CHAPTER 5A

Multiple Regression

General Considerations

Multiple regression is a very useful extension of simple linear regression in that we use several variables rather than just one to predict a value on a quantitatively measured criterion variable. It has become a very popular technique to employ in behavioral research. Many researchers believe that using more than one predictor can paint a more complete picture of how the world works than is permitted by simple linear regression because constructs in the behavioral sciences are believed to be multiply determined. Using only a single variable as a predictor will at best capture only one of those sources. In the words of one author (Thompson, 1991), multivariate methods such as multiple regression have accrued greater support in part because they "best honor the reality to which the researcher is purportedly trying to generalize" (p. 80).

Based on what we have already discussed regarding simple linear regression, it may be clear that multiple regression can be used for predictive purposes, such as estimating from a series of entrance tests how successful various job applicants might be. But the regression technique can also guide researchers toward explicating or explaining the dynamics underlying a particular construct by indicating which variables in combination might be more strongly associated with it. In this sense, the model that emerges from the analysis can serve an explanatory purpose as well as a predictive purpose.

As was true for simple linear regression, multiple regression generates two variations of the prediction equation, one in raw score form and the other in standardized form. These equations are extensions of the simple

linear regression models and thus still represent linear regression. We will contrast some differences between linear and nonlinear regression later in the chapter.

The Variables in a Multiple Regression Analysis

The variables in a multiple regression analysis fall into one of two categories: One of them is the variable being predicted; the others are used as the basis of prediction. We discuss each in turn.

The Variable Being Predicted

The variable that is the focus of a multiple regression design is the one being predicted. In the regression equation, as we have already seen for simple linear regression, it is designated as an upper case Y_{pred}. This variable is known as the *criterion variable* but is often referred to as the *dependent variable* in the analysis. It needs to have been assessed on one of the quantitative scales of measurement.

The Variables Used as Predictors

The variables used as predictors comprise a set of measures designated with upper case *X*s and known as the *predictor variables* or the *independent variables* in the analysis.

You are probably aware that in many research design courses, the term independent variable is reserved for a variable in the context of an experimental study. Some of the differences in the typical nature of independent variables in experimental and regression studies are listed in Table 5a.1.

Multiple Regression Research

If the research problem is expressed in a form that either specifies or implies prediction, multiple regression becomes a viable candidate for the design. Here are some examples of research objectives that imply a regression design:

- Wanting to predict one variable from a combined knowledge of several others
- Wanting to determine which variables of a larger set are better predictors of some criterion variable than others
- Wanting to know how much better we can predict a variable if we add one or more predictor variables to the mix

Table 5a.1 Some Differences in How Independent Variables are Treated in Experimental and Regression Studies

Independent Variables in Experimental Study	*Independent Variables in Regression Study*
Often actively manipulated but can also be an enduring (e.g., personality) characteristic of research participants.	Usually an enduring (e.g., personality) characteristic of research participants.
Uncorrelated so long as cells in the design have equal sample sizes; as cells contain increasingly unequal sample sizes the independent variables become more correlated.	All else equal, we would like them to be uncorrelated, but they should be correlated to some extent if that more appropriately reflects the relationships in the population.
Usually nominal (qualitatively measured) variables.	Usually quantitatively measured variables.
Usually coded into a relatively few levels or categories.	Usually fully continuous if possible.

- Wanting to examine the relationship of one variable with a set of other variables
- Wanting to statistically explain or account for the variance of one variable using a set of other variables

The goal of multiple regression is to produce a model in the form of a linear equation that identifies the best weighted combination of independent variables in the study to optimally predict the criterion variable. Its computational procedure conforms to the ordinary least squares solution; the solution or model describes a line for which the sum of the squared differences between the predicted and actual values of the criterion variable is minimal. These differences between the predictions we make with the model and the actual observed values are the prediction errors. The model can be thus thought of as representing the function that minimizes the sum of the squared errors. When we say that the model is fitted to the data to "best" predict the dependent variable, what we technically mean is that the sum of squared errors has been minimized.

Because the model configures the predictors together to maximize prediction accuracy, the specific weight (contribution) assigned to each independent variable in the model is relative to the other independent variables in the analysis. Thus, we can say only that considering this particular

set of variables, this one variable is able to predict the criterion to such and such an extent. In conjunction with a different set of independent variables, the predictive prowess of that variable may turn out to be quite different.

It is possible that variables not included in the research design could have made a substantial difference in the results. Some variables that could potentially be good predictors may have been overlooked in the literature review, measuring others may have demanded too many resources, and still others may not have been amenable to the measurement instrumentation available to researchers at the time of the study. To the extent that potentially important variables were omitted from the research, the model is said to be incompletely specified and may therefore have less external validity than is desirable.

Because of these concerns, we want to select the variables for inclusion in the analysis based on as much theoretical and empirical rationale as we can bring to bear on the task. It is often a waste of research effort to realize after the fact that a couple of very important candidate predictors were omitted from the study. Their inclusion would have produced a very different dynamic and likely would have resulted in a very different model than we have just obtained.

The Regression Equations

Just as was the case for simple linear regression, the multiple regression equation is produced in both raw score and standardized score form. We discuss each in turn.

The Raw Score Equation

The multiple regression raw score equation is an expansion of the raw score equation for simple linear regression. It is as follows:

$$Y_{\text{pred}} = \mathbf{a} + \text{b}_1 X_1 + \text{b}_2 X_2 + \cdots + \text{b}_n X_n$$

In this equation, Y_{pred} is the predicted score on the criterion variable, the *X*s are the predictor variables in the equation, and the bs are the weights or coefficients associated with the predictors. These b weights are also referred to as *partial regression coefficients* (Kachigan, 1986) because each reflects the relative contribution of its independent variable when we are statistically controlling for the effects of all the other predictors. Because this is a raw score equation, it also contains a constant, shown as a in the equation (representing the *Y* intercept).

All the variables are in raw score form in the equation even though the metrics on which they are measured could vary widely. If we were predicting early success in a graduate program, for example, one predictor may very well be average GRE performance (the mean of the verbal and quantitative subscores), and the scores on this variable are probably going to be in the 500 to 700 range. Another variable may be grade point average, and this variable will have values someplace in the middle to high 3s on a 4-point grading scale. We will say that success is evaluated at the end of the first year of the program and is measured on a scale ranging from the low 50s to the middle 70s (just to give us three rather different metrics for our illustration here).

The b weights computed for the regression equation are going to reflect the raw score values we have for each variable (the criterion and the predictor variables). Assume that the results of this hypothetical study show the b weight for grade point average to be about 5 and for GRE to be about .01 with a *Y* intercept value of 46.50. Putting these values into the equation would give us the following prediction model:

$$Y_{\text{pred}} = 46.50 + (5)\ (\text{gpa}) + (.01)\ (\text{GRE})$$

Suppose that we wished to predict the success score of one participant, Erin, based on her grade point average of 3.80 and her GRE score of 650. To arrive at her predicted score, we place her values into the variable slots in the equation. Here is the prediction:

$$\begin{aligned} Y_{\text{pred}} &= 46.50 + (5)\ (\text{gpa}) + (.01)\ (\text{GRE}) \\ Y_{\text{pred Erin}} &= 46.50 + (5)\ (\text{gpa}_{\text{Erin}}) + (.01)\ (\text{GRE}_{\text{Erin}}) \\ Y_{\text{pred Erin}} &= 46.50 + (5)\ (3.80) + (.01)\ (650) \\ Y_{\text{pred Erin}} &= 46.50 + (19) + (6.50) \\ Y_{\text{pred Erin}} &= 72 \end{aligned}$$

This computation allows you to see, to some extent, how the b weights and the constant came to achieve their respective magnitudes. Although they are all interdependent, we will arbitrarily start with the constant of 46.50 as a given. Analogous to simple linear regression, this value would be Erin's predicted success score if her GRE and grade point average were both zero, a situation, obviously, that would not exist empirically. This value of 46.50 is in the regression equation simply to make the predicted value work out properly.

Now, recall that success, the *Y* variable, ranges between 52 and 75. So how do you obtain Erin's predicted score in the low 70s given a constant of

46.5? Well, grade point average must be in the high 3s, so the b weight for it will have to be high enough for the result of the multiplication to add a decent number to 46.50. On the other hand, Erin's GRE score is mid-600. To predict a 72 in combination with grade point average, the GRE value has to be substantially stepped down, and you need a multiplier considerably less than 1 to make that happen.

Because the variables are assessed on different metrics, it follows that you cannot see from the b weights which independent variable is the stronger predictor in this model. Some of the ways by which you can evaluate the relative contribution of the predictors to the model will be discussed shortly.

The Standardized Equation

The multiple regression standardized score equation is an expansion of the standardized score equation for simple linear regression. It is as follows:

$$Y_{z\,\text{pred}} = \beta_1 X_{z1} + \beta_2 X_{z2} + \cdots + \beta_n X_{zn}$$

Everything in this equation is in standardized score form. Unlike the situation for the raw score equation, all the variables are now measured on the same metric—the mean and standard deviation for all the variables (the criterion and the predictor variables) are 0 and 1, respectively.

In the standardized equation, $Y_{z\,\text{pred}}$ is the predicted z score of the criterion variable. Each predictor variable (each X in the equation) is associated with its own weighting coefficient symbolized by β and called a beta weight, standardized regression coefficient, or beta coefficient, and just as was true for the b weights in the raw score equation, they are also referred to as partial regression coefficients. These coefficients usually compute to a decimal value, but they can exceed the range of ±1 if the predictors are correlated enough between themselves.

Each βX combination represents the z score of a predictor and its associated beta weight. With the equation in standardized form, the Y intercept is zero and is therefore not shown.

We can now revisit the example used above where we predicted success in graduate school based on grade point average and GRE score. Here is that final equation but this time in standard score form:

$$Y_{z\,\text{pred}} = \beta_1 X_{z1} + \beta_2 X_{z2} + \cdots + \beta_n X_{zn}$$

$$Y_{z\,\text{pred}} = (.48)\ (\text{gpa}_z) + (.22)\ (\text{GRE}_z)$$

We can also apply this standardized regression equation to individuals in the sample—for example, Erin. Within the sample used for this study, assume that Erin's grade point average of 3.80 represents a z score of 1.80 and that her GRE score of 650 represents a z score of 1.70. We can thus solve the equation as follows:

$$
\begin{aligned}
Y_{z\,\text{pred}} &= \beta_1 X_{z1} + \beta_2 X_{z2} + \cdots + \beta_n X_{zn} \\
Y_{z\,\text{pred}} &= (.48)\ (\text{gpa}_z) + (.22)\ (\text{GRE}_z) \\
Y_{z\,\text{pred Erin}} &= (.48)\ (\text{gpa}_{z\,\text{Erin}}) + (.22)\ (\text{GRE}_{z\,\text{Erin}}) \\
Y_{z\,\text{pred Erin}} &= (.48)\ (1.80) + (.22)\ (1.70) \\
Y_{z\,\text{pred Erin}} &= (.864) + (.374) \\
Y_{z\,\text{pred Erin}} &= 1.24
\end{aligned}
$$

The Variate in Multiple Regression

As was discussed in the overview at the very start of the book, multivariate procedures typically involve building, developing, or solving for a weighted combination of variables. This combination is called a *variate*. In the case of multiple regression, we are dealing with a variate made up of a weighted combination of the predictors or independent variables in the analysis. The variate in this instance is the entity on the right side of the multiple regression equation.

Although the variate may not be a measured variable, it is still important in the context of multiple regression. It is often possible to view this variate as representing some underlying dimension or construct (i.e., a latent variable). In the preceding example where we were predicting success in graduate school, the variate might be interpreted as "academic aptitude" indexed by the linear combination of grade point average and GRE score. From this perspective, indicators of academic aptitude were selected by the researchers to be used in the study. They then used the regression technique to shape the most effective academic aptitude variate to predict success in graduate school.

