4

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

4.1 INTRODUCTION

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

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

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

concerns, in turn, may raise questions about the conclusions that are reached about independent variables based on confidence intervals or significance tests. But, even more important, violations of assumptions can point to problems in the specification of the regression model and provide valuable clues that can lead to a revision of the model, yielding an even greater understanding of the data. In other cases, violations of assumptions may point to complexities in the data that require alternative approaches to estimating the original regression model. We present a number of graphical and statistical approaches that help diagnose violations of assumptions and introduce potential remedies for these problems.

The themes of data exploration/visualization coupled with careful checking of assumptions are familiar ones in statistics. Yet, diffusion of these themes to the behavioral sciences has been uneven, with some areas embracing and some areas neglecting them. Some areas have primarily emphasized hypothesis testing, confirmatory data analysis. Yet, as Tukey (1977) emphasized, "**Today, exploratory and confirmatory can—and should—proceed side by side**" (p. vii, bold in original). Some areas such as econometrics and some areas of sociology have emphasized careful checking of assumptions, whereas some areas of psychology have been more willing to believe that their results were largely immune to violations of assumptions. The increasing availability of both simple methods for detecting violations and statistical methods for addressing violations of assumptions decreases the force of this latter belief. Proper attention to the assumptions of regression analysis leads to benefits. Occasionally, gross errors in the conclusions of the analysis can be avoided. More frequently, more precise estimates of the effects of interest can be provided. And often, proper analyses are associated with greater statistical power, helping researchers detect their hypothesized effects (e.g., Wilcox, 1998). We hope to encourage researchers in those areas of the behavioral sciences that have overlooked these powerful themes of data exploration/visualization and assumption checking to begin implementing them in their everyday data analysis.

We defer until later in the book consideration of two other issues that arise in multiple regression with real data sets. First is the existence of *outliers*, unusual cases that are far from the rest of the data. In some cases, the existence of a few outliers, even one, can seriously jeopardize the results and conclusions of the regression analysis. Second is multicollinearity, the problem of high redundancy between the IVs first introduced in Section 3.8.4. Multicollinearity leads to imprecise estimates of regression coefficients and increased difficulty in interpreting the results of the analysis. Both of these issues receive detailed consideration in Chapter 10.

Boxed Material in the Text

Finally, the structure of Chapter 4 adds a new feature. We adopt a strategy of putting some material into boxes to ease the presentation. The material in the boxes typically contains technical details that will be of primary interest to the more advanced reader. The boxes provide supplementation to the text; readers who are new to regression analysis can skip over the boxed material without any loss of continuity. Boxed material is set apart by bold lines; boxes appear in the section in which the boxed material is referenced. Readers not interested in the technical details may simply skip the boxed material.

4.2 SOME USEFUL GRAPHICAL DISPLAYS OF THE ORIGINAL DATA

In regression analysis, the analyst should normally wish to take a careful look at both the original data and the residuals. In this section, we present graphical tools that are particularly useful in examining the original data. Reflecting the ease of producing graphical displays on

modern personal computers, our focus here is on using graphical tools as a fundamental part of the data-analysis process. These tools help display features of the distributions of each of the variables as well as their joint distributions, providing initial clues about the likely outcomes and potential problems of the regression analysis. Graphical displays can provide a more complete and more easily understandable portrayal of the data than typically reported summary statistics. Graphical displays also do not depend on assumptions about the form of the distribution or the nature of the relationship between the independent and dependent variables. These themes will be further considered in Section 4.4, where we introduce graphical tools for the examination of the residuals.

We begin Section 4.2 with univariate displays that describe single variables, then consider the scatterplot, which shows the relationship between two variables, and finally the scatterplot matrix, which simultaneously displays each possible pair of relationships between three or more variables. The basics of several of these graphical displays will be familiar to many readers from introductory statistics courses and Chapters 2 and 3 of this book. What will be new are several enhancements that can increase the information available from the plot. We will also use many of these displays as building blocks for graphical examination of data in later sections of this chapter.

4.2.1 Univariate Displays

Univariate displays present a visual representation of aspects of the distribution of a single variable. Data are typically portrayed in textbooks as having a normal or bell-shaped distribution. However, real data can be skewed, have multiple modes, have gaps in which no cases appear, or have *outliers*—atypical observations that do not fit with the rest of the data. Because each of the univariate displays portrays different aspects of the data (Fox, 1990; Lee & Tu, 1997), it is often helpful to examine more than one display to achieve a more complete understanding of one's data.

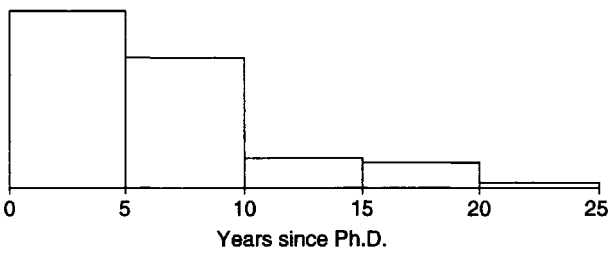
Frequency Histograms

Perhaps the most commonly used univariate display is the frequency histogram. In the frequency histogram, the scores on the variable are plotted on the x axis. The range of the scores is broken down into a number of intervals of equal width called *bins*. The number of scores in each interval, the frequency, is represented above the interval. Frequency histograms provide a rough notion of the central tendency, variability, range, and shape of the distribution of each separate variable. They also have some ability to identify outliers.

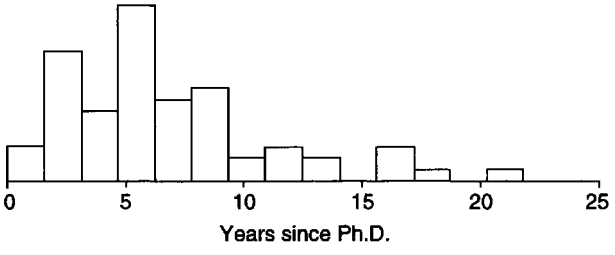
Figure 4.2.1 presents two different histograms of the variable years since Ph.D. for the faculty salaries data set ($n = 62$) originally presented in Table 3.5.1. Panel (A) uses 5 bins. This histogram depicts a single mode around 2.5 years since the Ph.D. (the midpoint of the interval) with a right skewed distribution. Panel (B) uses 20 bins. Here, the distribution still appears to be right skewed, but this histogram suggests that there may be two modes at about 2.5 years and 6 years post Ph.D., and that there may be an outlier at about 21 years post Ph.D. The shape of the distribution is a bit less apparent and gaps in the distribution (scores with a frequency of 0) are now more apparent.

The comparison of the two panels illustrates that histograms derived from the same data can give different impressions with respect to the distribution of the underlying data and the presence of outliers in the data set. The exact shape of the distribution will depend on two decisions made by the analyst or the statistical package generating the histogram: (1) The number of *bins*, which are intervals of equal width (e.g., 0–4, 5–9, 10–14, etc.), and (2) the range of the data represented by the histogram. With respect to (2), some histograms represent

(A) Five bins.



(B) Twenty bins.



Note: The two histograms portray the same data. The two graphical displays give somewhat different impressions of the underlying distribution. The gaps and outlying values in the data are distinct with 20 bins.

FIGURE 4.2.1 Histograms of years since Ph.D.

the range of possible scores, some represent the actual range of data, and some represent a convenient range (e.g., 0–25 years in the present case rather than the actual range of the data, 1–21 years). Statistical packages use one of several algorithms[1] to calculate the number of bins to be displayed, none of which assure that the frequency histogram will optimally represent all of the features of the distribution. With complicated distributions, it can often be useful to vary the number of bins and to examine more than one histogram based on the data. A large number of bins should be used to identify outliers. Some statistical packages now permit the analyst to easily vary the number of bins using a simple visual scale (a "slider") with the histogram being continuously redisplayed as the number of bins is changed.

Stem and Leaf Displays

Closely related to the histogram is the stem and leaf display. This display is of particular value in data sets ranging from about 20 to 200 cases. The stem and leaf display is the *only* graphical display that retains the values of the original scores so that the original data can be precisely reconstructed. Otherwise, the strengths and limitations of the stem and leaf display closely parallel those of the histogram presented above (see also Fox, 1990, pp. 92–93 for a more detailed discussion).

---

[1]One popular algorithm is Sturgis' rule—the number of bins is $1 + \log_2(N)$. For example, for $N = 62$, $1 + \log_2(62) = 1 + 5.95 \approx 7$ bins would be used. The interval width is chosen as the (maximum−minimum score + 1)/(number of bins). If for the 62 cases the highest score were 21 and the lowest score were 1, Sturgis' rule produces an interval width of $(21 - 2 + 1)/7 = 2.86$. When graphs are constructed by hand, easily interpretable interval widths are often chosen (e.g., interval width = 1, 2, 5, 10). For example, if interval width = 5 were chosen, the intervals would be 0–4, 5–9, 10–14, 15–19, and 20–24.

Stem	Leaves
0–4	11122333333333444444
5–9	55555555556666667777777888899999
10–14	0011133
15–19	6668
20–24	1

Note: Each stem represents the interval. Leaves represent the last digit of values of the observations within the interval. The number of times the digit is repeated corresponds to the frequency of the observation.

FIGURE 4.2.2 Stem and leaf display of years since Ph.D.

To illustrate, Fig. 4.2.2 presents a stem and leaf display for the years since Ph.D. data ($n = 62$) presented in Table 3.5.1. To construct this display, an interval width is initially chosen. We have chosen an interval width of 5 so that the display will be similar to the histogram with 5 bins presented in Fig. 4.2.1, Panel (A). The stem indicates the range of scores that fall in the interval.[2] The lowest interval (0–4) includes scores from 0 to 4 and the second from the highest interval (15–19) includes scores from 15 to 19. *Leaves* provide information about the exact values that fall in each interval. The frequency of occurrence of each score is indicated by repeating the value. For example, the lowest interval indicates no scores with a value of 0, three scores with a value of 1, two scores with a value of 2, nine scores with a value of 3, and six scores with a value of 4. The second to highest interval (15–19) has no scores with a value of 15, three scores with a value of 16, no scores with a value of 17, one score with a value of 18, and no scores with a value of 19.

