165 - 194

# CHAPTER 7

# MULTIPLE REGRESSION

Up to this point, we have focused our attention on statistical analysis techniques that investigate the existence of differences between groups. In this chapter, we begin to redirect the focus of our attention to a second grouping of advanced/multivariate techniques—those that describe and test the existence of predictable relationships among a set of variables. Our discussion will include a brief review of simple linear regression, followed by an in-depth examination of multiple regression.

## *SECTION 7.1 PRACTICAL VIEW*

### Purpose

Regression analysis procedures have as their primary purpose the development of an equation that can be used for *predicting* values on some DV for all members of a population. (A secondary purpose is to use regression analysis as a means of *explaining* causal relationships among variables, the focus of Chapter 8.) The most basic application of regression analysis is the bivariate situation, to which the reader was undoubtedly exposed in an introductory statistics course. This is often referred to as *simple linear regression*, or just *simple regression*. Simple regression involves a single IV and a single DV. The goal of simple regression is to obtain a linear equation so that we can predict the value of the DV if we have the value of the IV. Simple regression capitalizes on the correlation between the DV and IV in order to make specific predictions about the DV (Sprinthall, 2000). The correlation tells us how much information about the DV is contained in the IV. If the correlation is perfect (i.e., $r = \pm 1.00$), the IV contains everything we need to know about the DV, and we will be able to perfectly predict one from the other. However, this is seldom, if ever, the case.

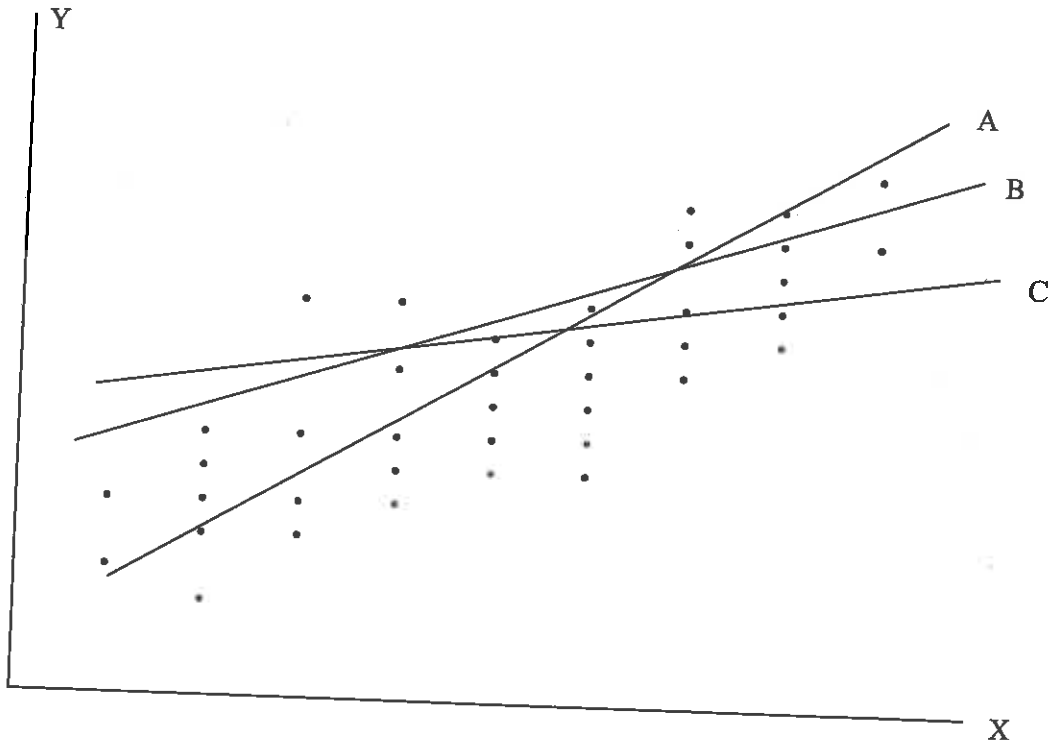
The idea behind simple regression is that we want to obtain the equation for the best-fitting line through a series of points. If we were to view a bivariate scatterplot for our fictitious IV (X) and DV (Y), we could then envision a line drawn through those points. Theoretically, an infinite number of lines could be drawn through those points (see Figure 7.1). However, only one of these lines would be the best-fitting line. Regression analysis is the means by which we determine the best-fitting line, called the ***regression line***.

The regression line is the single straight line that lies closest to all points in a given scatterplot—this line is sometimes said to pass through the *centroid* of the scatterplot (Sprinthall, 2000). In order to make predictions, three important facts about the regression line must be known:

(1) the extent to which points are scattered around the line,
(2) the slope of the regression line, and
(3) the point at which the line crosses the *Y*-axis (Sprinthall, 2000).

These three facts are so important to regression that they serve as the basis for the calculation of the regression equation itself. The extent to which the points are scattered around the line is typically indicated by the degree of relationship between the IV (*X*) and the DV (*Y*). This relationship is measured by a correlation coefficient (e.g., the Pearson correlation, symbolized by *r*)—the stronger the relationship, the higher the degree of predictability between *X* and *Y*. (You will see in Section 7.3 just how important *r* is to the regression equation calculation.) The slope of the regression line can greatly affect prediction (Sprinthall, 2000). The degree of slope is determined by the amount of change in *Y* that accompanies a unit change (i.e., one point, one inch, one degree, etc.) in *X*. It is the slope that largely determines the predicted values of *Y* from known values for *X*. Finally, it is important to determine exactly where the regression line crosses the *Y*-axis (this value is known as the *Y*-intercept). Said another way, it is crucial to know what value is expected for *Y* when $X = 0$.

**Figure 7.1** Bivariate Scatterplot Showing Several Possible Regression Lines.




These three facts we have just discussed actually define the regression line. The regression line is essentially an equation that expresses *Y* as a function of *X* (Tate, 1992). The basic equation for simple regression is:

$$Y = bX + a \qquad \text{(Equation 7.1)}$$

where *Y* is the predicted value for the DV, *X* is the known raw score value on the IV, *b* is the slope of the regression line, and *a* is the *Y*-intercept. The significant role that both the slope and *Y*-intercept play in the regression equation should now be apparent to the reader. Oftentimes, you will see the above equation presented in the following analogous, although more precise, form:

$$\hat{Y} = B_0 + B_1 X_1 + \hat{e}_i \qquad \text{(Equation 7.2)}$$

where $\hat{Y}$ is the predicted value for the DV, $X$ is the raw score value on the IV, $B_1$ is the slope of the regression line, and $B_0$ is the $Y$-intercept. We have added one important term, $\hat{e}_i$, in Equation 7.2. This is the symbol for the errors of prediction, also referred to as the ***residuals***. As previously mentioned, unless we have a perfect correlation between the IV and DV, the predicted values obtained by our regression equation will also be less than perfect—that is, there will be some errors. The residuals constitute an important measure of those errors and are essentially calculated as the difference between the actual value and predicted value for the DV (i.e., $\hat{e}_i = y_i - \hat{y}$).

Let us return momentarily to the concept of the best-fitting line (see Figure 7.2). The reason that we obtain the best-fitting line as our regression equation is because we mathematically calculate the line with the smallest amount of total squared error. This is commonly referred to as the ***least squares solution*** (Stevens, 1992; Tate, 1992) and actually provides us with values for the constants in the regression equation, $B_1$ and $B_0$ (also known as the ***regression coefficients*** ***(B)***, ***beta coefficients*** or ***beta weights*** ***($\beta$)***) that minimize the sum of squared residuals—that is, $\Sigma(y_i - \hat{y})^2$ is minimized. In other words, the *total* amount of prediction error, both positive and negative, is as small as possible, giving us the best mathematically achievable line through the set of points in a scatterplot.

*Multiple regression* is merely an extension of simple linear regression involving more than one IV, or predictor variable. This technique is used to predict the value of a single DV from a weighted, linear combination of IVs (Harris, 1998). A multiple regression equation appears similar to its simple regression counterpart except that there are more coefficients, one for the $Y$-intercept and one for each of the IVs:

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + \ldots + B_kX_k + \hat{e}_i \qquad \text{(Equation 7.3)}$$

where there is a corresponding $B$ coefficient for each IV ($X_k$) in the equation and the best linear combination of weights and raw score $X$ values will again minimize the total squared error in our regression equation.