Based on the previous example, the academic aptitude variate is built to do the best job possible to predict a value on a variable. That variable is the predicted success score. Note that the result of applying the multiple regression equation—the result of invoking the linear composite of the predictor variables, the variate—is the predicted success score and not the actual success score. For most of the cases in the data file, the predicted and the actual success scores of the students will be different. The model minimizes these differences; it does not eliminate them. Thus, the variable "predicted

success score" and the variable "actual success score" are different variables, although we certainly hope that they are reasonably related to each other. The variate that we have called academic aptitude generates the predicted rather than the actual value of the success score.

A Range of Regression Methods

The main work done in multiple regression is to build the prediction equation. This involves generating the weighting coefficients—the b weights for the raw score equation and the beta coefficients for the standardized equation—as well as the *Y* intercept for the raw score equation.

Several different methods are available to researchers to build the variate or linear function; these can be organized into two groups or classes. One subset of methods relies exclusively on statistical decision-making criteria built into the computer programs to decide at which point in the process which predictors are to be entered into the equation. These are ordinarily called, as a class, *statistical* methods. The most popular of these statistical methods include the standard, forward, backward, and stepwise methods although others (not covered here), such as max *R*-squared and min *R*-squared, have been developed as well. In using these methods, researchers permit the computer program to autonomously carry out the analyses.

The other subset of methods calls for the researchers to determine which predictors are to be entered into the regression equation at each stage of the analysis. Thus, the researcher rather than the computer program assumes control of the regression procedure. These researcher-based decisions regarding order of entry are typically derived from the theoretical model with which the researchers are working.

The Standard (Simultaneous) Regression Method

The *standard regression method*, also called the *simultaneous* or the *direct method,* is what most authors refer to if they leave the method unspecified. It is the most widely used statistical method. Under this method, all the predictors are entered into the equation in a single "step" (stage in the analysis). The standard method provides a full model solution in that all the predictors are part of it.

The idea that these variables are entered into the equation simultaneously is true only in the sense that the variables are entered in a single step. But that single step is not at all simple and unitary; when we look inside this step, we will find that the process of determining the weights for independent variables is governed by a complex strategy.

The Example to Be Used

Rather than referring to abstract predictors and some amorphous dependent variable to broach this topic, we will present the standard regression method by using an example with variables that have names and meaning. To keep our drawings and explication manageable, we will work with a smaller set of variables than would ordinarily be used in a study conceived from the beginning as a regression design. Whereas an actual regression design might typically have from half a dozen to as many as two dozen or more variables as potential predictors, we will use a simplified example of just three predictors for our presentation purposes.

We have taken our variables from a larger study in which we collected data from 420 college students. The dependent variable we use for this illustration is self-esteem as assessed by Coopersmith's (1981) Self-Esteem Inventory. Two of the predictors we use for this illustration are Tellegen's (1982) measures of the number of positive and negative affective behaviors a person ordinarily exhibits. The third independent variable represents scores on the Openness scale of the NEO Five-Factor Personality Inventory (Costa & McCrae, 1992). Openness generally assesses the degree to which respondents appear to have greater aesthetic sensitivity, seek out new experiences, and are aware of their internal states.

It is always desirable to initially examine the correlation matrix of the variables participating in a regression analysis. This gives researchers an opportunity to examine the interrelationships of the variables, not only between the dependent variable and the independent variables but also between the independent variables themselves.

Table 5a.2 displays the correlation matrix of the variables in our example. We have presented it in "square" form where the diagonal from upper left to lower right (containing the value 1.000 for each entry) separate the matrix into two redundant halves. As can be seen, the dependent variable of self-esteem is moderately correlated with both positive and negative positive affect but is only modestly correlated with openness. It can also be

Table 5a.2 Correlation Matrix of the Variables in the Regression Analysis

	Esteem	*PosAfect*	*NegAfect*	*NeoOpen*
Esteem	1.000	.555	−.572	.221
PosAfect	.555	1.000	−.324	.221
NegAfect	−.572	−.324	1.000	−.168
NeoOpen	.221	.221	−.168	1.000

seen that positive and negative affect correlate more strongly with each other than either does with openness.

Building the Regression Equation

The goal of any regression procedure is to predict or account for the variance of the criterion variable. In this example, that variable is self-esteem. At the beginning of the process, before the predictors are entered into the equation, 100% of the variance of self-esteem is unaccounted for. This is shown in Figure 5a.1. The dependent variable of self-esteem is in place, and the predictors are ready to be evaluated by the regression procedure.

On the first and only step of the standard regression procedure, all the predictors are entered as a set into the equation. But to compute the weighting coefficients (b weights for the raw score equation and beta weights for the standardized equation), the predictors must be individually evaluated. To accomplish this, and this is the essence of standard regression, each predictor's weight is computed as though it had entered the equation last.

The idea of treating each predictor as if it was the last to enter the model is to determine what predictive work it can do over and above the prediction attributable to the rest of the predictors. In this manner, standard regression focuses on the unique contribution that each independent variable makes to the prediction when combined with all the other predictors.

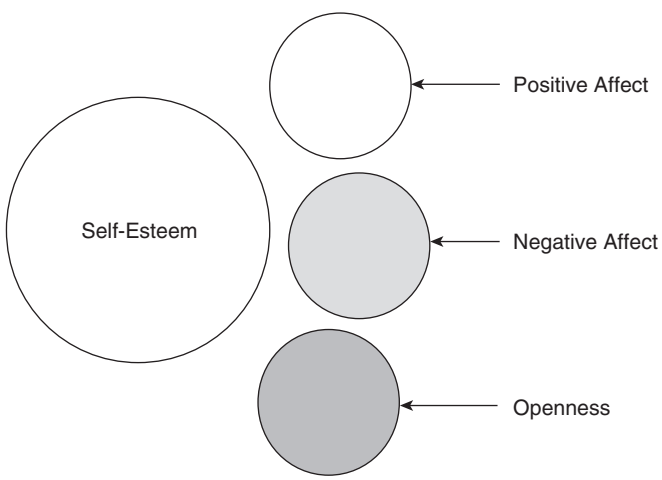

Figure 5a.1 Self-Esteem Dependent Variable Prior to Regression Analysis

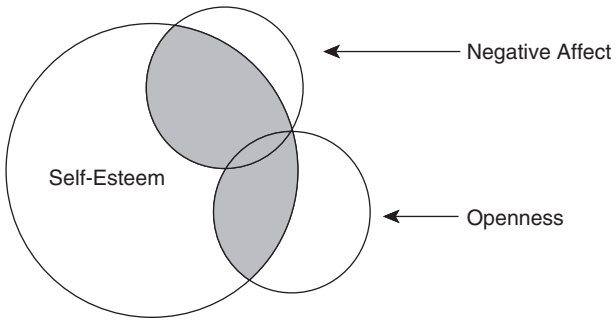

Figure 5a.2 Self-Esteem Variance Accounted for by Simultaneous Entering of Negative Affect and Openness Predictors

The way in which standard regression assesses the unique contribution of each independent variable is the key to understanding the standard method, and we will go through the process here.

The Squared Multiple Correlation

We can demonstrate the dynamics of assessing the unique contribution of each independent variable by focusing on how one of these predictors—say, positive affect—is evaluated. In determining the weight that positive affect will receive in the regression equation, the program momentarily places the other predictors (negative affect and openness) in the equation. This is illustrated by the diagram in Figure 5a.2.

Negative affect and openness are both entered into the equation simultaneously. Their relationship to the dependent variable, self-esteem, is shown in the Venn diagram in Figure 5a.2. Two features of this depiction are important to note at this point.

First, this diagram still represents a correlation. If the criterion variable was shown with just a single predictor, you would immediately recognize a representation of the Pearson (or any bivariate) correlation. The shaded area would show the strength of the correlation, and its magnitude would be indexed by r^2.

The relationship shown in Figure 5a.2 is more complex than that. Three variables, not two, are involved in the relationship. Specifically, we are looking at the relationship of the criterion (self-esteem) to two predictors (negative affect and openness). When we have three or more variables involved in the relationship, we can no longer use the Pearson correlation coefficient to quantify the magnitude of that relationship—the Pearson r can index the

degree of relationship only when two variables are being considered. The correlation coefficient we need to call on to quantify the degree of a more complex relationship is known as the *multiple correlation.* It is symbolized as an uppercase italic *R.*

A multiple correlation coefficient indexes the association of one variable with a set of other variables, and the *squared multiple correlation* (R^2), sometimes called the *coefficient of multiple determination,* tells us the strength of this complex relationship.

In Figure 5a.2, the shaded area—the overlap of negative affect and openness with self-esteem—represents the R^2 for that relationship. This R^2 value can be thought of in a way analogous to r^2; that is, it can be thought of in terms of explained or accounted-for variance. In this case, we are explaining the variance of self-esteem.

The R^2 value represents one way to evaluate the model. Larger values mean that the model has accounted for greater amounts of the variance of the dependent variable. How large an R^2 it takes to say that you have accounted for a "large" percentage of the variance depends on the theoretical context within which the research has been done as well as prior research in the topic area.

The second feature important to note in Figure 5a.2 is that negative affect and openness overlap with each other. In Venn diagram format, an overlap of the variables indicates a correlation between them. Here, the two predictors do overlap but not by all that much (they correlate –.17). The degree to which they correlate affects the beta weights these variables are assigned in the regression equation, so the correlations of the predictors become a matter of some interest to researchers using a regression design.

The Partial Correlation and Covariance

With these two other variables in the equation for the moment, we are ready to evaluate the contribution of positive affect. The criterion variable or dependent variable is the focus of the multiple regression design. It is therefore the variance of the dependent variable that we want to account for or predict, and our goal is to account for as much of it as possible with our set of independent variables. We face an interesting but subtle feature of multiple regression in its efforts to maximize the amount of dependent variable variance that we can account for. In the context of multiple regression, our predictors must account for separate portions—rather than the same portion—of the dependent variable's variance. This is the key to understanding the regression process.

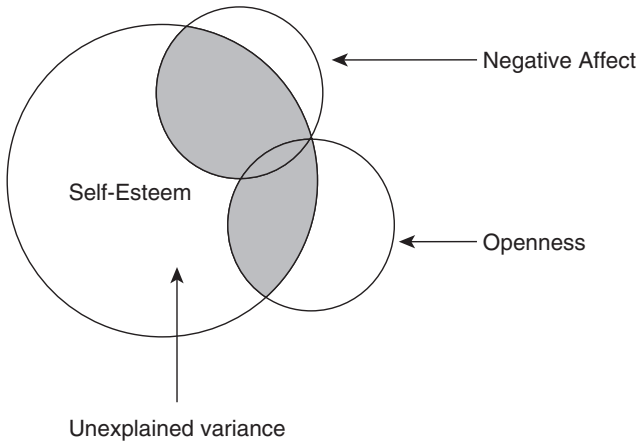

Figure 5a.3 Outlined Area of Unexplained Residual Variance for Self-Esteem Known as $(1 - R^2)$

With negative affect and openness already in the model, and thus already accounting for variance the amount of which is indexed by R^2, positive affect, as the last variable to enter, must target the variance that remains—the *residual* variance—in self-esteem. This is shown in Figure 5a.3, which is the same diagram that was shown in Figure 5a.2 except that we have added a couple of features to it. The shaded area in Figure 5a.3 is the variance of the dependent variable (self-esteem) explained by the two independent variables. It is indexed by R^2. The remaining portion of the dependent variable variance is, by definition, not accounted for by these two predictors. It is shown in the diagram as the blank space in the circle representing self-esteem, and its value must be $1 - R^2$. That is, it is the residual variance of self-esteem after negative affect and openness have performed their predictive (correlational) work.

