CHAPTER 5A

Multiple Regression Analysis

5A.1 General Considerations

Multiple regression analysis, a term first used by Karl Pearson (1908), is an extremely useful extension of simple linear regression in that we use several quantitative (metric) or dichotomous variables in combination rather than just one such variable to predict or explain the value of a quantitatively measured criterion (outcome/dependent) variable. Most researchers believe that using more than one predictor or potentially explanatory variable can paint a more complete picture of how the world works than is permitted by simple linear regression because behavioral scientists generally believe that behavior, attitudes, feelings, and so forth are determined by multiple variables rather than just one. Using only a single variable as a predictor or explanatory variable as is done in simple linear regression will at best capture only one of those sources. In the words of one author (Thompson, 1991), multivariate methods such as multiple regression analysis 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 job applicants might perform on the job. 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 analysis generates two variations of the prediction equation, one in raw score or unstandardized form and the other in standardized form (making it easier for researchers to compare the effects of predictor variables that are assessed on different scales of measurement). These equations are extensions of the simple linear regression models and thus still represent linear regression, that is, they are still linear equations but use multiple variables as predictors. The main work done in multiple regression analysis is to build the prediction equation. This primarily involves generating the weighting coefficients—the b (unstandardized) coefficients for the raw score equation and the beta (standardized) coefficients for the standardized equation. This prediction model informs us that if we weight each of the predictors as the statistical analysis has indicated, then we can minimize our error in predicting the dependent variable.

5A.2 Statistical Regression Methods

The regression procedures that we cover in this chapter are known as *statistical regression methods*. The most popular of these statistical methods include the *standard*, *forward*, *backward*, and *stepwise* methods, although others (not covered here), such as the Mallows Cp method (e.g., Mallows, 1973) and the maxi *R* squared and mini *R* squared (see Freund & Littell, 2000), have been developed as well. Using such a label that includes the term "statistical" may seem a little odd (of course regression is a statistical procedure), but the label is meant to communicate something rather important but subtle regarding the analysis procedures. The reason for calling the procedures "statistical regression methods" is to emphasize that once the researchers identify the variables to be used as predictors, they relinquish all control of the analysis to the mathematical algorithms in carrying out the analysis.

In statistical regression procedures, the mathematical procedures determine the optimal weighting for each of the predictors as a set that will minimize the amount of prediction error. Researchers cannot, for example, propose that this particular variable be given "priority" by allowing it to do its prediction work ahead of the other variables in the set. Although they were actively making decisions about what constructs were to be used as predictors, how to measure those constructs, who was to be sampled and how, and so on, now that they are on the brink of analyzing their data the researchers must passively wait for the software to generate its output and inform them of the way the software has deemed it best to weight each of the variables in the prediction model and/or which variables were granted "priority" in the analysis by the algorithm (assuming that such "priority" was permitted by the regression method).

This relinquishing of complete control when using the statistical regression methods is not necessarily a bad thing in that we are trying to maximize the predictive work of our variables. However, as we have become increasingly familiar and comfortable with more complex alternative methods in which researchers take on more active roles in shaping the prediction model, the use of these statistical methods has given way to alternatives that call for more researcher input into the process of building the model; many of these methods are covered in Chapters 6A and 6B as well as in Chapters 12A through 14B.

5A.3 The Two Classes of Variables in a Multiple Regression Analysis

The variables in a multiple regression analysis fall into one of two categories: One category comprises the variable being predicted and the other category subsumes the variables that are used as the basis of prediction. We briefly discuss each in turn.

5A.3.1 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* or *outcome 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.

5A.3.2 The Variables Used as Predictors

The predictors comprise a set of measures designated in the regression equation with upper case *X*s and are known as *predictor variables* or *independent variables* in the analysis. In many research design

courses, the term *independent variable* is reserved for a variable in the context of an experimental study, but the term is much more generally applied because ANOVA (used for the purpose of comparing the means of two or more groups or conditions) and multiple regression are just different expressions of the same general linear model (see Section 5A.5). In the underlying statistical analysis, whether regression or ANOVA, the goal is to predict (explain) the variance of the dependent variable based on the independent variables in the study.

Talking about independent and dependent variables can get a bit confusing when the context is not made clear. In one context (that of the general linear model), predicting the variance of the dependent variable is what the statistical analysis is designed to accomplish. This is the case whether the research (data collection) design is ANOVA or regression.

In the context of the research methodology underlying the data collection process itself, experimental studies are distinguished from regression or correlation studies by the procedures used to acquire the data. Some of the differences in the typical nature of independent variables in experimental and regression studies within this methodology and data collection context are listed in Table 5a.1. For example, in experimental studies, independent variables are often categorical and are manipulated by the researchers and dependent variables can be some sort of behaviors measured under one or more of the treatment conditions. However, independent variables may also be configured after the fact in correlation designs (e.g., we may define different groups of respondents to a survey medical treatment satisfaction based on the class of medication patients were prescribed) rather than be exclusively based on manipulated conditions. In regression designs, it is usual for all of the variables (the variable to be predicted as well as the set of predictor variables) to be measured in a given "state of the system" (e.g., we administer a battery of personality inventories, we ask employees about their attitudes on a range of work satisfaction issues, we extract a set of variables from an existing archival database). To minimize the potential for confusion, our discussion will remain in the context of the statistical analysis; should we refer to the methodological context, we will make that explicit.

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 but could be changeable over time (e.g., attitudes).
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 categorical and coded into a relatively few levels or categories.	Usually quantitative, measured on summative response, interval, or ratio scales.

5A.4 Multiple Regression Research

5A.4.1 Research Problems Suggesting a Regression Approach

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

- We want to predict one variable from a combined knowledge of several others.
- We want to determine which variables of a larger set are better predictors of some criterion variable than others.
- We want to know how much better we can predict a variable if we add one or more predictor variables to the mix.
- We want to examine the relationship of one variable with a set of other variables.
- We want to statistically explain or account for the variance of one variable using a set of other variables.

5A.4.2 The Statistical Goal in a Regression Analysis

The statistical goal of multiple regression analysis is to produce a model in the form of a linear equation that identifies the best weighted linear combination of independent variables in the study to optimally predict the criterion variable. That is, in the regression model—the statistical outcome of the regression analysis—each predictor is assigned a weight. Each predictor for each case is multiplied by that weight to achieve a product, and those products are summed together with the constant in the raw score model. The final sum for a given case is the predicted score on the criterion variable for that case.

The weights for the predictor variables are generated in such a way that, across all of the cases in the analysis, the predicted scores of the cases are as close to their actual scores as is possible. Closeness of prediction is defined in terms of the ordinary least squares solution. This strategy underlying the solution or model describes a straight-line function 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 equation or model and the actual observed values are the prediction errors. The model thus can be 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.

5A.4.3 The Regression Weights For the Predictors

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 when considering this particular set of variables, this one variable is weighted in the model to such and such an extent. In conjunction with a different set of independent variables, the weight assigned to that variable may turn out to be quite different.

This "relativity" of the variable weights has a couple of implications for the interpretation of the results. One implication is to recognize that the weight is not a feature of the variable per se but simply describes the particular role that it has played in this one analysis in combination with these other

specific variables predicting this particular outcome variable. Even a variable that can substantially predict the outcome variable in isolation may have received a very low weight in the multiple regression analysis because its prediction work might be redundant with one or more other predictors in the model.

A second implication for the interpretation of the results is that we tend to focus on how well the model as a whole performed. This is typically thought of in terms of the amount of variance of the outcome variable that is explained by the model as described in Section 5A.9.

5A.4.4 Fully Specifying the Regression Model

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 for them to be included, and still others may not have been amenable to the measurement instrumentation available to researchers at the time of the study. However, our working assumption in interpreting the results of the regression analysis is that the model is *fully specified*, that is, that we have captured all of the important variables that are predictive of our outcome variable. With this assumption in place, we can draw inferences about the phenomenon we are studying from the results of our analysis. 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 this assumption, 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 potentially produced a very different dynamic and would likely have resulted in a very different model than what we have just obtained.

5A.5 The Regression Equations

The regression equation, representing the prediction model, is perhaps the most straightforward expression of the *general linear model* that was introduced more than two centuries ago by Adrien-Marie Legendre in 1805 (Stigler, 1990) in which a weighted linear composite of a set of variables is used to predict the value of some variable. For multiple regression analysis, what is predicted is a single variable but it is possible to predict the value of a weighted linear composite of another set of variables as we do in canonical correlation analysis (see Chapter 7A).

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.

5A.5.1 The Raw Score Model

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

$$Y_{\text{pred}} = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

In this equation, Y_{pred} is the predicted score on the criterion variable, the Xs 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, 1991) because each reflects the relative contribution of its independent variable when we are statistically controlling for the effects of all the other predictors. Each *b* coefficient informs us of how many units (and in what direction) the predicted *Y* value will increment for a 1-unit change in the corresponding *X* variable (we will show this by example in a moment), statistically controlling for the other predictors in the model. Because this is a raw score equation, it also contains a constant representing the *Y* intercept, shown as *a* in the equation.

All the variables are in raw score form in the model 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 test performance (the mean of the verbal and quantitative subscores), and the scores on this variable are probably going to be in the 150 to 165 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 a low 50 to a high of 75 (just to give us three rather different metrics for our illustration here).

The *b* coefficients computed for the regression model 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* coefficient for grade point average to be 7.00 and for GRE to be about .50 with a *Y* intercept value of –40.50. Thus, controlling for the effects of GRE, a 1-unit increase in grade point average (e.g., the difference between 3.0 and 4.0) is associated with a 7-point increase (because of the positive sign in the model) in the predicted success criterion variable. Likewise, controlling for the effects of grade point average, a 1-unit increase in GRE score is associated with a 0.50-point increase in the predicted success criterion variable. Putting these values into the equation would give us the following prediction model:

$$Y_{\text{pred}} = -40.50 + (7)(\text{gpa}) + (.5)(\text{GRE})$$

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 160. To arrive at her predicted score, we place her values into the variable slots in the raw score form of the regression equation. Here is the prediction:

$$\begin{aligned} Y_{\text{pred}} &= -40.50 + (7)(\text{gpa}) + (.5)(\text{GRE}) \\ Y_{\text{pred Erin}} &= -40.50 + (7)(\text{gpa}_{\text{Erin}}) + (.5)(\text{GRE}_{\text{Erin}}) \\ Y_{\text{pred Erin}} &= -40.5 + (7)(3.80) + (.5)(160) \\ Y_{\text{pred Erin}} &= -40.50 + (26.6) + (80) \\ Y_{\text{pred Erin}} &= 66.10 \end{aligned}$$

We therefore predict that Erin, based on her grade point average and GRE score, will score a little more than 66 on the success measure. Given that the range on the success measure is from 50 to 75, it would appear, at least on the surface, that Erin would be predicted to have performed moderately successfully. We would hope that this level of predicted performance would be viewed in a favorable light by the program to which she was applying.

This computation allows you to see, to some extent, how the *b* coefficients and the constant came to achieve their respective magnitudes. We expect a success score between 50 and 75. One predictor is grade point average, and we might expect it to be someplace in the middle 3s. We are therefore likely to have a partial regression weight much larger than 1.00 to get near the range of success scores. But the GRE scores are probably going to be in the neighborhood of 150 plus or minus, and these may need to be lowered by the equation to get near the success score range by generating a partial regression weight likely to be less than 1.00. When the dust settles, the weights overshoot the range of success scores, requiring the constant to be subtracted from their combination.

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

5A.5.2 The Standardized Model

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

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

Everything in this model is in standardized (*z*) 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 the lowercase Greek letter 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 between 0 and 1, but they can exceed the range of ±1 if the predictors are correlated enough between themselves (an undesirable state of affairs known as collinearity and multicollinearity (see Section 5A.21) that should be avoided either by removing all but one of the highly correlated predictors or by combining them into a single composite variable).

Each term in the equation represents the *z* score of a predictor and its associated beta coefficient. 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 model, but this time in standard score form:

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

$$Y_{z\,\text{pred}} = (.31)(\text{gpa}_z) + (.62)(\text{GRE}_z)$$

Note that the beta weights, because they are based on the same metric, can now be compared to each other. We will treat this topic in more detail in Section 5A.14, but for now note that the beta coefficient for the GRE is greater than the beta coefficient for grade point average. Thus, in predicting academic success, the underlying mathematical algorithm gave greater weight to the GRE than to grade point average.

We can also apply this standardized regression model 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 160 represents a z score of 1.25. We can thus solve the equation as follows:

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

$$Y_{z\,\text{pred}} = (.31)\ (\text{gpa}_z) + (.62)\ (\text{GRE}_z)$$

$$Y_{z\,\text{pred Erin}} = (.31)\ (\text{gpa}_{z\ \text{Erin}}) + (.62)\ (\text{GRE}_{z\,\text{Erin}})$$

$$Y_{z\,\text{pred Erin}} = (.31)\ (1.80) + (.62)\ (1.25)$$

$$Y_{z\,\text{pred Erin}} = (.558) + (.775)$$

$$Y_{z\,\text{pred Erin}} = 1.33$$

We therefore predict that Erin, based on her grade point average and GRE score, will score about 1.33 *SD* units above the mean on the success measure. This predicted standardized outcome score of 1.33 is equivalent to Erin's raw (unstandardized) predicted outcome score of 66.10.