Let us consider a concrete example: Suppose we wanted to determine the extent to which we could predict female life expectancy from a set of predictor variables for a selected group of countries throughout the world. The predictor variables we have selected include percent urban population; gross domestic product per capita; birthrate per 1,000; number of hospital beds per 10,000; number of doctors per 10,000; number of radios per 100; and number of telephones per 100. In our analysis, we would be looking to obtain the regression coefficients for each IV that would provide us with the best linear combination of IVs—and their associated weights—in order to predict, as accurately as possible, female life expectancy. The regression equation predicting female life expectancy is as follows:

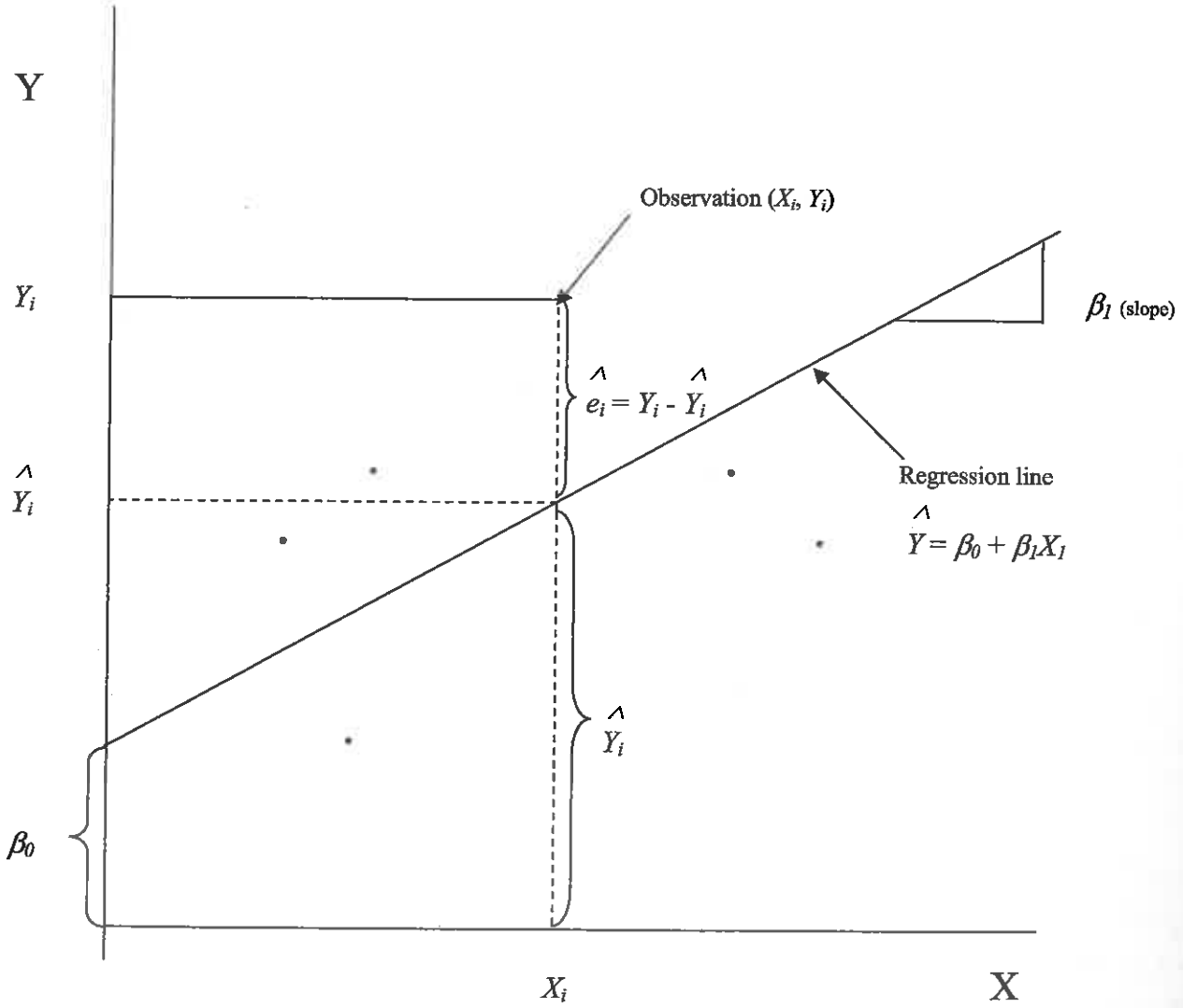
$$\textit{Female life exp.} = B_{urban}X_{urban} + B_{GDP}X_{GDP} + B_{birthrate}X_{birthrate} + B_{beds}X_{beds} + B_{docs}X_{docs} + B_{radios}X_{radios} + B_{phones}X_{phones} + \hat{e}_i$$

We will return to this example in greater detail a bit later in the chapter, but first there are several important issues related to multiple regression that warrant our attention.

A first issue of interest is a set of measures unique to multiple regression. Another way of looking at the previously mentioned concept of the minimization of total error is to consider multiple regression as a means of seeking the linear combination of IVs that *maximally* correlate with the DV (Stevens, 1992). This maximized correlation is called the ***multiple correlation*** and is symbolized by $R$. The multiple correlation is essentially equivalent to the Pearson correlation between the actual, or observed, val-

ues and the predicted values on the DV (i.e., $R = r_{y\hat{y}i}$). Analogous to our earlier interpretation of the Pearson correlation, the multiple correlation tells us how much information about a DV (e.g., female life expectancy) is contained in the combination of IVs (e.g., percent urban population, gross domestic product, birthrate, number of hospital beds, number of doctors, number of radios, and number of telephones). In multiple regression, there is a test of significance ($F$-test) to determine whether the relationship between the set of IVs and the DV is large enough to be meaningful.

**Figure 7.2** Graphical Representation of a Linear Regression Model and the Least Squares Criterion.




You may recall from an earlier course in statistics a term called the ***coefficient of determination***, or $r^2$. For the Pearson $r$, this value was interpreted as the proportion of one variable in the pair that can be explained (or accounted for) by the other variable. In multiple regression, $R^2$ is also called the coefficient of determination and has a similar interpretation. The coefficient of determination is the propor-

tion of DV variance that can be explained by the combination of the IVs (Levin & Fox, 2000; Sprinthall, 2000). In our example, an obtained value for $R^2$ would be interpreted as the proportion of variability in female life expectancy that could be accounted for by the combination of the seven predictor variables. If one multiplies this value by 100, $R^2$ becomes the percentage of explained variance (Sprinthall, 2000).

A second issue is one that has an associated word of caution, which we will address momentarily. The issue at hand is that of ***multicollinearity***. Multicollinearity is a problem that arises when there exists moderate to high intercorrelations among predictor variables (IVs) to be used in a regression analysis. (Recall from Chapter 1 that the opposite of multicollinearity is *orthogonality*, or complete independence among variables.) The underlying problem of multicollinearity is that if two variables are highly correlated, they are essentially containing the same—or at least much of the same—information and are therefore measuring the same thing (Sprinthall, 2000). Not only does one gain little by adding to a regression analysis variables that are measuring the same thing, but multicollinearity can cause real problems for the analysis itself. Stevens (1992) points out three reasons why multicollinearity can be problematic for researchers.

(1) Multicollinearity severely limits the size of $R$ since the IVs are "going after" much of the same variability on the DV.
(2) When trying to determine the importance of individual IVs, multicollinearity causes difficulty because individual effects are confounded due to the overlapping information.
(3) Multicollinearity tends to increase the variances of the regression coefficients, which ultimately results in a more unstable prediction equation.

Multicollinearity should be addressed by the researcher prior to the execution of the regression analysis. The simplest method for diagnosing multicollinearity is to examine the correlation matrix for the predictor variables, looking for moderate to high intercorrelations. However, it is preferable to use one of two statistical methods to assess multicollinearity. First, tolerance statistics can be obtained for each IV. ***Tolerance*** is a measure of collinearity among IVs, where possible values range from 0 to 1. A value for tolerance close to zero is an indication of multicollinearity. Typically, a value of 0.1 serves as the cutoff point—if the tolerance value for a given IV is less than 0.1, multicollinearity is a distinct problem (Norusis, 1998). A second method is to examine values for the *variance inflation factor* for each predictor. The variance inflation factor (VIF) for a given predictor "indicates whether there exists a strong linear association between it and all remaining predictors" (Stevens, 1992). The VIF is defined by the quantity $1/(1-R_j^2)$ and is obtainable on most computer programs. Although there is no steadfast rule of thumb, values of VIF that are greater than 10 are generally cause for concern (Stevens, 1992).

There are several methods for combating multicollinearity in a regression analysis; two of the most straightforward are presented here. The simplest method is to delete the problematic variable from the analysis (Sprinthall, 2000). If the information in one variable is being "captured" by another, no real information is being lost by deleting one of them. A second approach is to combine the variables involved so as to create a single measure that addresses a single construct, thus deleting the repetition (Stevens, 1992). One might consider this approach for variables with intercorrelations of .80 or higher. Several other approaches to dealing with multicollinear relationships exist, but they are beyond the scope of this text. If interested, the reader is advised to pursue the discussion in Stevens (1992).

A third issue of great importance in multiple regression is the method of specifying the regression model; in other words, determining or selecting a good set of predictor variables. Keeping in mind that the goal of any analysis should be to achieve a *parsimonious* solution, we want to select IVs that will give us an efficient regression equation without including "everything under the sun." Initially, one of the most efficient methods of selecting a group of predictors is to rely on the researcher's substantive

knowledge (Stevens, 1992). Being familiar with and knowledgeable about your population, sample, and data will provide you with meaningful information about the relationships among variables and the likely predictive power of a set of predictors. Furthermore, for reasons we will discuss later, a recommended ratio of subjects to IVs (i.e., *n/k*) of at least 15 to 1 will provide a reliable regression equation (Stevens, 1992). Keeping the number of predictor variables low tends to improve this ratio, since most researchers do not have the luxury of increasing their sample size at will, which would be necessary if one were to continue to add predictors to the equation.

