 80% *CI* $= 93 \pm 112$, from -19 to 205. For citations, the 80% *me* $= 1.3(57.5) = 74.6$, so the 80% *CI* $= 202 \pm 74.6$, from 127 to 277. Using $t_c = 2$ as the multiplier, the 95% *me* for B for publications is $2(85.9) = 172$, so the 95% *CI* is $93 \pm 172 = -79$ to 265. For citations, the 95% *me* is $2(57.5) = 115$ and the 95% *CI* for B for citations is 202 ± 115, $= 87$ to 317.

One may use the SE to determine the bounds within which we can assert with a chosen level of confidence that the population β falls much as we did in Chapter 2 for its zero-order analog, r. There, in Section 2.8.2, we initially used the exact t values for the available degrees of freedom. Using that method, for the 95% confidence interval, the *margin of error*, $me_\beta = t(SE_\beta)$, where, for $df = n - k - 1 = 62 - 4 - 1 = 57$, $t = 2.002$ (Appendix Table A). The *me* for a β_i is 2.002 ($SE_{\beta i}$). The standard errors for publications and citations are, respectively, .124 and .102, and the *margins of error* are .248 and .204, so the 95% confidence interval for β for publications is $.134 \pm .248$, from $-.11$ to .38, and the 95% confidence interval for β for citations is $.357 \pm .204$, from .15 to .56.

3.6.2 Confidence Intervals for R^2

The *CI*s that follow for R^2 and differences between independent R^2s are from Olkin and Finn (1995). They are based on large-sample theory and will yield adequate approximations for $df > 60$.

We have found that for our sample of 62 faculty members, our four IVs yield an R^2 of .5032. The variance error of R^2 is given by

$$SE^2_{R^2} = \frac{4R^2(1-R^2)^2(n-k-1)^2}{(n^2-1)(n+3)}. \tag{3.6.3}$$

Substituting,

$$SE^2_{R^2} = \frac{4(.5032)(.4968^2)(57^2)}{(62^2-1)(65)} = .006461.$$

Therefore the standard error, $SE_{R^2} = \sqrt{.00646} = .080$.

95% confidence intervals using exact t values are routinely reported in the literature. Alternatively, one may opt to use some other probability, such as 80%, as providing reasonable bounds for the purposes of the study. In recognition of the fairly rough approximation provided by any of these limits, one may use the approximate constant multipliers (t_c) of the *SE*s for the desired degree of inclusion of Section 2.8.2:

CI	99%	95%	80%	2/3
t_c	2.6	2	1.3	1

The 80% *me* for $R^2 = 1.3(.0804) = .1045$, so the approximate 80% *CI* is $.503 \pm .104$, from .40 to .61. (The 95% *me* $= 2(.0804) = .161$, so the approximate 95% *CI* for R^2 is $.503 \pm .161$, from .34 to .66.)

3.6.3 Confidence Intervals for Differences Between Independent R^2s

For our running example of 62 cases (University *V*), we found the $R^2_V = .5032$ for the $k = 4$ IVs. For the same IVs in University *W*, where $n = 143$, assume that $R^2_W = .2108$. The difference is $.5032 - .2108 = .2924$. Since these are different groups that were independently sampled, we can find *CI*s and perform null hypothesis significance tests on this difference, using the *SE* of the difference. As we have seen for other statistics, this is simply the square root of the sum of the SE^2s of the two R^2s. We found the SE^2 for *V* to be .006461 in the previous section, and assume we find the SE^2 for *W* to be .003350. Substituting,

$$SE_{R^2_V - R^2_W} = \sqrt{SE_{R^2_V} + SE_{R^2_W}} = \sqrt{.006461 + .003350} = \sqrt{.006811} = .0825. \tag{3.6.4}$$

The approximate 95% *me* $= 2(.0825) = .1650$, so the approximate 95% *CI* for a *nil* hypothesis significance test $= .2924 \pm .1650$, from .13 to .46. Since the 95% *CI* does not include 0, the difference between the universities' R^2s is significant at the $\alpha = .05$ level.

3.6.4 Statistical Tests on Multiple and Partial Coefficients

In Chapter 2 we presented statistical inference methods for the statistics of simple regression and correlation analysis, that is, when only two variables are involved. As we have seen, the test

of the null hypothesis that R^2 is zero in the population can be accomplished by examination of the lower confidence limit for the desired alpha level (e.g., the 95% two-tailed *CI*). Equivalently, the statistic F may be determined as

$$F = \frac{R^2(n-k-1)}{(1-R^2)k} \tag{3.6.5}$$

with $df = k$ and $n-k-1$.