5A.6 The Variate in Multiple Regression

Multivariate procedures typically involve building, developing, or solving for a weighted linear combination of variables. This weighted linear combination is called a *variate*. The variate in this instance is the entity on the right side of the multiple regression equation composed of the weighted independent (predictor) variables.

Although the variate is a weighted linear composite of the measured variables in the model, it is often possible to view this variate holistically as representing some underlying dimension or construct—that is, to conceive of it as 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 weighted linear combination of grade point average and GRE score. From this perspective, the indicators of academic aptitude—grade point average and GRE score—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 model—the result of invoking the linear weighted 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, but it cannot 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 (we will see in Section 13A.4 that the structural equation used in structural equation modeling predicts the actual Y value because the prediction error is included as a term in the model).

5A.7 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 regression method unspecified. It is currently the most widely used of the statistical methods. 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 statistical step or block. But that single step is not at all simple and unitary; when we look inside this step, we find that the process of determining the weights for independent variables is governed by a coherent but complex strategy.

5A.7.1 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 a dozen or more variables as potential predictors, we will use a simplified example of just three predictors for our presentation purposes.

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 Watson, Clark, and Tellegen's (1988) measures of the relative frequency of positive and negative affective behaviors a person typically 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.

5A.7.2 Correlations of the Variables

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.

In examining the correlation matrix, we are looking for two features primarily. First, we want to make sure that no predictor is so highly correlated with the dependent variable as to be relatively interchangeable with it; correlations of about .70 and higher would suggest that such a predictor might best be entered in the first block of a hierarchical analysis (see Section 6A.2) or not included in the analysis rather than proceed with the standard regression analysis that we cover here. Second, we want to make sure that no two predictors are so highly correlated that they are assessing the same underlying construct; again, correlations of about .70 and higher would suggest that we might want to either remove one of the two or combine them into a single composite variable before performing a standard regression analysis.

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

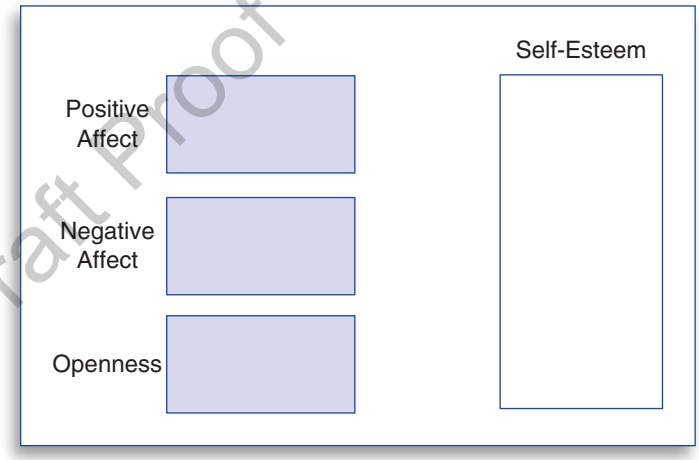
	Self-Esteem	Positive Affect	Negative Affect	Openness
Self-Esteem	1.000	.555	-.572	.221
Positive Affect	.555	1.000	-.324	.221
Negative Affect	-.572	-.324	1.000	-.168
Openness	.221	.221	-.168	1.000

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) separates the matrix into two redundant halves. As can be seen, the dependent variable of self-esteem is moderately correlated with both Positive Affect and Negative Affect but is only modestly correlated with Openness. It can also be seen that Positive Affect and Negative Affect correlate more strongly with each other than either does with Openness.

5A.7.3 Building the Regression Equation

The goal of any regression procedure is to predict or account for as much of the variance of the criterion variable as is possible using the predictors at hand. In this example, that dependent 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 unexplained. 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.

Figure 5a.1 Predictors Assembled Prior to the Regression Analysis

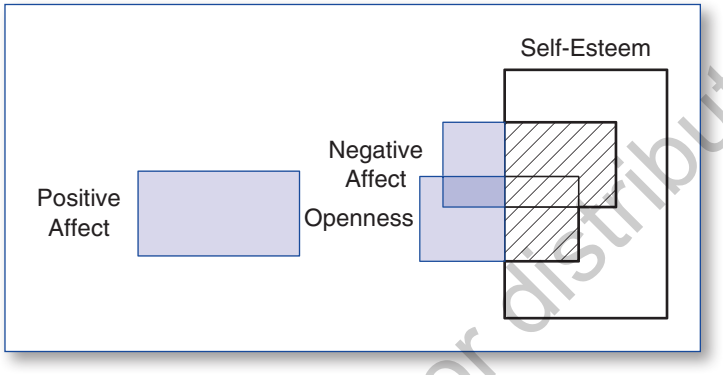


On the first and only step of the standard regression procedure, all the predictors are entered as a set into the equation. But to compute the weighting coefficients (*b* coefficients for the raw score equation and beta coefficients for the standardized equation), the predictors must be individually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

The purpose 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 *statistically controlling* for all the other predictors. These other predictors thus behave as a set of *covariates* in the analysis in that the predictive work that they do as a set is allowed to account for the variance of the dependent variable before the predictive work of a given predictor is evaluated. Because these other predictors are afforded the opportunity to perform their predictive work before the given predictor, we say that we have statistically controlled for these other predictors. Each predictor is evaluated in turn in this manner, so that the regression coefficient obtained for any predictor represents the situation in which all of the other predictors have been statistically controlled (the other predictors have played the role of covariates).

Figure 5a.2 Evaluating the Predictive Power of Positive Affect



This process is illustrated in Figure 5a.2. To evaluate the effectiveness of Positive Affect, the variables of Negative Affect and Openness are inserted as a set into the model. Negative Affect and Openness thus take on the role of covariates. Together, Negative Affect and Openness have accounted for some of the variance of Self-Esteem (shown as diagonal lines in Figure 5a.2).

With Negative Affect and Openness in the equation for the moment, we are ready to evaluate the contribution of Positive Affect. The criterion variable or dependent variable (Self-Esteem here) is the focus of the multiple regression design. It is therefore the variance of Self-Esteem that we want to account for or predict, and our goal is to explain as much of it as possible with our set of independent variables. We face an interesting but subtle feature of multiple regression analysis in its efforts to maximize the amount of dependent variable variance that we can explain. In the context of multiple regression analysis, 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. With Negative Affect and Openness already in the model, and thus already accounting for some of the variance of Self-Esteem, Positive Affect, as the last variable to enter, must target the variance in Self-Esteem that remains—the *residual* variance of Self-Esteem. The blank portion of Self-Esteem's rectangle in Figure 5a.2 represents the unaccounted-for (residual) portion of the variance of Self-Esteem, and this is what the regression procedure focuses on in determining the predictive weight Positive Affect will achieve in the model.

After the computations of the *b* and beta coefficients for Positive Affect have been made, it is necessary to evaluate another one of the predictors. Thus, Positive Affect and another predictor (e.g., Negative Affect) are entered into the equation, and the strategy we have just outlined is repeated for the remaining predictor (e.g., Openness). Each independent variable is put through this same process until the weights for all have been determined. At the end of this complex process (which is defined as a single "step" or "block" despite its complexity), the final weights are locked in and the results of the analysis are printed.

5A.8 Partial Correlation

Most statistical software packages, such as IBM SPSS, routinely compute and have available for output other statistical information in addition to the regression weights and the constant. One such statistic is the *partial correlation*. In the context of our present discussion, this is a good place to broach that subject.

As the term implies, a partial correlation is a correlation coefficient. Everything that we have described about correlation coefficients (e.g., the Pearson *r*) applies equally to the particular coefficient known as a partial correlation. But as the term also implies, a partial correlation is special. When you think of the Pearson *r*, you envision an index that captures the extent to which a linear relationship is observed between one variable and another variable. A partial correlation describes the linear relationship between one variable and a part of another variable. Specifically, *a partial correlation is the relationship between a given predictor and the residual variance of the dependent variable* when the rest of the predictors have been entered into the model. We discuss this in somewhat more detail in Section 5A.10.2.

Consider the situation depicted in Figure 5a.2. Negative Affect and Openness have been entered into the model (for the moment) so that we can evaluate the effectiveness of Positive Affect. The diagonal lines in Figure 5a.2 show the variance of Self-Esteem that is accounted for by Negative Affect and Openness; the remaining blank area shows the residual variance of Self-Esteem. The partial correlation associated with Positive Affect is 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 describes the linear 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 over and above the predictive work done by the set of predictors already in the model (how much of the residual variance of Self-Esteem it can explain), the software starts the process of computing the weight that Positive Affect will receive in the model.

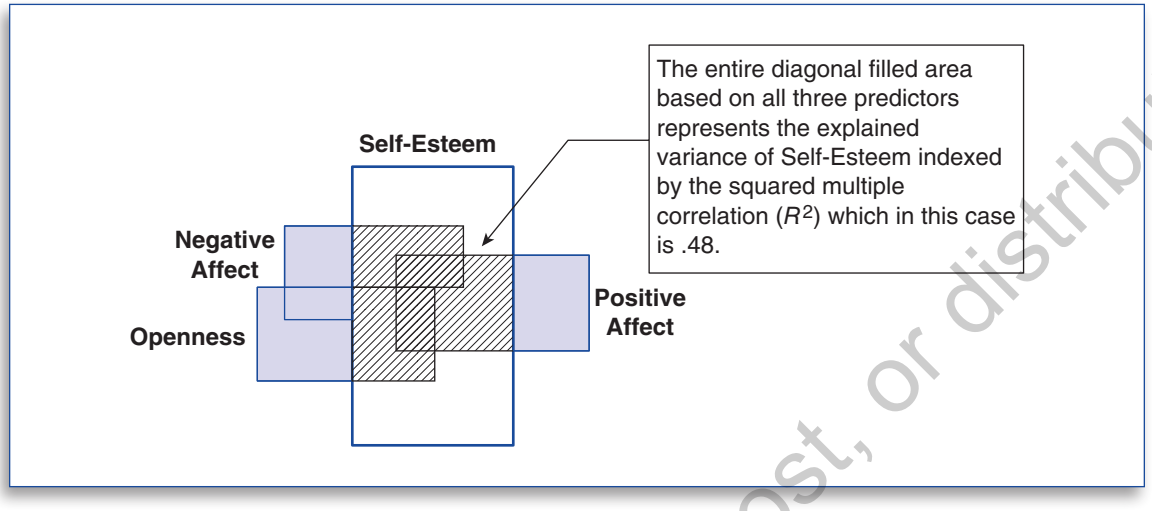
5A.9 The Squared Multiple Correlation

Assume that each of the three predictors has been evaluated in a single but obviously complex step or block, so that we know their weights and can construct the model. We will discuss the specific numerical results shortly, but first we need to cover three additional and important concepts and their associated statistical indexes: the squared multiple correlation, the squared semipartial correlations, and the structure coefficients.

The first of these is the squared multiple correlation, symbolized by R^2 and illustrated in Figure 5a.3. All three predictors are now in the model, and based on our discussion in the last two sections, it likely makes sense to you that all three variables in combination are accounting for the amount of Self-Esteem variance depicted by the entire filled (shaded) area in Figure 5a.3.

You are already familiar with the idea that the degree of correlation between two variables can be pictured as overlapping figures (we have used squares to conform to the pictorial style of path analysis and structural equation modeling that we cover in the later chapters, but introductory statistics and research methods texts tend to use circles). For the case of the Pearson *r* (or any bivariate correlation),

Figure 5a.3 All Three Predictors Are Now in the Model



the shaded or overlapping area would show the strength of the correlation, and its magnitude would be indexed by r^2.

The relationship shown in Figure 5a.3 is more complex than what is represented by r^2, but it is conceptually similar. Three variables are drawn as overlapping with Self-Esteem. Nonetheless, this still represents a correlation, albeit one more complex than a bivariate correlation. Specifically, we are looking at the relationship between the criterion variable (Self-Esteem) on the one hand and the three predictors (Positive Affect, Negative Affect, and Openness) on the other hand. When we have three or more variables involved in the relationship (there are four here), we can no longer use the Pearson correlation coefficient to quantify the magnitude of that linear 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 a *multiple correlation* coefficient. It is symbolized as an uppercase italic *R*. That said, the Pearson *r* is really the limiting case of *R*, and thus the bulk of what we have said about *r* applies to *R*.

A multiple correlation coefficient indexes the degree of linear association of one variable (the outcome variable in the case of multiple regression analysis) with a set of other variables (the predictor variables in the case of multiple regression analysis), and the *squared multiple correlation* (R^2), sometimes called the *coefficient of multiple determination*, tells us the strength of this complex relationship, that is, it tells us how much variance of the outcome variable is explained by the set of predictor variables. In Figure 5a.3, the diagonally shaded area—the overlap of Positive Affect, Negative Affect, and Openness with Self-Esteem—represents the R^2 for that relationship. 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.