We have outlined that scallop-like shape showing the unexplained variance of self-esteem with negative affect and openness in the equation by using a heavy line to make it easier to see and have labeled it as "unexplained variance" in Figure 5a.3. In evaluating the contribution of positive affect, the predictor currently under consideration, it is this residual variance of self-esteem that positive affect must target. The question becomes how much of this residual variance can positive affect correlate with on its own.

However strange this sounds, we are talking about the correlation between positive affect and the residual variance of self-esteem when the effects of negative affect and openness have been statistically removed, controlled, or "partialled out." Such a correlation is called a *partial correlation.*

A partial correlation addresses the relationship between two variables when the effects of other variables have been statistically removed from one of them. In this sense, the variables already in the model are conceived of as *covariates* in that their effects are statistically accounted for prior to evaluating the relationship of positive affect and self-esteem.

Once the regression procedure has determined how much positive affect can contribute to the set of predictors already in the model (how much of the residual variance of self-esteem it can explain), the computer starts the process of computing the weight that positive affect will receive in the model. We will not get into that computational process here. Instead, we have presented the situation depicting the results of those computations in Figure 5a.4. In this figure, we have added the positive affect variable into the predictor set. The "prediction work" that it does is shown in a darker cross-hatched fill. Note that some of the prediction supplied by positive affect is not accomplished by any other variable and that "other" of what positive affect predicts for self-esteem is also predicted by negative affect.

Repeating the Process for the Other Predictors

After the computation of the b and beta weights for positive affect have been made, it is necessary to evaluate another one of the predictors. Thus, positive affect and another predictor are entered into the equation, and the strategy we have just outlined is repeated for the remaining predictor. Each

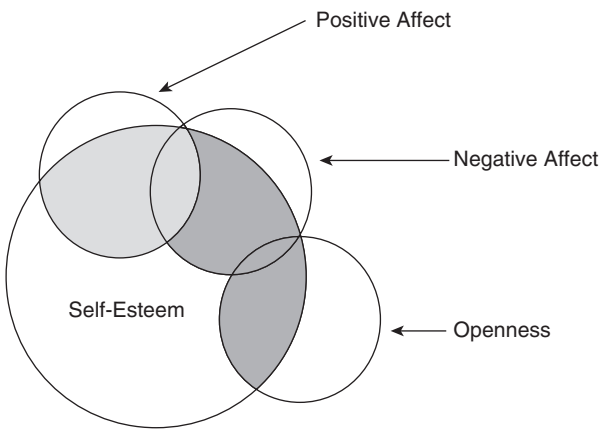

Figure 5a.4 Additional Self-Esteem Variance Accounted for After Entering Positive Affect

independent variable is put through this same process until the weights for all have been determined. At the end of this complex process, the final weights are locked in and the results of the analysis are printed.

We also know the value of R^2 with all the variables in the equation. This final R^2 tells us how much variance of the dependent variable is accounted for by the full regression model. By subtracting that value from 1 $(1 - R^2)$, we can also ascertain how much of the dependent variable's variance remains unexplained; this is the residual variance of the dependent variable after the regression model has accomplished its predictive work. Obviously, adding the value of the coefficient of multiple determination (R^2) to the residual variance $(1 - R^2)$ results in a value of 1.00; this subsumes 100% of the variance of the dependent variable.

The Squared Semipartial Correlation

A Venn diagram suggesting the final solution is shown in Figure 5a.5. We say "suggesting" because even with as small a set as three independent variables, it is difficult to draw all the relationships between them in only two dimensions (we have not captured the correlation between positive affect and openness). As a result, such a pictorial representation is at best an approximation to the full mathematical solution, which we will present in the next section of this chapter.

Despite the shortcoming of using the Venn diagram here, we can still point out an important element of the solution. Note that we have used two different types of shading in the figure, cross-hatching and slanted-line fill. The total filled-in area, combining across all fill portions, represents the total amount of self-esteem variance explained by the regression model, a quantity indexed by R^2.

In Figure 5a.5, the darker cross-hatched areas are associated with explained variance resulting from the overlapping of predictors. Positive and negative affect, for example, explain a common portion of self-esteem variance, which is shown by the dark cross-hatched area between them.

The slanted-line areas are components of explained variance unique to a single predictor; that is, there is no overlap with the other predictors in those regions. This uniquely explained variance is indexed by another correlation statistic known as the *squared semipartial correlation.* It represents the extent to which variables do independent predictive work when combined with the other predictors in the model. Such correlations are, therefore, strongly tied to the specific regression model and may not necessarily generalize if any of these predictors are combined with a different set of predictors in a subsequent study.

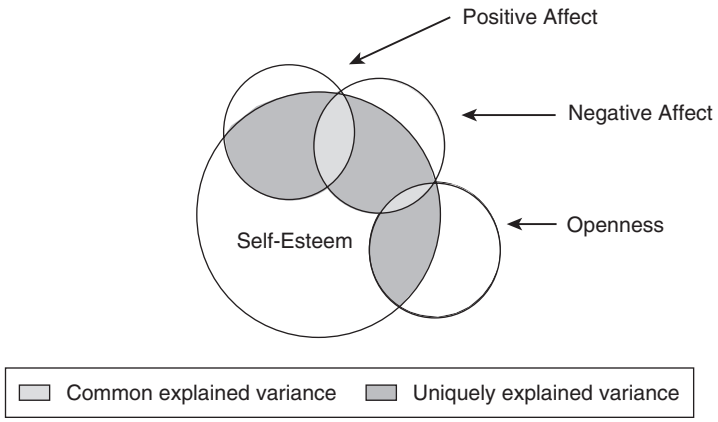

Figure 5a.5 Depiction of Both Common and Uniquely Explained Variance

We can also evaluate how well the model works by examining the squared semipartial correlations (Tabachnick & Fidell, 2001b). With the squared semipartial correlations, you are looking directly at the unique contribution of each predictor within the context of the model, and clearly, independent variables with larger squared semipartial correlations are making a larger unique contribution.

There are some limitations in using this statistic to compare the contributions of the predictors. The unique contribution of each variable in multiple regression is very much a function of the correlations of the variables used in the analysis. It is quite likely, as we stated earlier, that within the context of a different set of predictors, the unique contributions of these variables would change, perhaps substantially. Of course, this argument is true for the beta coefficients as well.

Based on this line of reasoning, one could put forward the argument that it would therefore be extremely desirable to select predictors in a multiple regression design that are not at all correlated between themselves but are highly correlated with the criterion variable. In such a fantasy scenario, the predictors would account for different portions of the dependent variable's variance, the squared semipartial correlations would be substantial, and the overlap of the predictors in Venn diagram format would be minimal.

This argument may have a certain appeal at first glance, but it is not a viable strategy for both practical and theoretical reasons. On the practical side, it would be difficult or perhaps even impossible to find predictors in many research arenas that are related to the criterion variable but at the same time are not themselves at least moderately correlated. On the theoretical side, it is desirable that the correlations between the predictors in a

research study are representative of those relationships in the population. All else equal, to the extent that variables are related in the study as they are in the outside world, the research results may be said to have a certain degree of external validity.

The consequence of moderate or greater correlation between the predictors is that the unique contribution of each independent variable may be relatively small in comparison with the total amount of explained variance of the prediction model, because the predictors in such cases may overlap considerably with each other. Comparing one very small semipartial value with another even smaller semipartial value is often not a productive use of your time and runs the risk of yielding distorted or inaccurate conclusions.

Structure Coefficients

In our discussion of the variate, we emphasized that there was a difference between the predicted value and the actual score that individuals obtained on the dependent variable. Our focus here is on the predicted score, which is the value of the variate for the particular values of the independent variables substituted in the model. The *structure coefficient* is the bivariate correlation between a particular independent variable and the predicted (not the actual) score (Dunlap & Landis, 1998). Each predictor is associated with its own structure coefficient.

The numerical value of the structure coefficient is not contained in the output of SPSS but is easy to compute with a hand calculator using the following information available in the printout:

$$\text{Structure Coefficient} = \frac{r_{\text{IV} \times \text{DV}}}{R}$$

where $r_{\text{IV} \times \text{DV}}$ is the Pearson correlation between the given predictor and the actual (measured) dependent variable and R is the multiple correlation. The structure coefficient indexes the correlation between the predictor and the variate; stronger correlations indicate that the predictor is a stronger reflection of the construct underlying the variate.

Summary of the Solution for the Standard Regression Method Example

The results of the regression procedure for our simplified example are displayed in Table 5a.3. For each predictor, we have shown its Pearson correlation (r) with the dependent variable, its raw (b) and standardized (beta) regression weighting coefficients, the amount of self-esteem variance it has

Table 5a.3 Summary of the Example for Multiple Regression

Variable	*r*	*b*	*beta*	*Squared Semipartial Correlation*	*Structure Coefficients*	*t*
Positive affect	.55	2.89	.40	.14	.80	10.61*
Negative affect	−.57	−2.42	−.43	.16	−.82	−11.50*
Openness	.22	.11	.06	.00	.32	1.64
Constant (*Y* intercept)		56.66				

*$p < .01$.

uniquely explained (squared semipartial correlation), its structure coefficient, and the *t* value associated with each regression weight. We will discuss each in turn. The constant (the *Y* intercept) is shown in the last line of the table.

The Regression Equations

Using the raw and standardized regression weights and the *Y* intercept shown in Table 5a.3, we have the elements of the two regression equations. We produce them below.

The raw score equation is as follows:

$$\text{Self-esteem}_{\text{pred}} = 56.66 + (2.89)(\text{pos affect}) - (2.42)(\text{neg affect}) + (.06)(\text{open})$$

The standardized equation is as follows:

$$\text{Self-esteemz}_{\text{pred}} = (.40)(\text{pos affect}_z) - (.43)(\text{neg affect}_z) + (.06)(\text{open}_z)$$

Variables in the Equation

The predictor variables are shown in the first column of the table. This represents a complete solution in the sense that all the independent variables are included in the final equation regardless of how much they contribute to the prediction model. Such a solution is considered *atheoretical* because all the variables that were originally assessed are included in the final solution.

R^2 and Adjusted R^2

The shaded areas in Figure 5a.5 (the slanted-line areas together with the darker cross hatched areas) represent the total amount of variance accounted for by the prediction model. The computer printout shows the actual value for R^2. In the present case, this turned out to be .48 rounded to two decimal places. Thus, the three predictors in this particular weighted linear combination were able to explain about 48% of the variance of self-esteem.

SPSS also prints an adjusted R^2 value, which essentially tries to take into account a bit of error inflation in the regular R^2 value. Because it is a human endeavor, there is always some error of measurement associated with anything we assess. If this error is random, as we assume it to be, then some of that measurement error will actually be in the direction of enhanced prediction. Multiple regression, however, is unable to distinguish between this chance enhancement (i.e., blind luck from the standpoint of trying to achieve the best possible R^2) and the real predictive power of the variables. So it uses everything it can to maximize prediction—it generates the b and beta weights from both true and error sources combined.

The problem for us is that in another sample the random dictates of error will operate differently, and if the old weighting coefficients are applied to the new sample, they will be less effective than they were in the original sample. This overprediction is more of a problem when we have relatively small sample sizes and relatively more variables in the analysis. As sample size reaches more acceptable proportions (20 or more cases per predictor), the inflation of R^2 becomes that much less of an issue. Nonetheless, virtually every statistical program computes an adjusted value for R^2. These programs attempt to extract from the computed R^2 value some portion of it to which we can ascribe error and then subtract that out. We recommend that you report the adjusted R^2 value in addition to the uncorrected value.