F may also be computed (or provided as computer output) as a function of raw scores in the classic analysis of variance format. As we saw in the one-IV case, the total sample variance of Y may be divided into a portion accounted for by the IV, which is equal to the variance of the estimated $\hat{Y}$ values, $sd^2_{\hat{Y}}$, and a portion not associated with the IV, the "residual" or "error" variance, $sd^2_{Y-\hat{Y}}$. Similarly, the sum of the squared deviations about the mean of Y may be divided into a sum of squares (SS) due to the regression on the set of IVs, and a residual sum of squares. When these two portions of the total are divided by their respective *df*, we have the mean square (MS) values necessary for determining the F values, thus

$$\text{regression MS} = \frac{\text{regression SS}}{k} = \frac{R^2\sum y^2}{k},$$
$$\text{residual or error MS} = \frac{\text{residual SS}}{n-k-1} = \frac{(1-R^2)\sum y^2}{n-k-1}. \tag{3.6.6}$$

When F is expressed as the ratio of these two mean squares, we obtain

$$F = \frac{\text{regression MS}}{\text{residual MS}} = \frac{R^2\sum y^2/k}{(1-R^2)\left(\sum y^2\right)/(n-k-1)}. \tag{3.6.7}$$

Canceling the $\sum y^2$ term from the numerator and denominator and simplifying, we obtain Eq. (3.6.5).

Let us return to our running example of academic salaries. The four independent variables produced $R^2 = .5032$. Because there were 62 faculty members, by Eq. (3.6.5),

$$F = \frac{.5032(62-4-1)}{(1-.5032)4} = 14.43$$

for $df = 4, 57$.

Turning to the tabled F values for $\alpha = .01$ (Appendix Table D.2), we find an F value (by interpolation) of 3.67 is necessary for significance. Because the obtained F value exceeds this value, we conclude that the linear relationship between these four IVs and salary is not likely to be zero in the population.

As previously noted, sr_i, pr_i, and β_i differ only with regard to their denominators. Thus none can equal zero unless the others are also zero, so it is not surprising that they must yield the same t_i value for the statistical significance of their departure from zero. It should also be clear that because B_i is the product of β and the ratio of standard deviations, it also can equal zero only when the standardized coefficients do. Thus, a single equation provides a test for the significance of departures of all the partial coefficients of X_i from zero. They either are, or are not, all significantly different from zero, and to exactly the same degree.

$$t_i = sr_i\sqrt{\frac{n-k-1}{1-R^2}}, \tag{3.6.8}$$

t_i will carry the same sign as sr_i and all the other partial coefficients for that variable.

For example, let us return to the running example where the obtained R^2 of .5032 was found to be significant for $k = 4$ and $n = 62$. The sr_is for the four IVs were, respectively, .278, .101, −.046, and .328.

Determining their t values we find

$$t_{Time} = .278\sqrt{\frac{62-4-1}{1-.5032}} = 2.98$$

$$t_{Publications} = .101\sqrt{\frac{62-4-1}{1-.5032}} = 1.08$$

$$t_{Sex} = -.046\sqrt{\frac{62-4-1}{1-.5032}} = -.49$$

$$t_{Citations} = .328\sqrt{\frac{62-4-1}{1-.5032}} = 3.51.$$

Looking these values up in the t table (Appendix Table A) for 57 *df*, we find that time since Ph.D. and number of citations are significant at the .01 level but publications and sex are not significant at the .05 level. We conclude that time and citations both make unique (direct) contributions to estimating salary. We may *not* reject the nil hypothesis that sex and publications have no unique (direct) relationship to salary in the population once the effects of time and citations are taken into account.

It is quite possible to find examples where R^2 is statistically significant but none of the tests of significance on the individual IVs reaches the significance criterion for rejecting the nil hypothesis. This finding occurs when the variables that correlate with Y are so substantially redundant (intercorrelated) that none of the unique effects (βs) is large enough to meet the statistical criterion (see Chapter 10 for a more extensive discussion of this problem). On the other hand, it may also happen that one or more of the t tests on individual variables does reach the criterion for significance although the overall R^2 is not significant. The variance estimate for the regression based on k IVs is divided by k to form the numerator of the F test for R^2, making of it an average contribution per IV. Therefore, if most variables do not account for more than a trivial amount of Y variance they may lower this average (the mean square for the regression) to the point of making the overall F not significant in spite of the apparent significance of the separate contributions of one or more individual IVs. In such circumstances, we recommend that such IVs *not* be accepted as significant. The reason for this is to avoid spuriously significant results, the probability of whose occurrence is controlled by the requirement that the F for a set of IVs be significant before its constituent IVs are t tested. This, the "protected t test," is part of the strategy for statistical inference that is considered in detail in Chapter 5.

3.7 STATISTICAL PRECISION AND POWER ANALYSIS

3.7.1 Introduction: Research Goals and the Null Hypothesis

Almost every research effort is an attempt to estimate some parameter in some population. In the analyses described in this book, the parameters in question are represented by multiple and partial regression and correlation coefficients. Traditionally the behavioral sciences have focused almost entirely on the issue of the simple presence and direction of a partial regression coefficient, or the confidence that there is some correlation between a dependent variable and a set of independent variables in the population. Thus the statistical tests have generally been

focused on the null (nil) hypothesis that the population coefficient is zero. Although this is sometimes a useful question and thus an appropriate research goal, its limitations in advancing the progress of science have been recognized in articles as well as in an organized effort to change the focus of research reports (Wilkinson & the APA Task Force on Statistical Inference, 1999).