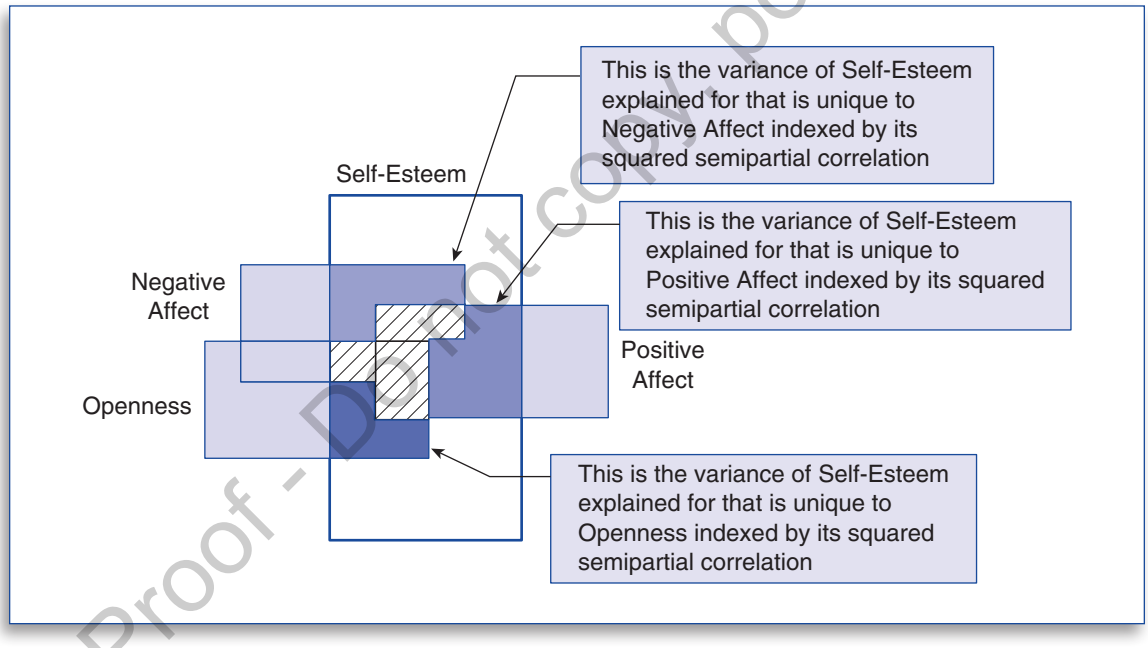
The second feature important to note in Figure 5a.3 is that the three predictors overlap with each other, indicating that they correlate with one another (as we documented in Table 5a.2). The degree to which they correlate affects the partial regression weights these variables are assigned in the regression equation, so the correlations of the predictors become a matter of some interest to researchers using a regression design.

5A.10 The Squared Semipartial Correlation

5A.10.1 The Squared Semipartial Correlation Itself

We have redrawn with a bit of a variation in Figure 5a.4 the relationship between the three predictors and the dependent variable previously shown in Figure 5a.3. In Figure 5a.4, we have distinguished the variance of Self-Esteem that is uniquely associated with only a single predictor by using a solid fill and have retained the diagonal fill area to represent variance that overlaps between two or all of the predictors. The amount of variance explained *uniquely* by a predictor is indexed by another correlation statistic known as the *squared semipartial correlation*, often symbolized as sr^2. It represents the extent to which each variable does independent predictive work when combined with the other predictors in the model. Each predictor is associated with a squared semipartial correlation. The semipartial correlation describes the linear relationship between a given predictor and the variance of the dependent variable.

Figure 5a.4 A Depiction of the Squared Semipartial Correlations



Note. The solid fill represents self-esteem variance accounted for that is unique to each predictor; the diagonal fill represents self-esteem variance accounted for common to two or more predictors.

5A.10.2 The Difference Between the Squared Semipartial Correlation and the Squared Partial Correlation

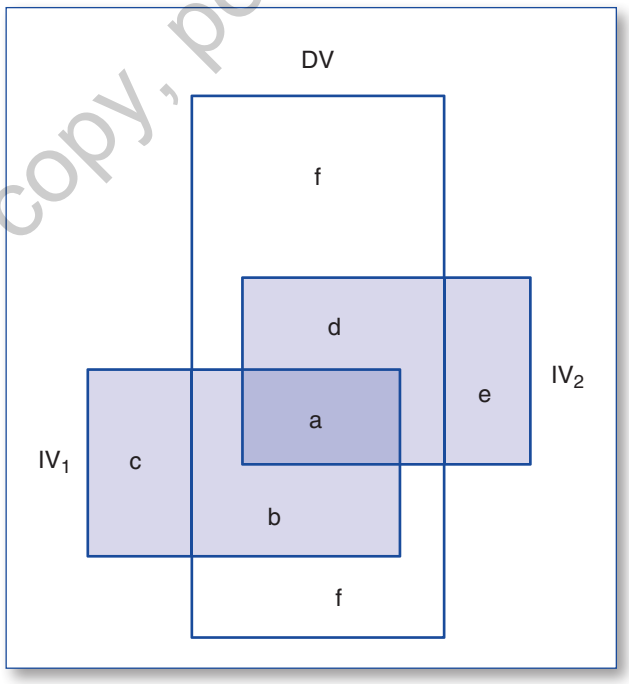
Distinguishing between the squared semipartial and squared partial correlations is subtle but very important because these represent descriptions of two similar but different relationships between each predictor and the dependent variable. To simplify our discussion, we have drawn in Figure 5a.5 only two generic predictor (independent) variables (IV_1 and IV_2) for a given generic dependent variable.

In Figure 5a.5, the lowercase letters identify different areas of the variance of the predictor and dependent variables so that we may contrast the squared semipartial correlation coefficient with the squared partial correlation coefficient. These different variance areas are as follows:

- Area **a** is the variance of the dependent variable that is explained but cannot be attributed uniquely to either predictor.
- Area **b** is the variance of the dependent variable that is explained uniquely by IV_1.
- Area **c** is the variance of IV_1 that is not related to the dependent variable.
- Area **d** is the variance of the dependent variable that is explained uniquely by IV_2.
- Area **e** is the variance of IV_2 that is not related to the dependent variable.
- Area **f** (the blank area of the dependent variable labeled twice in the figure) is the variance of the dependent variable that is not explained by either predictor (it is the residual variance of the dependent variable once the model with two predictors has been finalized).

Consider Area **b** in Figure 5a.5, although an analogous analysis can be made for Area **d**. This is the variance of the dependent variable that is explained only (uniquely) by IV_1. Because we are dealing with squared correlations that are interpreted as a percent of variance explained, we must compute a proportion or percent, that is, we must compute the value of a ratio between two variances. Area **b** is the conceptual focus of both the squared semipartial correlation as well as the squared partial correlation for IV_1; because it is the focus of the proportion we calculate, it must be placed in the numerator in the computation for both the squared semipartial correlation and the squared partial correlation. Stated in words, both the squared semipartial correlation and the squared partial correlation associated with IV_1 each describe the percentage of variance attributable to the unique contribution of IV_1.

Figure 5a.5 Contrasting the Squared Semipartial and Squared Semipartial Correlations for IV_1



Note. The squared semipartial correlation is computed as **b/(a + b + d + f)** and the squared partial correlation is computed as **b/(b + f)**.

The difference between the two indexes is what representation of the variance is placed in the denominator of the ratio. The denominator of a ratio provides the frame of reference. In the present situation, we wish to know the percentage of variance attributable to the unique contribution of IV_1 with respect to one of two frames of reference:

- The frame of reference used in computing the squared *semipartial* correlation is the total variance of the dependent variable. In Figure 5a.5, the denominator would be **a** + **b** + **d** + **f**. Thus, the computation of the squared semipartial correlation for IV_1 is **b**/(**a** + **b** + **d** + **f**).

What we obtain is the percent of variance of the dependent variable that is associated with the unique contribution of IV_1.

- The frame of reference used in computing the squared *partial* correlation between the predictor and the dependent variable is only that portion of the variance of the dependent variable remaining when the effects of the other predictors have been removed (statistically removed, nullified). In Figure 5a.5, the denominator would be **b** + **f**. Thus, the computation of the squared partial correlation for IV_1 is **b**/(**b** + **f**). What we obtain is the percent of variance of the dependent variable not predicted by the other predictor(s) that is associated with the unique contribution of IV_1.

Given that the frame of reference for the squared partial correlation contains only Areas **b** and **f** whereas the frame of reference for the squared semipartial correlation contains Areas **b**, **f**, **a**, and **d**, it follows that the denominator for the squared semipartial correlation will always be larger than the denominator for the squared partial correlation (unless the other areas have a zero value). Because (in pictorial terms) we are asking about the relative size of Area **b** for both squared correlations, the upshot of this straightforward arithmetic is that the squared partial correlation will almost always have a larger value than the squared semipartial correlation. This explanation may be summarized as follows:

- The squared semipartial correlation represents the proportion of variance of the dependent variable uniquely explained by an independent variable when the other predictors are taken into consideration.
- The squared partial correlation is the amount of explained variance of the dependent variable that is incremented by including an independent variable in the model that already contains the other predictors.

When the regression model is finally in place, as it is in Figure 5a.5, our interest in the squared partial correlation fades because it was more useful in constructing the model, and our interest shifts to the squared semipartial correlation. Thus, when examining Figure 5a.5 or the numerical results of the regression analysis, what interests us is the variance of the dependent variable that is uniquely explained by each predictor, that is, we are interested in the squared semipartial correlation associated with each predictor (in Figure 5a.5, that would be Areas **b** and **d** with respect to the total variance of the dependent variable).

5A.10.3 The Squared Semipartial Correlation and the Squared Multiple Correlation

In Figure 5a.5, the squared multiple correlation (R^2) can be visualized as the proportion of the total variance of the dependent variable covered by Areas **a**, **b**, and **d**. That is, the squared multiple correlation takes into account not only the unique contributions of the predictors (Areas **b** and **d**) but also the overlap between them (Area **a**). The squared semipartial correlations focus only on the unique contributions of the predictors (Areas **b** and **d**).

Note that the sum of Area **b** (the variance of the dependent variable uniquely explained by IV_1) and Area **d** (the variance of the dependent variable uniquely explained by IV_2) does not cover all of the shaded area (it does not take into account Area **a**). Translated into statistical terms, the sum of the squared semipartial correlations—the total amount of variance uniquely associated with individual

predictors—does not equal the squared multiple correlation. The reason for this is that the predictor variables are themselves correlated and to that extent will overlap with each other in the prediction of the dependent variable that they are attempting to predict.

Generally, we can tell how well the model fits the data by considering the value of the squared multiple correlation. But we can also evaluate how well the model works on an individual predictor level by examining the squared semipartial correlations (Tabachnick & Fidell, 2013b). With the squared semipartial correlations, we 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 larger unique contributions.

There are some limitations in using squared semipartial correlations to compare the contributions of the predictors. The unique contribution of each variable in multiple regression analysis is very much a function of the correlations of the variables used in the analysis. It is quite likely 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 partial regression 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 themselves be substantial, the overlap of the predictors would be minimal, and the sum of the squared semipartial correlations would approximate the value of the squared multiple correlation.

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 most 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 a researcher's time and runs the risk of yielding distorted or inaccurate conclusions.

5A.11 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 a structure coefficient.

A structure coefficient represents the correlation between an individual variable that is a part of the variate and the weighted linear combination itself. Stronger correlations indicate that the predictor is a stronger reflection (indicator, gauge, marker) of the construct underlying the variate. Because

the variate is a latent variable, a structure coefficient can index how well a given predictor variable can serve as an indicator or marker of the construct represented by the variate. This feature of structure coefficients makes them extremely useful in multivariate analyses, and we will make considerable use of them in the context of canonical correlation analysis (Chapters 7A and 7B), principal components and factor analysis (Chapters 10A and 10B), and discriminant function analysis (Chapters 19A and 19B).

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

$$\text{Structure coefficient} = \frac{r_{xi\,y}}{R}$$

where $r_{xi\,y}$ is the Pearson correlation between the given predictor (x_i) and the actual (measured) dependent variable, and R is the multiple correlation.

5A.12 Statistical Summary of the Regression Solution

There are two levels of the statistical summary of the regression solution, a characterization of the effectiveness of the overall model and an assessment of the performance of the individual predictors. Examining the results for the overall model takes precedence over dealing with the individual predictors—if the overall model cannot predict better than chance, then there is little point in evaluating how each of the predictors fared. We discuss evaluating the overall model in Section 5A.13 and examining the individual predictor variables in Section 5A.14.