The adjusted R^2 is a statistical estimate of the *shrinkage* we would observe if we were to apply the model to another sample. We can instead approach the issue from an empirical perspective through the processes of either *cross-validation* or *double cross-validation.* To perform a cross-validation, we ordinarily divide a large sample in half (into two subsamples) by randomly selecting the cases to be assigned to each. We then compute our regression analysis on one subsample and use those regression weights to predict the criterion variable of the second "hold-back" sample. The R^2 difference tells us the degree of predictive loss we have observed. We can also correlate the predicted scores of the hold-back sample with their actual scores; this can be thought of as the *cross-validation coefficient.*

Double cross-validation can be done by performing the cross-validation process in both directions—that is, performing the regression analysis on each subsample and applying the results to the other. In a sense, you obtain two estimates of shrinkage rather than one. If the shrinkage is not excessive, and there are few guidelines as to how to judge this, you can then perform an analysis on the combined sample and report the double cross-validation results to let readers know how generalizable your model is.

In the present example, the adjusted R^2 value for this analysis is rounded to .48, giving us virtually the same value as the unadjusted R^2 (the actual R^2 was .48257 and the adjusted R^2 was .47883). That such little adjustment was made is probably a function of the sample size to number of variables ratio we used and the fact that we used a very small predictor set.

By virtue of our sample size ($N = 420$) the R^2 of .48 obtained here is clearly statistically significant. However, SPSS tests the efficacy of the model by an analysis of variance. In this case, we can say that these three independent variables in combination significantly predicted self-esteem, $F(3, 416) = 129.32$, $p < .05$, $R^2 = .48$, adjusted $R^2 = .48$. This information is part of the printout as we will see in Chapter 5B.

We should also consider the magnitude of the R^2 obtained here. One would ordinarily think of .48 as reasonably substantial, and you should not be terribly disappointed with R^2s considerably less than this in your own study. In the early stages of a research project or when studying a variable that may be complexly determined (e.g., rate of spread of an epidemic, recovery from a certain disease), very small but statistically significant R^2s may be cause for celebration by a research team.

Pearson Correlations With the Criterion Variable

The second numerical column in Table 5a.3 shows the simple Pearson correlations between self-esteem and each of the predictors. We have briefly described the correlations earlier. For present purposes, you can see that the correlations between self-esteem and positive affect and openness are positive. This was the case because each of these variables are scored in the positive direction—higher scores mean that respondents exhibit more positive affective behaviors and that they are more open to new or interesting experiences, respectively. Because higher scores on the self-esteem scale indicate greater positive feelings about oneself, it is not surprising that these two predictors are positively correlated with it. On the other hand, negative affect is negatively correlated with self-esteem. This is also not surprising in that individuals who exhibit more negative affective behaviors are typically those who have lower levels of self-esteem.

b and Beta Coefficients

The b and beta coefficients in Table 5a.3 show us the weights that the variables have been assigned at the end of the equation-building process. The b weights are tied to the metrics on which the variables are measured and are therefore difficult to compare with one another. But with respect to their own metric, they are quite interpretable. The b weight for positive affect, for example, is 2.89. We may take it to mean that when the other variables are controlled for, an increase of 2.89 points on the positive affect measure is, on average, associated with a 1-point gain in self-esteem.

Table 5a.3 also shows the *Y* intercept for the linear function. This value of 56.66 would need to be added to the weighted combination of variables in the raw score equation to obtain the predicted value of self-esteem for any given research participant.

The beta weights for the independent variables are also shown in Table 5a.3. Here, all the variables are in *z*-score form and thus their beta weights, within limits, can be compared with each other. We can see from Table 5a.3 that positive and negative affect have beta weights of similar magnitudes and that openness has a very low beta value. Thus, in achieving the goal of predicting self-esteem to the greatest possible extent (to minimize the sum of the squared prediction errors), positive and negative affect are given much more weight than openness.

The Case for Using Beta Coefficients to Evaluate Predictors

Some authors (e.g., Cohen, Cohen, West, & Aiken, 2003; Pedhazur, 1982, 1997; Pedhazur & Schmelkin, 1991) have cautiously argued that at least under some circumstances, we may be able to compare the beta coefficients with each other. That is, on the basis of visual examination of the equation, it may be possible to say that predictors with larger beta weights contribute more to the prediction of the dependent variable than those with smaller weights.

It is possible to quantify the relative contribution of predictors using beta weights as the basis of the comparison. Although Kachigan (1986) has proposed examining the ratio of the squared beta weights to make this comparison, that procedure may be acceptable only in the rare situation when those predictors whose beta weights are being compared are uncorrelated (Pedhazur & Schmelkin, 1991). In the everyday research context, where the independent variables are almost always significantly correlated, we may simply compute the ratio of the actual beta weights (Pedhazur, 1982, 1997; Pedhazur & Schmelkin, 1991), placing the larger beta weight in the numerator of the ratio. This ratio reveals how much more one independent variable contributes to prediction than another within the context of the model.

This comparison could work as follows. If we wanted to compare the efficacy of negative affect (the most strongly weighted variable in the model) with the other (less strongly weighted) predictors, we would ordinarily limit our comparison to only the statistically significant ones. In this case, we would compare negative affect only with positive affect. We would therefore compute the ratio of the beta weights (negative affect / positive affect) without carrying the sign of the beta through the computation. This is shown below:

$$\frac{\text{negative affect}}{\text{positive affect}}$$

$$\frac{-.43}{.40} = 1.075$$

Based on this ratio (although we could certainly see this just by looking at the beta weights themselves), we would say that negative and positive affect make approximately the same degree of contribution to the prediction of self-esteem in the context of this research study with the present set of variables.

Concerns With Using the Beta Coefficients to Evaluate Predictors

We indicated above that even when authors such as Pedhazur (1982, 1997; Pedhazur & Schmelkin, 1991) endorse the use of beta coefficient ratios to evaluate the relative contribution of the independent variables within the model, they usually do so with certain caveats. Take Pedhazur (1997) as a good illustration:

> Broadly speaking, such an interpretation [stating that the effect of one predictor is twice as great as the effect of a second predictor] is legitimate, but it is not free of problems because the Beta[s] are affected, among other things, by the variability of the variable with which they are associated. (p. 110)

Thus, beta weights may not be generalizable across different samples.

Another concern regarding using beta coefficients to evaluate predictors is that beta weight values are partly a function of the correlations between the predictors themselves. That is, a certain independent variable may predict the dependent variable to a great extent in isolation, and one would

therefore expect to see a relatively high beta coefficient associated with that predictor. Now place another predictor that is highly correlated with the first predictor into the analysis and all of a sudden the beta coefficients of both predictors can plummet. The first predictor's relationship with the dependent variable has not changed in this scenario, but the presence of the second correlated predictor could seriously affect the magnitude of the beta weight of the first. This "sensitivity" of the beta weights to the correlations between the predictors, reflected in the beta values, places additional limitations on the generality of the betas and thus their use in evaluating or comparing predictive effectiveness of the independent variables.

Recommendations for Using Betas

We do not want to leave you completely hanging at this point in our treatment, so we will answer some obvious questions. Should you use the beta weights to assess the relative strengths of the predictors in your own research? Yes. Should that be the only index you check out? No. The structure coefficients and the squared semipartial correlations should be examined as well.

Positive Versus Negative Weights

The positive and negative regression weights of the predictors reflect the nature of their respective correlations with the dependent variable. This makes sense when you recall that we are predicting self-esteem. The regression model tells us that greater levels of self-esteem will be predicted by the combination of more positive affect and openness and less negative affect. Thus, we should be adding the contribution of positive affect and openness but subtracting the contribution of negative affect in predicting self-esteem.

Squared Semipartial Correlations

The fourth column of Table 5a.3 displays the squared semipartial correlations for each predictor. These correlations are shown in the SPSS printout as "part correlations" and appear in the printout in their nonsquared form. This statistic indexes the variance accounted for uniquely by each predictor in the full model. What is interesting here, and this is pretty typical of multiple regression research, is that the sum of these squared semipartial correlations is less than the R^2. That is, .14, .16, and .00 add up to .30 and not to the R^2 of .48.

The reason these squared semipartial correlations do not add to the value of R^2 is that the independent variables overlap (are correlated) with

each other. Here, 30% of the variance is accounted for uniquely by the predictors, whereas (by subtraction) 18% of the accounted for variance is handled by more than one of them. We therefore have some but not a huge amount of redundancy built into our set of predictors.

Using the squared semipartial correlations as a gauge of the relative strength of the predictors results in an evaluation similar to the one we made based on comparing the beta weights. From this perspective, positive and negative affect are approximately tied in their unique contribution to the prediction model under the present research circumstances.

The Structure Coefficients

The next-to-last column in Table 5a.3 shows the structure coefficients. These needed to be hand calculated because SPSS does not provide them. For each independent variable in the table, we divided the Pearson r representing the correlation of the independent variable and the dependent variable (shown in the second numerical column) by the multiple correlation. To illustrate, the square root of .48 (the R^2) is approximately .69. For positive affect's structure coefficient, we divided .55 by .69 to obtain approximately .80.

The structure coefficients indicate that positive and negative affect are reasonably highly correlated with predicted self-esteem and so are very reasonable instances of (they correlate reasonably highly with) the variate. In this example, using the structure coefficients as a basis to compare the contribution of the predictors presents the same picture as those painted by the beta weights and the squared semipartial correlations. Such consistency, however, is not always obtained.

Beta coefficients and structure coefficients differ in at least two important ways.

1. A beta coefficient associated with its predictor reflects the correlations of that predictor with the other predictors in the analysis. A structure coefficient does not take into account the degree to which that predictor correlates with the other predictors.
2. Beta weights can exceed the range of ±1 when the predictors are correlated with each other. Many researchers have a problem interpreting beta weights greater than unity. Structure coefficients are bounded by the range ±1 because they are correlation coefficients, thus making them pretty clearly interpretable.

Our recommendations are consistent with what we offered above for beta weights. We concur with Thompson and Borrello (1985) that the

structure coefficients are a useful companion index of relative predictor contribution. Pedhazur (1982) notes that structure coefficients will show the same pattern of relationships as the preregression correlations of the predictors and the criterion. Because of this, Pedhazur is not convinced of the utility of structure coefficients. In our view, by focusing on the correlation between the predictor and the variate, we believe that structure coefficients may add a nuance to the interpretation of the regression analysis that we think is worthwhile.

t tests

SPSS tests the significance of each predictor in the equation using *t* tests. The null hypothesis is that a predictor's weight is effectively equal to zero when the effects of the other predictors are taken into account. This means that when the other predictors act as covariates and this predictor is targeting the residual variance, according to the null hypothesis the predictor is unable to account for a statistically significant portion of it; that is, the partial correlation between the predictor and the criterion variable is not significantly different from zero. And it is a rare occurrence when every single independent variable turns out to be a significant predictor. The *t* tests shown in the last column of Table 5a.3 inform us that only positive and negative affect are statistically significant predictors in the model; even with our large sample size, openness does not receive a strong enough weight to reach that touchstone.

Step Methods of Building the Model

Step methods of building the regression equation that we briefly cover here are the forward method, the backward method, and the stepwise method. These methods construct the model one variable at a time rather than all at once as the standard method does. The primary goal of these step methods is to build a model with only the "important" predictors in it. The methods differ primarily in how they determine the importance of the predictors.

The Forward Method

In the forward method, rather than placing all the variables in the equation at once, we add independent variables to the equation one variable or step at a time. At each step, we enter the particular variable that adds the most predictive power at that time. If we were working with the set of variables we used to illustrate the standard regression method, negative

affect would be entered first. We know this because, with no variables in the model at the start and building the model one variable at a time, the variable correlating most strongly with self-esteem would be entered first.