Stem and leaf displays must be presented using a fixed-width typefont (e.g., Courier). The leading digit of the leaves (for example, 1 for 10–14 and 2 for 20–24) is dropped so that each number is represented as the same size in the display. If the stem and leaf display is rotated 90° counterclockwise so that the numbers form vertical columns, then the display depicts the same distribution as a histogram with the same number of bins. However, stem and leaf displays have the advantage of also representing the exact numerical values of each of the data points.

Smoothing: Kernel Density Estimates

A technique known as smoothing provides the foundation for an excellent visual depiction of a variable's underlying general frequency distribution. Often in the behavioral sciences the size of our samples is not large enough to provide a good depiction of distribution in the full population. If we were to take several samples from the same population and construct histograms or stem and leaf displays, there would be considerable variation in the shape of the distribution from sample to sample. This variation can be reduced by averaging adjacent data points prior to constructing the distribution. This general approach is called *smoothing*.

The simplest method of smoothing is the *running average* in which the frequencies are averaged over several adjacent scores. To illustrate, Table 4.2.1 presents the subset of data from Table 3.5.1 for the six faculty members who have values between 12 and 20 years

[2]The apparent limits of the interval are shown (e.g., 0–4 for first interval). Recall from introductory statistics that when the data can be measured precisely, values as low as −0.5 or up to 4.50 could fall in the first interval. These values are known as the real limits of the interval.

TABLE 4.2.1
Weights Based on the Bisquare Distribution for $X = 16$

	(A) $d = 4$			(B) $d = 3$		
X_i	f	$\left(\frac{X_i - X_C}{d}\right)$	W_i	f	$\left(\frac{X_i - X_C}{d}\right)$	W_i
20	0	1.0	0.00	0	1.33	0.00
19	0	0.75	0.19	0	1.00	0.00
18	1	0.50	0.56	1	0.67	0.31
17	0	0.25	0.88	0	0.33	0.79
16*	3	0.00	1.00	3	0.00	1.00
15	0	−0.25	0.88	0	−0.33	0.79
14	0	−0.50	0.56	0	−0.67	0.31
13	2	−0.75	0.19	2	−1.00	0.00
12	0	−1.00	0.00	0	−1.33	0.00

Note: X_i is the score, f is the frequency of the score, X_C is the location of the center of the smoothing window, d is the bandwidth distance, and W_i is the weight given to a score of X_i. W_i is calculated from the bisquare distribution. (A) provides the weights when the bandwidth distance = 4; (B) provides the weights when the bandwidth distance = 3. X_C is a score of 16 in this example, which is marked by an asterisk.

since the Ph.D. We identify a smoothing window over which the averaging should occur. The *smoothing window* is a symmetric range of values around a specified value of the variable. We identify a score of interest which establishes the center of the smoothing window, X_C. We then identify a bandwidth distance, d, which represents the distance from the center to each edge of the smoothing window. The width of the smoothing window is then $2d$. For our illustration of the calculation of a running average, we arbitrarily choose $X_C = 16$ and $d = 2$. The running average is calculated using all scores that fall in the smoothing window between $X_C - d$ and $X_C + d$. In our example the smoothing window only includes the scores from 14 to 18, so the width of the smoothing window is 4. For $X_C = 16$, the score marked by an asterisk in Table 4.2.1, we would average the frequencies for scores of 14, 15, 16, 17, and 18, so for $X_C = 16, f_{\text{avg}} = (0 + 0 + 3 + 0 + 1)/5 = 0.8$. For $X_C = 17$, we would average the frequencies for frequencies 15, 16, 17, 18, and 19, so for $X_C = 17, f_{\text{avg}} = (0 + 3 + 0 + 1 + 0) = 0.8$. Running averages are calculated in a similar fashion for each possible score in the distribution—we simply let X_C in turn equal the value of each possible score.

In practice, a more complex smoothing method known as the *kernel density estimate* is typically used because this method provides an even more accurate estimate of the distribution in the population. The kernel density estimate is based on a *weighted* average of the data. Within the smoothing window, the scores that lie close to X_C, the center of the smoothing window, are given a relatively high weight. Scores that are further from X_C characterize the smoothing window less well and are given a lower weight. Scores that lie outside the smoothing window are given a weight of 0. This method of smoothing results in a density curve, a continuous function whose height at any point estimates the relative frequency of that value of X_C in the population. The height of the density curve at any point is scaled so that the total area under the curve will be 1.0. Unlike the previous topics we have considered in this book, kernel density estimation requires very intensive calculation and is consequently only performed on a computer. Box 4.2.1 shows the details of the calculation for interested readers.

BOX 4.2.1
Inside the Black Box: How Kernel Density Estimates Are Calculated

To illustrate how a kernel density estimate is created at a single point, consider $X_C = 16$ in our distribution of years since Ph.D. in Table 4.2.1. We arbitrarily choose the bandwidth distance $d = 4$ so that the width of the smoothing window $= 8$. Recall that we wish to give scores at the center of the smoothing window a high weight and scores further from the center of the interval lower weights. Several different weight functions will achieve this goal; the bisquare weight function presented in Eq. (4.2.1) is commonly used.

(4.2.1) Bisquare $$W_i = \left[1 - \left(\frac{X_i - X_C}{d}\right)^2\right]^2 \quad \text{when} \quad |X_i - X_C| \leq d.$$

$$W_i = 0 \quad \text{when} \quad |X_i - X_C| > d.$$

Table 4.2.1 presents the values used in the calculation of the weights and shows the desired pattern of high weights at the center of the smoothing window and lower weights for scores further from the center of the smoothing window. For example, for $X_i = 17$,

$$W_i = \left[1 - \left(\frac{17 - 16}{4}\right)^2\right]^2 = \left[1 - (0.25)^2\right]^2 = (1 - 0.0625)^2 = 0.88.$$

Returning to our kernel density estimate, we can calculate its height at any value of X_C using the following equation:

(4.2.2) $$\text{height at } X_C = \frac{1}{nd} \sum_{i=1}^{n} f W_i.$$

In Eq. (4.2.2), the height of the density curve is calculated at our chosen value of X_C, here 16, n is the number of cases in the *full* sample (here, 62 cases), d is bandwidth distance, and f is the frequency of cases at X_i. W_i is the weight given to each observation at X_i. The weight for X_1 is determined by the value of the bisquare function applied to $(X_i - X_C)/d$.

Let us apply Eq. (4.2.2) to the data in Table 4.2.1A for $d = 4$. The score is given in column 1, the frequency in column 2, $(X_i - X_C)/d$ in column 3, and the weight from the bisquare weight function (Eq. 4.2.1) in column 4. Beginning at $X = 12$ and continuing to $X = 20$, we get

$$\text{height} = \frac{1}{(62)(4)}[(0)(.00) + (2)(.19) + 0(.56) + 0(.88) + 3(1.0) + 0(.88) + 1(.56) + 0(.19) + (0)(.00)] = .03.$$

In practice, the statistical package calculates the value of the height for *every* possible value of X_C over the full range of X, here $X = 1$ to $X = 21$, and produces the kernel density estimate, which is a smooth curve that estimates the distribution of X in the population.

In Table 4.2.1B, we also provide the values of W_i if we choose a smaller bandwidth, here $d = 3$. As can be seen, the weight given to each observed score declines more quickly as we move away from the X_C of interest. For example, when $d = 3$, the weight for a score of 18 is 0.31, whereas when $d = 4$ the weight for a score of 18 is 0.56. The analyst controls the amount of smoothing by selecting the bandwidth d, with larger values of d producing more smoothing.

The central problem for the analyst is to decide how big the width of the smoothing window should be. Some modern statistical packages allow the analyst to change the size of the smoothing window using a visual scale (slider), with the resulting kernel density estimate being continuously redisplayed on the computer screen. Simple visual judgments by the analyst are normally sufficient to choose a reasonable window size that provides a good estimate of the shape of the distribution in the population. Figure 4.2.3 provides an illustration of the effect of choosing different bandwidth distances for the full years since Ph.D. data ($n = 62$) originally presented in Table 3.5.1. Figure 4.2.3(A) depicts bandwidth = 10, in which oversmoothing has occurred and features of the data are lost (e.g., the mode near the score of 5 is not distinct). Figure 4.2.3(B) depicts bandwidth = 1.5, in which too little smoothing has occurred and the distribution is "lumpy." Figure 4.2.3(C), with bandwidth = 4, provides a smooth curve that appears to depict the major features of the data. Good kernel density plots include enough detail to capture the major features of the data, but not to the extent of becoming "lumpy." Figure 4.2.3(D) shows a kernel density estimate as an overlay over a histogram so that both depictions of the same data set are available.

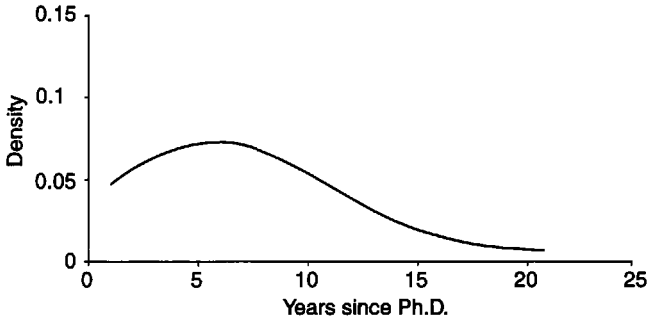
The great strength of kernel density plots is their ability to depict the underlying distribution of the data in the population. However, kernel density estimates do not reproduce the information about the data in the original sample and do not clearly portray unusual observations or gaps in the distribution over which the scores have a frequency of 0.