Once a set of predictors has been selected, there are several methods by which they may be incorporated into the regression analysis and subsequent equation. Tabachnick and Fidell (1996) identify three such strategies: standard multiple regression, sequential multiple regression, and stepwise multiple regression. (The reader should recall—and possibly *revisit*—the discussion of standard and sequential analyses as presented in Chapter 1.) It should be noted that decisions about model specification can and do affect the nature of the research questions being investigated. In ***standard multiple regression***, all IVs are entered into the analysis simultaneously. The effect of each IV on the DV is assessed as if it had been entered into the equation after all other IVs had been entered. Each IV is then evaluated in terms of what it adds to the prediction of the DV, as specified by the regression equation (Tabachnick & Fidell, 1996).

In ***sequential multiple regression***, sometimes referred to as ***hierarchical multiple regression***, a researcher may want to examine the influence of several predictor IVs in a specific order. Using this approach, the researcher specifies the order in which variables are entered into the analysis. Substantive knowledge, as previously mentioned, may lead the researcher to believe that one variable may be more influential than others in the set of predictors and that variable is entered into the analysis first. Subsequent variables are then added in order to determine the specific amount of variance they can account for, above and beyond, what has been explained by any variables entered prior (Aron & Aron, 1999). Individual effects are assessed at the point of entry of a given variable (Tabachnick & Fidell, 1996).

Finally, ***stepwise multiple regression***, also sometimes referred to as ***statistical multiple regression***, is often used in studies that are exploratory in nature (Aron & Aron, 1999). The researcher may have a large set of predictors and may want to determine which specific IVs make meaningful contributions to the overall prediction. There are essentially three variations of stepwise regression, listed and described below:

(1) ***Forward selection*** — The bivariate correlations among all IVs and the DV are calculated. The IV that has the highest correlation with the DV is entered into the analysis first. It is assessed in terms of its contribution (in terms of $R^2$) to the DV. The next variable to be entered into the analysis is the IV that contributes most to the prediction of the DV, after partialing out the effects of the first variable. This effect is measured by the increase in $R^2$ ($\Delta R^2$) due to the second variable. This process continues until, at some point, predictor variables stop making significant contributions to the prediction of the DV. It is important to remember that once a variable has been entered into the analysis, it remains there (Stevens, 1992; Pedhazur, 1982).

(2) ***Stepwise selection*** — Stepwise selection is a variation of forward selection. It is an improvement over the previous method in that, at each step, tests are performed to determine the significance of each IV already in the equation as if it were to enter last. In other words, if a variable entered into the analysis is measuring much of the same construct as another, this reassessment may determine that the first variable to enter may no longer contribute anything to the overall analysis. In this procedure, that variable would then be dropped out of the analysis.

Even though it was at one time a "good" predictor, in conjunction with others, it may no longer serve as a substantial contributor (Pedhazur, 1982).

(3) ***Backward deletion*** — The initial step here is to compute an equation with all predictors included. Then, a significance test (a partial $F$-test) is conducted for every predictor, as if each were entered last, in order to determine the level of contribution to overall prediction. The smallest partial $F$ is compared to a preselected "$F$ to remove" value. If the value is less than the "$F$ to remove" value (not significant), that predictor is removed from the analysis and a new equation with the remaining variables is computed, followed by another test of the resulting smallest partial $F$. This process continues until only significant predictors remain in the equation (Stevens, 1992).

It is important to note that both sequential and stepwise approaches to regression contain a distinct advantage over standard multiple regression—one variable is added at a time and each is continually checked for significant improvement to prediction. However, the important difference between these two is that sequential regression orders and adds variables based on some *theory or plan by the researcher*; whereas, in stepwise regression, those decisions are being made by a computer based *solely on statistical analyses* (Aron & Aron, 1999). Sequential regression should be used in research based on theory or some previous knowledge; stepwise regression should be used where exploration is the purpose of the analysis.

A fourth issue of consequence in multiple regression is that of model validation, sometimes called *model cross-validation*. A regression equation is developed in order to be able to predict DV values for individuals in a population, but remember that the equation itself was developed based only on a sample from that population. The multiple correlation, $R$, will be at its maximum value for the sample from which the equation was derived. If the predictive power drops off drastically when applied to an independent sample from the same population, the regression equation is of little use since it has little or no generalizability (Stevens, 1992). If the equation is not predicting well for other samples, it is not fulfilling its designed and intended purpose.

In order to obtain a reliable equation, substantial consideration must be given to the sample size ($n$) and the number of predictors ($k$). As mentioned earlier, a recommended ratio of these two factors is about 15 subjects for every predictor (Stevens, 1992). This results in a equation that will cross-validate with relatively little loss in its ability to predict. Another recommendation for this ratio is identified by Tabachnick and Fidell (1996). The simplest rule of thumb they offer is that $n \geq 50 + 8k$, for testing multiple correlations, and $n \geq 104 + k$, for testing individual predictors. They suggest calculating $n$ both ways and using the larger value.

Cross-validation can be accomplished in several ways. Ideally, one should wait a period of time, select an independent sample from the same population, and test the previously obtained regression equation (Tatsuoka, 1988). This is not always feasible, so an alternative would be to split the original sample into two "subsamples." Then one subsample can be used to develop the equation, while the other is used to cross-validate it (Stevens, 1992). Of course, this would only be feasible if one had a large enough sample, based on the criteria set forth above.

A final issue of importance in regression is the effect that outliers can have on a regression solution. Recall that regression is essentially a maximization procedure (i.e., we are trying to maximize the correlation between observed and predicted DV scores). Because of this fact, multiple regression can be very sensitive to extreme cases. One or two outliers have been shown to adversely affect the interpretation of regression analysis results (Stevens, 1992). It is therefore recommended that outliers be identified and dealt with appropriately prior to running the regression analysis. This is typically accomplished

by initial screenings of boxplots, but more precisely with the statistical procedure known as *Mahalanobis distance* (as described in Chapter 3).

One final note regarding multiple regression. There does exist a multivariate version of multiple regression (i.e., multivariate multiple regression), but it is so similar in its approach and conduct that it will not be discussed in detail in this text. Basically, multivariate multiple regression involves the prediction of several DVs from a set of predictor IVs. This procedure is a variation of multiple regression in that the regression equations realized are those that would be obtained if each DV were regressed *separately* on the set of IVs. The actual correlations among DVs in the analysis are ignored (Stevens, 1992).

## Sample Research Questions

Building on the example we began discussing in the previous section, we can now specify the research questions to be addressed by our multiple regression analysis. The methods by which the regression model is developed often dictates the type of research question(s) to be addressed. For example, if we were entering all seven IVs from our data set into the model, the appropriate research questions would be:

(1) Which of the seven predictor variables (i.e., percent urban population, GDP, birthrate, number of hospital beds, number of doctors, number of radios, and number of telephones) are most influential in predicting female life expectancy? Are there any predictor variables that do not contribute significantly to the prediction model?
(2) Does the obtained regression equation resulting from a set of seven predictor variables allow us to reliably predict female life expectancy?

However, if we were using a stepwise method of specifying the model, the revised questions would be:

(1) Which of the possible seven predictor variables (i.e., percent urban population, GDP, birthrate, number of hospital beds, number of doctors, number of radios, and number of telephones) are included in an equation for predicting female life expectancy?
(2) Does the obtained regression equation resulting from a subset of the seven predictor variables allow us to reliably predict female life expectancy?

## *Section 7.2 Assumptions and Limitations*

In multiple regression, there are actually two sets of assumptions—assumptions about the raw scale variables and assumptions about the residuals (Pedhazur, 1982). With respect to the raw scale variables, the following conditions are assumed:

(1) The independent variables are fixed (i.e., the same values of the IVs would have to be used if the study were to be replicated).
(2) The independent variables are measured without error.
(3) The relationship between the independent variables and the dependent variable is linear (in other words, the regression of the DV on the combination of IVs is linear).

The remaining assumptions concern the residuals. Recall again from Chapter 3 that *residuals*, or *prediction errors*, are the portions of scores not accounted for by the multivariate analyses. Meeting these assumptions is necessary in order to achieve the best linear estimations (Pedhazur, 1982). These assumptions are:

(4) The mean of the residuals for each observation on the dependent variable over many replications is zero.
(5) Errors associated with any single observation on the dependent variable are independent of (i.e., not correlated with) errors associated with any other observation on the dependent variable.
(6) The errors are not correlated with the independent variables.
(7) The variance of the residuals across all values of the independent variables is constant (i.e., homoscedasticity of the variance of the residuals).
(8) The errors are normally distributed.