The precision of any statistic is identified by its standard error and the associated confidence interval. Its statistical power is the probability of rejecting the null hypothesis when it is false. Both are determined as a function of three elements, the size of the effect in the population, the *df* which are determined primarily by the sample size, and the chosen margin of error or alpha level. Thus, it is appropriate to view statistical power as a special case of the more general issue of the precision of our estimates.

In this section we extend consideration of these issues, which were introduced in Chapter 2, to the multiple independent variable case. Although we review the steps necessary for the hand computation of power and precision, and provide the necessary Appendix tables, we recommend the use of a contemporary user-friendly computer program such as Sample Power (SPSS) or Power and Precision (Borenstein, Cohen, and Rothstein, 2001), which will facilitate the user's appreciation of the interaction among the determinants of statistical power. The emphasis of this presentation is an understanding of the influences that contribute to the precision and power of any study, so that each investigator can make appropriate choices of such parameters in planning or evaluating research.

3.7.2 The Precision and Power of R^2

As noted earlier, both precision and power are determined as a function of the effect size, the sample size, and the selected probability parameter. For simplicity let us begin with the assumption that we will be using 95% *CI* or, equivalently for the special case of the nil hypothesis, the .05 significance criterion. As we plan our study, the question is what precision and power will we have for a given proposed n, or for each of a set of alternative ns. The effect size that is relevant is the population R^2.

Precision

Suppose that we anticipate that the population R^2 as estimated by a set of six IVs is about .2. The sample size that we have budgeted for is 120 cases. Application of Eq. (3.6.3) gives us the SE_{R^2} and tells us that an empirical estimate of this population value (which would average .24 in a sample of this size; see the section on shrinkage) would have an 80% *CI* of .16 – .32. Our substantive theory will be needed to guide us in the judgment as to whether this *CI* is so large that it fails to contribute an increment to our knowledge about these phenomena. If it is judged that it is too large, there are two possible remedies. The simple, but often expensive and sometimes infeasible one is to increase the sample size. If this is possible the precision can be recomputed with a new n.

An alternative method of increasing the *df* for precision (and power) is to reduce the number of IVs from the proposed six to a smaller number, if that will result in no material loss of effect size or critical information. The effective n in these equations is not the actual sample size but rather the *df*, which is $n - k - 1$. If some of the variables are substantially correlated it may be that they can be usefully consolidated. If the loss to R^2 is small enough, a recomputation of the *CI* may demonstrate adequate precision.

The selected *me* can also be altered. As we argued earlier, it is often the case that an 80% *CI*, or even a *CI* that yields 2 to 1 odds of including the parameter, may be adequate for the scientific purposes of the investigation.

For an illustration, let us return to our academic salary example. Suppose that we were interested in examining another university department for the same issues. This department has 34 current faculty members. We anticipate that this department is a member of the same population, and that the population R^2 will be about the same as it was in the department represented by our current data, where we found $R^2 = .503$ (80% $CI = .40 - .61$, Section 3.6.2). We find that in the proposed department the 80% *CI*, given a .5 R^2 in the population, would be on average .38 – .74. If this is too large to be informative, and we do not feel that using the 2/1 odds rule to generate narrower *CI* would serve our purpose, there is little point in carrying out the study. Once again, this *SE* was developed for large samples, so caution must be used in applying it to small samples.

Power Analysis

As we noted earlier, statistical power analysis is concerned with the special case of determining the probability that the sample value will be significantly different from some hypothesized value, typically one of no effect such as a zero R^2. This is a special case because when the *CI* does not include this null value the statistical criterion has been met (at the α criterion used to determine the *CI*). Again, one employs the appendix tables (or more conveniently a computer program) by selecting the expected population R^2, the proposed sample n, the number of predictor variables, and the significance criterion, and determining (or reading out) the probability that the sample *CI* will not include the null value. Although more complete tables for this purpose are provided in J. Cohen (1988), this can also be accomplished by following these steps:

1. Set the significance criterion to be used, α. Provision is made in the appendix for $\alpha = .01$ and $\alpha = .05$ in the L tables (Appendix Tables E.1 and E.2).
2. Determine the population effect size *ES* for $R^2 =$

$$f^2 = \frac{R^2}{1 - R^2}. \tag{3.7.1}$$

3. Determine L by

$$L = f^2(n - k - 1) \tag{3.7.2}$$

4. Determine the power by finding the row corresponding to the *df* in the selected appendix table, locating an L as close as possible to the computed value, and looking up the column to determine the estimated power.

For example, in the case noted earlier of the new department with 34 faculty members, if the population value is similar to our computed one (.50), the ES $= .503/1 - .503 = 1.012$, and $L = 1.012\,(34 - 3 - 1) = 30.36$. Looking in Appendix Table E.1 at $k_B = df = 3$, we find that the computed L is larger than the value in the last column and thus the probability of finding the sample R^2 to be greater than zero with $\alpha = .01$ is at least $\beta = .99$. On the other hand, the reader may confirm that if the relationship were more in the range that is typical of many behavioral studies—for example, a population R^2 of .2—even using the less conservative $\alpha = .05$, our chances of finding the sample value to be statistically significant are only slightly better than 50-50.