5A.13 Evaluating the Overall Model

The overall model is represented by the regression equation. There are two questions that we address in evaluating the overall model, one involving a somewhat more complex answer than the other:

- Is the model statistically significant?
- How much variance does the model explain?

5A.13.1 Is the Model Statistically Significant?

The simpler of the two questions to answer concerns the statistical significance of the model. The issue is whether the predictors as a set can account for a statistically significant amount of the variance of the dependent variable. This question is evaluated by using an ANOVA akin to a one-way between subjects design (see Chapter 18A), with the single "effect" in the ANOVA procedure being the regression model itself. The degrees of freedom associated with the total variance and its partitions are as follows:

- The degrees of freedom for the total variance are equal to $N - 1$, where N is the sample size.
- The degrees of freedom for the regression model (the effect) is equal to the number of predictor variables in the model that we symbolize here as v.
- The degrees of freedom for the error term are equal to $(N - 1) - v$; that is, it is equal to the total degrees of freedom minus the number of predictors in the model.

The null hypothesis tested by the *F* ratio resulting from the ANOVA is that prediction is no better than chance in that the predictor set cannot account for any of the variance of the dependent variable. Another way to express the null hypothesis is that $R^2 = 0$. If the *F* ratio from the ANOVA is statistically significant, then it can be concluded that the model as a whole accounts for a statistically significant percentage of the variance of the dependent variable (i.e., $R^2 > 0$). In our example analysis using Positive Affect, Negative Affect, and Openness to predict Self-Esteem, the effect of the regression model was statistically significant, $F(3, 416) = 129.32$, $p < .001$, $R^2 = .48$, adjusted $R^2 = .48$. We would therefore conclude that the three independent variables of Positive Affect, Negative Affect, and Openness in combination significantly predicted Self-Esteem.

5A.13.2 The Amount of Variance Explained by the Model: R^2 and Adjusted R^2

5A.13.2.1 Variance Explained: The Straightforward Portion of the Answer

The more complex of the two questions to answer concerns how much variance of the dependent variable the model explains. We can answer this question at one level in a straightforward manner for the moment by examining the value of R^2. In our example, the value for R^2, shown in a separate table in the IBM SPSS output, was .483. The straightforward answer to the question, then, is that the three predictors in this particular weighted linear combination were able to explain about 48% of the variance of Self-Esteem.

We should also consider the magnitude of the obtained R^2. One would ordinarily think of .483 as reasonably substantial, and most researchers should not be terribly disappointed with R^2s considerably less than this. 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, multicultural sensitivity), very small but statistically significant R^2s may be cause for celebration by a research team. Just as we suggested in Sections 2.8.2 for eta square, in the absence of any context R^2 values, .10, .25, and .40 might be considered to be small, medium, and large strengths of effect, respectively (Cohen, 1988); however, we conduct our research within a context, and so the magnitude of the effect must be considered with respect to the theoretical and empirical milieu within which the research was originally framed.

5A.13.2.2 Variance Explained: R^2 Is Somewhat Inflated

At another level, the answer to the question of how much variance is explained by the regression model has a more complex aspect to it. The reason it is more complex is that the obtained R^2 value is actually somewhat inflated. Two major factors drive this inflation:

- The inevitable presence of error variance.
- The number of predictors in the model relative to the sample size.

We can identify two very general and not mutually incompatible strategies that can be implemented (Darlington, 1960; Yin & Fan, 2001) to estimate the amount of R^2 inflation (as we will see in a moment, the ordinary terminology focuses on R^2 *shrinkage*, the other side of the R^2 inflation coin); one set of strategies is empirical and is focused on error variance; another set of strategies is analytic and is focused on the number of predictors with respect to sample size.

5A.13.2.3 Variance Explained: Empirical Strategies for Estimating the Amount of R^2 Shrinkage

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 working in the direction of enhanced prediction. The multiple regression algorithm, 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 raw and standardized regression weights from both true and error sources combined. By drawing information from both true score variance and error variance because it cannot distinguish between the two sources in fitting the model to the data, multiple regression procedures will overestimate the amount of variance that the model explains (Cohen et al., 2003; Yin & Fan, 2001).

The dynamics of this problem can be understood in this way: In another sample, the random dictates of error will operate differently, and if the weighting coefficients obtained from our original regression analysis are applied to the new sample, the model will be less effective than it appeared to be for the original sample, that is, the model will likely yield a somewhat lower R^2 than was originally obtained. This phenomenon is known by the term R^2 *shrinkage*. R^2 shrinkage is more of a problem when we have relatively smaller sample sizes and relatively more variables in the analysis. As sample size and the number of predictors reach more acceptable proportions (see Section 5A.13.2.4), the shrinkage of R^2 becomes that much less of an issue.

Empirical strategies estimating the amount of R^2 shrinkage call for performing a regression analysis on selected portions of the sample, an approach generally known as a *resampling* strategy. In the present context, resampling can be addressed through procedures such as *cross-validation*, *double cross-validation*, and the use of a *jackknife* procedure.

To perform a cross-validation procedure, we ordinarily divide a large sample in half (into two equal-sized subsamples, although we can also permit one sample to be larger than the other) by randomly selecting the cases to be assigned to each. We then compute our regression analysis on one subsample (the larger one if unequal sample sizes were used) and use those regression weights to predict the criterion variable of the second "hold-back" sample (the smaller one if unequal sample sizes were used). 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 is known 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, we obtain two estimates of shrinkage rather than one, that is, we can obtain two cross-validation coefficients. If the shrinkage is not excessive, and there are few guidelines as to how to judge this, we can then perform an analysis on the combined sample and report the double cross-validation results to let readers know the estimated generalizability of the model.

The *jackknife* procedure was introduced by Quenouille (1949) and is currently treated (e.g., Beasley & Rodgers, 2009; Carsey & Harden, 2014; Chernick, 2011; Efron, 1979; Fuller, 2009; Wu, 1986) in the context of *bootstrapping* (where we draw with replacement repeated subsamples from our sample to estimate the sampling distribution of a given parameter). The jackknife procedure is also called, for what will be obvious reasons, a *leave-one-out* procedure.

Ordinarily, we include all of the cases with valid values on the dependent and independent variables in an ordinary regression analysis. To apply a jackknife procedure, we would temporarily

remove one of those cases (i.e., leave one of those cases out of the analysis), say Case A, perform the analysis without Case A, and then apply the regression coefficients obtained in that analysis to predict the value of the dependent variable for Case A. We then "replace" Case A back into the data set, select another case to remove, say Case B, repeat the process for Case B, and so on, until all cases have had their *Y* score predicted.

The comparison of these jackknifed results with those obtained in the ordinary (full-sample) analysis gives us an estimate of how much shrinkage in explained variance we might encounter down the road. IBM SPSS does not have a jackknife procedure available for multiple regression analysis, but that procedure is available for discriminant function analysis (see Section 19A.10.4). IBM SPSS does, however, offer a bootstrapping add-on module (IBM, 2013b) but we do not cover it in this book; coverage of this topic can be found in Chernick (2011), Davison and Hinkley (1997), and Efron and Tibshirani (1993).

5A.13.2.4 Variance Explained: Analytic Strategies for Estimating the Amount of R^2 Shrinkage

In addition to capitalizing on chance, R^2 can also be mathematically inflated by having a relatively larger number of predictors relative to the size of the sample, that is, R^2 can be increased simply by increasing the number of predictors we opt to include in the model for a given sample size (e.g., Stuewig, 2010; Wooldridge, 2009; Yin & Fan, 2001). The good news here is that statisticians have been able to propose ways to "correct" or "adjust" the obtained R^2 that takes into account both the sample size and the number of predictors in the model. When applied to the obtained R^2, the result is known as an *adjusted* R^2. This adjusted R^2 provides an estimate of what the R^2 might have been had it not been inflated by the number of predictors we have included in the model relative to our sample size. All of the major statistical software packages report in their output an adjusted R^2 value in addition to the observed R^2.

To further complicate this scenario, there are two different types of adjusted R^2 values that are described in the literature, and there is no single way to compute either of them. Some of these formulas (e.g., Olkin & Pratt, 1958; Wherry, 1931) are intended to estimate the population value of R^2, whereas others (e.g., Browne, 1975; Darlington, 1960; Rozeboom, 1978) are intended to estimate the cross-validation coefficient (Yin & Fan, 2001). Yin and Fan (2001) made a comparison of the performance of 15 such formulas (some estimating the population value of R^2 and others estimating the cross-validation coefficient), and Walker (2007) compared four estimates of the cross-validation coefficient in addition to a bootstrap procedure.

Our interest here is with the adjusted R^2 as an estimate of the population value of R^2. The algorithm that is used by IBM SPSS to compute its adjusted R^2 value (IBM, 2013a) and three other well-known formulas, the Wherry formulas 1 and 2 and the Olkin–Pratt formula, are presented in Figure 5a.6. They all contain the following three elements:

- The obtained value of R^2 from the multiple regression analysis.
- The sample size, designated as *N*.
- The number of predictor variables in the model, which we are designating as *v*.

Although these formulas are somewhat different and when solved will lead to somewhat different adjusted values of R^2, the estimates appear to be relatively close to each other from the standpoint of researchers. For example, let us assume that investigators conducting a hypothetical research study

Figure 5a.6 The IBM SPSS Formula, the Wherry Formulas 1 and 2, and the Okin–Pratt Formula for Computing the Adjusted R^2 as an Estimate of the Population R^2

IBM SPSS:
$$\text{Adjusted } R^2 = R^2 - \frac{(1 - R^2)(v)}{N - (v + 1)}$$

Wherry – 1:
$$\text{Adjusted } R^2 = 1 - \frac{N}{N - v}(1 - R^2)$$

Wherry – 2:
$$\text{Adjusted } R^2 = 1 - \frac{N - 1}{N - v}(1 - R^2)$$

Olkin – Pratt:
$$\text{Adjusted } R^2 = 1 - \left[\frac{(N - 3)(1 - R^2)}{N - v - 1}\right]\left[1 + \frac{(2)(1 - R^2)}{N - v + 1}\right]$$

Note. In the formulas to complete the adjusted R^2 value, R^2 is the obtained value from the analysis, N is the sample size, and v is the number of predictor variables in the model.

used a sample size of 100 (what we would regard as too small an N) and eight predictor variables (an excessive number of predictors for that sample size), and they obtained an R^2 of .25. The resulting adjusted R^2 values from each of the formulas are as follows:

- IBM SPSS algorithm: .184066
- Wherry Formula 1: .1847827
- Wherry Formula 2: .1929349
- Olkin–Pratt formula: .1876552

Considering that these values hover around .18 or .19, the adjusted R^2 values appear to be approximately three quarters the magnitude of the observed R^2 value. In our view, this is a considerable amount of estimated shrinkage.

We can contrast this hypothetical result with our worked example. In our worked example, the adjusted R^2 value was .479 (shown as part of the IBM SPSS output), giving us virtually the same value as the unadjusted R^2. That such little adjustment was made to R^2 is most likely a function of the sample size to number of variables ratio we used (the analysis was based on a sample size of 420 cases; with just three predictors, our ratio was 140:1).

Yin and Fan (2001) suggested that good quality estimates of adjusted R^2 values were obtained with most of the R^2 estimation formulas they evaluated when the ratio of sample size to number of predictors was 100. In research environments where there are a limited number of potential cases that may be recruited for a study (e.g., a university, hospital, or organizational setting), such a ratio may be difficult to achieve. In more restrictive settings where we must accept pragmatic compromises to get the research done, our recommendations are that the sample size should generally be no less than about 200 or so cases and that researchers use at least 20 but preferably 30 or more cases per predictor.

5A.14 Evaluating the Individual Predictor Results

We have performed the analysis for the regression design that we have been discussing, and portions of the output for the predictor variables is summarized in Table 5a.3. For each predictor, we show its unstandardized (*b*) and standardized (beta) regression weighting coefficients, the *t* value associated with each regression weight, the Pearson correlation (*r*) of each predictor with the dependent variable, the amount of Self-Esteem variance each predictor has uniquely explained (squared semipartial correlation), and the structure coefficient associated with each predictor. The constant is the *Y* intercept of the raw score model and is shown in the last line of the table, and the R^2 and the adjusted R^2 values are shown in the table note.

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

Variables in Model	b	Beta	t	r	Squared Semipartial Correlation	Structure Coefficient
Positive Affect	2.89	.40	10.61*	.55	.14	.80
Negative Affect	-2.42	-.43	-11.50*	-.57	.16	-.82
Openness	0.11	.06	1.64	.22	.00	.32
Constant (Y intercept)	56.66					

Note. Dependent variable is Self-Esteem. R^2 = .48, Adjusted R^2 = .48
*p < .05.

5A.14.1 Variables in the Model

The predictor variables are shown in the first column of Table 5a.3. This represents a complete solution in the sense that all the independent variables are included in the final equation regardless of how much (or how little) they contribute to the prediction model.

5A.14.2 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. These are produced below.