In the forward method, once a variable is entered into the model, it remains in the model. For the next step, the variable with the highest partial correlation (the correlation between the residual variance of self-esteem and each remaining predictor with negative affect as a covariate) is entered if that partial correlation is statistically significant. In this case, we will assume that positive affect would be entered.

This process is repeated for each remaining predictor with the variables in the model all acting as covariates. We would find, with negative and positive affect in the model, that openness would not be entered; that is, it would not account for a significant amount of the residual variance. Because that is the entire set of predictors, the forward procedure would stop at the end of the second step.

The Backward Method

The backward method works, not by adding significant variables to the equation but, rather, by removing nonsignificant predictors from it one step at a time. The very first action performed by the backward method is the same one used by the standard method; it enters all the predictors into the equation regardless of their worth. But whereas the standard method stops here, the backward method is just getting started.

The model with all the variables in it is now examined, and the significant predictors are marked for retention. Nonsignificant predictors are then evaluated and the most expendable of them—the one whose loss would least significantly decrease the R^2—is removed from the equation. A new model is built in the absence of that one independent variable and the evaluation process is repeated. Once again, the most expendable independent variable is removed. This removal process and equation-reconstruction process continues until there are only significant predictors remaining in the equation. In our example, openness would have been removed at the first opportunity. It is probable that the method would have stopped at that point because both remaining predictors would almost certainly have been significant predictors.

Backward Versus Forward Solutions

Backward regression does not always produce the same model as forward regression even though it probably would have in our simplified

example. Here is why: Getting into the equation in the forward method requires predictors to meet a more stringent criterion than variables being retained in the equation in the backward method. This creates a situation in which it is more difficult to get into the equation than to remain in it. Stringency or difficulty is defined statistically by the alpha or probability level associated with entry and removal.

Predictors earn their way into the equation in the forward method by significantly predicting variance of the dependent variable. The alpha level governing this entry decision is usually the traditional .05 level. By most standards, this is a fairly stringent criterion. When we look for predictors to remove under the backward method, the alpha level usually drops to .10 as the default in most programs. This means that a predictor needs to be significant at only .10 (not at .05) to retain its place in the equation. Thus, an independent variable is eligible to be removed from the equation at a particular step in the backward method if its probability level is greater than .10 (e.g., $p = .11$) but it will be retained in the equation if its probability level is equal to or less than .10 (e.g., $p = .09$).

The consequences of using these different criteria for entry and removal affects only those variables whose probabilities are between the entry and removal criteria. To see why this is true, first consider variables that are not within this zone.

- If a variable does not meet the standard of $p = .10$, it is removed from the equation. This variable would also by definition not meet the .05 alpha level criterion for entry either, so there is no difference in the outcome for this predictor under either criterion—it is not going to wind up in the equation in either the forward or backward methods.
- If a variable does meet the .05 criterion, it will always be allowed entry to the equation and will certainly not be removed by the backward method; again, there is no difference in outcome for such a predictor under either method.

Variables with probability levels between these two criteria are in a more interesting position. Assume that we are well into the backward process and at this juncture the weakest predictor is one whose probability is .08. This variable would not have been allowed into the equation by the forward method if it were considered for entry at this point because to get in it would have to meet a .05 alpha level to achieve statistical significance.

However, under the backward method, this variable was freely added to the equation at the beginning, and the only issue here is whether it is to be

removed. When we examine its current probability level and find it to be .08, we determine that this predictor is "significant" at the .10 alpha level. It therefore remains in the equation. In this case, the model built under the backward model would incorporate this predictor but the model built under the forward method would have excluded it.

The Stepwise Method

The stepwise method of building the multiple regression equation is essentially a composite of the forward and backward methods. The stepwise and forward methods act in the same fashion until we reach the point where a third predictor is added to the equation. The stepwise method therefore begins with an empty equation and builds it one step at a time. Once a third independent variable is in the equation, the stepwise method invokes the right to remove an independent variable if that predictor is not earning its keep.

Predictors are allowed to be included in the equation if they significantly ($p = .05$) add to the predicted variance of the dependent variable. With correlated independent variables, as we have seen, the predictors in the equation admitted under a probability level of .05 can still overlap with each other. This is shown in Figure 5a.6.

In Figure 5a.6, predictor *J* was entered first, *K* was entered second, and *L* was entered third. We are poised at the moment when *L* joined the equation. Note that between predictors *J* and *L,* there is very little work that can be attributed uniquely to *K*. At this moment, the squared semipartial correlation associated with *K* (showing its unique contribution to the prediction model) is quite small.

In the forward method, the fact that *K*'s unique contribution has been substantially reduced by *L*'s presence would leave the procedure unfazed because it does not have a removal option available to it. But this is the stepwise method, and it is prepared to remove a predictor if necessary. When the amount of residual variance that *K* now accounts for is examined, let's presume that it is not significant at the removal criterion of .10 (say its *p* value is .126). *K* is thus judged to no longer be contributing effectively to the prediction model, and it is removed from the equation. Of course, as more predictors are entered into the equation, the gestalt could change dramatically, and *K* might very well be called on to perform predictive duties later in the analysis.

We have just described the reason that the entry criterion is more severe than the removal criterion. It can be summarized as follows. If getting into the equation was easier than getting out, then variables removed at one step

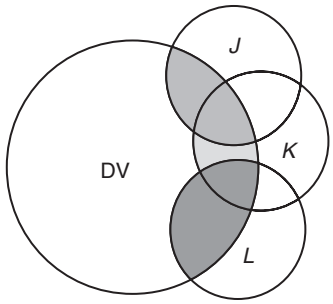

Figure 5a.6 Unique Contribution of Variable *K* Is Reduced by Variable *J* and Variable *L*

might get entered again at the next step because they might still be able to achieve that less stringent level of probability needed for entry. There is then a chance that the stepwise procedure could be caught in an endless loop where the same variable kept being removed on one step and entered again on the next. By making entry more exacting than removal, this conundrum is avoided.

Evaluation of the Statistical Methods

Benefits of the Statistical Methods

The primary advantage of using the standard model is that it presents a complete picture of the regression outcome to researchers. If the variables were important enough to earn a place in the design of the study, then they are given room in the model even if they are not adding very much to the R^2. That is, on the assumption that the variables were selected on the basis of their relevance to theory or at least on the basis of hypotheses based on a comprehensive review of the existing literature on the topic, the standard model provides an opportunity to see how they fare as a set in predicting the dependent variable.

The argument for using the stepwise method is that we end up with a model that is "lean and mean." Each independent variable in it has earned the right to remain in the equation through a hard, competitive struggle. The argument for using the forward and backward methods is similar to one used by those advocating the stepwise method. The forward and backward methods give what their users consider the essence of the solution by excluding variables that add nothing of merit to the prediction.

Criticisms of the Statistical Methods

One criticism of all the statistical methods is that independent variables with good predictive qualities on their own may be awarded very low weight in the model. This can happen because their contribution is being evaluated when the contributions of the other predictors have been partialled out. Such "masking" of potentially good predictors can lead researchers to draw incomplete or improper conclusions from the results of the analysis. One way around this problem is for the research to exercise some judgment in which variables are entered at certain points in the analysis, and this is discussed in the section titled "Researcher-Controlled Methods." This issue is also related to multicollinearity, a topic that we discuss later in the chapter.

The step methods have become increasingly less popular over the years as their weaknesses have become better understood and as research-controlled methods have gained in popularity. Tabachnick and Fidell (2001b), for example, have expressed serious concerns about this group of methods, especially the stepwise method, and they are not alone. Here is a brief summary of the interrelated drawbacks of using this set of methods.

- These methods, particularly the stepwise method, may need better than the 20 to 1 ratio of cases to independent variables because there are serious threats to external validity (Tabachnick & Fidell, 2001b). That is, the model that is built may overfit the sample because a different sample may yield somewhat different relationships (correlations) between the variables in the analysis, and that could completely change which variables were entered into the equation.
- The statistical criteria for building the equation identify variables for inclusion if they are better predictors than the other candidates. But "better" could mean "just a tiny bit better" or "a whole lot better." One variable may win the nomination to enter the equation, but the magnitude by which the variable achieved that victory may be too small to matter to researchers.
- If the victory of getting into the equation by one variable is within the margin of error in the measurement of another variable, identifying the one variable as a predictor at the expense of the other may obscure viable alternative prediction models.
- Variables that can substantially predict the dependent variable may be excluded from the equations built by the step methods because some other variable or combination of variables does the job a little bit better. It is conceivable that several independent variables taken together may predict the criterion variable fairly well, but step procedures consider only one variable at a time.

Balancing the Value of All the Statistical Methods of Building the Model

The standard method works well if you have selected the independent variables based on theory or empirical research findings and wish to examine the combined predictive power of that set of predictors. But because they are functioning in combination, the weights of the predictors in the model are a function of their interrelationships; thus, you are not evaluating them in isolation or in subsets. The standard method will allow you to test hypotheses about the model as a whole; if that is your goal, then that's what you should use.

The stepwise methods are intended to identify which variables should be in the model on purely statistical grounds. Such an atheoretical approach is discouraged by many researchers. On the other hand, there may be certain applications where all you want is to obtain the largest R^2 with the fewest number of predictors, recognizing that the resulting model may have less external validity than desired. Under these conditions, some researchers may consider using a step method.

Before one decides that one of the statistical procedures is to be used, it is very important to consider a researcher-controlled method of performing the regression analysis. Although it does require more thoughtful decision making rather than just entering the variables and selecting a statistical method, the flexibility it affords and the control it offers more than compensate for the effort it takes to run such analyses.

Researcher-Controlled Methods of Building the Model

Researcher-controlled regression methods are really variations on a theme. In all cases, it is the researchers who specify the order of entry of predictors into the equation. The main issue that researchers face is to determine how many variables are instructed to enter the equation at any one time. Several labels are applied to variations of researcher control: *sequential analysis, covariance analysis, hierarchical analysis,* and *block-entry analysis* are among the most common labels.

What makes this approach different from the statistical methods described above is that instead of the computer program using statistical criteria to make such entry decisions, the researchers determine which variables they would like to propose as covariates. Selection of covariates should have a solid rational basis in that the decision should be based on a particular theory, or covariate selection should rest on a solid empirical basis in which the research literature has shown the need to take into

account the relationship(s) between the criterion variable and one or more of the predictors.

For example, suppose we are interested in predicting performance on the nursing multiple choice licensing examination. Specifically, suppose that we want to determine the extent to which the time spent in various activities during their internships combines together to predict candidates' exam scores. Further suppose that we believe there is enough variation in the reading skill of licensing candidates to want to statistically control for the effects of reading skill on license exam performance in evaluating those internship experiences.

If we have a measure of reading skill in addition to the time-spent survey results for each individual, we will conduct the regression analysis so that the reading variable is the first to enter the equation and thus we will use this variable as a covariate. That forces the analysis to assign whatever variance in test scores that it can to reading skill. We then enter the internship variables simultaneously to account for whatever variance remains after reading skill does its predictive (covariance) work. Causal hypotheses and mediating variables, which this example is on the borderline of addressing, can be directly examined through the technique of path analysis, described in Chapter 14A.

Other possibilities for order of entry exist because we have now taken control of the process. We can, for example, enter the predictors of our choice into the equation one at a time. Once again, researchers should determine the order of entry on the basis of some theory or at least on some empirical basis, but as long as not too many orders are chosen, it may be possible to test some interesting hypotheses. The main advantage of entering one variable at a time is to give precedence to predictors entered earlier over predictors entered later. As you can imagine, doing such a sequential analysis is a delicate matter. Several independent variables may actually account for the same component of the dependent variables variance, but only the earlier entered ones will actually get the credit for doing so. This sort of hierarchical analysis works well with more developed theories.