Assumptions 1, 2, and 4 are largely research design issues. We will focus our attention on assumptions 3, 5, and 6—which address the issue of linearity—and assumptions 7 and 8—which address homoscedasticity and normality, respectively.

## Methods of Testing Assumptions

There are essentially two approaches to testing the assumptions in multiple regression (Tabachnick & Fidell, 1996). The first approach involves the routine pre-analysis data screening procedures that have been discussed in the preceding several chapters. As a reminder, linearity can be assessed through examination of the various bivariate scatterplots. Normality is evaluated in similar fashion, as well as through the assessment of the values for skewness, kurtosis, and Kolmogorov-Smirnov statistics. Finally, homoscedasticity is assessed by interpreting the results of Box's M Test.

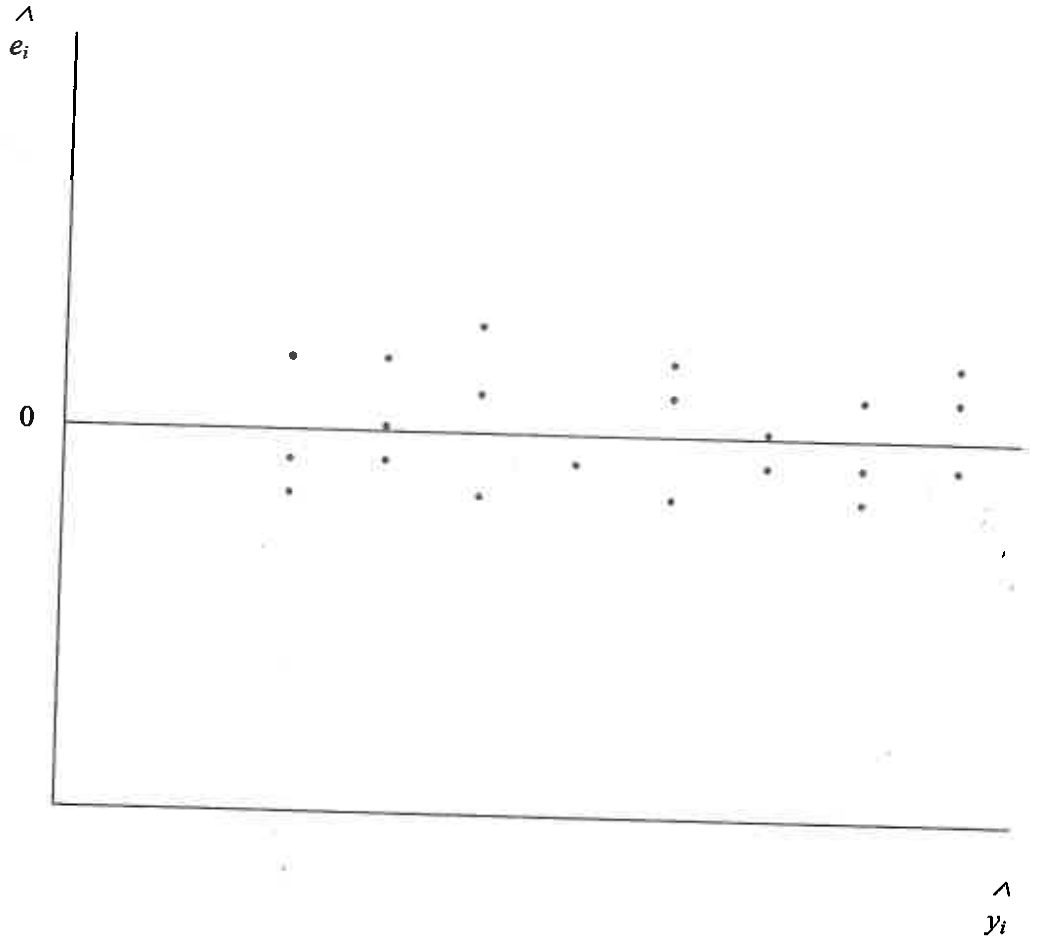
The alternative approach to the routine procedure is to examine the residuals scatterplots. These scatterplots resemble bivariate scatterplots in that they are plots of values on the combination of two "variables"—in this case, these are the predicted values of the DV ($\hat{y}$) and the standardized residuals or prediction errors ($\hat{e}_i$). Examination of these residual scatterplots provides a test of *all three* of these crucial assumptions (Tabachnick & Fidell, 1996). If the assumptions of linearity, normality, and homoscedasticity are tenable, we would expect to see the points cluster along the horizontal line defined by $\Lambda_i = 0$, in a somewhat rectangular pattern (see Figure 7.3).

Any systematic, differential patterns or clusters of points are an indication of possible model violations (Tabachnick & Fidell, 1996; Stevens, 1992). Examples of residuals plots depicting violations of the three assumptions are shown in Figure 7.4. (*It is important to note that the plots shown in this figure are idealized and have been constructed to show clear violations of assumptions. A word of caution—with real data, the patterns are seldom this obvious.*) If the assumption of linearity is tenable, we would expect to see a relatively straight line relationship among the points in the plot. This typically appears as a rectangle (Tabachnick & Fidell, 1996), as depicted in Figure 7.3. However, as shown in Figure 7.4(a), the points obviously appear in a nonlinear pattern. In fact, this example is so extreme as to depict a clearly curvilinear pattern. This is an unmistakable violation of the assumption of linearity.

If the assumption of normality is defensible, we would expect to see an even distribution of points both above and below the line defined by $\hat{e}_i = 0$. In Figure 7.4(b), there appears to be a clustering of points the farther we move both above and below that reference line, indicating a non-normal (in this case, bimodal) distribution of residuals (Tate, 1992).

Finally, Figure 7.4(c) shows a violation of the assumption of homoscedasticity. If this assumption is tenable, we would expect to see the points dispersed evenly about the reference line—again, defined by $\hat{e}_i = 0$—across all predicted values for the DV. In Figure 7.4(c), notice that the width is very narrow at small predicted values for the DV; however, the width increases rapidly as the predicted DV value increases. This is a clear indication of heteroscedasticity, or a lack of constant variance.

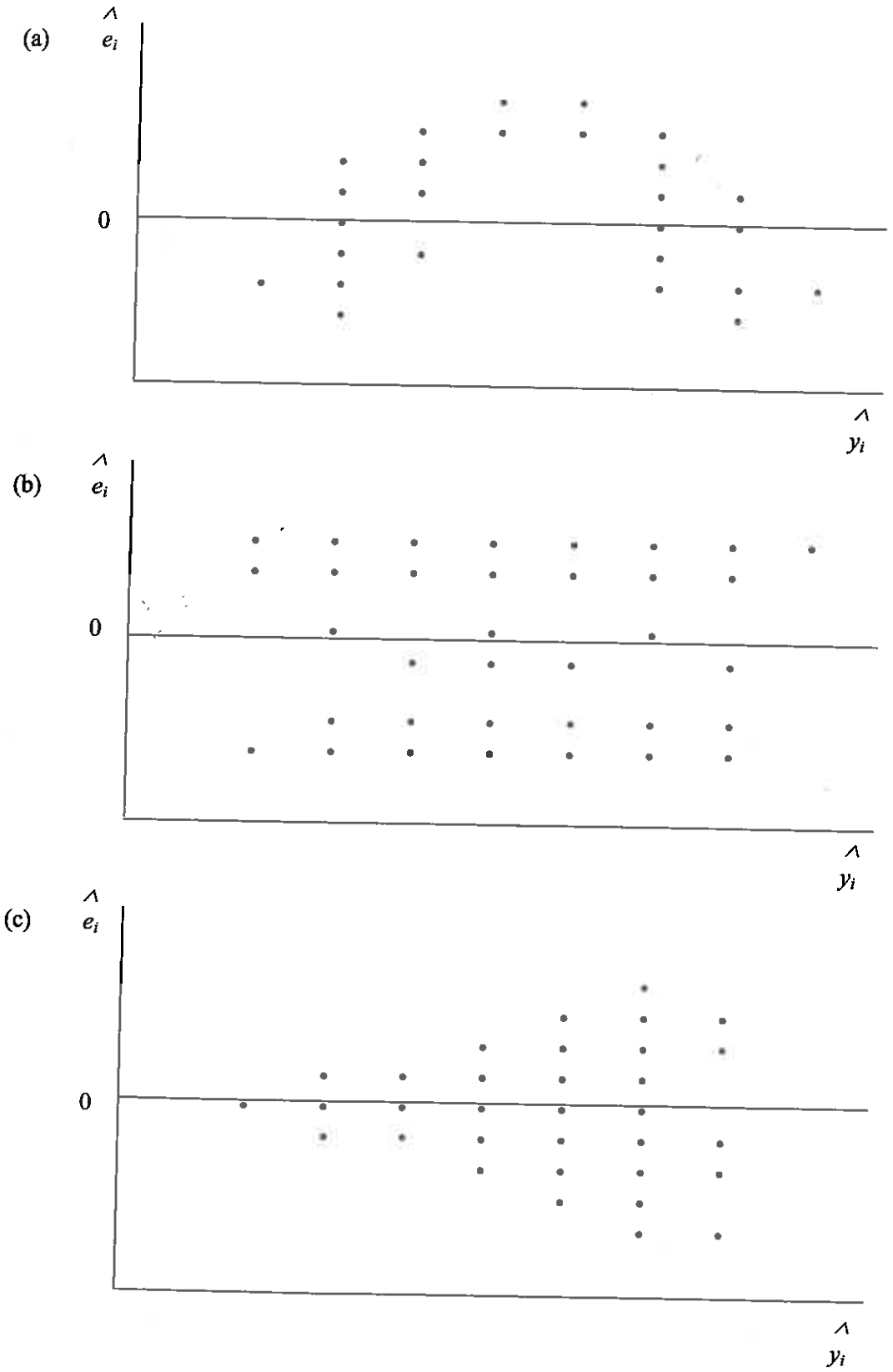
**Figure 7.3** Residuals Plot of Standardized Residuals ($e_i$) Versus Predicted Values ($y_i$) When Assumptions Are Met.