The raw score equation is as follows:

$$\text{Self-Esteem}_{\text{pred}} = 56.66 + (2.89)(\text{pos affect}) - (2.42)(\text{neg affect}) + (.11)(\text{open})$$

The standardized equation is as follows:

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

5A.14.3 t Tests

IBM SPSS tests the significance of each predictor in the model 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* test output shown in Table 5a.3 informs us that only Positive Affect and Negative Affect are statistically significant predictors in the model; Openness does not receive a strong enough weight to reach that touchstone.

5A.14.4 b 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 that to mean that when the other variables are controlled for, an increase of 1 point on the Positive Affect measure is, on average, associated with a 2.89-point gain in Self-Esteem. Likewise, the *b* weight of −2.42 for Negative Affect would mean that, with all of the other variables statistically controlled, every point of increase on the Negative Affect measure (i.e., greater levels of Negative Affect) corresponds to a lower score on the Self-Esteem measure of 2.42 points.

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.

5A.14.5 Beta Coefficients

5A.14.5.1 General Interpretation

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 Affect 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 Affect and Negative Affect are given much more relative weight than Openness.

5A.14.5.2 The Case for Using Beta Coefficients to Evaluate Predictors

Some authors (e.g., Cohen et al., 2003; Pedhazur, 1997; Pedhazur & Schmelkin, 1991) have cautiously argued that, at least under some circumstances, we may be able to compare the beta coefficients of the predictors 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 (1991) 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, 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 (.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 Affect was 1.075 times a more potent predictor in this model. Translated to ordinary language, we would say that Negative Affect 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.

5A.14.5.3 Concerns With Using the Beta Coefficients to Evaluate Predictors

We indicated above that even when authors such as Pedhazur (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 is 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 the use of 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 (because each is evaluated with the other treated as a covariate). 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 places additional limitations on the generality of the betas and thus their use in evaluating or comparing predictive effectiveness of the independent variables.

The sensitivity of a beta coefficient associated with a given predictor to the correlation of that predictor with other predictors in the model can also be manifested in the following manner: If two or three very highly correlated predictors were included in the model, their beta coefficients could exceed a value of 1.00, sometimes by a considerable margin. Ordinarily, researchers would not include very highly correlated variables in a regression analysis (they would either retain only one or create a single composite variable of the set), but there are special analyses where researchers cannot condense such a set of variables (see Section 6B.2, for an example); in such analyses, researchers focus on R^2 or (depending on the analysis) the change in R^2 and ignore these aberrant beta coefficient values.

Recognizing that the value of a beta coefficient associated with a variable is affected by, among other factors, the variability of the variable, the correlation of the variable with other predictors in the model, and the measurement error in assessing the variable, Jacob Cohen (1990) in one of his classic articles titled "Things I Have Learned (So Far)" went so far as to suggest that in many or most situations, simply assigning unit or unitary weights (values of 1.00) to all significant predictors can result in at least as good a prediction model as using partial regression coefficients to two decimal values as the weights for the variables. Cohen's strategy simplifies the prediction model to a yes/no decision for each potential predictor, and although it is not widely used in regression studies, it is a strategy that is commonly used in connection with factor analysis where a variable is either included with a unitary weight or not included when we construct the subscales that are used to represent a factor (see Section 10B.14).

5A.14.5.4 Recommendations for Using Betas

We do not want to leave you completely hanging at this point in our treatment, so we will answer the obvious questions. Should you use the beta weights to assess the relative strengths of the predictors in your own research? Yes, although we have considerable sympathy with the wisdom expressed by Cohen (1990) of using unit weights. Should beta coefficients be the only index you check out? No. The structure coefficients and the squared semipartial correlations should be examined as well. And, ultimately, using the raw regression weights to inform us of how much of a change in the dependent variable is associated with a unit difference in the predictor, given that all of the other predictors are acting as covariates, will prove to be a very worthwhile interpretation strategy.

5A.14.6 Pearson Correlations of the Predictors With the Criterion Variable

The fourth 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, we can see that the correlations between Self-Esteem and Positive Affect and Openness are positive. This was the case because each of these variables is scored in the positive direction—higher scores mean that respondents exhibit greater levels of self-esteem and more positive affective behaviors and that they are more open to new or interesting experiences. 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 because those individuals who exhibit more negative affective behaviors are typically those who have lower levels of self-esteem.

5A.14.7 Squared Semipartial Correlations

The next to last column of Table 5a.3 displays the squared semipartial correlations for each predictor. These correlations are shown in the IBM 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, the predictors uniquely account for 30% of the variance, 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 is another gauge of relative predictor strength in the model. From this perspective, Positive Affect and Negative Affect are approximately tied in their unique contribution to the prediction model under the present research circumstances, whereas Openness is making no contribution on its own.

5A.14.8 The Structure Coefficients

The last column in Table 5a.3 shows the structure coefficients, an index of the correlation of each variable with the weighted linear combination (the variate or prediction model). These coefficients needed to be hand calculated (see Section 5A.11) because IBM 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 fourth numerical column) by the value of the multiple correlation. To illustrate this computation for Positive Affect, we divide its Pearson correlation with Self-Esteem (.55) by the value of *R* (the square root of R^2); thus, .55 is divided by the square root of .483, or .55/.69 = approximately .80.

The structure coefficients indicate that Positive Affect and Negative Affect 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. We would use these structure coefficients to interpret the variate; in this example, we would say that in the context of this predictor set, the affect levels of individuals best predict self-esteem. Note that in the everyday world more than affect levels unquestionably predict self-esteem but, because we used only three predictors in this study, our conclusions are limited. Such a limitation is generally true for multiple regression analysis, in that we can draw our conclusions only on the variables in the study, and the variable set we used may not be inclusive of all the potential determiners of our outcome variable (we may not be able to realistically *fully specify* all of the potentially viable predictors).

Such consistency in interpretation between the interpretations based on the structure coefficients and the beta coefficients as we saw in this example is not always obtained. Beta coefficients and structure coefficients differ in at least two important ways.

- 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.
- Beta coefficients can exceed the ordinary range of ±1 when the predictors are relatively highly correlated with each other. Many researchers are not keen on seeing, much less interpreting, beta weights greater than unity. However, structure coefficients are absolutely bounded by the range ±1 because they are correlation coefficients, thus making them always 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. Unlike the beta coefficients and the squared semipartial correlations, structure

coefficients are not affected by the correlations between the predictors although, as is true for all of the regression statistics, the structure coefficients could change substantially if a different set of predictors happened to be used.

Pedhazur (1997) notes that structure coefficients will show the same pattern of relationships as the Pearson correlations of the predictors and the criterion. Because of this, Pedhazur is not convinced of the utility of structure coefficients. Nonetheless, 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. Furthermore, it is common practice to make extensive and regular use of structure coefficients in other multivariate analyses where our focus is on interpreting a variate (e.g., factor analysis, discriminant function analysis, canonical correlation analysis), and so it makes sense to include regression under that umbrella as well.

5A.15 Step Methods of Building the Model

The step methods of building the regression equation that we briefly cover here are part of the class of statistical regression methods, and it will be clear from our descriptions just how the software programs are in charge of the decision process for selecting the ordering of the predictors as the model is built. We cover here the forward method, the backward method, and the stepwise method. These methods construct the model one step 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, although importance is still relative to the set of predictors that are participating in the analysis. The methods differ primarily in how they arrange the steps in entering or removing variables from the model.

5A.16 The Forward Method

In the forward method of multiple regression analysis, rather than placing all the variables in the model at once, we add independent variables to the model one variable at a time. Thus, each step corresponds to a single variable absorbed into the model. At each step, we enter the particular variable that adds the most predictive power at that time, with the proviso that the variable accounts for a statistically significant amount of the unexplained variance of the dependent variable (i.e., that its partial correlation is statistically significant). Most software applications use an alpha of .05 to define statistical significance. 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 the outcome variable (Self-Esteem in our example) would be entered first (assuming statistical significance).

In the forward method, once a variable is entered into the model, that variable remains permanently in the model. It may seem odd for us to say this, but permanent membership in the model is not necessarily true for variables entered into the model for the other step methods we discuss. For the next step in the forward method, the remaining variables are evaluated and the variable with the highest statistically significant partial correlation (the correlation between the residual variance of Self-Esteem and that additional predictor) is entered provided that the partial correlation is statistically significant. In this case, Positive Affect would join Negative Affect as a predictor in the model.

This process of adding variables to the model is repeated for each remaining predictor with the variables in the model all acting as covariates. We would find, with Negative Affect 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 accounted for by Negative Affect and Positive Affect (i.e., it would not be associated with a statistically significant partial correlation coefficient). Once the next variable in line (the best of the remaining predictors) fails to reach the entry criterion for entry into the model, the forward procedure ends with however many variables already in the model and the remaining variables not included in the model. In our example, the forward procedure would stop at the end of the second step and Openness would remain on the sidelines.

5A.17 The Backward Method

The backward method works not by adding significant variables to the model but, rather, by removing nonsignificant predictors from the model 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 on the next step. 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 (with the requirement that it is not statistically significantly contributing to R^2) is removed. This removal process and model reconstruction process continues until there are only statistically significant predictors remaining in the equation. In our example, Openness would have been removed at the first opportunity. The backward method would have stopped at that point because both remaining variables would have been significant predictors.

5A.18 Backward Versus Forward Solutions

Backward regression does not always produce the same model as forward regression even though it would have done so in our simplified example. Here is why: Being entered into the equation in the forward method requires predictors to meet a more stringent criterion than variables being retained in the model in the backward method. This creates a situation in which it is more difficult to get into the model than to remain in it. The alpha or probability level associated with entry and removal defines stringency or difficulty statistically.

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 alpha 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 (the removal criterion needs to be somewhat less stringent than the entry criterion in order to avoid a logic glitch in the entry-removal decision process—see Section 5A.19). 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

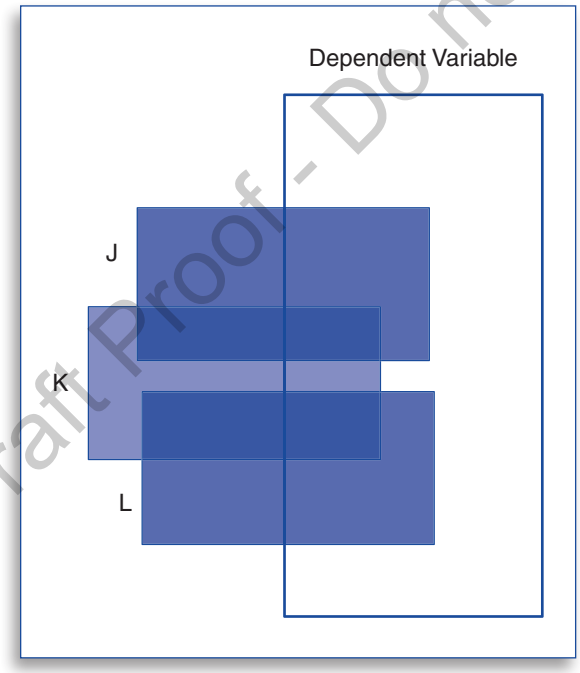
method if its probability level is greater than .10 (e.g., $p = .11$), but it will be retained in the equation if its probability level is equal to or less than .10 (e.g., $p = .09$).

The consequences of using these different criteria for entry and removal affect 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 criterion of .05, 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 statistically significant at the .10 alpha level. It therefore remains in the equation. In this case, the model built under the backward method would incorporate this predictor, but the model built under the forward method would have excluded it.

Figure 5a.7 The Unique Contribution of Variable *K* Is Reduced by the Addition of Variable *L*



5A.19 The Stepwise Method

The stepwise method of building the multiple regression equation is a fusion 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 model and builds it one step at a time. Once a third independent variable is added to the model, 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 model if they significantly ($p \leq .05$) 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.7.

In Figure 5a.7, predictor *J* was entered first, *K* was entered second, and *L* was just entered as the third predictor. We are poised at the moment when *L* joined the equation. Note that between predictors *J* and *L*, there is very little predictive 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 unique variance that *K* now accounts for is examined with variables *J* and *L* acting as covariates, 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. 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 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.

5A.20 Evaluation of the Statistical Methods

5A.20.1 Benefits of the Standard Method

The primary advantage of using the standard method 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.

5A.20.2 Benefits of the Step Methods

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. This same argument applies when considering the forward and backward methods. 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.

5A.20.3 Criticisms of the Statistical Methods as a Set

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 statistically controlled. 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 researchers to exercise some judgment in which variables are entered at certain points in the analysis, and this is discussed in Chapter 6A. This issue is also related to multicollinearity, a topic that we discuss in Section 5A.21.

The step methods have become increasingly less popular over the years as their weaknesses have become better understood and as hierarchal methods and path approaches have gained in popularity. Tabachnick and Fidell (2013b), 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 40 to 1 ratio of cases to independent variables because there are serious threats to external validity (Cohen et al., 2003, p. 162). 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 model.
- 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 model 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.
- There is a tendency using the step methods to "overfit" the data (Lattin, Carroll, & Green, 2003). Briefly, this criticism suggests that variables are chosen for inclusion in the model "based on their ability to explain variance in the sample that may or may not be characteristic of the variance in the population by capitalizing unduly on error, chance correlation, or both" (Lattin et al., 2003, p. 51).