Boxplots

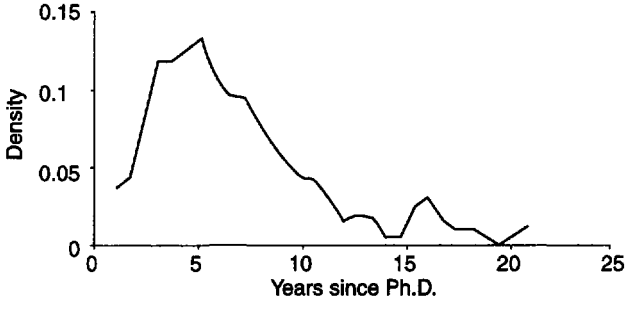
Another commonly used univariate display is the boxplot, also known as the "box and whiskers plot." Figure 4.2.4 displays a boxplot of the years since Ph.D. data. Note that the values of the scores, here years since Ph.D., are represented on the y axis. The center line of the box at a value of 6 is the median of the distribution. The upper edge of the box at a value of about 8.5 is the third quartile, Q_3 (75% of the cases fall at or below this value); the lower edge of the box at a value of about 4 is the first quartile, Q_1 (25% of the cases fall at or below this value). The semi-interquartile range, *SIQR*, is $(Q_3 - Q_1)/2 = (8.5 - 4)/2 = 2.25$, or half the distance between the upper and lower edges of the box. The *SIQR* is a measure of the variability of the distribution that is useful when the distribution is skewed. When there are no outlying observations, the two vertical lines (termed whiskers) extending from the box represent the distance from Q_1 to the lowest score (here, 1) and Q_3 to the highest score in the distribution. Values of any outlying scores are displayed separately when they fall below $Q_1 - 3SIQR$ or above $Q_3 + 3SIQR$. If the distribution were normal, these scores would correspond to the most extreme 0.35% of the cases. In Fig. 4.2.4 the value of the horizontal line corresponding to the end of the top whisker is $Q_3 + 3SIQR$ (here, about 14). The open circles above this line corresponding to 16, 18, and 21 years are outlying cases. If one whisker is long relative to the other or there are outlying values on only one side of the distribution, the distribution will be skewed in that direction. Figure 4.2.4 depicts a positively skewed distribution.

The boxplot provides a good depiction of the skewness of the distribution and clearly identifies the existence, but not the frequency, of outlying observations in the data set. When multiple cases occur at a single outlying score (here at $X = 16$, $f = 3$), a problem known as *overplotting* occurs. Standard computer programs plot one case on top of the other, and the multiple cases cannot be discerned. The boxplot also provides clear information about the range and the median of the data in the sample. However, the boxplot does not portray the existence of more than one mode or gaps in the distribution over which the scores have a frequency of 0.

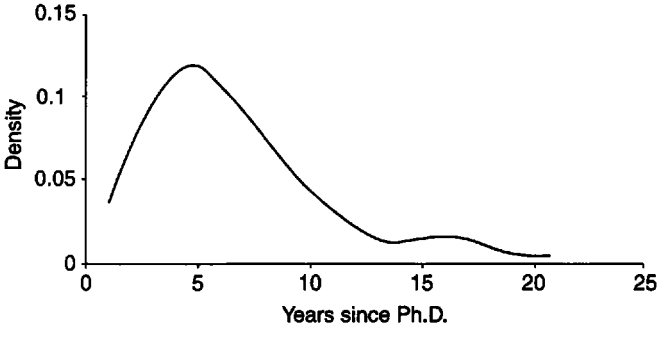
(A) Width = 10: too much smoothing.



(B) Width = 1.5: too little smoothing.



(C) Width = 4: appropriate smoothing.



(D) Histogram with kernel density superimposed (combines Fig. 4.2.1B with Fig. 4.2.3C).

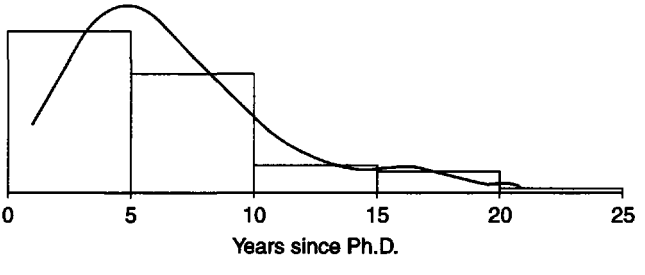


FIGURE 4.2.3 Kernel density estimates: years since Ph.D.

Note: From bottom to top, the first four horizontal lines in the figure represent the lowest score (year = 1), the first quartile (years = 4), the median (years = 6), and the third quartile (Q_3, years = 8.5). Because there are high outliers in the data set, the top line represents $Q_3 + 3SIQR$ (years = 14). The outlying cases are plotted as separate points (here, at years = 16, 18, and 21).

FIGURE 4.2.4 Boxplot of years since Ph.D.

Comparisons With the Normal Distribution

Researchers may wish to compare the univariate distribution of their sample data with a normal distribution. The regression model that we emphasize in this volume makes no assumption about the distribution of the independent or dependent variables. However, as we will present in Section 4.3, this model does make the assumption that the residuals are normally distributed. Normal curve overlays and normal q-q plots permitting comparison of the residuals with the normal distribution are presented in Section 4.4.6. These plots can be applied to the original data as well. For example, these plots can be particularly useful in the context of structural equation modeling with latent variables (see Chapter 12), a technique which assumes that each measured variable has a normal distribution.

4.2.2 Bivariate Displays

We will often wish to examine the relationship between two variables. These can be two independent variables, X_1 and X_2, or one of the independent variables and the dependent variables, for example, X_1 and Y. As we have already seen in Chapters 2 and 3, scatterplots present an excellent way of doing this. In a scatterplot, variable 1 is plotted on the x axis and variable 2 is plotted on the y axis. Note that scatterplots are not symmetric: Unless $sd_Y = sd_X$, the scatterplot of variable 2 (on the y axis) versus variable 2 (on the x axis) will have a different appearance than the scatterplot of variable 1 (on the y axis) versus variable 2 (on the x axis). Recall from Chapter 2 that the regression of Y on X and the regression of X on Y are different unless the standard deviations of X and Y are equal.

Figure 4.2.5(A) presents a scatterplot of years since Ph.D. versus salary. In this figure the salaries are large numbers, so they are printed in scientific notation, a useful method of compactly representing large numbers.[3] The minimum value on the y axis is $2e + 04 = 2 \times 10^4 = \$20{,}000$. The maximum value is $1e + 05 = 1 \times 10^5 = 1 \times 100{,}000 = \$100{,}000$. Scatterplots help us detect whether the relationship between each X and Y is linear or takes some other form.

One special problem in interpreting scatterplots occurs when one of the variables is categorical. For example, we may wish to compare the years since Ph.D. of male and female faculty members in our sample. This relationship is depicted in Fig. 4.2.5(B), where males and females are represented by values of 0 and 1, respectively. A problem arises with the graph because cases having the same value on Y, here years since Ph.D., are plotted on top of each other and cannot be discerned—another instance of overplotting.[4] This problem can be reduced by adding a small random value to each case's score on X, which helps spread out the points, a technique known as *jittering*. Figure 4.2.5(C) presents the same data after the points have been jittered. Note that points at the same value of Y have been spread out, making it much easier to get a sense of the distribution of cases within each group.

Our understanding of the X–Y relationship can be improved by superimposing lines on the scatterplot. Figure 4.2.5(D) returns to the data depicted in Fig. 4.2.5(A) and superimposes the best fitting regression line, $\hat{Y} = 1379X + 45{,}450$. This plot suggests that a straight line provides a good characterization of the data. Figure 4.2.5(E) superimposes a *lowess* fit line[5] representing the best nonparametric fit of the X–Y relationship. Lowess is a method of producing a smooth line that represents the relationship between X and Y in a scatterplot. The lowess method makes no assumptions about the form of the relationship between X and Y. It follows the trend in the data instead of superimposing a line representing a linear or some other specified mathematical relationship. If this relationship is linear in the population, the lowess line should look like a very rough approximation of a straight line. Perhaps an apt metaphor is that it will look like a young child's freehand drawing of a straight line. To provide a contrast, Fig. 4.2.5(F) illustrates a case in which a straight line does not characterize the data (we will consider nonlinear relationships in Chapter 6). Lowess is very computer intensive and is calculated only using a computer. For interested readers we have presented the details of how lowess is calculated in Box 4.2.2.

As with kernel density estimates, the central problem for the analyst is to choose an appropriate value of the smoothing parameter α. This parameter represents the proportion of the data that is included in the smoothing window. Higher values of α produce more smoothing. Good lowess lines include enough detail to capture the major features of the data, but not to the extent of becoming "lumpy." Simple visual judgments by the analyst are normally sufficient to choose a reasonable value of α that provides a good estimate of the relationship between X and Y. Figure 4.2.5(F), (G), and (H) provide an illustration of the effect of choosing different values of α depicting three different lowess fits to the nonlinear data that will be presented in Chapter 6. Figure 4.2.5(G), with $\alpha = .05$, does not provide enough smoothing, and the X–Y relationship appears "lumpy." Figure 4.2.5(H), with $\alpha = .95$, provides too much smoothing such that the lowess line exceeds the observed values of Y for the lowest values of X. In contrast, Fig. 4.2.5(F)

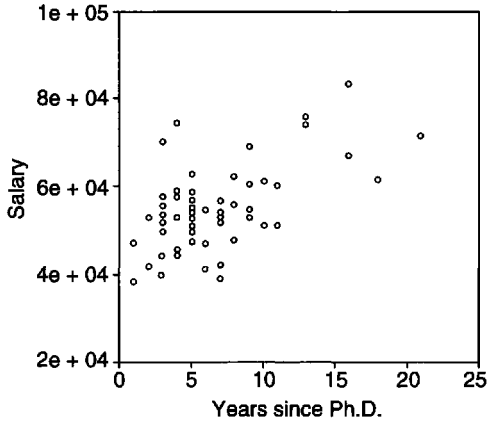
---

[3]Another useful method for the present example would be to divide each person's salary by 1,000. The values then represent salary in thousands of dollars.