Residuals scatterplots may be examined *in place of* the routine pre-analysis data screening or *following* those procedures (Tabachnick & Fidell, 1996). If examination of the residuals scatterplots is conducted instead of the routine procedures—and if no violations are evident, no outliers exist, there are sufficient number of cases, and there is no evidence of multicollinearity—then one can be safe in interpreting that single regression run on the computer. However, if the initial residuals scatterplots do not look "clean," then further data screening using the routine procedures is warranted (Tabachnick & Fidell, 1996). In many cases, this may involve the transformation of one or more variables in order to meet the assumptions. If a curvilinear pattern appears, one possible remedy is to use a polynomial (i.e., nonlinear) model (Stevens, 1992), which is beyond the scope of this book.

In cases that involve moderate violations of linearity and homoscedasticity, one should be aware that these violations merely weaken the regression analysis, but do not invalidate it (Tabachnick & Fidell, 1996). Furthermore, moderate violations of the normality assumption may often be ignored—especially with larger sample sizes—since there are no adverse effects on the analysis (Tate, 1992). It may still be possible to proceed with the analysis, depending on the subjective judgments of the researcher. Unfortunately, however, there are no rules to explicitly define that which constitutes a "moderate" violation. In reality, we would probably be justified in expecting some slight departures from the "ideal" situation, as depicted in Figure 7.3, due to sampling fluctuations (Tate, 1992).

**Figure 7.4** Residuals Plots Showing Violations of (a) Linearity, (b) Normality, and (c) Homoscedasticity.



## *SECTION 7.3 PROCESS AND LOGIC*

### The Logic Behind Multiple Regression

You will recall from your previous exposure to simple regression that the statistical calculations basically involve the determination of the constants $a$ and $b$. The slope of the line (i.e., $b$) is first calculated by multiplying the correlation coefficient between $X$ and $Y$—recall we discussed earlier in this chapter the important role played by the correlation between $X$ and $Y$—by the standard deviation of $Y$ and dividing that term by the standard deviation of $X$:

$$b = \frac{(r)(SD_Y)}{(SD_X)} \qquad \text{(Equation 7.4)}$$

The constant $a$ (the $Y$-intercept) is then calculated in the following manner:

$$a = \bar{Y} - b\bar{X} \qquad \text{(Equation 7.5)}$$

There are analogous equations for the multivariate regression situation, although they appear slightly more ominous and, therefore, will not be shown here. Recall from Equation 7.3 that in multiple regression there are *at least* two regression coefficients (specifically, the slope coefficients $B_1$ and $B_2$) that must be calculated. The calculations mirror Equation 7.4; the only substantial difference is that they incorporate a concept known as partial correlation. *Partial correlation* is a measure of the relationship between an IV and DV, holding all other IVs constant. For example, the calculated value for $B_1$ tells us how much of a change in $Y$ can be expected for a given change in $X_1$ when the effects of $X_2$ are held constant (Sprinthall, 2000).

The other main calculation in multiple regression is the determination of the value for $R^2$ and its associated significance test. Recall that $R^2$ is a measure of variance accounted for in the DV by the predictors. One can think of this as being similar to analysis of variance, in that we must partition the sum of squares variability. In regression analysis, we separate the total variability into variability due to regression (see Equation 7.6) and variability about the regression, also known as the sum of squares residual (see Equation 7.7).

$$SS_{reg} = \Sigma(\hat{y}_i - \bar{y})^2 \qquad \text{(Equation 7.6)}$$

$$SS_{res} = \Sigma(y_i - \hat{y})^2 \qquad \text{(Equation 7.7)}$$

The total sum of squares is simply the sum of these two terms and is symbolized by $\Sigma(y_i - \bar{y})^2$. The squared multiple correlation is then calculated by dividing the sum of squares due to regression ($SS_{reg}$) by the sum of squares total ($SS_{tot}$):

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \qquad \text{(Equation 7.8)}$$

The standard *F*-test from analysis of variance can be written making some simple algebraic substitutions:

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)} \qquad \text{(Equation 7.9)}$$

where $k$ and $n - k - 1$ are the appropriate degrees of freedom for the numerator and denominator, respectively (Stevens, 1992). From this point, the significance of the obtained value for $R^2$ can be tested using the standard *F*-test criteria, or by simply examining the associated *p* value from the computer printout. This then tells the researcher whether the set of IV predictor variables is accounting for, or explaining, a statistically significant amount of variance in the DV.

## Interpretation of Results

Interpretation of multiple regression focuses on determining the adequacy of the regression model(s) that has been developed. Conducting multiple regression typically generates output that can be divided into three parts: model summary, ANOVA, and coefficients. Our discussion on how to interpret regression results will address these three parts. The first part of the regression output, model summary, displays several multiple correlation indices—multiple correlation ($R$), squared multiple correlation ($R^2$), and adjusted squared multiple correlation ($R^2_{adj}$)—all of which indicate how well an IV or combination of IVs predicts the criterion variable (DV). The multiple correlation ($R$) is a Pearson correlation coefficient between the predicted and actual scores of the DV. The squared multiple correlation ($R^2$) represents the degree of variance accounted for by the IV or combination of IVs. Unfortunately, $R$ and $R^2$ typically overestimate their corresponding population values especially with small samples; thus $R^2_{adj}$ is calculated to account for such bias. Change in $R^2$ ($\Delta R^2$) is also calculated for each step and represents the change in variance that is accounted for by the set of predictors once a new variable has been added to the model. Change in $R^2$ is important since it is used to determine which variables significantly contribute to the model, or in the case of a stepping method, which variables are added or removed from the model. If a stepping method is used, the model summary will present these statistics for each model or step that is generated.

The ANOVA table presents the *F*-test and corresponding level of significance for each step or model generated. This test examines the degree to which the relationship between the DV and IVs is linear. If the *F*-test is significant, then the relationship is linear and therefore the model significantly predicts the DV.

The final part of the output is the coefficients table that reports the following: unstandardized regression coefficient ($B$), the standardized regression coefficient (beta or $\beta$), *t* and *p* values, and three correlation indices. The unstandardized regression coefficient ($B$), also known as the *partial regression coefficient*, represents the slope weight for each variable in the model and is used to create the regression equation. $B$ weights also indicate how much the value of the DV changes when the IV increases by 1 and the other IVs remain the same. A positive $B$ specifies a positive change in the DV when the IV increases, whereas a negative $B$ indicates a negative change in the DV when the IV increases. Since it is difficult to interpret the relative importance of the predictors when the slope weights are not standardized, beta weights ($\beta$) or standardized regression coefficients are often utilized to create a prediction equation for the standardized variables. Beta weights are based upon *z*-scores with a mean of 0 and standard deviation of 1. The coefficients table also presents *t* and *p* values, which indicate the significance of the $B$ weights, beta weights, and the subsequent part and partial correlation coefficients. Actu-

ally, three correlation coefficients are displayed in the coefficients table. The zero-order correlation represents the bivariate correlation between the IV and DV. The partial correlation coefficient indicates the relationship between the IV and DV after partialing out all other IVs. The part correlation, rarely used when interpreting the output, represents the correlation between the DV and IVs after partialing only one of the IVs.

The final important statistic in the coefficient table is tolerance, which is a measure of multicollinearity among the IVs. Since the inclusion of IVs that are highly dependent upon each other can create an erroneous regression model, determining which variables account for a high degree of common variance in the DV is critical. Tolerance is reported for all the IVs included and excluded in the generated model. This statistic represents the proportion of variance in a particular IV that is not explained by its linear relationship with the other IVs. Tolerance ranges from 0 to 1, with 0 indicating multicollinearity. Typically, if tolerance of an IV is less than .1, the regression procedure should be repeated without the violating IV.

As one can see, there is a lot to interpret when conducting multiple regression. Since tolerance is an indicator of the appropriateness of IVs utilized in the regression, this statistic should be interpreted first. If some IVs violate the tolerance criteria, regression should be conducted again without the violating variables. If the value for tolerance is acceptable, one should proceed with interpreting the model summary, ANOVA table, and table of coefficients.