When the expected power is unacceptably low it may be increased by increasing the *df* (mainly by increasing n) or by lowering the selected value of α. The first two steps used to determine the $n*$ required for a desired power and R^2 are as shown earlier. L is located for

the row corresponding to the *df* and column corresponding to the desired power. Then $n*$ is determined by

$$n* = \frac{L}{f^2} + k + 1. \tag{3.7.1}$$

In the proposed example of population $R^2 = .20$, so $ES = f^2 = R^2/(1-R^2) = .2/(1-.2) = .25$, if we desire power = .80 we will need $L = 10.90$ (Appendix Table E.2, with 3 *df*) so that $n* = (10.90/.25) + 3 + 1 = 48$.

Sometimes the effect size can be increased by changes in the sampling strategy (for example, by selecting more extreme subjects), by improvement of measures (increases in their reliability and validity), or by altering the experimental protocol to produce a stronger experimental manipulation. These methods of enhancing power are likely to be especially positive for the scientific payoff of a study, and thus may often be recommended as the first alterations to be considered.

Although it is much to be preferred that the substantive theory and prior research determine the expected population value of R^2, some rules of thumb have been suggested for the use of researchers who are unable to provide more appropriate values. Values of .02, .13, and .26 have been proposed as potentially useful estimates of small, medium, and large effect sizes for the population R^2. These values should probably be adjusted upward by the researcher who intends to use more than a few IVs.

3.7.3 Precision and Power Analysis for Partial Coefficients

Precision

As noted earlier, partial coefficients for a given IV share the same numerator, the exception being the raw unit regression coefficient for which the ratio of standard deviations of Y and that IV also appears. When the units employed for B are meaningful, the *CI* for B will provide the most useful information about the precision of the expected sample values. (See Chapter 5 for a discussion of methods of improving the utility of measure units.) When the units are not meaningful, precision is usually referenced to β as a function of its *SE*.[10]

For example, again using our academic salary illustration, we are interested in the value of the gender difference in salary in departments from some other academic field than that represented by our current data. We would like to be able to assess the sex difference with a *me* of $1000. The researcher may know that about 30% of the faculty members in these departments are women; thus the *sd* of sex will be about $\sqrt{.30(.70)} = .458$ in the proposed study. The *sd* of faculty salaries may be determined from administration records or estimated from the current study as about $8000. Using Eq. (3.6.1), rearranging, and solving for $SE_B = me/2 = \$500$, we find

$$(n - k - 1)(\$500)^2 = \frac{\$8000^2(1 - .40)}{.21(1 - .10)}$$

$$n - k - 1 = 948,$$

so that we will need nearly a thousand cases. If, on the other hand, we were content with a *me* representing about 2/1 odds of including the population value (so that we could tolerate a SE_B of $1000), a sample of about 230 would suffice.

---

[10] We do not provide *CI*s for *sr* or sr^2, which are asymmetrical and complex.

Suppose, however, that for the research that we are planning we have no reasonable precedent for estimating B, previous research having used different measures of this construct than the one we are planning to employ. In this case we may use the β obtained in these studies to estimate the value expected in the planned study, and appropriately adjust for correlations with other IVs.

Statistical Power of Partial Coefficients

As we have noted, partial coefficients have a common test of statistical significance. Therefore they also have in common the statistical power to reject a false null (nil) hypothesis that the population value is zero. In the case of statistical power, however, it is convenient to define the effect size as the increment in R^2 attributable to a given IV, that is, its sr^2. As we noted earlier in the chapter, sr differs from β by the square root of its tolerance, the proportion of its variance that is independent of other predictors. As noted previously and discussed in Chapter 10, other things being equal, the *SE*s of the partial effects of an IV, and thus imprecision in their estimates, are generally increased by increases in correlation with other IVs.

In order to calculate the power of the proposed study to reject the null hypothesis that the partial coefficients are nonzero, one enters the first row of the power tables (or, preferably, a computer program) for the selected significance criterion with the L determined from the proposed n and the estimated proportion of Y variance that is uniquely accounted for by the IV in question. If it should happen that the investigator can more readily estimate B or β, these coefficients can be converted to sr^2 providing that the multiple correlation of the IV in question with the other IVs and, in the case of B, S_Y, and S_i or their ratio, can be estimated.

To illustrate, suppose that we want to have 90% power to detect a sex difference in salary under the same assumptions as in the previous example. Using Eq. (3.5.8) to convert from B to sr, we estimate

$$sr = \$3000\left(\frac{\sqrt{.21}}{\$8000}\right)\left(\sqrt{1-.3}\right) = .144,$$

and $sr^2 = .02$. These parameters may be looked up in the computer program. Alternatively, we may compute the $ES =$

(3.7.2)
$$f^2 = \frac{sr^2}{1-R^2}.$$

If $R^2 = .20$ the $ES = .02/.80 = .025$. Looking up L in Appendix Table E.2 for row $k_B = 1$ and column $\beta = .90$ we find that $L = 10.51$, and applying this to Eq. (3.7.1), we find that we will need 422 cases to have a 90% chance of rejecting the null hypothesis at $\alpha = .05$. Further calculation will show that we will need $n* = 510$ if we wish to reject the null hypothesis at the .01 level. If this number is too large, we may reconsider whether we can be content with 80% power.[11]

In general it is likely to be more practically and theoretically useful to examine the consistency of the new data against some non-nil value. For example, it might be decided that any discrepancy as large as \$1000 in annual salary (net of the effects attributable to other causes) would be unmistakably material to the people involved. In such a case the difference between our estimated population value (\$3000) and this value is only \$2000, so we re-enter the equation with this value.