[4]Some programs (e.g., SPSS) print different symbols (e.g., 1, 2, 3) representing the number of cases that are overplotted.

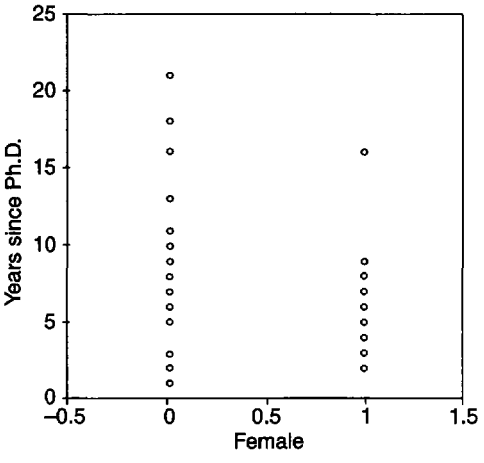
[5]Lowess has become an acronym for *lo*cally *we*ighted *s*catterplot *s*moother. In his original writings describing the technique, Cleveland (1979) used the term *loess*.

(A) Basic scatterplot: salary vs. years since Ph.D.



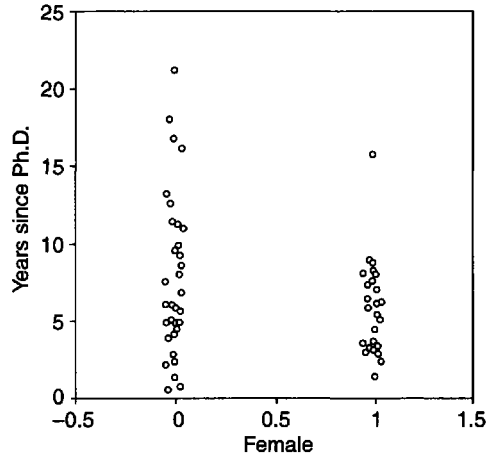
Note: Each point represents one case ($n = 62$). Salary is presented in scientific notation (e.g., 2e + 04 = $20,000).

(B) Basic scatterplot: years since Ph.D. vs. female.



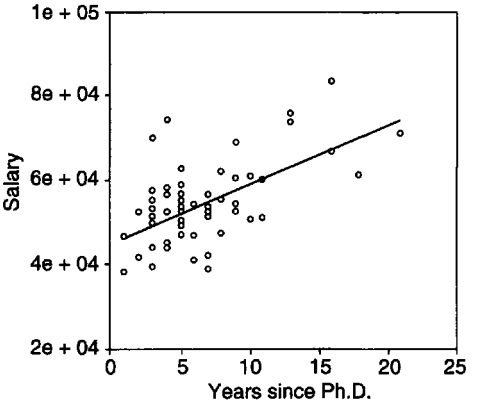
Note: 0 = male; 1 = female. Some points are overplotted.

(C) Jittered scatterplot: years since Ph.D. vs. female.



Note: Each case is now distinct following jittering. In jittering a small random value is added to or subtracted from each case's score on the categorical variable (here, 0 or 1).

(D) Salary vs. years since Ph.D. superimposed regression line.



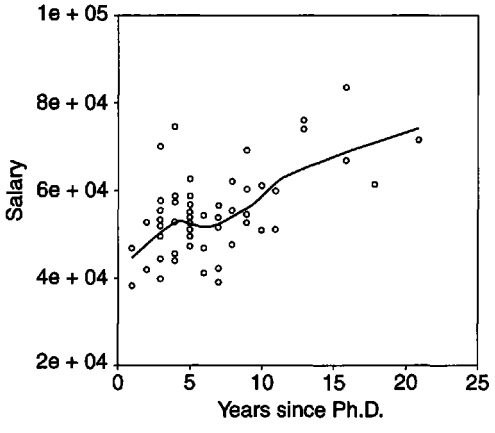
Note: Regression line is $\hat{Y} = 1379X + 45{,}450$.

FIGURE 4.2.5 Scatterplots and enhanced scatterplots.

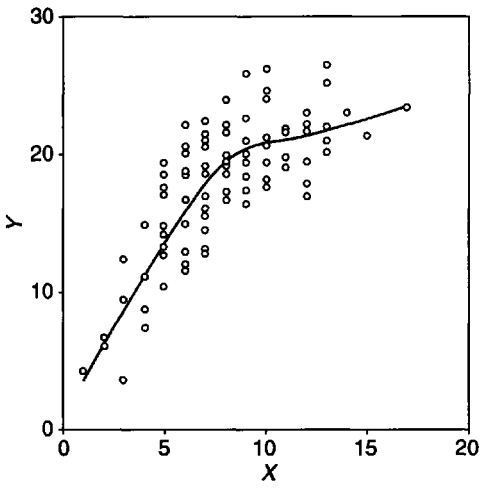
with $\alpha = .60$ produces a smooth line that appears to adequately represent the X–Y relationship.

Analysts normally choose values of α that range between about 0.25 and 0.90. Some modern computer programs are beginning to allow the analyst to vary the value of α using a slider, with the original data and the lowess line being continuously redisplayed on the computer screen. Other programs (e.g., SPSS 10) require the analyst to specify a series of different

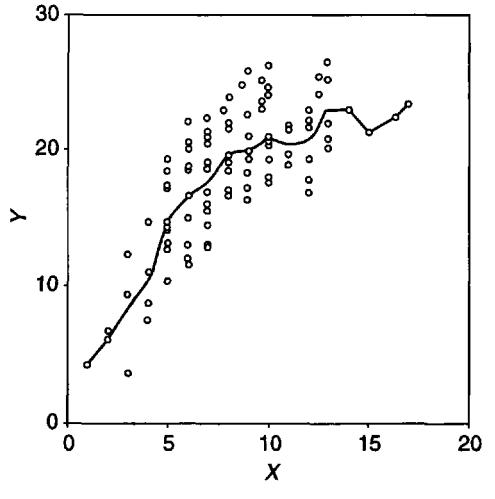
(E) Superimposed lowess fit: salary vs. years since Ph.D.



(F) Superimposed lowess fit $\alpha = 0.6$: nonlinear relationship.



(G) Lowess fit $\alpha = .05$: too little smoothing.



(H) Lowess fit $\alpha = .95$: too much smoothing.

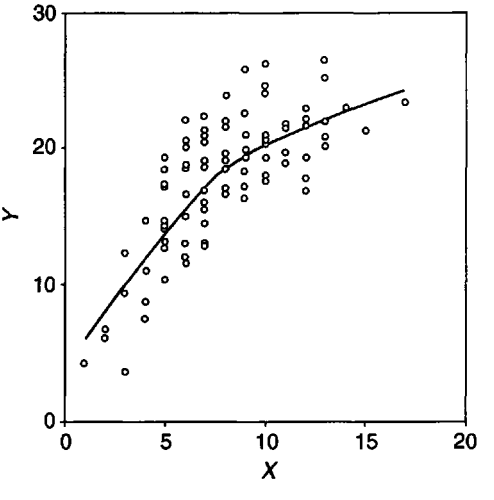


FIGURE 4.2.5 (Continued)

values of α and compare the results. Smaller values of α can be chosen as sample size increases because more cases will be included in each window that is used in the calculation of the lowess line. In addition, a smaller value of α should be chosen if the relationship between X and Y has many peaks and valleys. The lowess line provides a good overall summary of the relationship between X and Y. However, the lowess estimates at the ends of the distribution of X (where the data are often sparse) may be far less precise than in the center.

BOX 4.2.2
Inside the Black Box: How Lowess Lines Are Computed

As an illustration of the computation of lowess, consider again the data for the 62 cases originally presented in Table 3.5.1. We wish to plot a lowess curve for the relationship between years since Ph.D. on the x axis and salary on the y axis. We illustrate the calculation of the predicted lowess value using only one value of X. In practice, the computer calculates the values for hundreds of different possible values of X within the actual range of the data.

Suppose we wished to compute the predicted lowess value of Y corresponding to an arbitrarily chosen value of time since Ph.D. of 16. We must first define a symmetric smoothing window centered around our central score $X_C = 16$. Rather than using a fixed bandwidth, lowess defines its smoothing window so that there is a constant number of cases in the smoothing window, n_{window}. The smoothing parameter α specifies the proportion of cases that are to be used, $n_{\text{window}} = \alpha n$. The smoothing window becomes narrower in regions where there are many cases and wider in regions where data are sparse.

We will arbitrarily[6] choose a value of $\alpha = .1$, which leads to $(62)(.1) = 6$ cases being included in the smoothing window. We identify those 6 cases that are closest to $X_C = 16$ regardless of whether they are higher or lower in value. In Table 4.2.1, we see that a symmetrical smoothing window from $X = 13$ to $X = 19$ contains 6 cases. For this smoothing window, $d = 3$. We then calculate the bisquare function weights for these cases using Eq. (4.2.1). The resulting bisquare weights are shown in column 7 of Table 4.2.1 for $d = 3$. Once again, scores at the center of the window ($X_C = 16$) are given a weight of 1, whereas scores further from the center of the window have a lower weight.