Let us now apply this process to our example. Since we will utilize the Forward stepping method, our research question is more exploratory in nature: Which IVs (% urban population [*urban*]; gross domestic product per capita [*gdp*]; birthrate per 1,000 [*birthrat*]; hospital beds per 10,000 [*hospbed*]; doctors per 10,000 [*docs*]; radios per 100 [*radio*]; and phones per 100 [*phone*]) are predictors of female life expectancy? Data were first screened for missing data and outliers and then examined for test assumptions. Outliers were identified by calculating Mahalanobis distance in a preliminary **Regression** procedure (see Chapter 3 for SPSS "How To"). **Explore** was then conducted on the newly generated Mahalanobis variable (*mah_1*) to determine which cases exceeded the chi square ($\chi^2$) criteria (See Figure 7.5). Using a chi square table, we found the critical value of chi square at $p<.001$ with $df=8$ to be 26.125. Case #83 exceeds this critical value and so was deleted from our analysis. Linearity was then analyzed by creating a scatterplot matrix (see Figure 7.6). Scatterplots display nonlinearity for the following variables: *gdp, hospbed, docs, radio,* and *phone*. These variables were transformed by taking the natural log of each. The reader should note that the data set already includes these transformations as *lngdp, lnbeds, lndocs, lnradio,* and *lnphone*. A scatterplot matrix (see Figure 7.7) with the transformed variables displays elliptical shapes that indicate linearity and normality. Univariate normality was also assessed by conducting **Explore**. Histograms and normality tests (see Figure 7.8) indicate some non-normal distributions; however, the distributions are not extreme. Multivariate normality and homoscedasticity were examined through the generation of a residuals plot within another preliminary **Regression** (see Chapter 3 for SPSS "How To"). The residuals plot is somewhat scattered but again is not extreme (see Figure 7.9). Thus, multivariate normality and homoscedasticity will be assumed.

**Regression** was then conducted using the Forward method. The three major parts of the output—model summary, ANOVA table, and coefficient table—are presented in Figures 7.10 –7.12, respectively. Tolerance among the IVs is adequate since coefficients for all IVs included and excluded are above .1 (see Figure 7.12). Since the Forward method was utilized, only some of the IVs were entered into the model. The model summary (see Figure 7.10) indicates that three of the seven IVs were entered into the model. For the first step, *lnphone* was entered as it accounted for the most unique vari-

ance in the DV ($R^2$=.800). The variables of *birthrat* and *lndocs* were entered in the next two steps, respectively, creating a model that accounted for 86.9% of the variance in female life expectancy. The ANOVA table (see Figure 7.11) presents the $F$-test for each step/model. The final model significantly predicts the DV, $F(3, 102)=226.50$, $p<.001$. The table of coefficients (see Figure 7.12) is then utilized to create a prediction equation for the DV. The following equation is generated using the $B$ weights.

$$Female\ life\ expectancy = 2.245X_{lnphone} - .241X_{birthrat} + 2.172X_{lndocs} + 68.159$$

If we utilize the beta weights, we develop the following equation for predicting the standardized DV.

$$Z_{Female\ life\ expectancy} = .394\ Z_{lnphone} - .288\ Z_{birthrat} + .306\ Z_{lndocs}$$

Bivariate and partial correlation coefficients should also be noted in the coefficients table.

**Figure 7.5** Outliers for Mahalanobis Distance.

**Extreme Values**

| | | | Case Number | Value |
|---|---|---|---|---|
| MAH_1 | Highest | 1 | 83 | 36.89903 |
| | | 2 | 72 | 20.98981 |
| | | 3 | 19 | 18.35770 |
| | | 4 | 99 | 18.26913 |
| | | 5 | 81 | 17.51827 |
| | Lowest | 1 | 26 | 2.41654 |
| | | 2 | 107 | 2.96129 |
| | | 3 | 102 | 3.26488 |
| | | 4 | 108 | 3.37951 |
| | | 5 | 6 | 3.38166 |

Case #83 exceeds $\chi 2$ critical value.

## Writing Up Results

The summary of multiple regression results should always include a description of how variables have been transformed or cases deleted. Typically, descriptive statistics (e.g., correlation matrix, means and standard deviations for each variable) are presented in tables unless only a few variables are analyzed. The reader should note that our example of a results summary will not include these descriptive statistics due to space limitations. The overall regression results are summarized in the narrative by identifying the variables in the model: $R^2$, $R^2_{adj}$, $F$ and $p$ values with degrees of freedom. If a step approach has been utilized, you may want to report each step ($R^2$, $R^2_{adj}$, $R^2$ change, and level of significance for change) within a table. Finally, you may want to report the B weight, beta weight, bivariate correlation coefficients, and partial correlation coefficients of the predictors with the DV in a table. If you do not present these coefficients in a table, you may want to report the prediction equation, either standardized or unstandardized. The following results statement applies the results presented in Figures 7.10 – 7.12.

> Forward multiple regression was conducted to determine which independent variables (% urban population [urban); gross domestic product per capita [gdp]; birthrate per 1,000 [birthrate]; hospital beds per 10,000 [beds]; doctors per 10,000 [docs]; radios per 100 [radios]; and phones per 100 [phones]) were the predictors of female life expectancy. Data screening led to the elimination of one case. Evaluation of linearity led to the natural log transformation of gdp, beds, docs, radios, and phones. Regression results indicate an overall model of three predictors (phone, birthrate, and docs) that significantly predict female life expectancy, $R^2$=.869, $R^2_{adj}$=.866, $F(3, 102,)=226.50$, $p<.001$. This model accounted for 86.9% of variance in female life expectancy.

A summary of the regression model is presented in Table 1. In addition, bivariate and partial correlation coefficients between each predictor and the dependent variable are presented in Table 2.

**Table 1** Model Summary

| Step | $R$ | $R^2$ | $R^2_{adj}$ | $\Delta R^2$ | $F_{chg}$ | $p$ | $df_1$ | $df_2$ |
|---|---|---|---|---|---|---|---|---|
| 1. Phones | .894 | .800 | .798 | .800 | 416.03 | <.001 | 1 | 104 |
| 2. Birthrate | .921 | .849 | .846 | .049 | 33.52 | <.001 | 1 | 103 |
| 3. Doctors | .932 | .869 | .866 | .020 | 15.92 | <.001 | 1 | 102 |

**Table 2** Coefficients for Final Model

| | $B$ | $\beta$ | $t$ | Bivariate $r$ | Partial $r$ |
|---|---|---|---|---|---|
| Phones per 100 | 2.245 | .394 | 5.078* | .894 | .449 |
| Birthrate per 1,000 | –.241 | –.288 | –4.263* | –.861 | –.389 |
| Doctors per 10,000 | 2.172 | .306 | 3.990* | .881 | .367 |

Note: * Indicates significance at p<.001.