[11] One reason we like computer programs for determining power and needed sample sizes is that they quickly train the researcher to appreciate how statistical power is closely linked to α, *ES*, *df*, and *n*, which may lead to improvements in judgments and strategy on these issues.

Once again, for those investigators who absolutely cannot come up with a more substantively or theoretically based estimate of the population effect size, some rules of thumb are sometimes useful. These values are usually expressed in terms of the proportion of the Y variance that is not explained by other variables that is explained by X_i. Small effects may be defined as 2% of the unexplained variance, medium effects as 15% of the unexplained variance, and large effects as 35%. As we will see in Chapter 5, these values are relatively large when we are talking about a single IV, and it may be at least as appropriate to use the values of r given in Chapter 2 as small, medium, and large, when one is examining a single sr_i.

Several other topics in power analysis are presented in Chapter 5, following the exposition of power analysis when multiple sets of IVs are used. Among the issues discussed there are determination of power for a given n, reconciling different $n*$s for different hypotheses in a single analysis, and some tactical and other considerations involved in setting effect size and power values.

3.8 USING MULTIPLE REGRESSION EQUATIONS IN PREDICTION

One use of MRC is for prediction, literally forecasting, with only incidental attention to explanation. Although we have emphasized the analytic use of MRC to achieve the scientific goal of explanation, MRC plays an important role in several behavioral technologies, including personnel and educational selection, vocational counseling, and psychodiagnosis. In this section we address ourselves to the accuracy of prediction in multiple regression and some of its problems.

3.8.1 Prediction of *Y* for a New Observation

The standard error of estimate, $SE_{Y-\hat{Y}}$, as we have seen, provides us with an estimate of the magnitude of error that we can expect in estimating $\hat{Y}$ values over sets of future $X_1, X_2, \ldots, X_k$ values that correspond to those of the present sample. Suppose, however, we wish to determine the standard error and confidence intervals of a *single* estimated $\hat{Y}_O$ from a new set of observed values $X_{1O}, X_{2O}, \ldots, X_{kO}$. In Section 2.8.2 we saw that the expected magnitude of error increases as the X_i values depart from their respective means. The reason for this should be clear from the fact that any discrepancy between the sample estimated regression coefficients and the population regression coefficients will result in larger errors in $\hat{Y}_O$ when X_i values are far from their means than when they are close.

Estimates of the standard error and confidence intervals for $\hat{Y}_O$ predicted from known *values* $X_{1i}, X_{2i}, \ldots, X_{ki}$ is given by

(3.8.1)
$$sd_{Y_O-\hat{Y}_O} = \frac{SE_{Y-\hat{Y}}}{\sqrt{n}}\sqrt{n+1+\sum\frac{z_{iO}^2}{1-R_i^2}-2\sum\frac{\beta_{ij}z_{iO}z_{jO}}{1-R_i^2}},$$

where the first summation is over the k IVs, the second over the $k(k-1)/2$ pairs of IVs (i.e., $i < j$) expressed as standard scores, β_{ij} is the β for estimating X_i from X_j, holding constant the remaining $k-2$ IVs, and R_i^2 is literally $R^2_{i.12\ldots(i)\ldots k}$. Although at first glance this formula appears formidable, a closer examination will make clear what elements affect the size of this error. The $SE_{Y-\hat{Y}}$ is the standard error of estimate, and as in the case of a single IV, we see that increases in it and/or in the absolute value of the IV (z_{iO}) will be associated with larger error. The terms that appear in the multiple IV case that did not appear in the Eq. (2.8.3) for the single

variable case (β_{ij} and R_i^2) are functions of the relationships among the independent variables. When all independent variables are uncorrelated (hence all β_{ij} and all R_i^2 equal zero), we see that the formula simplifies and $sd_{Y_o-\hat{Y}_o}$ is minimized (for constant $SE_{Y-\hat{Y}}$, n, and z_{iO} values).

It is worth emphasizing the distinction between the validity of the significance tests performed on partial coefficients and the accuracy of such coefficients when used in prediction. In analytic uses of MRC, including formal causal analysis, given the current level of theoretical development in the behavioral and social sciences, the information most typically called upon is the significance of the departure of partial coefficients from zero and the sign of such coefficients. The significance tests are relatively robust to assumption failure, particularly so when n is not small. Using the regression equation for prediction, on the other hand, requires applying these coefficients to particular individual variable values for which the consequence of assumption failure is likely to be much more serious.