Block-entry analysis, entering subsets of variables in a sequential manner, is a variant of this general researcher-controlled methodology in that one enters a set of variables rather than a single variable at a particular stage of the analysis. For example, if we have variables *K, L, M, N, O, P, Q,* and *R* as predictors of some criterion variable, we might wish to enter variables *L, P,* and *R* together on a single step in the analysis. These variables are therefore entered simultaneously (as described under the standard regression model) where the effects of the other variables (and any variables already in the equation) have been partialled out before the contribution of that variable is computed. One can also enter blocks of variables and single variables at various stages throughout the entire process.

Block-entry analysis can serve at least two research functions. First, as mentioned earlier in a criticism of the step procedures, several variables in combination may predict better than any one of them taken in isolation. Entering variables as a set (block) allows researchers an opportunity to explore that possibility. Second, some variables in a study may either naturally relate to each other or may all pertain to a general area of the content domain and so may lend themselves to be entered as a block. For example, in predicting the strength of certain symptom patterns, one may want to enter physical or medical variables before the more purely psychological variables.

In addition to exploring the theoretical consequences of varying the order of entry of the predictors and in addition to determining the result of using certain variables as covariates, several other issues can be broached by using a sequential form of regression analyses. Here are two examples:

1. A very "expensive" variable achieved substantial weight in the model. To collect data on this predictor might take a great deal of time, trouble, funding, or some combination of these. It may be worthwhile to ask if any variable on the sidelines could do almost as good a job but work for cheaper research wages.
2. A set of variables received negligible weights in the model, but these are easy to measure (e.g., they may be subscales of a single inventory). Similar measures might have been weighted substantially but could be more difficult to work with. It may make sense to investigate the R^2 consequences of replacing the latter with the former.

Outliers

As discussed in Chapter 3A and 3B, outliers are extreme scores on either the criterion or the predictor variables. They are typically thought of as being anomalous values, often three or more standard deviation units from their respective means, that suggest possible problems with the measurement instrument, the way the responses were recorded or transcribed, or the participants' membership in the population that was presumably sampled.

The presence of outliers can adversely affect (distort) the results of the analysis. This distortion takes several different forms (Darlington, 1990). As one example, consider the use of the least squares rule. Because this line-fitting procedure calls for minimizing the squared distance between each data point and the regression line, data points that are extremely far removed from the mainstream have a rather disproportionate influence in determining where the regression line is best placed. That is, because the square of a large distance is extremely large, the regression line is drawn

closer to the outlier to keep that squared distance as small as possible. This is done, of course, at a sacrifice—the regression line no longer coincides with the best-fitting location for all the other data points (excluding the outlier). For this reason, most statisticians suggest that outliers should be deleted prior to data analysis.

Researchers should also consider the possibility that the participants whose scores are defined as outliers might actually have something in common. For example, if most of the outliers represent older participants in a sample that contained a good mix of ages, then age may suddenly become an important variable to study.

Collinearity and Multicollinearity

Collinearity is a condition that exists when two predictors correlate very strongly; *multicollinearity* is a condition that exists when more than two predictors correlate very strongly. Note that we are talking about the relationships between the predictor variables only and not about correlations between each of the predictors and the dependent variable.

Regardless of whether we are talking about two predictors or a set of three or more predictors, multicollinearity can distort the interpretation of multiple regression results. For example, if two variables are highly correlated, then they are largely confounded with one another; that is, they are essentially measuring the same characteristic, and it would be impossible to say which of the two was the more relevant. Statistically, because the standard regression procedure controls for all the other predictors when it is evaluating a given independent variable, it is likely that neither predictor variable would receive any substantial weight in the model. This is true because at the time the procedure evaluates one of these two predictors, the other is (momentarily) already in the equation accounting for almost all the variance that would be explained by the first. The irony is that each on its own might very well be a good predictor of the criterion variable. With both variables in the model, the R^2 value will be appropriately high; if the goal of the research is to maximize R^2, then multicollinearity might not be an immediate problem.

When the research goal is to understand the interplay of the predictors and not simply to maximize R^2, multicollinearity can cause several problems in the analysis. One problem caused by the presence of multicollinearity is that the values of the regression coefficients of the highly correlated independent variables are distorted. Often, they are quite low and may even fail to achieve statistical significance. A second problem is that the standard errors of the regression weights of those multicollinear predictors can be

inflated, thereby enlarging their confidence intervals, sometimes to the point where they contain the zero value. If that is the case, you could not reliably determine if increases in the predictor are associated with increases or decreases in the criterion variable. A third problem is that if multicollinearity is sufficiently great, certain internal mathematical operations (e.g., matrix inversion) are disrupted and the statistical program comes to a screeching halt.

Identifying collinearity or multicollinearity requires researchers to examine the data in certain ways. A high correlation is easy to spot when considering only two variables. Just examine the Pearson correlations between the variables in the analysis as a prelude to multiple regression. Two variables that are very strongly related should raise a "red flag." As a general rule of thumb, we recommend that two variables correlated in the middle .7s or higher should probably not be used together in a regression or any other multivariate analysis. Allison (1999b) suggests that you "almost certainly have a problem if the correlation is above .8, but there may be difficulties that appear well before that value" (p. 64).

One common cause of multicollinearity is researchers using subscales of an inventory as well as the full inventory score as predictors. Depending on how the subscales have been computed, it is possible for them in combination to correlate almost perfectly with the full inventory score. You should use either the subscales or the full inventory score, but not all of them in the analysis. Another common cause of multicollinearity is including in the analysis variables that assess the same construct. You should either drop all but one of them from the analysis or consider the possibility of combining them in some fashion if it makes sense. For example, you might combine height and weight to form a measure of body mass. As another example, you might average three highly correlated survey items; exploratory factor analysis, discussed in Chapter 12A, can be used to help determine which variables you might productively average together without losing too much information. A less common cause of an analysis failing because of multicollinearity is placing into the analysis two measures that are mathematical transformations of each other (e.g., number of correct and incorrect responses; time and speed of response); researchers should use only one of these measures.

Multicollinearity is much more difficult to detect when it is some (linear) combination of variables that produces a high multiple correlation in some subset of the predictor variables. We would worry if that correlation reached the mid .8s but Allison (1999b, p. 141) gets concerned if those multiple correlations reached into the high .7s (R^2 of about .60). Many statistical programs will allow you to compute multiple correlations for different combinations of variables so that you can examine them. Thus, you

can scan these correlations for such high values and take the necessary steps to attempt to fix the problem.

Most regression programs have what is called a *tolerance* parameter that tries to protect the procedure from multicollinearity by rejecting predictor variables that are too highly correlated with other independent variables. Conceptually, tolerance is the amount of a predictor's variance not accounted for by the other predictors ($1 - R^2$ between predictors). Lower tolerance values indicate that there are stronger relationships (increasing the chances of obtaining multicollinearity) between the predictor variables. Allison (1999b) cautions that tolerances in the ranges of .40 are worthy of concern; tolerance values in the range of .1 are problematic (Myers, 1990; Stevens, 2002).

A related statistic is the *variance inflation factor* (VIF), which is computed as 1 dividend by tolerance. A VIF value of 2.50 is associated with a tolerance of .40 and is considered problematic by Allison (1999b); a VIF value of 10 is associated with a tolerance of .1 and is considered problematic by Myers (1990) and Stevens (2002).

Suppressor Variables

Suppressor variables, when included in regression equations increase R^2, but they accomplish this feat in a somewhat different way from what we have already discussed. Suppressor variables often are not correlated particularly strongly with the criterion variable itself. Rather, they are correlated in a special way with one or more of the other predictor variables, and that is where they do their job.

A suppressor variable works its magic by correlating with what is usually thought of as a source of error in another predictor (Darlington, 1990). Pedhazur (1982) describes it well in saying that by correlating with the error in another predictor, the suppressor variable helps purify that predictor and thereby enhances its predictive power. Often, under these conditions, this suppressor variable will then be given a negative weight in the equation (assuming that the predictor it is partially suppressing is positively correlated with the dependent variable). Tabachnick and Fidell (2001b) have provided the following rubric to help identify a suppressor variable:

- The correlation between it and the criterion variable is substantially smaller than its beta weight.

or

- Its Pearson correlation with the criterion and its beta weight have different signs (p. 149).

Other signs that you may have a suppressor variable in the equation are offered by Pedhazur (1982):

- It may have a near-zero correlation with the criterion variable but yet it is a significant predictor in the regression model.
- It may have little or no correlation with the criterion variable but is correlated with one or more of the predictors. (pp. 104–105)

Conceptualizing how a suppressor variable does its work is not an easy matter. Guilford's example (Guilford & Fruchter, 1978) is better than most, and we will present it to you as a way to exemplify this somewhat elusive concept.

J. P. Guilford, one of the great pioneers in measurement and psychometrics, did research for the Air Force during World War II to develop selection procedures for pilots. He speaks of a vocabulary test that slightly negatively correlated with success in pilot training. His research team had also used a reading test in the study (the trainees read passages and answered questions about them), which turned out to correlate positively with success in pilot training.

At first, it must have seemed odd that a vocabulary test slightly negatively correlated with pilot success but that a reading test correlated positively with pilot training success. But Guilford soon realized that the reading test correlated positively with pilot success not because it measured some verbal skill but rather because of the content it presented to the trainees. This content tapped into their experience with mechanical devices and their ability to visualize information contained in the passages.

According to Guilford's appraisal of the situation, the score these trainees received on the reading test was a function of three factors—mechanical experience, visualization, and, of course, verbal comprehension—only two of which (mechanical experience and visualization) were viable predictors of pilot success. The third factor measured by the reading test, verbal comprehension, did not predict success, something he already knew from the results of his vocabulary test. This third factor in the reading test actually represented, from the standpoint of predicting training success, error variance.

Now consider the relationship between the vocabulary test and the reading test. Roughly speaking, the vocabulary test was a measure of verbal comprehension or something that correlated highly with it. From this perspective, the reading test and the vocabulary test share common variance. But not just any variance, mind you. The variance they share is common to what they both measure—verbal comprehension or its kin.

How a suppressor variable works lies, in this example, with what happens when you make use of the correlation between the vocabulary and reading tests. If the vocabulary test was placed in the regression model together with the reading test, even though it could not directly predict pilot success, it would have the opportunity to correlate with that portion of the reading test assessing verbal comprehension. By virtue of that correlation, it would account for (statistically control for or negate) that portion of the reading test's variance attributable to verbal comprehension and thus make the reading test a better predictor than it would be in the absence of the vocabulary test. All that we would need to do is subtract that error variance out.

Based on this reasoning, Guilford (Guilford & Fruchter, 1978) tells us that the "combination of a vocabulary test with the reading test, with a negative weight for the vocabulary test [to subtract out this variance accounted for by the vocabulary test], would have improved predictions [of pilot success] over those possible with the reading test alone" (p. 182). That is, including the vocabulary test would have accounted for and subtracted out the variance due to verbal comprehension in the reading test (which was not contributing to the prediction of success anyhow), freeing up the other components of the reading test (the parts assessing mechanical experience and visualization) to more purely predict success in pilot training. In this context, the vocabulary test would have operated as a suppressor variable.

Linear and Nonlinear Regression: Completely Linear Models

The general regression model that we have been discussing thus far in this chapter is one form of a linear model. For the purposes of this book, we can distinguish between three types of regression models: a form we will call *completely linear,* another form of linear model called *intrinsically linear* (Pedhazur, 1982), and a form of nonlinear model called *intrinsically nonlinear* (Pedhazur, 1982) or *general curve fitting* (Darlington, 1990).

In the completely linear model, both the variables specified by the model (the dependent and the independent variables) as well as the parameters (the coefficients and the intercept) are in their "regular" form. We then multiply each variable by its weight and add the results of that multiplication together (adding the constant in the raw score equation) to obtain the predicted value of the criterion variable.