Lowess now estimates a regression equation $\hat{Y} = B_1X + B_0$ for the six cases in the smoothing window. The bisquare function weights are used in determining the fit of the regression line. A method known as weighted least squares regression (to be presented in Section 4.5) is used, which gives cases further from the center of the smoothing window less importance in determining the fit of the regression line. Once the regression equation for the six cases in the smoothing window is determined, the value of $\hat{Y}$ for the lowess line is calculated by substituting the value of the center point, here $X_C = 16$, into the equation. The computer program then calculates the value of $\hat{Y}$ for a large number of different values over the full range of X, here $X = 1$ to $X = 21$. That is, the window is moved along X and a large number (e.g., 100–200) of values of X serve in turn as X_C. For each value of X_C, those six cases that fall closest to X_C are used to define the width of the smoothing window and to compute each separate local regression equation. The value of X_C is then substituted into the local regression equation, giving the lowess value of $\hat{Y}$ corresponding to the value of X_C. The $\hat{Y}$ values are connected to produce the lowess line.

Other, more complicated variants of lowess fitting may be used when the relationship between X and Y is very complicated (e.g., there are sharp peaks and valleys in the X–Y relationship) or when there are extreme outlying observations on Y. Accessible introductions to these more advanced lowess techniques[7] can be found in Cleveland (1993) and Fox (2000a).

[6]This value is chosen only to simplify our presentation of the computations. In practice, a much larger value of α (e.g., .6 to .8) would be chosen for these data.

[7]Alternatives to lowess such as cubic splines and kernel smoothers (Fox, 2000a; Silverman, 1986) exist that can also be used to produce lines representing good nonparametric fits of the X–Y relationship.

4.2.3 Correlation and Scatterplot Matrices

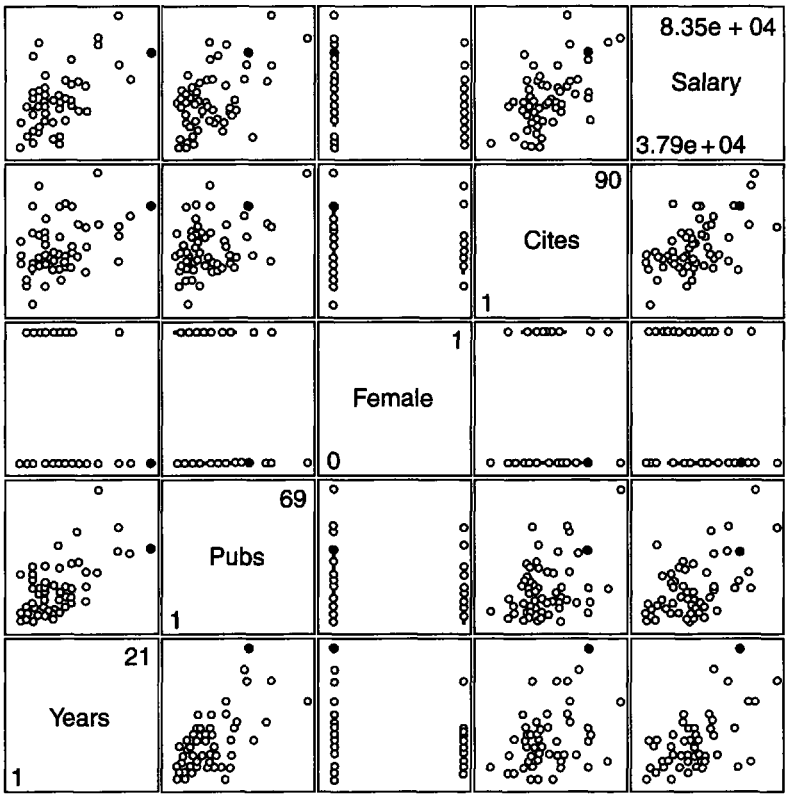
Correlation matrices present a convenient method of summarizing the correlations between each pair of predictors as well as the correlation between each predictor and the dependent variable, thus providing considerable information on the direction and magnitude of the *linear* relationships among the variables. Table 4.2.2 presents the correlation matrix for the five variables in our faculty salary example presented in Section 3.5. To illustrate, row 2, column 1, presents the correlation between publications and years since Ph.D. ($r = .65$) and row 4, column 5 presents the correlation between citations and salary ($r = .55$). We see from the correlation matrix that, although years, publications, and citations are all strongly related to salary, some of the independent variables are themselves highly intercorrelated so that they may not provide substantial unique prediction of salary over and above that of the other independent variables.

An improvement on the correlation matrix is the scatterplot matrix, which provides a graphical display of the scatterplot for each pair of variables. The scatterplot matrix corresponding to the faculty salary data is presented in Fig. 4.2.6. Each row and column of the scatterplot matrix forms a cell. Within each cell is a scatterplot depicting the relationship between two of the variables. The variable identified in the row is depicted on Y and the variable identified in the column is depicted on X. For example, row 1, column 1 depicts the regression of salary on the y axis on years since Ph.D. on the x axis. This cell is identical to the scatterplot presented in Fig. 4.2.5(A). Row 1, column 4 depicts the regression of salary on the y axis on number of citations on the x axis. Row 2, column 3 depicts the regression of number of citations on the y axis on female (male = 0; female = 1) on the x axis. The present illustration identifies each of the variables and its range on the minor diagonal of the matrix going from the lower left to upper right corner. Some versions of scatterplot matrices present a histogram of each variable on the diagonal (e.g., SYSTAT, which terms this graphical display a SPLOM, Tukey's term for *scatterplot matrix*).

The scatterplot matrix provides a compact visual summary of the information about the relationship between each pair of variables. First, the analyst can make visual judgments about the nature of the relationship between each pair of variables. Unlike the correlation matrix, which only represents the direction and magnitude of the linear relationship, the scatterplot matrix helps the analyst visualize possible nonlinear relationships between two variables. Strong nonlinear relationships between either two independent variables or an independent and a dependent variable suggest that the linear regression model discussed in previous chapters may not be appropriate. In addition, cases in which the variance of Y is not constant but changes as a function of the value of X can also be observed. Any of these problems would lead the analyst to consider some of the remedies considered later in Section 4.5. Second, the panels of the scatterplot matrix can be linked in some personal computer–based statistical packages.

TABLE 4.2.2
Correlation Matrix for Faculty Salary Example

	Years	Publications	Sex	Citations	Salary
Years	1.00	0.65	0.21	0.37	0.61
Publications	0.65	1.00	0.16	0.33	0.51
Sex	0.21	0.16	1.00	0.15	0.20
Citations	0.37	0.33	0.15	1.00	0.55
Salary	0.61	0.51	0.20	0.55	1.00



Note: Each cell of the scatterplot matrix represents a separate scatterplot. For example, row 4, column 1 has years since Ph.D. on the x axis and number of publications on the y axis. The dark point that appears in each cell of the scatterplot matrix is the case with the highest value on time since Ph.D. (years = 21).

FIGURE 4.2.6 Faculty salary example: scatterplot matrix.

Individual cases can then be selected and highlighted in *all* panels of the matrix. This feature is particularly important in the examination of outlying cases. For example, in Fig. 4.2.6, we have highlighted the faculty member (the dark filled circle) who has completed the largest number of years of service—21 years since the Ph.D. From the other panels of the scatterplot, we see that this person is a male who has a relatively high number of publications and a high number of citations. This feature of being able to link single or multiple cases across different panels of a scatterplot matrix can be very useful in understanding data sets involving several variables[8] (Cleveland, 1993).

CH04EX01

[8]Computer code is only provided for the base system for each computer package. Not all of the features described in this chapter are presently available in each of the packages. The graphical capabilities of each of the packages is changing rapidly, with new features coming on line every few months. SAS users may wish to investigate the additional capabilities of SAS/INSIGHT. ARC, an outstanding freeware regression and interactive graphics package, is described in Cook and Weisberg (1999) and is downloadable from the School of Statistics, University of Minnesota, Twin Cities: http://stat.umn.edu/arc/

4.3 ASSUMPTIONS AND ORDINARY LEAST SQUARES REGRESSION

All statistical procedures including multiple regression require that assumptions be made for their mathematical development. In this section we introduce the assumptions underlying the linear regression models presented in the previous chapters. Of importance, violation of an assumption may potentially lead to one of two problems. First and more serious, the estimate of the regression coefficients may be biased. *Bias* means that the estimate based on the sample will not on average equal the true value of the regression coefficient in the population. In such cases, the estimates of the regression coefficients, R^2, significance tests, and confidence intervals may all be incorrect. Second, only the estimate of the standard error of the regression coefficients may be biased. In such cases, the estimated value of the regression coefficients is correct, but hypothesis tests and confidence intervals may not be correct.

Violations of assumptions may result from problems in the data set, the use of an incorrect regression model, or both. Many of the assumptions focus on the residuals; consequently, careful examination of the residuals can often help identify problems with regression models. In Section 4.4, we present both graphical displays and statistical tests for detecting whether each of the assumptions is met. We particularly focus on graphical displays because they can detect a wider variety of problems than statistical tests. We then provide an introduction in Section 4.5 to some remedial methods that can produce improved estimates of the regression coefficients and their standard errors when the assumptions underlying multiple regression are violated.

4.3.1 Assumptions Underlying Multiple Linear Regression

We focus on the basic multiple linear regression equation with k predictors originally presented in Chapter 3,

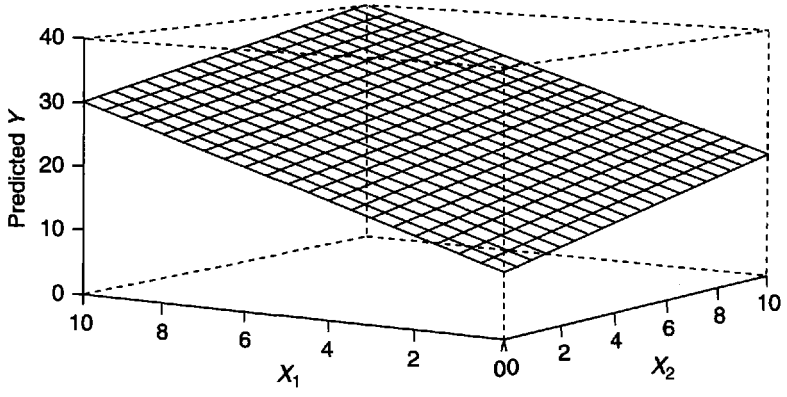
$$Y = B_1X_1 + B_2X_2 + B_3X_3 + \cdots + B_kX_k + B_0 + e. \tag{4.3.1}$$

The assumptions presented here and their effects on estimates of regression coefficients and their standard errors also apply to most of the more complex regression models discussed in later chapters.