As an illustration, let us examine the scatterplot matrix (SPLOM)[12] for our running example of academic salaries. Figure 3.8.1 provides the scatterplot for each pair of variables, including the predicted salary and the residual. As can be seen, the original distributions of years and publications are not as symmetrical as is the distribution of salary. Probably as a consequence, the residuals above the mean $\hat{Y}$ appear to have a somewhat higher variance than those below the mean (the reader may check to determine that this is indeed the case). The variance of the residuals otherwise looks passably normal (as indeed they should, because this example was generated to meet these assumptions in the population). Failure of the homoscedasticity assumption may not be serious enough to invalidate tests of statistical significance, but it still could invalidate actual prediction if based on the assumption of equal error throughout the distribution.

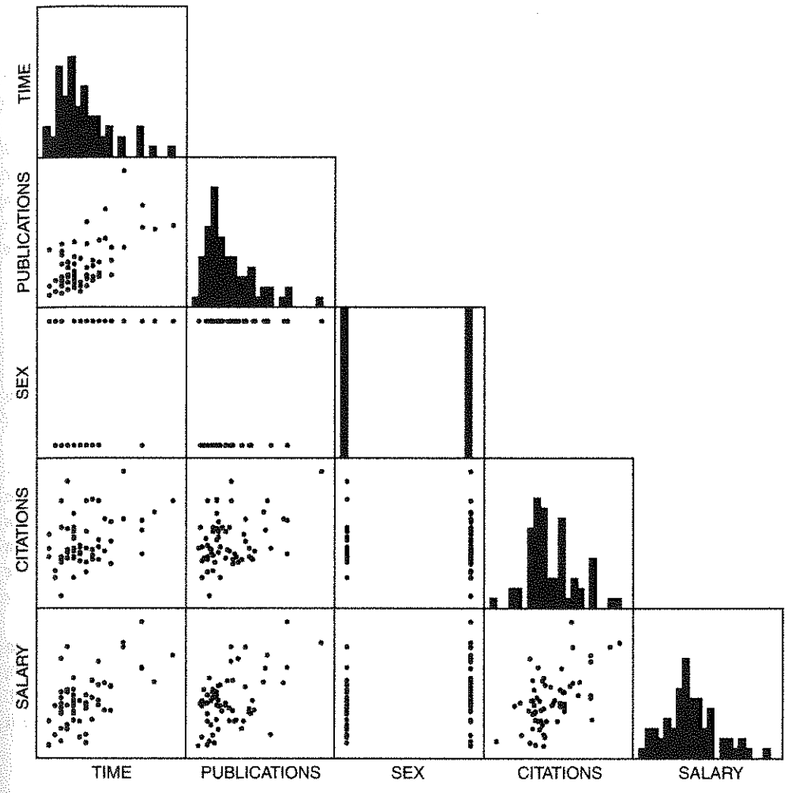
3.8.2 Correlation of Individual Variables With Predicted Values

Further insight may be gained by noting that regardless of the sign, magnitude, or significance of its partial regression coefficient, the correlation between X_i and the $\hat{Y}$ determined from the entire regression equation is

$$r_{\hat{Y}i} = \frac{r_{Yi}}{R_{Y.123\ldots k}}. \tag{3.8.2}$$

Thus it is invariably of the same sign and of larger magnitude than its zero-order r with Y. (See values at the bottom of Fig. 3.8.1 for reflection of this in our running example.) Reflection on this fact may help the researcher to avoid errors in interpreting data analyses in which variables that correlate materially with Y have partial coefficients that approach zero or are of opposite sign. When partial coefficients of the X_i approximate zero, whatever linear relationship exists between X_i and Y is accounted for by the remaining independent variables. Because neither its zero-order correlation with Y nor its (larger) correlations with $\hat{Y}$ is thereby denied, the interpretation of this finding is highly dependent on the substantive theory being examined. Even without a full causal model, a weak theoretical model may be employed to sort out the probable meaning of such a finding. One theoretical context may lead to the conclusion that the true causal effect of X_i on Y operates fully through the other IVs in the equation. Similarly, when the B_{Yi} and r_{Yi} are of opposite sign, X_i and one or more of the remaining IVs are in a suppressor relationship. Although it is legitimate and useful to interpret the partialed relationship, it is also important to keep in mind the zero-order correlations of X_i with Y (and hence with $\hat{Y}$).

[12]The fact that we have not previously introduced this graphical aid should not be taken to deny an assertion that such a matrix is probably the first step in analyzing a data set that should be taken by a competent data analyst (see Chapter 4). The figures along the diagonal reflect the distribution of each variable.



	Y = Salary	Time since Ph.D.	Sex	No. of publications	No. of citations
Correlation with $\hat{Y}$	.709	.857	.283	.713	.775
Correlation with residual	.705	0.00	0.00	0.00	0.00

FIGURE 3.8.1 Scatterplot matrix for the academic salary example.