5A.20.4 Balancing the Value of All the Statistical Methods of Building the Model

The standard method works well if we 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, we are not evaluating them in isolation or in subsets. The standard method will allow us to test hypotheses about the model as a whole; if that is the goal, then that is what should be used.

The step methods are intended to identify which variables should be in the model on purely statistical grounds. Many researchers discourage such an atheoretical approach. On the other hand,

there may be certain applications where all we 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 alternative methods of performing the regression analysis. Although they do require more thoughtful decision making rather than just entering the variables and selecting a statistical method, the flexibility and potential explanatory power they afford more than compensate for the effort it takes to run such analyses. Some of these regression procedures are discussed in Chapter 6A and the path model approach is broached in Chapters 12A through 15B.

5A.21 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 when 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. On the positive side, with both variables in the model, the R^2 value will be appropriately high, and 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 standardized regression coefficients of the highly correlated independent variables are distorted, sometimes exceeding the ordinarily expected range of ±1. 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, we 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 analysis. 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 (1999a) 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 that researchers may use 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. We strongly advise users to employ 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. Researchers 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, we might combine height and weight to form a measure of body mass. As another example, we might average three highly correlated survey items; principal components analysis and exploratory factor analysis, discussed in Chapters 10A and 10B, can be used to help determine which variables might productively be averaged together without losing too much information. Further, related measures may be able to be used as indicators of a latent variable that can then be placed into a structure equation model (see Chapters 14A and 14B).

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 the measures rather than both of them.

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 (1999a, p. 141) gets concerned if those multiple correlations reached into the high .7s (R^2 of about .60). Many statistical software programs will allow us to compute multiple correlations for different combinations of variables so that we can examine them. Thus, we can scan these correlations for such high values and take the necessary steps to attempt to fix the problem.

Most regression software packages 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 (1999a) cautions that tolerances in the range of .40 are worthy of concern; other authors have suggested that tolerance values in the range of .1 are problematic (Myers, 1990; Pituch & Stevens, 2016).

A related statistic is the *variance inflation factor* (VIF), which is computed as 1 divided by tolerance. A VIF value of 2.50 is associated with a tolerance of .40 and is considered problematic by Allison (1999a); a VIF value of 10 is associated with a tolerance of .1 and is considered problematic by Cohen et al. (2003), Myers (1990), and Pituch and Stevens (2016).

5A.22 Recommended Readings

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

Berry, W. D. (1993). *Understanding regression assumptions* (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-92). Newbury Park, CA: SAGE.

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

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*, 161–182.

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

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (3rd ed.). Hoboken, NJ: Wiley & Sons.

Fox, J. (1991). *Regression diagnostics.* Newbury Park, CA: SAGE.

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

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

Lopez, R. P., & Guarino, A. J. (2011). Uncertainty and decision making for residents with dementia. *Clinical Nursing Research*, *20*, 228–240.

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

Montgomery, D. C., Peck, E. A., & Vining, G. G (2012). *Introduction to linear regression analysis* (5th ed.). Hoboken, NJ: Wiley & Sons.

Pardoe, I. (2012). *Applied regression modeling* (2nd ed.). Hoboken, NJ: Wiley & Sons.

Pituch, K. A., & Stevens, J. P. (2016). *Applied statistics for the social sciences* (6th ed.). New York, NY: Routledge.

Schafer, W. D. (1991). 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.* Beverly Hills, CA: SAGE.

Sherry, A., & Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment*, *84*, 37–48.

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

Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the *Journal of Counseling & Development*. *Journal of Counseling & Development*, *82*, 107–110.

Weisberg, S. (2013). *Applied linear regression analysis* (4th ed.). Hoboken, NJ: Wiley & Sons.

CHAPTER 5B

Multiple Regression Analysis Using IBM SPSS

This chapter will demonstrate how to perform multiple linear regression analysis with IBM SPSS first using the standard method and then using the stepwise method. We will use the data file **Personality** in these demonstrations.

5B.1 Standard Multiple Regression

5B.1.1 Analysis Setup: Main Regression Dialog Window

For purposes of illustrating standard linear regression, assume that we are interested in predicting self-esteem based on the combination of negative affect (experiencing negative emotions), positive affect (experiencing positive emotions), openness to experience (e.g., trying new foods, exploring new places), extraversion, neuroticism, and trait anxiety. Selecting the sequence **Analyze ➔ Regression ➔ Linear** opens the **Linear Regression** main dialog window displayed in Figure 5b.1. From the variables list panel, we move **esteem** to the **Dependent** panel and **negafect**, **posafect**, **neoopen**, **neoextra**, **neoneuro**, and **tanx** to the **Independent(s)** panel. The **Method** drop-down menu will be left at its default setting of **Enter**, which requests a standard regression analysis (all of the predictors are entered into the model in a single step).

5B.1.2 Analysis Setup: Statistics Window

Selecting the **Statistics** pushbutton opens the **Linear Regression: Statistics** dialog window shown in Figure 5b.2. By default, **Estimates** in the **Regression Coefficients** panel is checked. This instructs IBM SPSS to print the value of the regression coefficient and related measures. We also retained the following defaults: **Model fit**, which provides *R* square, adjusted *R* square, the standard error, and an ANOVA table; **R squared change**, which is useful when there are multiple predictors that are being entered in stages so that we can see where this information is placed in the output; **Descriptives**, which provides the means and standard deviations of the variables as well as the correlations table; and **Part and partial correlations**, which produces the partial and semipartial correlations when multiple predictors are used. Clicking **Continue** returns us to the main dialog screen.

Figure 5b.1 Main Dialog Window for Linear Regression

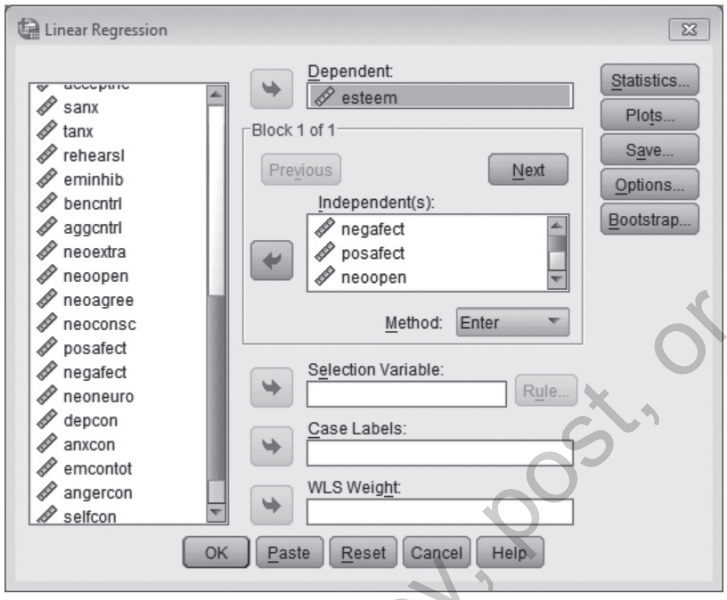


Figure 5b.2 The Linear Regression Statistics Window

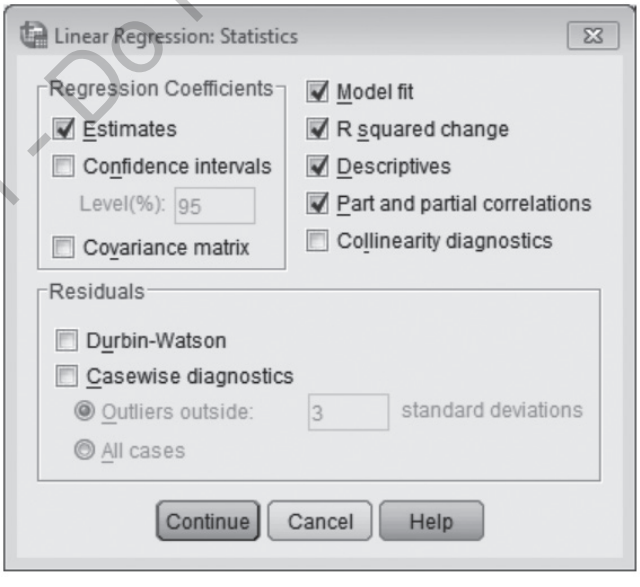
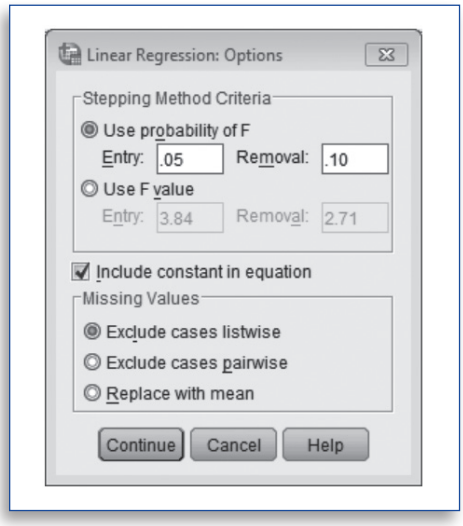


Figure 5b.3 The Linear Regression Options Window



5B.1.3 Analysis Setup: Options Window

Select the **Options** pushbutton; this displays the **Linear Regression: Options** dialog window shown in Figure 5b.3. The top panel is applicable if we were using one of the step methods, and we will discuss this in Section 5B.2. We have retained the defaults of including the *Y* intercept (the constant) in the equation and of excluding cases listwise. The choice **Exclude cases listwise** (sometimes called listwise deletion) means that all cases must have valid values on all of the variables in the analysis in order to be included; a missing value on even one of the variables is sufficient to exclude that case. Selecting this choice ensures that the set of variables, and thus the regression model, is based on the same set of cases. So long as there is relatively little missing data, this choice is best. Clicking **Continue** returns us to the main dialog box, and selecting **OK** produces the analysis.

5B.1.4 Analysis Output: Descriptives and Correlations

We examine the output of the analysis in the order we suggest that you proceed. Figure 5b.4 contains descriptive information. The upper table contains the means and standard deviations of the variables, and the lower table shows the square correlation matrix. The correlation results are divided into three major rows: the first contains the Pearson *r* values, the second contains the probabilities of obtaining those values if the null hypothesis was true, and the third provides sample size.

The dependent variable **esteem** is placed by IBM SPSS on the first row and column of the correlation table, and the other variables appear in the order we entered them into the analysis. The study represented by our data set was designed for a somewhat different purpose, so our choice of variables was a bit limited. Thus, the correlations of self-esteem with the predictor variables in the analysis are higher than we would ordinarily prefer, and many of the other variables are themselves likewise intercorrelated more than we would typically find in most studies. Nonetheless, the example is still useful for our purposes.

5B.1.5 Analysis Output: Omnibus Analysis

Figure 5b.5 displays the results of the analysis. The middle table shows the test of significance of the model using an ANOVA. There are **419** ($N - 1$) total degrees of freedom. With six predictors, the **Regression** effect has 6 degrees of freedom. The **Regression** effect is statistically significant, indicating that the combination of predictors explains more of the variance of the dependent variable than can be done by chance alone.

The upper table in Figure 5b.5 labeled **Model Summary** provides an overview of the results. Of primary interest are the **R Square** and **Adjusted R Square** values, which are **.607** and **.601**, respectively. We learn from these that the weighted combination of the predictor variables explained approximately 60% of the variance of self-esteem. The loss of so little strength in computing the **Adjusted R Square** value is primarily due to our relatively large sample size combined with a relatively small set of

Figure 5b.4 Descriptive Statistics and Correlations Output for Standard Regression

Descriptive Statistics

	Mean	Std. Deviation	N
esteem self-esteem: coopersmith	71.2952	20.95292	420
negafect negative affect: mpq	5.7786	3.73029	420
posafect positive affect: mpq	7.7048	2.90970	420
neoopen openness: neo	55.4607	10.90952	420
neoextra extraversion: neo	55.8324	11.28265	420
neoneuro neuroticism: neo	50.5394	11.14793	420
tanx trait anxiety: spielberger	38.3262	10.59431	420

Correlations

		esteem self-esteem: coopersmith	negafect negative affect: mpq	posafect positive affect: mpq	neoopen openness: neo	neoextra extraversion: neo	neoneuro neuroticism: neo	tanx trait anxiety: spielberger
Pearson Correlation	esteem self-esteem: coopersmith	1.000	-.572	.555	.221	.425	-.693	-.724
	negafect negative affect: mpq	-.572	1.000	-.324	-.168	-.218	.712	.713
	posafect positive affect: mpq	.555	-.324	1.000	.221	.528	-.441	-.528
	neoopen openness: neo	.221	-.168	.221	1.000	.051	-.227	-.183
	neoextra extraversion: neo	.425	-.218	.528	.051	1.000	-.347	-.375
	neoneuro neuroticism: neo	-.693	.712	-.441	-.227	-.347	1.000	.809
	tanx trait anxiety: spielberger	-.724	.713	-.528	-.183	-.375	.809	1.000
Sig. (1-tailed)	esteem self-esteem: coopersmith	.	.000	.000	.000	.000	.000	.000
	negafect negative affect: mpq	.000	.	.000	.000	.000	.000	.000
	posafect positive affect: mpq	.000	.000	.	.000	.000	.000	.000
	neoopen openness: neo	.000	.000	.000	.	.148	.000	.000
	neoextra extraversion: neo	.000	.000	.000	.148	.	.000	.000
	neoneuro neuroticism: neo	.000	.000	.000	.000	.000	.	.000
	tanx trait anxiety: spielberger	.000	.000	.000	.000	.000	.000	.
N	esteem self-esteem: coopersmith	420	420	420	420	420	420	420
	negafect negative affect: mpq	420	420	420	420	420	420	420
	posafect positive affect: mpq	420	420	420	420	420	420	420
	neoopen openness: neo	420	420	420	420	420	420	420
	neoextra extraversion: neo	420	420	420	420	420	420	420
	neoneuro neuroticism: neo	420	420	420	420	420	420	420
	tanx trait anxiety: spielberger	420	420	420	420	420	420	420

predictors. Using the standard regression procedure where all of the predictors were entered simultaneously into the model, **R Square Change** went from zero before the model was fitted to the data to **.607** when all of the variables were simultaneously entered.