**Figure 7.6** Scatterplot Matrix for Original IVs and DV.

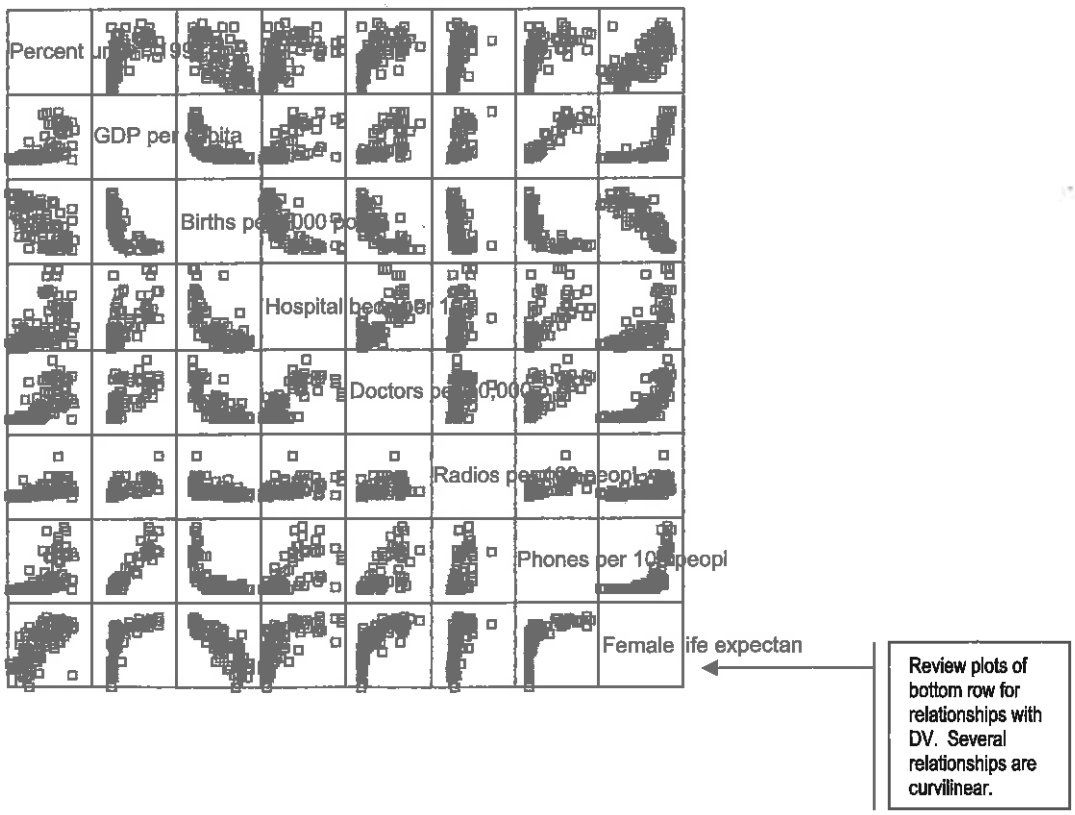


**Figure 7.7** Scatterplot Matrix of Transformed IVs with DV.

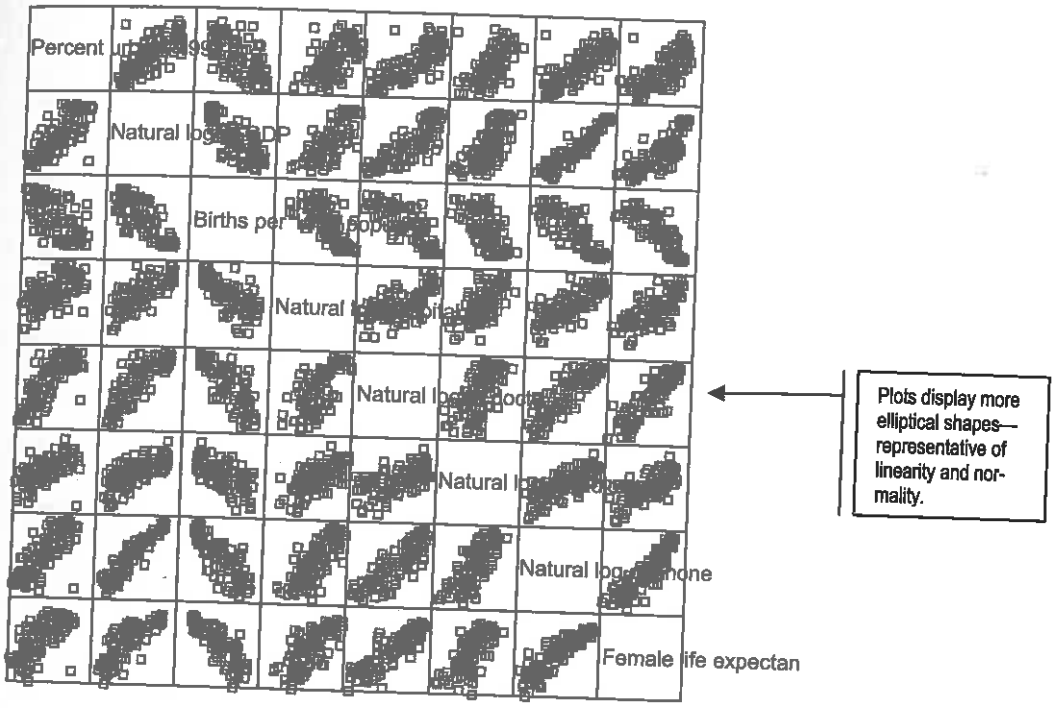



**Figure 7.8** Test of Normality.

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | |
|---|---|---|---|
| | Statistic | df | Sig. |
| URBAN | .089 | 107 | .037 |
| LNGDP | .096 | 107 | .016 |
| BIRTHRAT | .132 | 107 | .000 |
| LNBEDS | .059 | 107 | .200* |
| LNDOCS | .138 | 107 | .000 |
| LNRADIO | .092 | 107 | .026 |
| LNPHONE | .087 | 107 | .047 |

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Indicates that most distributions are non-normal.

**Figure 7.9** Residuals Plot.

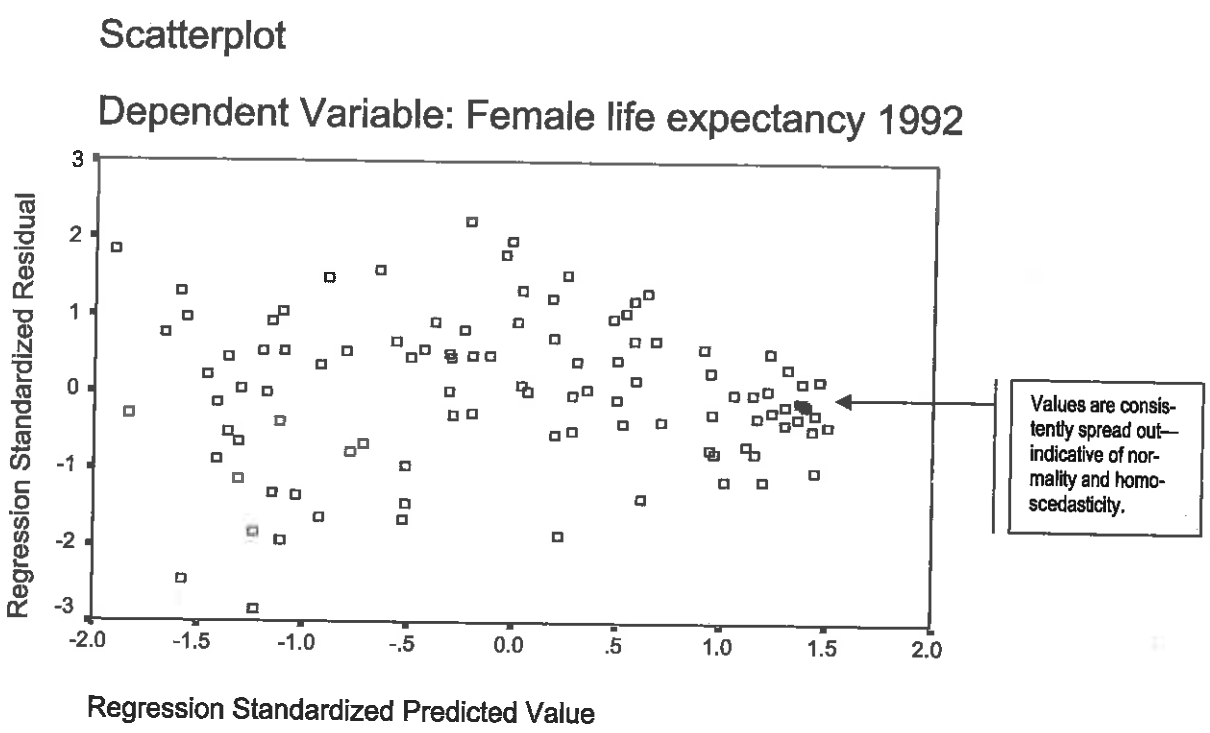



**Figure 7.10** Model Summary Table for Female Life Expectancy.

**Model Summary[d]**

| 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 | .894[a] | .800 | .798 | 4.98 | .800 | 416.027 | 1 | 104 | .000 |
| 2 | .921[b] | .849 | .846 | 4.34 | .049 | 33.516 | 1 | 103 | .000 |
| 3 | .932[c] | .869 | .866 | 4.06 | .020 | 15.919 | 1 | 102 | .000 |

a. Predictors: (Constant), LNPHONE

b. **Predictors: (Constant), LNPHONE, BIRTHRAT**

c. Predictors: (Constant), LNPHONE, BIRTHRAT, LNDOCS

d. Dependent Variable: LIFEEXPF

Represents each step in the model building.

**Figure 7.11** ANOVA Summary Table.

**ANOVA[d]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 10299.784 | 1 | 10299.784 | 416.027 | .000[a] |
| | Residual | 2574.782 | 104 | 24.758 | | |
| | Total | 12874.566 | 105 | | | |
| 2 | Regression | 10931.921 | 2 | 5465.960 | 289.808 | .000[b] |
| | Residual | 1942.645 | 103 | 18.861 | | |
| | Total | 12874.566 | 105 | | | |
| 3 | Regression | 11194.184 | 3 | 3731.395 | 226.497 | .000[c] |
| | Residual | 1680.382 | 102 | 16.474 | | |
| | Total | 12874.566 | 105 | | | |

a. Predictors: (Constant), LNPHONE

b. Predictors: (Constant), LNPHONE, BIRTHRAT

c. Predictors: (Constant), LNPHONE, BIRTHRAT, LNDOCS

d. Dependent Variable: LIFEEXPF

Indicates that the final model is significant in predicting the DV.

## *Section 7.4 Sample Study and Analysis*

This section provides a complete example of the process of conducting multiple regression. This process includes the development of research questions and hypotheses, data screening methods, test methods, interpretation of output, and presentation of results. The example utilizes the data set, *country.sav* from the Web site that accompanies this book (see p. *xi*).