3.8.3 Cross-Validation and Unit Weighting

Several alternatives to regression coefficients for forming weighted composites in prediction have been proposed (Darlington, 1978; Dawes, 1979; Green, 1977; Wainer, 1976). Although β weights are guaranteed to produce composites that are most highly correlated with z_Y (or Y) in the sample on which they are determined, other weights produce composites (call them u_Y) that are almost as highly correlated in that sample. "Unit weighting", the assignment of the weights of +1 to positively related, −1 to negatively related, and 0 to poorly related IVs are popular candidates—they are simple, require no computation, and are not subject to sampling error (Green, 1977; Mosteller & Tukey, 1977; Wainer, 1976). For our running example on

academic salary, we simply add (that is, we use weights of +1.0) the z scores of each subject for time since Ph.D., publications, and citations, and subtract (that is, use weights of −1.0) each score for female to produce the composite u_Y for each subject. We find that u_Y correlates .944 with the β-weighted $\hat{z}_Y$ (or $\hat{Y}$), and therefore (not surprisingly) .670 with z_Y (or Y), only modestly lower than the .709 (= R_Y) of $\hat{z}_Y$ with z_Y (or Y).

However, the real question in prediction is not how well the regression equation determined for a sample works on that sample, but rather how well it works in the population or on other samples from the population. Note that this is *not* the estimate of the population $\hat{\rho}^2_Y$, (i.e., the "shrunken" value given in Eq. 3.5.5), but rather an estimate of the "cross-validated" $r^2_{Y\hat{Y}}$ for each sample's β applied to the other sample, which is even more shrunken and which may be estimated by

$$\hat{R}^2 = 1 - (1 - R^2)\frac{(n+k)}{(n-k)} \tag{3.8.3}$$

(Rozeboom, 1979). $\hat{R}^2$ answers the relevant question, "If I were to apply the *sample* regression weights to the population, or to another sample from the population, for what proportion of the Y variance would my thus-predicted Y values account?"

For our running example with $n = 62$, our sample regression equation yields $\hat{R}^2 = 1 - (1 - .5032)(62 + 4)/(62 - 4) = .4347$, so $\hat{R} = .659$. We found earlier, however, that the unit-weighted composite for the cases we have yielded an $r = .670$, greater than $\hat{R}$. Now this value is subject to sampling error (so is $\hat{R}$), but *not* to shrinkage, because it does not depend on unstable regression coefficients. As far as we can tell, unit weights would do as well or better in prediction for these data than the sample's standardized regression weights based on only 62 cases.

Unit weights have their critics (Pruzek & Fredericks, 1978; Rozeboom, 1979). For certain patterns of correlation (suppression is one) or a quite large $n : k$ ratio (say more than 20 or 30), unit weights may not work as well in a new sample as the original regression coefficients will. An investigator who may be in such a circumstance is advised to compute $\hat{R}$ and compare it with the results of unit weighting in the sample at hand.

3.8.4 Multicollinearity

The existence of substantial correlation among a set of IVs creates difficulties usually referred to as the problem of multicollinearity. Actually, there are two distinct problems—the substantive interpretation of partial coefficients and their sampling stability.

Interpretation

We have already seen in Section 3.4 that the partial coefficients of highly correlated IVs analyzed simultaneously are reduced. Because the IVs involved lay claim to largely the same portion of the Y variance by definition, they cannot make much by way of unique contributions. Interpretation of the partial coefficients of IVs from the results of a simultaneous regression of such a set of variables that ignores their multicollinearity will necessarily be misleading.

Attention to the R^2_i of the variables may help, but a superior solution requires that the investigator formulate some causal hypotheses about the origin of the multicollinearity. If it is thought that the shared variance is attributable to a single central property, trait, or *latent variable*, it may be most appropriate to combine the variables into a single index or drop the more peripheral ones (Sections 4.5 and 5.7), or even to turn to a latent variable causal model (see Chapter 12). If, on the other hand, the investigator is truly interested in each of the

variables in its own right, analysis by a hierarchical procedure may be employed (Section 5.3). To be sure, the validity of the interpretation depends on the appropriateness of the hierarchical sequence, but this is preferable to the complete anarchy of the simultaneous analysis in which everything is partialed from everything else indiscriminately.

Sampling Stability

The structure of the formulas for SE_{B_i} (Eq. 3.6.1) and $SE_{\hat{\beta}_i}$ (Eq. 3.6.2) makes plain that they are directly proportional to $\sqrt{1/(1 - R^2_i)}$. A serious consequence of multicollinearity, therefore, is highly unstable partial coefficients for those IVs that are highly correlated with the others.[13] Concomitantly, the trustworthiness of individually predicted $\hat{Y}_O$ is lessened as the R^2_is for a set of IVs increase, as is evident from the structure of Eq. (3.6.1). Large standard errors mean both wide confidence intervals and a lessened probability of rejecting a null hypothesis (see Section 3.7). Chapter 10 discusses issues of multicollinearity in more detail.

3.9 SUMMARY

This chapter begins with the representation of the theoretical rationale for analysis of multiple independent variables by means of causal models. The employment of an explicit theoretical model as a working hypothesis is advocated for all investigations except those intended for simple prediction. After the meaning of the term cause is briefly discussed (Section 3.1.1) rules for diagrammatic representation of a causal model are presented (Section 3.1.2).