5B.1.6 Analysis Output: Individual Predictor Results

The bottom table in Figure 5b.5 labeled **Coefficients** provides the details of the results. The **Zero-order** column under **Correlations** lists the Pearson *r* for the dependent variable (self-esteem in this case) with each of the predictors. These values are the same as those shown in the correlation matrix of Figure 5b.4.

Figure 5b.5 The Results of the Standard Regression Analysis

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.779[a]	.607	.601	13.22887	.607	106.356	6	413	.000

a. Predictors: (Constant), tanx trait anxiety: spielberger, neoopen openness: neo, neoextra extraversion: neo, posafect positive affect: mpq, negafect negative affect: mpq, neoneuro neuroticism: neo

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	111675.183	6	18612.531	106.356	.000[a]
	Residual	72276.207	413	175.003		
	Total	183951.390	419			

a. Predictors: (Constant), tanx trait anxiety: spielberger, neoopen openness: neo, neoextra extraversion: neo, posafect positive affect: mpq, negafect negative affect: mpq, neoneuro neuroticism: neo

b. Dependent Variable: esteem self-esteem: coopersmith

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	96.885	7.294		13.283	.000			
	negafect negative affect: mpq	-.386	.264	-.069	-1.462	.144	-.572	-.072	-.045
	posafect positive affect: mpq	1.338	.293	.186	4.564	.000	.555	.219	.141
	neoopen openness: neo	.088	.062	.046	1.420	.156	.221	.070	.044
	neoextra extraversion: neo	.185	.069	.099	2.684	.008	.425	.131	.083
	neoneuro neuroticism: neo	-.477	.106	-.254	-4.505	.000	-.693	-.216	-.139
	tanx trait anxiety: spielberger	-.646	.117	-.326	-5.511	.000	-.724	-.262	-.170

a. Dependent Variable: esteem self-esteem: coopersmith

The **Partial** column under **Correlations** lists the partial correlations for each predictor as it was evaluated for its weighting in the model (the correlation between the predictor and the dependent variable when the other predictors are treated as covariates). For example, the partial correlation associated with **posafect** is **.219**. This represents the correlation between **posafect** and the residual variance of the self-esteem dependent variable when statistically controlling for the other predictor variables.

The **Part** column under **Correlations** lists the semipartial correlations for each predictor once the model is finalized. Squaring these values informs us of the percentage of variance of self-esteem that each predictor uniquely explains. For example, trait anxiety accounts uniquely for about 3% of the variance of self-esteem (−.170 * −.170 = .0289 or approximately .03) given the contributions of the other variables in the model.

The *Y* intercept of the raw score model is labeled as the **Constant** and has a value here of **96.885**. Of primary interest here are the unstandardized or raw (**B**) and standardized (**Beta**) coefficients, and their significance levels determined by *t* tests. With the exception of negative affect and openness, all of the predictors are statistically significant. As can be seen by examining the beta weights, trait anxiety followed by neuroticism followed by positive affect were all making relatively larger contributions to the prediction model.

The regression coefficients are *partial regression coefficients* because their values take into account the other predictor variables in the model; they inform us of the predicted change in the dependent variable for every unit increase in that predictor. For example, positive affect is associated with an unstandardized partial regression coefficient of **1.338** and signifies that, when controlling for the other predictors, for every additional point on the positive affect measure, we would predict a gain of **1.338** points on the self-esteem measure. As another example, neuroticism is associated with a partial regression coefficient of **–.477** and signifies that, when controlling for the other predictors, every additional point on the neuroticism measure corresponds to a decrement of **.477** points on the self-esteem measure.

This example serves to illustrate two important related points about multiple regression analysis. First, it is the model as a whole that is the focus of the analysis. Variables are treated akin to team players weighted in such a way that the sum of the squared residuals of the model is minimized. Thus, it is the set of variables in this particular (weighted) configuration that maximizes prediction—swap out one of these predictors for a new variable and the whole configuration that represents the best prediction can be quite different.

The second important point about regression analysis that this example illustrates, which is related to the first, is that a highly predictive variable can be "left out in the cold," being "sacrificed" for the "good of the model." Note that negative affect in isolation correlates rather substantially with self-esteem ($r = -.572$), and if it was the only predictor it would have a beta weight of –.572 (recall that in simple linear regression, the Pearson *r* is the beta weight of the predictor), yet in combination with the other predictors is not a significant predictor in the multiple regression model. The reason for it not being weighted substantially in the model is that one or more of the other variables in the analysis are accomplishing its predictive work. But the point is that just because a variable receives a modest weight in the model or just because a variable is not contributing a statistically significant degree of prediction in the model is not a reason to presume that it is itself a poor predictor.

It is also important to note that the IBM SPSS output does not contain the structure coefficients. These are the correlations of the predictors in the model with the overall predictor variate, and these structure coefficients help researchers interpret the dimension underlying the predictor model (see Section 5A.11). They are easy enough to calculate by hand, and we incorporate these structure coefficients into our report of the results in Section 5B.1.7. Structure coefficients are computed by dividing the Pearson correlation for the given variable by the value of the multiple correlation coefficient associated with the model (r/R). For example, the structure coefficient for **negafect** would be **–.572/.779** or –.734. This represents the correlation of each predictor with the predicted value of the dependent variable.

5B.1.7 Reporting Standard Multiple Regression Results

> Negative affect, positive affect, openness to experience, extraversion, neuroticism, and trait anxiety were used in a standard regression analysis to predict self-esteem. The correlations of the variables are shown in Table 5b.1. As can be seen, all correlations, except for the one between openness and extraversion, were statistically significant.
>
> *(Continued)*

(Continued)

The prediction model was statistically significant, $F(6, 413) = 106.356$, $p < .001$, and accounted for approximately 60% of the variance of self-esteem ($R^2 = .607$, adjusted $R^2 = .601$). Lower levels of trait anxiety and neuroticism, and to a lesser extent higher levels of positive affect and extraversion, primarily predicted self-esteem. The raw and standardized regression coefficients of the predictors together with their correlations with self-esteem, the squared semipartial correlations, and the structure coefficients are shown in Table 5b.2. Trait anxiety received the strongest weight in the model, followed by neuroticism and positive affect. With the sizeable correlations between the predictors, the unique variance explained by each of the variables indexed by the squared semipartial correlations was quite low. Inspection of the structure coefficients suggests that, with the possible exception of extraversion (whose correlation is still relatively substantial), the other significant predictors were strong indicators of the underlying (latent) variable described by the model, which can be interpreted as well-being.

Table 5b.1 Correlations of the Variables in the Analysis ($N = 420$)

Variable	2	3	4	5	6	7
1. Self-Esteem	−.572	.555	.221	.425	−.693	−.724
2. Negative Affect	--	−.324	−.168	−.218	.712	.713
3. Positive Affect		--	.221	.528	−.441	−.528
4. Openness			--	.051	−.227	−.183
5. Extraversion				--	−.347	−.375
6. Neuroticism					--	.809
7. Trait Anxiety						--

Note. All correlations except for the correlation between Openness and Extraversion were statistically significant ($p < .001$).

Table 5b.2 Standard Regression Results

Model	b	SE-b	Beta	Pearson *r*	sr^2	Structure Coefficient
Constant	96.885	7.294				
Negative Affect	−.386	.264	−.069	−.572	.002	−.734
Positive Affect*	1.338	.293	.186	.555	.020	.712
Openness	.088	.062	.046	.221	.002	.284
Extraversion*	.185	.069	.099	.425	.007	.546
Neuroticism*	−.477	.106	−.254	−.693	.019	−.890
Trait Anxiety*	−.646	.117	−.326	−.724	.029	−.929

Note. The dependent variable was Self-Esteem. $R^2 = .607$, Adjusted $R^2 = .601$.
sr^2 is the squared semi-partial correlation.
* $p < .05$.

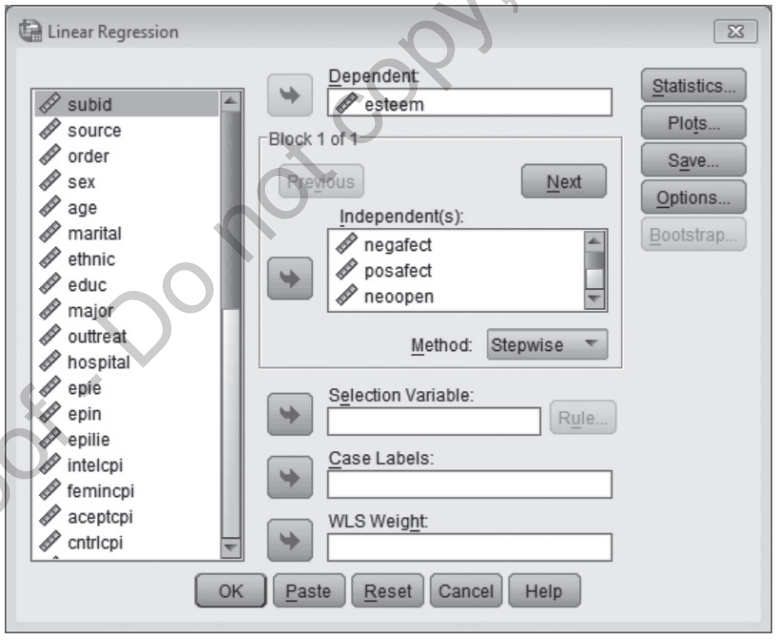
5B.2 Stepwise Multiple Regression

We discussed the forward, backward, and stepwise methods of performing a regression analysis in Chapter 5A. To illustrate how to work with these methods, we will perform a stepwise analysis on the same set of variables that we used in our standard regression analysis in Section 5B.1. We will use the data file **Personality** in this demonstration. In the process of our description, we will point out areas of similarity and difference between the standard and step methods.

5B.2.1 Analysis Setup: Main Regression Dialog Window

Select **Analyze ➔ Regression ➔ Linear**. This brings us to the **Linear Regression** main dialog window displayed in Figure 5b.6. From the variables list panel, we move **esteem** to the **Dependent** panel and **negafect**, **posafect**, **neoopen**, **neoextra**, **neoneuro**, and **tanx** to the **Independent(s)** panel. The **Method** drop-down menu contains the set of step methods that IBM SPSS can run. The only one you may not recognize is **Remove**, which allows a set of variables to be removed from the model together. Choose **Stepwise** as the **Method** from the drop-down menu as shown in Figure 5b.6.

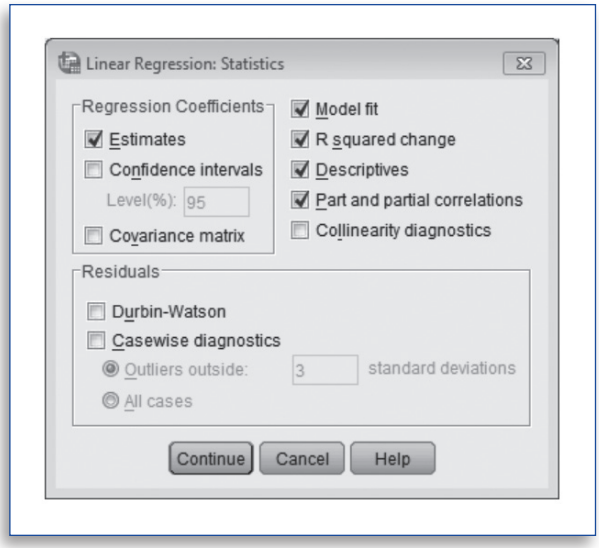
Figure 5b.6 Main Dialog Window for Linear Regression



5B.2.2 Analysis Setup: Statistics Window

Selecting the **Statistics** pushbutton brings us to the **Linear Regression: Statistics** dialog window shown in Figure 5b.7. We configure it in the same way as was discussed in Section 5B.1.2. Clicking **Continue** returns us to the main dialog box.