Correct Specification of the Form of the Relationship Between IVs and DVs

An important assumption of multiple regression is that we have properly specified the *form* or mathematical shape of the relationship between Y and each of the IVs in the population. In Chapters 2 and 3 we have consistently assumed that all relationships are linear (straight line). To illustrate the meaning of this assumption in the one IV case, imagine we could identify the set of all cases in the population with a specific value of X_1, for example $X_1 = 5$, and compute the mean of their Y scores. This mean is called the *conditional mean* of Y given X_1, $\mu_{Y|X_1}$. If the assumption of linearity is correct, then each of $\mu_{Y|X_1}$ values that resulted as X_1 took on different values would fall precisely on a straight line, $\hat{Y} = B_1X_1 + B_0$. The slope B_1 of the straight line will be constant across the full range of X_1.

These same ideas apply to regression equations with more than one IV. Figure 4.3.1 provides an illustration of the regression plane when there are two independent variables. We imagine selecting all cases in the population with a specified value of X_1 and a specified value of X_2. We calculate the conditional mean value of Y given the specified values of X_1 and X_2, $\mu_{Y|X_1,X_2}$. The conditional mean in the population must fall exactly on the regression plane



Note: The equation for the linear regression plane is $Y = B_0 + 2X_1 + 1X_2$. B_1 has the same value (here, $B_1 = 2$) regardless of the value of X_2. B_2 has the same value (here $B_2 = 1$) regardless of the value of X_1.

FIGURE 4.3.1 Linear regression plane with two IVs.

$\hat{Y} = B_1X_1 + B_2X_2 + B_0$ for all possible values of X_1 and X_2 if the regression model is properly specified.

To understand the meaning of the regression plane, look at Fig. 4.3.1. The intercept, $B_0 = 10$, is located at the front corner of the plane and represents the predicted value of Y when $X_1 = 0$ and $X_2 = 0$. The line representing $B_{Y1.2}$ is the edge of the plane above the axis labeled X_1. This line goes from $\hat{Y} = 10$ when $X_1 = 0$ to $\hat{Y} = 30$ when $X_2 = 10$. Recall that $B_{Y1.2}$ represents the predicted change in Y for a 1 unit change in X_1 when X_2 is held constant (here, $X_2 = 0$), so $B_{Y1.2} = 2$. The other lines in the plane that are parallel to this edge line represent the regression lines that result when X_2 takes on different values between 0 and 10 (0, 0.5, 1.0, . . . , 9.5, 10). Each of these lines has the identical slope, $B_{Y1.2} = 2$. Now look at the edge of the regression plane above the axis labeled X_2. The line representing $B_{Y2.1}$ is the edge of the plane where $X_1 = 0$. This line goes from $\hat{Y} = 10$ when $X_2 = 0$ to $\hat{Y} = 20$ when $X_2 = 10$. The slope of this regression line is $B_{Y2.1} = 1$. The other lines in the plane that are parallel to this edge line represent the regression lines that result as X_1 takes on different values between 0 and 10. Each of these lines has the same identical slope, $B_{Y2.1} = 1$. The condition illustrated here of having linear relationships between each of the IVs and the DV is known as *linearity in the variables.*

Not all IVs have linear relationships to the DV. Beginning with Chapter 6, we will consider regression models that specify a variety of nonlinear relationships between the IVs and DV. For example, if we have a curvilinear (quadratic) relationship in which Y is low at low values of X, high at moderate values of X, and low at high values of X, the relationship between Y and X cannot be properly represented with a linear regression equation. The slope of the curve will change as the value of X changes. Chapter 6 will consider how to build terms into the regression so that they properly specify the relationship between Y and each of the IVs.

When the form of the relationship between the IVs and the DV in the population is not properly specified, severe problems may result. The estimates of both the regression coefficients and standard errors may be biased, resulting in incorrect significance tests and incorrect confidence intervals. This conclusion applies to regression models that are linear in the variables

that we have discussed in Chapters 2 and 3. It also applies to regression models specifying nonlinear relationships between the IVs and DV that we will consider in later chapters.

Correct Specification of the Independent Variables in the Regression Model

The second assumption is related to the first but focuses on the IVs in the regression model. If we presume that the theory we are testing is correct, then correct specification implies that all variables identified by the theory are included in the model, that they are properly measured (see next section on measurement error), and that the form of the relationship between each IV and DV has been properly specified. If these conditions are met, then each of the IVs and the residuals will be independent in the population and the estimates of regression coefficients will be unbiased. Of course, if any of these assumptions is not correct, then the IVs and the residuals will not be independent and the estimates of the regression coefficients and standard errors may be biased. This result also implies that significance tests and confidence intervals will not be correct. We consider these issues in more detail in our discussion of specification error in Section 4.4.

No Measurement Error in the Independent Variable (Perfect Reliability)

A third closely related assumption is that each independent variable in the regression equation is assumed to be measured without error. Recall from Section 2.10.2 that reliability is defined as the correlation between a variable as measured and another equivalent measure of the same variable. When there is no error of measurement in X, the reliability $r_{XX} = 1.0$. In practice, measures in the behavioral sciences differ in the magnitude of their reliabilities. Measures of some demographic variables such as age and gender typically have very close to perfect reliabilities, measures of adult abilities such as IQ typically have reliabilities in the range of about .80 to .95, and measures of attitudes and personality traits typically have reliabilities in the range of about .70 to .90.

When the assumption of no measurement error in the independent variable (perfect reliability) is violated, we saw in Section 2.10.2 that the estimate of r_{YX} will be biased. When there is only one IV in the regression equation, all of the indices of partial relationship between X and Y including B, standardized β, sr, and pr will be attenuated (too close to 0 regardless of the sign). Otherwise stated, the strength of prediction, R^2, will always be underestimated. When there are two or more IVs that are not perfectly reliable, the value of each measure of partial relationship including B, standardized β, sr, and pr will most commonly be attenuated. However, there is no guarantee of attenuation given measurement error—the value of a specific measure of partial relationship may be too low, too high, or even on rare occasions just right. Thus, measurement error commonly leads to bias in the estimate of the regression coefficients and their standard errors as well as incorrect significance tests and confidence intervals. We include a more detailed presentation of the effects of unreliability in multiple regression in Box 4.3.1.

Constant Variance of Residuals (Homoscedasticity)

For any value of the independent variable X, the *conditional variance* of the residuals around the regression line in the population is assumed to be constant. Conditional variances represent the variability of the residuals around the predicted value for a specified value of X. Imagine we could select the set of cases that have a specified value of X in the population (e.g., $X = 5$). Each of these cases has a residual from the predicted value corresponding to

the specified value of X, $e_i = Y_i - \hat{Y}_i$. The variance of the set of residuals is the conditional variance given that X_i equals the specified value. These conditional variances are assumed to be constant across all values of X in the population. Otherwise stated, the variance of the residuals around the regression line is assumed to be constant regardless of the value of X. When the assumption of constant variance of the residuals regardless of the value of X is met,[9] this condition is termed *homoscedasticity*. When the variance changes as the value of X changes, this condition is termed *heteroscedasticity*. When there is heteroscedasticity, the estimates of the regression coefficients remain unbiased, but the standard errors and hence significance tests and confidence intervals will be incorrect. In practice, the significance tests and confidence intervals will be very close to the correct values unless the degree of nonconstant variance is large. A rule of thumb for identifying a large degree of nonconstant variance is that the ratio of the conditional variances at different values of X exceeds 10.

Independence of Residuals

The residuals of the observations must be independent of one another. Otherwise stated, there must be no relationship among the residuals for any subset of cases in the analysis. This assumption will be met in any random sample from a population. However, if data are clustered or temporally linked, then the residuals may not be independent. *Clustering* occurs when data are collected from groups. For example, suppose a set of groups such as university residence halls, high schools, families, communities, hospitals, or organizations are first selected, then a random sample is taken from each group. In such cases, the responses of any two people selected from within the same group (e.g., fraternity A) are likely to be more similar than when the two people are selected from two different groups (e.g., fraternity A; honors dorm B). Similarly, in designs that repeatedly measure the same person or group of persons over time, responses that are collected from the same person at adjacent points in time tend to be more similar than responses that are collected from the same person at more distant points in time. This issue commonly occurs in studies of single individuals (single-subject designs) or in panel studies in which a group of participants is measured on the independent and dependent variables at several time points. For example, if we measure stressful events and mood in a sample of college students each day for two months, the similarity of mood from one day to the next will be greater than the similarity of mood from one day to a day two weeks later. Nonindependence of the residuals does *not* affect the estimates of the regression coefficients, but it does affect the standard errors. This problem leads to significance tests and standard errors which are incorrect.

Normality of Residuals

Finally, for any value of the independent variable X, the residuals around the regression line are assumed to have a normal distribution. Violations of the normality assumption do not lead to bias in estimates of the regression coefficients. The effect of violation of the normality assumption on significance tests and confidence intervals depends on the sample size, with problems occurring in small samples. In large samples, nonnormality of the residuals does not lead to serious problems with the interpretation of either significance tests or confidence intervals. However, nonnormal residuals are often an important signal of other problems in the regression model (e.g., *misspecification*–using an incorrect regression model) and can help guide appropriate remedial actions.