### Problem

In the previous example, we identified predictors of female life expectancy. For this example, we will utilize the same IVs but change the DV to male life expectancy. In addition, the Enter method will be used, such that all IVs will be entered into the model. The following research question is generated to address this scenario:

> How accurately do the IVs [% urban population (*urban*); gross domestic product per capita (*gdp*); birthrate per 1,000 (*birthrat*); hospital beds per 10,000 (*hospbed*); doctors per 10,000 (*docs*); radios per 100 (*radios*); and phones per 100 (*phones*)] predict male life expectancy?

**Figure 7.12** Coefficients Tables for Variables Included and Excluded from Model.

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Correlations | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | 60.513 | .585 | | 103.493 | .000 | | | | | |
| | LNPHONE | 5.102 | .250 | .894 | 20.397 | .000 | .894 | .894 | .894 | 1.000 | 1.000 |
| 2 | (Constant) | 72.700 | 2.166 | | 33.564 | .000 | | | | | |
| | LNPHONE | 3.284 | .383 | .576 | 8.583 | .000 | .894 | .646 | .329 | .326 | 3.070 |
| | BIRTHRAT | -.325 | .056 | -.388 | -5.789 | .000 | -.861 | -.495 | -.222 | .326 | 3.070 |
| 3 | (Constant) | 68.159 | 2.322 | | 29.349 | .000 | | | | | |
| | LNPHONE | 2.245 | .442 | .394 | 5.078 | .000 | .894 | .449 | .182 | .213 | 4.696 |
| | BIRTHRAT | -.241 | .056 | -.288 | -4.263 | .000 | -.861 | -.389 | -.152 | .281 | 3.565 |
| | LNDOCS | 2.172 | .544 | .306 | 3.990 | .000 | .881 | .367 | .143 | .218 | 4.592 |

a. Dependent Variable: LIFEEXPF

Coefficients used to develop regression equation.

Coefficients used to develop a regression equation for standardized variables.

Tolerance statistics should be greater than .1.

**Excluded Variables[d]**

| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Tolerance | VIF | Minimum Tolerance |
| 1 | LNBEDS | .098[a] | 1.431 | .155 | .140 | .409 | 2.444 | .409 |
| | LNGDP | -.022[a] | -.166 | .868 | -.016 | .112 | 8.927 | .112 |
| | LNRADIO | .067[a] | 1.043 | .299 | .102 | .466 | 2.144 | .466 |
| | LNDOCS | .428[a] | 5.564 | .000 | .481 | .253 | 3.956 | .253 |
| | BIRTHRAT | -.388[a] | -5.789 | .000 | -.495 | .326 | 3.070 | .326 |
| | URBAN | .096[a] | 1.267 | .208 | .124 | .335 | 2.984 | .335 |
| 2 | LNBEDS | .017[b] | .274 | .785 | .027 | .386 | 2.587 | .259 |
| | LNGDP | -.159[b] | -1.372 | .173 | -.135 | .108 | 9.293 | .103 |
| | LNRADIO | .043[b] | .759 | .450 | .075 | .464 | 2.156 | .250 |
| | LNDOCS | .306[b] | 3.990 | .000 | .367 | .218 | 4.592 | .213 |
| | URBAN | .103[b] | 1.566 | .121 | .153 | .335 | 2.985 | .195 |
| 3 | LNBEDS | .013[c] | .229 | .819 | .023 | .386 | 2.588 | .184 |
| | LNGDP | -.153[c] | -1.412 | .161 | -.139 | .108 | 9.295 | 8.728E-02 |
| | LNRADIO | .047[c] | .900 | .370 | .089 | .464 | 2.157 | .176 |
| | URBAN | .003[c] | .049 | .961 | .005 | .280 | 3.573 | .175 |

a. Predictors in the Model: (Constant), LNPHONE

b. Predictors in the Model: (Constant), LNPHONE, BIRTHRAT

c. Predictors in the Model: (Constant), LNPHONE, BIRTHRAT, LNDOCS

d. Dependent Variable: LIFEEXPF

## Method

Data are screened to identify missing data and outliers and to evaluate the fulfillment of test assumptions. Outliers were identified by calculating Mahalanobis distance in a preliminary **Regression** procedure. **Explore** was then conducted on the newly generated Mahalanobis variable (*mah_1*) to determine which cases exceeded the chi square ($\chi^2$) criteria (see Figure 7.13). Using a chi square table, we found the critical value of chi square at $p<.001$ with $df=8$ to be 26.125. Cases #69, #72, and #67 exceed this critical value and so were deleted from the analysis. Linearity was then analyzed by creating a scatterplot matrix (see Figure 7.14). Scatterplots display nonlinearity for the following variables: *gdp, hospbed, docs, radios,* and *phones*. These variables were transformed by taking the natural log of each. The data set already includes these transformations as *lngdp, lnbeds, lndocs, lnradio,* and *lnphone*. A scatterplot of the transformed variables indicates linearity and normality (see Figure 7.15). Univariate normality was also assessed by conducting **Explore**. Histograms and normality tests (see Figure 7.16) indicate some non-normal distributions; however, the distributions are not too extreme. Multivariate normality and homoscedasticity were examined through the generation of a residuals plot within another preliminary **Regression**. The residuals plot is somewhat scattered but again is not extreme (see Figure 7.17). Thus, multivariate normality and homoscedasticity will be assumed. Multiple **Regression** was then conducted using the Enter method. See the section on SPSS "How To" for more details on how to generate the following output.

**Figure 7.13** Outliers for Mahalanobis Distance.

**Extreme Values**

| | | | Case Number | Value |
|---|---|---|---|---|
| MAH_1 | Highest | 1 | 67 | 50.81753 |
| | | 2 | 72 | 27.23509 |
| | | 3 | 69 | 26.69299 |
| | | 4 | 108 | 25.67930 |
| | | 5 | 83 | 21.00443 |
| | Lowest | 1 | 21 | 1.19032 |
| | | 2 | 40 | 1.29138 |
| | | 3 | 87 | 1.59614 |
| | | 4 | 53 | 1.73529 |
| | | 5 | 25 | 1.81816 |

Outliers exceed $\chi 2$ critical value.

## Output and Interpretation of Results

Figures 7.18 – 7.20 present the three primary parts of regression output: model summary, ANOVA table, coefficients table. Review of the tolerance statistics presented in the coefficients table (see Figure 7.20) indicate that all but one of the IVs were tolerated in the model. The model summary (see Figure 7.18) and the ANOVA summary (see Figure 7.19) indicate that the overall model of the seven IVs significantly predicts male life expectancy, $R^2=.845$, $R^2_{adj}=.834$, $F(7, 96,)=74.69$, $p<.001$. However, review of the beta weights in Figure 7.20 specify that only three variables, *birthrat* $\beta=-.241$, $t(96)=-3.02$, $p=.003$; *lndocs* $\beta=.412$, $t(96)=4.26$, $p<.001$; and *lnphone* $\beta=.548$, $t(96)=3.88$, $p<.001$, significantly contributed to the model. The reader should note that although the same three variables created the model for predicting female life expectancy despite a different method being utilized, the sig-

nificance of the model predicting male life expectancy is much lower since all seven variables were entered into the model.

**Figure 7.14** Scatterplot Matrix of Original IVs with DV.

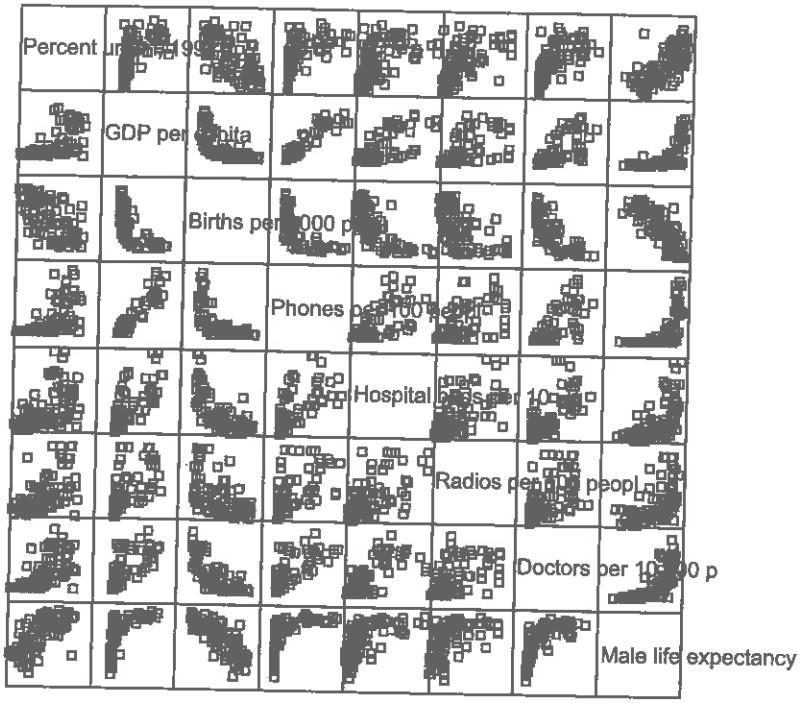



**Figure 7.15** Scatterplot Matrix of Transformed IVs with DV.