Figure 5b.7 The Linear Regression Statistics Window

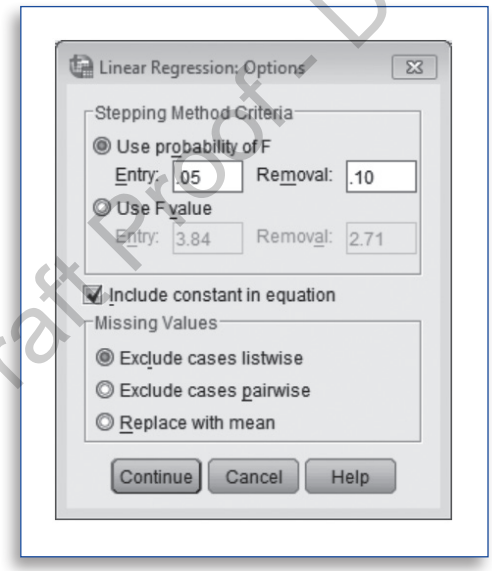


5B.2.3 Analysis Setup: Options Window

Selecting the **Options** pushbutton brings us to the **Linear Regression: Options** dialog window shown in Figure 5b.8. The top panel is now applicable as we are using the stepwise method. To avoid looping variables continually in and out of the model, it is appropriate to set different probability levels for **Entry** and **Removal**. The defaults used by IBM SPSS that are shown in Figure 5b.8 are common settings, and we recommend them. Remember that in the stepwise procedure, variables already entered into the model can be removed at a later step if they are no longer contributing a statistically significant amount of prediction.

Earning entry to the model is set at an alpha level of **.05** (e.g., a variable with a probability of .07 will not be entered) and is the more stringent of the two settings. But to be removed, a variable must have an associated probability of greater than **.10** (e.g., a variable with an associated probability of .12 will be removed but one with an associated probability of .07 will remain in the model). In essence, it is more difficult to get in than be removed. This is a good thing and allows the stepwise procedure to function. Click **Continue** to return to the main dialog window, and click **OK** to perform the analysis.

Figure 5b.8 The Linear Regression Options Window



5B.2.4 Analysis Output: Descriptives and Correlations

The descriptive statistics are identical to those presented in Section 5B.1.4, and readers are referred to the previous analysis for that output.

5B.2.5 Analysis Output: Omnibus Analysis

Figure 5b.9 displays the test of significance of the model using an ANOVA. The four ANOVAs that are reported correspond to four models, but don't let the terminology confuse you. The stepwise procedure adds only one variable at a time to the model as the model is "slowly" built. At the third step and beyond, it is also possible to remove a variable from the model (although that did not happen in our example). In the terminology used by IBM SPSS, each step results in a model, and each successive step modifies the older model and replaces it with a newer one. Each model is tested for statistical significance.

Figure 5b.9 Tests of Significance for Each Step in the Regression Analysis

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	96502.996	1	96502.996	461.281	.000[a]
	Residual	87448.394	418	209.207		
	Total	183951.390	419			
2	Regression	104086.724	2	52043.362	271.736	.000[b]
	Residual	79864.666	417	191.522		
	Total	183951.390	419			
3	Regression	109881.931	3	36627.310	205.712	.000[c]
	Residual	74069.460	416	178.052		
	Total	183951.390	419			
4	Regression	110934.239	4	27733.560	157.626	.000[d]
	Residual	73017.152	415	175.945		
	Total	183951.390	419			

a. Predictors: (Constant), tanx trait anxiety: spielberger

b. Predictors: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq

c. Predictors: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq, neoneuro neuroticism: neo

d. Predictors: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq, neoneuro neuroticism: neo, neoextra extraversion: neo

e. Dependent Variable: esteem self-esteem: coopersmith

Examining the output shown in Figure 5b.9 informs us that the final model was built in four steps; each step resulted in a statistically significant model. Examining the **df** column shows us that one variable was added during each step (the degrees of freedom for the **Regression** effect track this for us as they are counts of the number of predictors in the model). We can also deduce that no variables were removed from the model since the count of predictors in the model steadily increases from 1 to 4.

This deduction that no variables were removed is verified by the display shown in Figure 5b.10, which tracks variables that have been entered and removed at each step. As can be seen, trait anxiety, positive affect, neuroticism, and extraversion have been entered on Steps 1 through 4, respectively, without any variables having been removed on any step.

Figure 5b.10 Variables That Were Entered and Removed

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	tanx trait anxiety: spielberger	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	posafect positive affect: mpq	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	neoneuro neuroticism: neo	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	neoextra extraversion: neo	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: esteem self-esteem: coopersmith

Figure 5b.11 Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.724[a]	.525	.523	14.46398	.525	461.281	1	418	.000
2	.752[b]	.566	.564	13.83915	.041	39.597	1	417	.000
3	.773[c]	.597	.594	13.34360	.032	32.548	1	416	.000
4	.777[d]	.603	.599	13.26442	.006	5.981	1	415	.015

a. Predictors: (Constant), tanx trait anxiety: spielberger

b. Predictors: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq

c. Predictors: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq, neoneuro neuroticism: neo

d. Predictors: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq, neoneuro neuroticism: neo, neoextra extraversion: neo

Figure 5b.11, the **Model Summary**, presents the **R Square** and **Adjusted R Square** values for each step along with the amount of **R Square Change**. In the first step, as can be seen from the footnote beneath the **Model Summary** table, trait anxiety was entered into the model. The **R Square** with that predictor in the model was **.525**. Not coincidentally, that is the square of the correlation between trait anxiety and self-esteem ($-.724^2 = .525$) and is the value of **R Square Change**.

On the second step, positive affect was added to the model. The **R Square** with both predictors in the model was **.566**; thus, we gained .041 in the value of **R Square** (.566 − .525 = .041), and this is reflected in the **R Square Change** for that step. By the time we arrive at the end of the fourth step, our **R Square** value has reached **.603**. Note that this value is very close to but not identical to the R^2 value we obtained under the standard method (with the other statistically nonsignificant variables included in the model).

5B.2.6 Analysis Output: Individual Predictor Results

The **Coefficients** table in Figure 5b.12 provides the details of the results. Note that both the unstandardized and standardized regression coefficients are readjusted at each step to reflect the additional variables in the model. Ordinarily, although it is interesting to observe the dynamic changes taking place, we are usually interested in the final model. Note also that the values of the regression coefficients are different from those associated with the same variables in the standard regression analysis. That the differences are not huge is due to the fact that these four variables did almost the same amount of predictive work in much the same configuration as did the six predictors using the standard method. If economy of model were relevant, we would probably be very happy with the trimmed model of four variables replacing the full model containing six variables.

Figure 5b.13 addresses the fate of the remaining variables. For each step, IBM SPSS tells us which variables were not entered. In addition to tests of the statistical significance of each variable, we also see displayed the partial correlations. This information together tells us what will happen in the following step. For example, consider Step 1, which contains the five excluded variables. Positive affect has the highest partial correlation (**.294**), and it is statistically significant; thus, it will be the variable next entered on Step 2. On the second step, with four variables (of the six) being considered for inclusion, we see that neuroticism with a statistically significant partial correlation of **−.269** wins the struggle for entry next. By the time we reach the fourth step, there is no variable of the excluded set that has a statistically significant partial correlation for entry at Step 5; thus, the stepwise procedure ends after completing the fourth step.

Figure 5b.12 The Results of the Stepwise Regression Analysis

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	126.197	2.652		47.588	.000			
	tanx trait anxiety: spielberger	−1.432	.067	−.724	−21.477	.000	−.724	−.724	−.724
2	(Constant)	103.366	4.427		23.347	.000			
	tanx trait anxiety: spielberger	−1.183	.075	−.598	−15.742	.000	−.724	−.611	−.508
	posafect positive affect: mpq	1.722	.274	.239	6.293	.000	.555	.294	.203
3	(Constant)	114.095	4.665		24.459	.000			
	tanx trait anxiety: spielberger	−.706	.111	−.357	−6.386	.000	−.724	−.299	−.199
	posafect positive affect: mpq	1.679	.264	.233	6.364	.000	.555	.298	.198
	neoneuro neuroticism: neo	−.567	.099	−.302	−5.705	.000	−.693	−.269	−.177
4	(Constant)	105.775	5.751		18.392	.000			
	tanx trait anxiety: spielberger	−.698	.110	−.353	−6.345	.000	−.724	−.297	−.196
	posafect positive affect: mpq	1.384	.289	.192	4.793	.000	.555	.229	.148
	neoneuro neuroticism: neo	−.549	.099	−.292	−5.537	.000	−.693	−.262	−.171
	neoextra extraversion: neo	.167	.068	.090	2.446	.015	.425	.119	.076

a. Dependent Variable: esteem self-esteem: coopersmith

Figure 5b.13 The Results of the Stepwise Regression Analysis

Excluded Variables[e]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	negafect negative affect: mpq	−.112[a]	−2.350	.019	−.114	.492
	posafect positive affect: mpq	.239[a]	6.293	.000	.294	.721
	neoopen openness: neo	.091[a]	2.680	.008	.130	.967
	neoextra extraversion: neo	.179[a]	5.058	.000	.240	.859
	neoneuro neuroticism: neo	−.311[a]	−5.625	.000	−.266	.346
2	negafect negative affect: mpq	−.139[b]	−3.034	.003	−.147	.488
	neoopen openness: neo	.062[b]	1.872	.062	.091	.945
	neoextra extraversion: neo	.106[b]	2.777	.006	.135	.709
	neoneuro neuroticism: neo	−.302[b]	−5.705	.000	−.269	.346
3	negafect negative affect: mpq	−.061[c]	−1.296	.196	−.063	.434
	neoopen openness: neo	.038[c]	1.180	.239	.058	.928
	neoextra extraversion: neo	.090[c]	2.446	.015	.119	.704
4	negafect negative affect: mpq	−.070[d]	−1.487	.138	−.073	.432
	neoopen openness: neo	.047[d]	1.446	.149	.071	.918

a. Predictors in the Model: (Constant), tanx trait anxiety: spielberger

b. Predictors in the Model: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq

c. Predictors in the Model: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq, neoneuro neuroticism: neo

d. Predictors in the Model: (Constant), tanx trait anxiety: spielberger, posafect positive affect: mpq, neoneuro neuroticism: neo, neoextra extraversion: neo

e. Dependent Variable: esteem self-esteem: coopersmith

5B.2.7 Reporting the Stepwise Multiple Regression Results

Negative affect, positive affect, openness to experience, extraversion, neuroticism, and trait anxiety were used in a stepwise multiple regression analysis to predict self-esteem. The correlations of the variables are shown in Table 5b.1. As can be seen, all correlations except for the one between openness and extraversion were statistically significant.

A stepwise multiple regression procedure was performed to generate a parsimonious prediction model. The final model contained four of the six predictors and was reached in four steps with no variables removed. The model was statistically significant, $F(4, 415) = 157.626$, $p < .001$, and accounted for approximately 60% of the variance of self-esteem ($R^2 = .603$, adjusted $R^2 = .599$). Lower levels of trait anxiety and neuroticism, and to a lesser extent higher levels of positive affect and extraversion, predicted self-esteem. The raw and standardized regression coefficients of the predictors together with their correlations with self-esteem, their squared semipartial correlations, and their structure coefficients are shown in Table 5b.3. Trait anxiety received the strongest weight in the model, followed by neuroticism and positive affect; extraversion received the lowest of the four weights. With the sizeable correlations between the predictors, the unique variance explained by each of the variables indexed by the squared semipartial correlations was relatively low: trait anxiety, positive affect, neuroticism, and extraversion uniquely accounted for approximately 4%, 2%, 3%, and less than 1% of the variance of self-esteem. The latent factor represented by the model appears to be interpretable as well-being. Inspection of the structure coefficients suggests that trait anxiety and neuroticism were very strong indicators of well-being, positive affect was a relatively strong indicator of well-being, and extraversion was a moderate indicator of well-being.

Table 5b.3 Stepwise Regression Results

Model	b	SE-b	Beta	Pearson *r*	sr^2	Structure Coefficient
Constant	105.775	5.751				
Trait Anxiety	−.698	.110	−.353	−.724	.038	−.932
Positive Affect*	1.384	.289	.192	.555	.022	.714
Neuroticism*	−.549	.099	−.292	−.693	.029	−.892
Extraversion*	.167	.068	.090	.425	.006	.547

Note. The dependent variable was Self-Esteem. $R^2 = .603$, Adjusted $R^2 = .599$.
sr^2 is the squared semi-partial correlation.
* $p < .05$.