---

[9] In the multiple IV case, the variance of the residuals should not be related to any of the IVs or to $\hat{Y}$.

BOX 4.3.1
The Effects of Measurement Error in Two-Predictor Regression

Measurement error is a common and important problem in regression analysis. To understand more fully the potential effects of measurement error on the results of multiple regression analysis, it is very informative to study what happens when one of the variables is unreliable in the two-IV case. We focus initially on the partial correlation because its relation to the standardized effect size makes it useful in many applications.

Recall from Chapter 3 that we can calculate the partial correlation between Y and X_2 holding X_1 constant from the simple correlations using Eq. (3.3.11):

(3.3.11) $$pr_2 = r_{Y2\cdot1} = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{(1 - r_{Y1}^2)(1 - r_{12}^2)}}.$$

Note in Eq. (3.3.11) that X_1 is the IV that is being partialed. As it will turn out, the reliability of the IV being partialed is of particular importance.

To illustrate the use of Eq. (3.3.11), we can calculate $r_{Y2\cdot1}$ if $r_{12} = .3, r_{Y2} = .4$, and $r_{Y1} = .5$,

$$r_{Y2\cdot1} = \frac{.4 - (.5)(.3)}{\sqrt{(1 - .5^2)(1 - .3^2)}} = .30$$

We define $r_{Y_tX2_t\cdot X1_t}$ as the partial correlation of the true score Y_t with the true score X_{2_t} with the true score[10] X_{1_t} partialed out. If all of the variables have perfect reliability, $r_{Y2\cdot1}$ will be identical to $r_{Y_tX2_t\cdot X1_t}$.

Now suppose that one of the variables is measured with error. What would the partial correlation have been if the one fallible variable had been measured with perfect reliability? In Chapter 2 we showed that we can express a correlation between the true scores X_t and Y_t in terms of the reliabilities of X and Y and the correlation between the measured variables X and Y,

(2.10.5) $$r_{X_tY_t} = \frac{r_{XY}}{\sqrt{r_{XX}r_{YY}}}.$$

If measurement error occurs only in X, Eq. (2.10.5) simplifies to

(4.3.2a) $$r_{X_tY_t} = \frac{r_{XY}}{\sqrt{r_{XX}}};$$

if measurement error occurs only in Y, Eq. (2.10.5) simplifies to

(4.3.2b) $$r_{X_tY_t} = \frac{r_{XY}}{\sqrt{r_{YY}}}.$$

Consider first the effect of only having unreliability in X_2 on $r_{Y2\cdot1}$. Based on Eq. (4.3.2a) we know that $r_{Y_tX2_t} = r_{Y2}/\sqrt{r_{22}}$ and $r_{X1_tX2_t} = r_{12}/\sqrt{r_{22}}$. When we substitute these values into Eq. (3.3.11) and algebraically simplify the resulting expression, we find

(4.3.3) $$r_{YX2_t\cdot X1} = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{(1 - r_{Y1}^2)(r_{22} - r_{12}^2)}}.$$

(Continued)

[10]Recall from Section 2.10.2 that a true score is a hypothetical error-free score. True scores represent the mean score each individual would receive if he or she could be measured an infinite number of times.

Comparing Eq. (3.3.11) to Eq. (4.3.3), we see that the numerator does not change. However, the second term in the denominator changes from $(1 - r_{12}^2)$ to $(r_{22} - r_{12}^2)$. Because r_{22} is less than 1.0, if there is unreliability in X_2, the value of the denominator will always decrease. Thus, $r_{YX2_t \cdot X1}$ (after correction for measurement error) will always be larger in magnitude than $r_{Y2 \cdot 1}$ if there is unreliability in X_2. To illustrate, using the values $r_{12} = .3, r_{Y2} = .4$, and $r_{Y1} = .5$ from our previous numerical example, but now setting $r_{22} = .7$, we find

$$r_{YX2_t \cdot X1} = \frac{.4 - (.5)(.3)}{\sqrt{(1 - .5^2)(.7 - .3^2)}} = .37$$

as compared to $r_{Y2 \cdot 1} = .30$.

Following the same procedures, we can draw on Eq. (4.3.2b) to study the effect of unreliability only in Y on the partial correlation for the case in which X_1 and X_2 are both perfectly reliable. Paralleling the results observed for X_2, $r_{Y2 \cdot 1}$ will be attenuated relative to $r_{Y_t X2_t \cdot 1}$. Correcting for measurement error in X_2, Y, or both invariably leads to increases in the absolute value of the partial correlation.

Now we turn to unreliability only in X_1, the IV being partialed in the present example, and find a far more complex set of results. Following the same procedures, the partial correlation corrected for unreliability only in X_1 is

(4.3.4)
$$r_{Y2 \cdot 1_t} = \frac{r_{11} r_{Y2} - r_{Y1} r_{12}}{\sqrt{(r_{11} - r_{Y1}^2)(r_{11} - r_{12}^2)}}.$$

Unlike in our previous equations, both the numerator and the denominator are changed from Eq. (3.3.11). The same general finding of change in both the numerator and denominator also occurs when there is unreliability in X_1, X_2 and Y,

(4.3.5)
$$r_{Y_t 2_t \cdot 1_t} = \frac{r_{11} r_{Y2} - r_{Y1} r_{12}}{\sqrt{(r_{11} r_{YY} - r_{Y1}^2)(r_{11} r_{22} - r_{12}^2)}}.$$

The change in both the numerator and denominator in Eqs. (4.3.4) and (4.3.5) means that the effect of correcting for unreliability only in X_1 will depend on the specific values of r_{11}, r_{Y1}, r_{Y2}, and r_{12} in the research problem.

We illustrate in Table 4.3.1 the range of effects that unreliability in X_1, the variable being partialed, can have on measures of partial relationship presented in Section 3.2. In addition to partial correlations, we also report standardized regression coefficients. Recall that the standardized regression coefficient $\beta_{Y2.1}$ for the relation between Y and X_2 controlling for X_1 is

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

Again using the strategy of substituting Eq. (4.3.2a) into Eq. (3.2.4) and algebraically simplifying the results, we find that the standardized regression coefficient corrected for measurement error in X_1 is

(4.3.6)
$$\beta_{YX2 \cdot X1_t} = \frac{r_{Y2} r_{11} - r_{Y1} r_{12}}{r_{11} - r_{12}^2}$$

TABLE 4.3.1
Effects of Unreliability of a Partialed Variable (X_1)

Example	r_{Y2}	r_{Y1}	r_{12}	$r_{Y2\cdot1}$ (Eq. 3.3.11)	$r_{Y2.1_t}$ (Eq. 4.3.4)	$\beta_{Y2.1}$ (Eq. 3.2.4)	$\beta_{Y2\cdot X1_t}$ (Eq. 4.3.6)
1	.3	.5	.6	.00	−.23	.00	−.26
2	.5	.7	.5	.24	.00	.20	.00
3	.5	.7	.6	.14	−.26	.13	−.21
4	.5	.3	.6	.42	.37	.50	.50
5	.5	.3	.8	.45	.58	.72	1.83

Note: In each example, the reliabilities are $r_{11} = 0.70, r_{22} = 1.00, r_{YY} = 1.00$.

Once again, r_{11} appears in both the numerator and denominator of Eq. (4.3.6), meaning that measurement error in X_1 will have complex effects on the standardized regression coefficient. The effect of measurement error on standardized regression coefficients is an important issue in structural equation (path) models, which are considered in Chapter 12 and in O. D. Duncan (1975, chapter 9) and Kenny (1979, chapter 5).

Table 4.3.1 explores the effect of varying r_{11}, r_{Y1}, r_{Y2} on the partial correlation and standardized regression coefficients. In each case, there is only unreliability in X_1, with the value of r_{11} being set at .70, a value that is commonly cited as a minimum value for "acceptable" reliability of a measure. The value of $r_{Y2\cdot1}$ is computed using Eq. (3.3.11) and the value of $r_{Y2\cdot1_t}$ (corrected for measurement error) is computed using Eq. (4.3.4). The value of $\beta_{Y2.1}$ is calculated using Eq. (3.2.4) and the value of $\beta_{Y2\cdot1_t}$ is calculated using Eq. (4.3.8).

We focus on the results for the partial correlation in Table 4.3.1 (columns 5 and 6). In Example 1, measurement error in X_1 results in an observed partial r of 0, whereas the true partialed relationship ($r_{Y2\cdot1_t}$) is −.23. Thus, a real partial relationship is wiped out by measurement error in X_1, the variable being partialed. In Example 2, the converse occurs: An observed partial correlation, $r_{Y2\cdot1} = .24$, is actually 0 when the unreliability of the partialed IV is taken into account. Example 3 has the most dangerous implications. Here, an apparently positive partial correlation, $r_{Y2\cdot1} = .14$, turns out to be negative (and of larger magnitude) when corrected for measurement error. This result is not merely a mathematical curiosity of only academic interest. For example, Campbell and Erlebacher (1970) have strongly argued that incorrect conclusions were drawn about the effectiveness of the Head Start program because circumstances like those in Example 3 obtained. Example 4 illustrates for the partial correlation the most frequent outcome: Correction for measurement error in X_1 will lead to a decrease in the magnitude of the partial correlation. Finally, Example 5 illustrates a case in which the value of the partial correlation is increased after correction for measurement error. Note also in Example 5 the value of the standardized regression coefficient after correction for measurement error, $\beta_{YX2\cdot X1_t} = 1.83$ (see Table 4.3.1, column 8). The magnitude of the standardized regression coefficient substantially exceeds 1.0, indicating a potential problem. Note that in Example 5, $r_{12} = .8$, $r_{11} = .7$, and $r_{22} = 1.0$ so that from Eq. (4.3.2a) the value of $X1_tX2_t$ (corrected for measurement error) is .96. Such a result may mean that X_1 and X_2 are so highly related that their influence cannot be adequately distinguished (see Section 10.5 on multicollinearity for a discussion of this issue). Alternatively, the estimated value of $r_{11} = .70$ based on the sample data may be lower than the true value

(Continued)

of the reliability in the population. The results of Example 5 emphasize that the results of correction for unreliability must be undertaken cautiously, a point to which we will return in our presentation of remedies for measurement error in Section 4.5.3.