Bivariate linear regression analysis is extended to the case in which two or more independent variables (IVs), designated $X_i (i = 1, 2, \ldots, k)$ are linearly related to a dependent variable Y. As with a single IV, the multiple regression equation that produces the estimated $\hat{Y}$ is that linear function of the k IVs for which the sum over the n cases of the squared discrepancies of $\hat{Y}$ from Y, $\Sigma(Y - \hat{Y})^2$, is a minimum.

The regression equation in both raw and standardized form for two IVs is presented and interpreted. The standardized partial regression coefficients, β_i, are shown to be a function of the correlations among the variables; β_i may be converted to the raw score B_i by multiplying each by sd_Y/sd_i (Section 3.2).

The measures of correlation in MRC analysis include:

1. R, which expresses the correlation between Y and the best (least squared errors) linear function of the k IVs ($\hat{Y}$), and R^2, which is interpretable as the proportion of Y variance accounted for by this function (Section 3.3.1).
2. Semipartial correlations, sr_i, which express the correlation of X_i from which the other IVs have been partialed with Y. sr^2_i is thus the proportion of variance in Y uniquely associated with X_i, that is, the increase in R^2 when X_i is added to the other IVs. The ballantine is introduced to provide graphical representation of the overlapping of variance with Y of X_1 and X_2 (Section 3.3.2).
3. Partial correlations, pr_i, which give the correlation between that portion of Y not linearly associated with the other IVs and that portion of X_i that is not linearly associated with the other IVs. In contrast with sr_i, pr_i partials the other IVs from both X_i and Y. pr^2_i is the proportion of Y variance not associated with the other IVs that is associated with X_i (Section 3.3.3).

Each of these coefficients is exemplified, and shown to be a function of the zero-order correlation coefficients. The reader is cautioned that none of these coefficients provides a basis for a satisfactory Y variance partitioning scheme when the IVs are mutually correlated.

[13] This is the focus of the discussion in Section 4.5.

The alternative causal models possible for Y and two IVs are discussed, exemplified, and illustrated. The distinction between direct and indirect effects is explained, and models consistent with partial redundancy between the IVs are illustrated. Mutual suppression of causal effects will occur when any of the three zero-order correlations is less than the product of the other two (Section 3.4.1). Spurious effects and entirely indirect effects can be distinguished when the causal sequence of the IVs is known (Section 3.4.2).

The case of two IVs is generalized to the case of k IVs in Section 3.5. The use of the various coefficients in the interpretation of research findings is discussed and illustrated with concrete examples. The relationships among the coefficients are given.

Statistical inference with k IVs, including *SE*s and *CI*s for standardized and raw unit regression coefficients and R^2 are presented in Section 3.6. *CI*s for the difference between independent R^2s are shown as well as a series of statistical tests on multiple and partial coefficients.

Determination of the precision of expected findings from proposed investigations is described and illustrated. Statistical power analysis is shown to be a special case when the question is limited to a non-nil value of a multiple or partial coefficient (Section 3.7).

A range of prediction situations are described in Section 3.8, including the prediction of a value of Y for a newly observed case. Correlations among the predictors will affect the adequacy of the estimation of the individual coefficients and the stability of the model. It is shown that least squares estimation may not yield optimal prediction for future studies or cases.

4

Data Visualization, Exploration, and Assumption Checking: Diagnosing and Solving Regression Problems I

4.1 INTRODUCTION

In Chapters 2 and 3 we focused on understanding the basic linear regression model. We considered fundamental issues such as how to specify a regression equation with one, two, or more independent variables, how to interpret the coefficients, and how to construct confidence intervals and conduct significance tests for both the regression coefficients and the overall prediction. In this chapter, we begin our exploration of a number of issues that can potentially arise in the analysis of actual data sets. In practice, not all data sets are "textbook" cases. The purpose of the present chapter is to provide researchers with a set of tools with which to understand their data and to identify many of the potential problems that may arise. We will also introduce a number of remedies for these problems, many of which will be developed in more detail in subsequent chapters. We believe that careful inspection of the data and the results of the regression model using the tools presented in this chapter helps provide substantially increased confidence in the results of regression analyses. Such checking is a fundamental part of good data analysis.

We begin this chapter with a review of some simple graphical displays that researchers can use to visualize various aspects of their data. These displays can point to interesting features of the data or to problems in the data or in the regression model under consideration when it is applied to the current data. Indeed, Tukey (1977) noted that a graphical display has its greatest value "when it *forces* us to notice **what we never expected to see**" (p. v, italics and bold in original.) Historically, labor-intensive analyses performed by hand or with calculators served the function of providing researchers with considerable familiarity with their data. However, the simplicity of "point and click" analyses in the current generation of statistical packages has made it easy to produce results without any understanding of the underlying data or regression analyses. Modern graphical methods have replaced this function, producing displays that help researchers quickly gain an in-depth familiarity with their data. These displays are also very useful in comparing one's current data with other similar data collected in previous studies.

Second, we examine the assumptions of multiple regression. All statistical procedures including multiple regression make certain assumptions that should be met for their proper use. In the case of multiple regression, violations of these assumptions *may* raise concerns as to whether the estimates of regression coefficients and their standard errors are correct. These