If there is only one predictor in the model, as is the case in simple linear regression, the equation can be represented geometrically as a straight line in two-dimensional space (in a space defined with X and Y axes). With two

predictors in the model, the equation may be represented geometrically as a tilted plane in three-dimensional space (illustrated well by Darlington, 1990). If you think of a room in a house, the two predictor variables cover the floor (the width and the length) of the room and the criterion variable is mapped to the walls (height). Imagine a pitched and tilted ceiling to this room. This ceiling is the plane described by the regression equation. Although it has one more dimension to it than the straight line, it is still a completely linear model composed of flat, straight surfaces. Models with more than two predictors, even though we cannot easily picture them, are also linear in this sense.

Linear and Nonlinear Regression: Intrinsically Linear Models

Another class of linear models has the same basic structure of completely linear models in that (a) each variable has an associated coefficient, (b) we multiply each variable by its coefficient to obtain its weighted value, and (c) we add the results of that multiplication together (adding the constant in the raw score equation) to obtain the predicted value of the criterion variable.

The difference between completely linear models and intrinsically linear models is that in the latter, the variables themselves are not in their "regular" or raw form; rather, they have been "altered" in some manner. In this sense, we can say that the model is linear with respect to its parameters—in that the regression weights are still in "regular" form and we add the weighted variables together to obtain the predicted score—but that the model is nonlinear in its variables. We can think of an intrinsically linear model as one that combines variables that are themselves not linear in the best weighted linear combination to maximally predict the dependent variable.

Variables in an intrinsically linear model can be altered in several different nonlinear ways. We will consider three types of alterations here: transformations of dependent and independent variables, dummy coding of independent variables, and interactions between independent variables.

Transformations

A transformation is used either to bring the data more closely in line with the underlying assumptions of regression or because it makes more sense to frame the relationships between the predictor and criterion variables in terms of a transformed variable. We have already discussed a transformation to a standardized scale (a z score transformation) that is routinely performed by virtually every statistical program in computing a regression

solution, although this type of transformation still keeps the variables in linear form. Other transformations are generally applied to the dependent or criterion variable, and still other transformations are generally applied to the independent or predictor variables that convert them to a nonlinear form.

The most common transformation of the criterion variable (other than standardizing it) is to use its natural logarithm as the value to be predicted (Allison, 1999b). Such a transformation, which we described in Chapter 3A, may help to reduce the degree of heteroscedasticity in the measure.

Another transformation of the dependent variable is a logit transformation, provided that the criterion variable is a proportion (Allison, 1999b). If the criterion variable is symbolized by *Y,* then this transformation takes the following form:

$$\text{log of the expression } [Y \text{ divided by } (1-Y)\ ].$$

A common transformation of the predictor variables is to use a polynomial function in which the value of one or more of the independent variables is raised to a power (e.g., X^2). This is called *polynomial* or *curvilinear* regression. A second-order polynomial function, known as a quadratic function, has the predictor raised to the second power. We might use this transformation when we expect the criterion variable to first increase together with the predictor and then to decrease with further increases in values of the predictor variable. For example, using age as a predictor of physical agility, we would expect agility to increase up to some age level but to then show a decrease with further increases in age. A quadratic function has one "bend" in the curve. In contrast, a cubic function (a third-order polynomial in which a predictor X is raised to the power of 3) has two "bends" in the functions and is substantially more complex to interpret. Most researchers would not use a polynomial function in excess of third order.

An additional way to change a predictor variable is to subject it to log transformation. We would tend to use this type of transformation where we expect the criterion and predictor variables to increase together up to some point but then further expect the dependent variable to level off with further increases in the predictor. For example, in predicting income from education, we might expect that higher income levels are associated with increasingly more education up to some point but that more education would not alter income level beyond some point (Allison, 1999b).

Dummy Coding

As we indicated in Chapter 4A, it is appropriate to perform a Pearson correlation analysis with a dichotomously coded categorical variable. In the

context of multiple linear regression, we could use such a variable as an independent variable (or we could use it as the criterion variable in either a logistic regression analysis as described in Chapter 6A or in a discriminant function analysis as described in Chapter 7A). To include a dichotomously coded variable as a predictor, we need to assign arbitrary numerical codes to its categories; for example, we could use 1 (for the presence of some property) and 0 (for the absence of some property). This process is called *dummy coding.*

The interpretation of the results of the regression analysis—specifically, the coefficient associated with the dichotomously coded variable—is based on the difference between the two means when adjusted (statistically controlling) for the other predictors in the model. For example, suppose that we were using participation in high-risk sports as one of several predictors of aggressiveness. We code those who have participated in high-risk sports as 1 and those who have not as 0. In the regression analysis, we find that this predictor is significant with a b weight of 12.50. We may then interpret this value to indicate that, when controlling for the other predictors, those who participate in such sports (those coded as 1) have aggression scores on average 12.50 greater than those who have not participated in high-risk sports.

A regression analysis requires the arbitrary categorical coding yield interpretable results. Thus, we could not legitimately take a categorical variable with three or more levels (e.g., eastern, midwestern, and western regions); code the categories 1, 2, 3 (as examples); and include such a coded variable in the regression analysis. The statistical procedure would treat the values of 1, 2, and 3 as though they represented interval level measurement where 3 designated more of some quality than 2 and 2 designated more of some quality than 1. But because the categories are not quantitatively based (by definition), the results that you obtained will not make any sense. Think of it like this if you are unconvinced: If the categories were coded in all possible ways and you ran separate regression analyses for each coding scheme, you would get widely different regression results. This indicates that none of the coding schemes are appropriate.

At the same time, nominal variables with more than two categories often make interesting and useful predictors. We simply need to dummy code them appropriately to use them in a regression analysis. We do this by creating separate dummy variables to represent portions or levels of the nominal variable. Each dummy variable will be a dichotomous (0, 1) coding of a subcategory (level) of the main variable. Because the dummy variables need to be orthogonal to (independent of) each other, the number of levels of the variable we are allowed to use is one less than the number of categories we have. Thus, with the three geographical regions mentioned above, we can

create only two dummy variables (two regions coded 0 and 1) to represent the main variable.

The category excluded from the dummy coding is, in a certain sense, the focal point of the analysis; it is treated as the reference category with which the other categories will be compared. This is because the regression weights of the other categories are interpreted with respect to this reference category. We can illustrate this by selecting for the sake of this example the eastern region as the reference category. We would then dichotomously code the other two categories. If participants lived in the Midwest they would be coded as 1; if they lived elsewhere they would be coded as 0. A similar coding scheme would be used for Westerners: If they lived in the West, they would be coded as 1; if they lived elsewhere, they would be coded as 0.

Every participant receives a value on both of these dummy coded variables. If in the data file, the Midwest variable appeared first and the West variable appeared second, we would know that a case with the combined code 10 lived in the Midwest, a case with the combined code 01 lived in the West, and a case with the code 00 lived in the East. Note that this last code appears as a by-product of our coding scheme because with one less code than we have categories (we use only two dummy variables to handle three categories), we know that someone not falling into one of our two designated codes (someone without a 1 in one of the two fields) must be in last category. In this way, creating a variable for all but one category is sufficient to classify all the cases in the sample.

In the regression model, residing in the midwestern and western regions are each predictors with their own regression weights. In our interpretation, we use the eastern region, our reference category, as our base. Suppose that we were predicting the number of times individuals changed residences in the prior 5-year period and that region was one of many predictors. Assume that the raw regression weight for the midwestern region was –4.50 and that the raw regression weight for the western region was 9.94. We would then say, when controlling or adjusting for the other variables in the model, that midwestern participants moved on average 4.5 times less than Easterners (the negative b weight tells us that it is less) and that Westerners moved on average almost 10 times more than Easterners.

Selecting the category of the nominal variable to serve as the reference group should be based on statistical or methodological factors. From a statistical perspective, the reference group should be one that, all else equal, has a relatively large sample size (Allison, 1999b). This is because the mean of the reference group will be involved in all the comparisons and should therefore have as small a standard error as possible. From a methodological perspective, the reference group should be the one with which it makes sense to compare the other groups. If there is a “control” or “baseline”

group in the nominal variable, that category would present itself as a strong reference group candidate. If there is no such category, as was the case in the example of geographic regions, then the choice is rather arbitrary.

Moderator Variables and Interactions

Consider the case where we have two predictors, X_1 and X_2. At the end of the regression analysis, each predictor is associated with a regression weight. For example, b_1 is the coefficient of X_1. This coefficient is essentially an estimate of the slope of the function for X_1 controlling for X_2. An assumption underlying such an interpretation is that the value of b_1 is the same across the range of X_2.; that is, whether or not X_1 and X_2 are correlated, the regression function for X_1 is independent of X_2. (Aiken & West, 1991). This same reasoning can be applied to X_2.

Now suppose that this independence assumption is not true. Instead, the relationship between X_1 and the criterion variable differs for different levels of X_2. With the relationship betweenX_1 and the criterion depending on the level of X_2., the variable X_2. is thought of as a *moderator variable* in that we need to take into account the level of X_2. in describing the relationship between X_1 and the criterion. When this is the case, we say that X_1 and X_2. interact. Here are two examples of interactions that illustrate how different relationships between one predictor and a criterion variable might be expected at different levels of another predictor:

- As predictors of the degree of liberal attitudes held by people, we measure socioeconomic status and age. Let's look at socioeconomic status predicting liberal attitudes at two levels of age. Among younger people, we might find that higher levels of socioeconomic status predict more liberalism; among older people, we might find that higher levels of socioeconomic status predict less liberalism. If this was found, we would say that age and socioeconomic status interact in affecting (predicting) liberalism (Darlington, 1990).
- In predicting the self-assurance of managers, we use as independent variables how long participants have been managers as well as their actual managerial ability. Let's focus here on predicting self-assurance from the number of years the managers have held such a position but develop that prediction model separately for high and low ability levels. We might find that high-ability managers become more self-assured with increased time as a manager, but that managers with low ability become less self-assured the longer they have been in the position. Thus, time as manager and managerial ability would interact in predicting self-assurance (Aiken & West, 1991).

In the regression examples we discussed earlier in this chapter, the independent variables always "stood alone" with their regression weights either in their raw form or in some kind of transformed form. That is, we have always dealt with situations where the regression coefficient was the weight assigned to the predictor or some transformation of the predictor to represent one element of the regression model. Interactions involve estimating the coefficient (weight) of the product of two predictors. Usually, these two predictors are also included separately in the model. Thus, at minimum, we would have X_1, X_2, and $X_1 \times X_2$ as three predictors in the models. In this structure, each would be associated with its own regression weight. Our minimal model would then be as follows:

$$Y_{\text{pred}} = \mathbf{a} + \mathrm{b}_1 X_1 + \mathrm{b}_2 X_2 + \mathrm{b}_3 \mathrm{X}_1 X_2$$

In the above equation, the term X_1X_2 represents the interaction. The two terms preceding it in the equation containing the stand-alone predictors are known as the *main effects* of the variables; we thus can also address or speak to the main effects of X_1 and X_2.

In the regression solution, if the coefficient associated with the interaction was statistically significant we must be careful about interpreting the results of the stand-alone variables (the main effects). In the first example above, it would be inappropriate to speak of the slope of socioeconomic status in general (interpret the b coefficient associated with socioeconomic status) because the nature of the relationship would depend on age; in the second example, it would be inappropriate to speak of the overall slope of time as a manager (interpret the b coefficient associated with time in position) because the nature of the relationship depends on ability level.

When there is a significant interaction between predictors, it is necessary to explore its nature in more detail. Essentially, this means that you would examine the *simple slopes,* the slope of a predictor under different levels of the other predictor. To illustrate this for the first example, we would want to predict liberalism with socioeconomic status at different values of age. In some situations, we might have a theoretical or practical reason for selecting certain ages on which to focus. Under such circumstances, we should use those particular ages in the regression model to determine the slope (the regression coefficient) specific to those age levels. If there was no theoretical basis to select particular ages, Aiken and West (1991) appear to endorse the recommendation of Cohen et al. (2003) to use values corresponding to +1 standard deviation, the median, and –1 standard deviation; we would thus estimate the b coefficient at each of these three values in the distribution. In our example, if the mean age was 36 and the standard

deviation was 10, then the ages corresponding to these three locations would be 26, 36, and 46. We would then estimate the b coefficient for socioeconomic status at each of these three age levels. You should note that this means generating separate regression models to represent the relationship between the predictor (e.g., socioeconomic status) and criterion (e.g., liberalism) variables at each level (+1, 0, and −1 standard deviation units) of the moderator variable (e.g., age).

With an interaction in the model, we should interpret the main effects with great care. Allison (1999b), using an example of years of schooling and age predicting income, makes this clear:

> What you must always remember is that in models with interactions, the main-effect coefficients have a special (and often not very useful) meaning. The coefficient . . . for age . . . can be interpreted as the effect of age *when schooling is 0.* Similarly, the coefficient . . . for schooling can be interpreted as the effect of schooling *when age is 0.* . . . In general, whenever you have a product term in a regression model, you should not be concerned about the statistical significance (or lack thereof) of the main effects of the two variables in the product. That doesn't mean that you can delete the main effects from the model. Like the intercept in any regression equation, those terms play an essential role in generating correct predictions for the dependent variable. (p. 168)

If the interaction term is significant, then we must focus our attention on simplifying the interaction effect. That means examining the simple slopes. If the interaction is not significant, then you should perform another regression analysis without the interaction term being included. That model will include only the main effects and can be interpreted in the ways described earlier in this chapter.

We described the shape of the function in a two-predictor model to be a plane—the roof of a room serves as a suitable image. In a model that includes the interaction term of these two predictors, the surface can be thought of as "warped" (Darlington, 1990). Imagine a room with walls of different heights; represented geometrically, the roof that would be fitted to such room would represent the surface of an interaction.

It has been argued (e.g., Aiken & West, 1991; Darlington, 1990) that when interaction terms are included in the model, the predictor and the moderator variables should be *centered;* that is, the mean of the variable should be subtracted from each score on the variable to create a new (transformed) variable representing a deviation score.

Centering can reduce the chances of multicollinearity affecting the analysis. But its primary function is to facilitate the interpretation of the interaction (Aiken, 2005). The regression model that we ordinarily produce showing us a significant interaction represents the situation for the case where the predictor and moderator variables take on a value of zero. Yet it is very often the case that zero is not a possible value for these variables, and it is almost always the case that zero is not a representative value for these variables. For example, scores on a Likert-type summative response scale may have values of 1 through 5, and scores on many national administered standardized exams (e.g., GRE) have scores ranging from 200 to 800. For these measures, no one has achieved a valid score of zero.

To center the scores is to subtract the mean of the variable from each value yielding a deviation score. For example, if the mean GRE verbal score of the sample was 575, then a person whose original score was 600 would have a deviation score of +25. The mean of the predictor will thus have a transformed value of zero. When the predictor and moderator variables are centered in this manner, the ordinary regression solution will still show the prediction model appropriate for zero values of the predictor and moderator variables, but now this centered zero value is the mean of each distribution. The result of centering is that the regression model now represents the case for the typical score in the study.

In generating the regression lines for ±1 standard deviation from the mean, the predictor and moderator variables should be recentered twice—once at the value corresponding to +1 standard deviations and again at the value corresponding to −1 standard deviation. For each recentering operation, a regression analysis should be conducted so that each of these two functions can be obtained and the data points plotted.

Although centering is common practice, some authors have argued that it may not be worthwhile (Kromrey & Foster-Johnson, 1998) for linear regression. Our recommendation is to always center your predictor and moderator variables as Aiken and West (1991) have suggested. The idea here is that centering does not adversely affect the statistical analysis and, in our view, has the added advantage of facilitating your interpretation of the results in many circumstances.

Linear and Nonlinear Regression: Intrinsically Nonlinear Models

There is a whole set of models that do not take a linear form and thus cannot be analyzed through a procedure that uses ordinary least squares. These models are best handled by other curve-fitting techniques. They are

represented by equations of different forms. Darlington (1990) gives the following example:

$$Y_{\text{pred}} = \frac{(b_1X_1 + b_2X_2)}{(b_3X_3 + b_4X_4)}$$

Allison (1999b) gives an example as well:

$$Y_{\text{pred}} = 1 + A_x^B$$

One example of an intrinsically nonlinear model is when the dependent variable is a nominal variable. Such a situation is sufficiently relevant to the research conducted in the behavioral and social sciences that techniques have been developed to deal with such situations. Two analytic approaches to this nonlinear application, logistic regression and discriminant function analysis, will be presented in chapters 6A and B and 7A and B, respectively.

Canonical Correlation Analysis

In multiple regression, we form a variate of the independent variables to best predict the value of a single criterion measure. However, we are not limited necessarily to using a single dependent variable in our study. It is possible to assemble a set of dependent variables that can also be combined together in some weighted array (variate) whose value can be predicted by a weighted combination of independent variables. This is the realm of canonical correlation.

Canonical correlation analysis, also referred to as multivariate multiple regression (Lutz & Eckert, 1994), is a statistical test that assesses the relationship between two sets of continuous measured variables. Whereas one set may be considered as the predictor variate, the other set may be deemed the criterion, dependent, or outcome variate. These variates represent the weighted combination of the values on the various predictor variables that will correlate more highly with the criterion variate than any single predictor variable alone.

The advantage of using linear combinations of variables for both the predictor and criterion is that such a design increases the chances of discovering relationships that single variable designs could not capture. Canonical correlation in that sense is thus a potentially more powerful design than multiple regression, just as a multiple regression design is potentially more powerful than simple linear regression. This gain in power with canonical correlation is noteworthy, of course, to the extent that the

variables that are combined in the composite make theoretical sense (Benton, 1991). If the composite dependent variate does make sense—if it is interpretable in the context of the research problem—then it is possible to characterize canonical correlation as Cooley and Lohnes (1976) have done as "the simplest model that can begin to do justice to this difficult problem of scientific generalization" (p. 176).

Canonical correlation analysis is an exploratory statistical method (Tabachnick & Fidell, 2001b) as opposed to a confirmatory statistical procedure. Exploratory analyses are used as theory-generating procedures, whereas confirmatory analyses are treated as theory-testing procedures (Stevens, 2002). We need to be careful when we interpret and attempt to generalize results based on exploratory data analysis. Nunnally (1978) noted that exploratory methods are neither "a royal road to truth, as some apparently feel, nor necessarily an adjunct to shotgun empiricism, as others claim" (p. 371). Exploratory results must be viewed with caution partly because the relationships between the variates may not be replicated in other samples.

There is also the potential difficulty in the interpretation of the canonical function. Mulaik, James, Van Alstine, Bennett, Lind, and Stilwell (1989) suggested that one difficulty in interpretation comes about because researchers often lack prior knowledge about the underlying relationships between the variables; they therefore have no basis on which to make an interpretation of the result. It's one matter to predict a single measured variable from a set of independent variables—we do this informally regularly during our daily life in the normal course of social interaction when we take in information from a variety of sources to predict, say, how our friend liked a meal at the new Korean restaurant that just opened near campus. We fully expect that the set of predictor variables in this case is derived from quite different parts of our friend's life: the kind of food she ate as a child, the attitudes of her parents toward new food, her experiences with different types of restaurants when she started to date, how much traveling to different parts of the country (or abroad) she has done, and so on. Generally, the predictor variate, although interpretable, is often and appropriately composed of a diverse set of individual variables.

But it could be quite another matter to predict the value of a composite dependent variable unless that composite really represented a conceptual whole. We do have real-world experience in experiencing composite outcome variables, and it does make sense to us in those contexts. For example, we speak of a friend being "supportive" of us in when we experience a difficult time. "Support" is a judgment we make based on several factors relating to our friend's reactions to us, such as willingness to listen to us talk, being physically present, saying certain reassuring things, offering solutions to

problems, and maybe giving us a pat on the shoulder or a hug. And these behaviors are probably weighted in our minds (the things we need the most we probably weigh more when we figure out how supportive our friend was). If all the outcome measures that we used in our research were that tightly melded together to form variates that made such intuitive sense, perhaps some of the potential difficulties in interpreting the canonical function would be of less concern. But partly because canonical correlation is an exploratory analysis and is therefore combining variables in the dependent variate that may not have been combined before and that may therefore not produce a variate that is easily assimilated, canonical analysis is often viewed by some "as a last-ditch effort, to be used when all other higher-level techniques have been exhausted" (Hair, Anderson, Tatham, & Black, 1998, p. 444).

Added to this concern about potential interpretation matters, it is also possible that the measured variables employed for the predictor or criterion variables may represent different dimensions. To the extent that this is true, it is possible that more than one linear function relating the predictors and the criterion could emerge. Thus, it is possible that more than one solution to canonical correlation may be put forward, adding an extra layer of complexity to such a design. Although these functions are determined sequentially and are uncorrelated with one another (Stevens, 2002), it is vital that researchers have a reasonably thorough understanding of the content domain they are studying so that they can interpret multiple and independent functional relationships between the independent and dependent variates.

Canonical correlation, although never having been used as frequently as some other techniques such as multiple regression, tends to be used even less often today. In contemporary research, structural equation modeling has gradually replaced canonical correlation analysis (Maruyama, 1998). Structural equation modeling inherently determines statistical significance of the canonical function coefficients and structure coefficients, which is not easily accomplished in conventional canonical analysis (Thompson, 1984). We discuss the topics of structural equation modeling in the last three sets of chapters in this book.

Recommended Readings

Berk, R. A. (2003). *Regression analysis: A constructive critique.* Thousand Oaks, CA: Sage.

Berry, W. D. (1993). *Understanding regression assumptions: Vol 92. Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage.

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

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

Draper, N. R., Guttman, I., & Lapczak, L. (1979). Actual rejection levels in a certain stepwise test. *Communications in Statistics, A8,* 99–105.

Green, S. A. (1991). How many subjects does it take to do a multiple regression analysis. *Multivariate Behavioral Research, 26,* 499–510.

Hardy, M. A. (1993). *Regression with dummy variables.* Thousand Oaks, CA: Sage.

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin, 117,* 348–357.

Kahane, L. H. (2001). *Regression basics.* Thousand Oaks, CA: Sage.

Lorenz, F. O. (1987). Teaching about influence in simple regression. *Teaching Sociology, 15,* 173–177.

McClelland, G. H. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390.

Park, C., & Dudycha, A. (1974). A cross-validation approach to sample size determination. *Journal of the American Statistical Association, 69,* 214–218.

Schafer, W. D. (1991a). Reporting hierarchical regression results. *Measurement and Evaluation in Counseling and Development, 24,* 98–100.

Schafer, W. D. (1991b). Reporting nonhierarchical regression results. *Measurement and Evaluation in Counseling and Development, 24,* 146–149.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide.* Newbury Park, CA: Sage.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed., Chapter 12: Canonical Correlation). Hillsdale, NJ: Erlbaum.

Thompson, B. (1989). Why won't stepwise methods die? *Measurement and Evaluation in Counseling and Development, 21,* 146–148.

Weiss, D. J. (1972). Canonical correlation analysis in counseling psychology research. *Journal of Counseling Psychology, 19,* 241–252.