We have shown that measurement error in the dependent variable leads to bias in all standardized measures of partial relationship including *sr*, *pr*, and standardized $\beta_{Y2.1}$. Of importance, unlike measurement error in the independent variables, it does *not* lead to bias in the values of the *unstandardized* regression coefficients. Measurement error in the dependent variable does not affect the slope of the regression line, but rather only leads to increased variability of the residuals around the regression line (see, e.g., O. D. Duncan, 1975). This increase in the variability of the residuals means that confidence intervals will increase in size and the power to reject a false null hypothesis will be decreased.

4.3.2 Ordinary Least Squares Estimation

In the simplest case of one-predictor linear regression, we fit a straight line to the data. Our goal is to choose the best values of the intercept B_0 and the slope B_1 so that the discrepancy between the straight line and the data will be as small as possible. Carefully drawing a straight line through the data by hand can do a pretty good job of achieving this in the one-predictor case. However, we would like to have a more objective mathematical method of identifying the regression coefficients that would yield the best fitting straight line. The "obvious" method of doing this, examining the sum of the differences between the observed and predicted values of Y, $\Sigma(Y_i - \hat{Y}_i) = \Sigma e_i$, does not work because the sum of the residuals will always be 0 regardless of the values of B_0 and B_1 that are chosen. Several different mathematical methods could be used, but by far the most commonly used method is *ordinary least squares* (*OLS*). OLS seeks to minimize the sum of the squared differences between the observed and predicted squares of Y. That is, in the one-predictor case B_1 and B_0 are chosen so that

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - B_1 X_i - B_0)^2$$

is the smallest possible value. In the multiple predictor case, $B_0, B_1, \ldots, B_k$ are chosen so that $\Sigma(Y_i - \hat{Y}_i)^2$ is the smallest possible value. All equations we have considered so far in this book are based on the OLS method. A formal mathematical derivation of OLS estimation is given in Appendix 1.

In the previous section, the assumptions we presented were those associated with OLS estimation. When these assumptions are met, the OLS estimates of the population regression coefficients have a number of important and desirable properties.

1. They are unbiased estimates of each true regression coefficient in the population. If many samples were selected, the mean of the sample regression coefficients for B_0 and B_1 would equal the values of the corresponding regression coefficients in the population. The expected value of B_j, $E(B_j)$, will equal the corresponding regression coefficient β_j in the population.
2. They are consistent. The standard errors of each regression coefficient will get smaller and smaller as sample size increases.
3. They are efficient. No other method of estimating the regression coefficients will produce a smaller standard error. Small standard errors yield more powerful tests of hypotheses.

Taking these properties together, OLS is described as the *B*est *L*inear *U*nbiased *E*stimator (BLUE). However, when the assumptions of OLS regression are not met, these properties may not always not hold. When there are violations of assumptions, the values of the regression coefficients, their standard errors, or both may be biased. For example, as we will show in our

consideration of outliers (unusual observations) in Chapter 10, one very extreme data point for which the squared difference between the observed and predicted scores, $(Y_i - \hat{Y}_i)^2$, is very large may be too influential in the computation of the values of the regression coefficients.

In cases in which the assumptions of OLS regression are violated, we may need to use alternative approaches to the analysis. Three different general approaches may be taken. First, the analyst may build terms into the OLS regression model so that the form of the relationship between each IV and DV more adequately represents the data. Second, the analyst may be able to improve the data by deleting outlying observations or by transforming the data so that the assumptions of OLS regression are not so severely violated. Third, the analyst may consider using an alternative to OLS estimation that is more robust to the specific problem that has been identified. After considering methods for detecting violations of assumptions in Section 4.4, we will see examples of each of these approaches in subsequent sections of this chapter and Chapter 10.

4.4 DETECTING VIOLATIONS OF ASSUMPTIONS

A goal in regression analysis is that the model under consideration will account for all of the meaningful systematic variation in the dependent variable Y. Residuals ("errors") represent the portion of each case's score on Y that cannot be accounted for by the regression model, $e_i = Y_i - \hat{Y}_i$ for case i. If substantial *systematic* variation remains in the residuals, this suggests that the regression model under consideration has been misspecified in some way. Residuals magnify the amount of remaining systematic variation so that careful use of graphical displays of residuals can be very informative in detecting problems with regression models. We will also briefly present some formal statistical tests, but we will emphasize graphical displays because they make minimal assumptions about the nature of the problem.

4.4.1 Form of the Relationship

In current practice, most regression models specify a linear relationship between the IVs and the DV. Unless there is strong theory that hypothesizes a particular form of nonlinear relationship, most researchers begin by specifying linear regression models like those we considered in Chapters 2 and 3. However, there is no guarantee that the form of the relationship will in fact be linear. Consequently, it is important to examine graphical displays to determine if a linear relationship adequately characterizes the data.

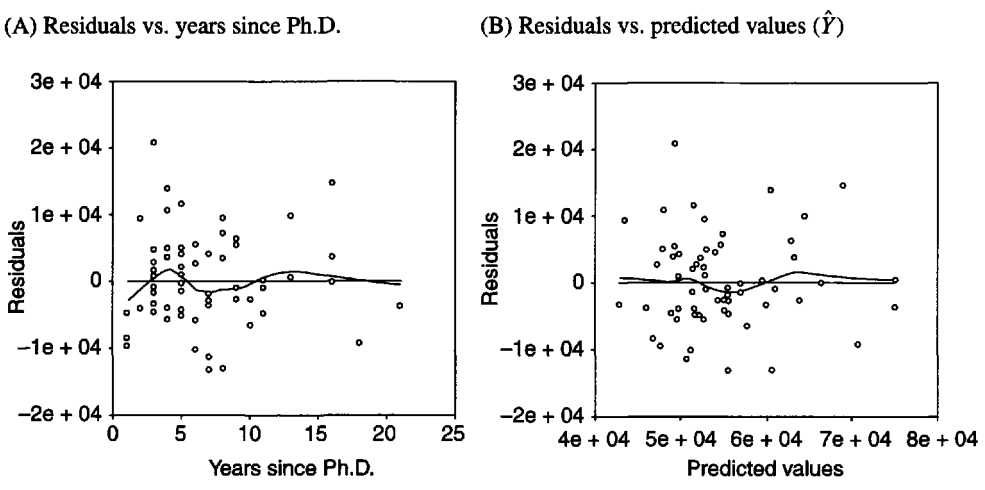
As we saw in Section 4.2, we can construct a separate scatterplot for the dependent variable (Y) against each independent variable (X) and superimpose linear and lowess curves to see if the relationship is linear. Even more revealing, we can plot the residuals on the y axis separately against each IV ($X_1, X_2, \ldots, X_k$) and against the predicted variable ($\hat{Y}$). The residuals will magnify any deviation from linearity so that nonlinear relationships will become even more apparent.

Returning to the salary data presented in Table 3.5.1, recall that the regression model using all four independent variables years, number of publications, gender, and number of citations was

$$\hat{Y} = 857 \text{ years} + 93 \text{ publications} - 918 \text{ female} + 202 \text{ citations} + 39{,}587.$$

We plot the residuals against each measured independent variable and against the predicted values, and look for evidence of nonlinearity.

Figure 4.4.1(A) is a scatterplot of the residuals from the regression equation against one of the IVs, years since Ph.D. The horizontal line identifies the point where the residuals are



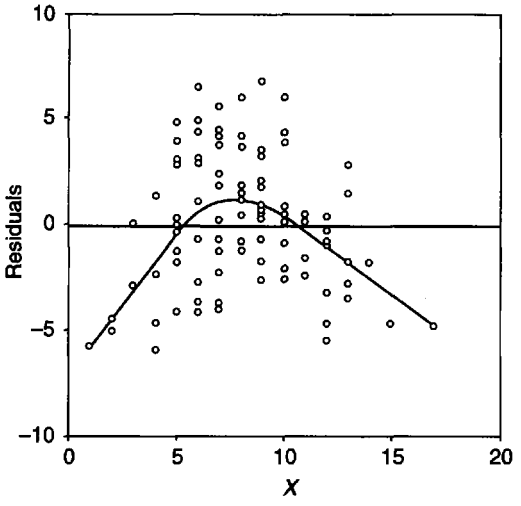
Note: The horizontal line corresponds to a value of 0 for the residuals (0-line). The lowess line is also shown. No systematic relationship between the residuals and either years or $\hat{Y}$ is indicated.

FIGURE 4.4.1 Scatterplots of salary data.

0 (i.e., *0-line*), where the predicted and observed values of Y are identical. If the form of the relationship is properly specified, then the mean of the residuals should be 0 regardless of the value of the IV. The curved line is the lowess fit. Recall that the lowess fit line follows the general trend of the data. If the form of the relationship is properly specified, then the lowess fit line should not exhibit any large or systematic deviations from the 0-line. In the present example, the lowess line generally follows the 0-line, suggesting that the relationship between X_1 (years since Ph.D.) and Y (salary) approximates linearity. Plots of the residuals against number of publications and of the residuals against number of citations (not depicted) also do not show any evidence of deviations from linear relationships. Plots of the residuals against female (gender) will not be informative about linearity because female is a nominal (qualitative) variable. Finally, Fig. 4.4.1(B) shows the scatterplot of the residuals against the predicted values ($\hat{Y}$) with superimposed zero and lowess fit lines. These scatterplots support the specified linear relationship between each of the independent variables and the outcome variable.

In contrast, consider the data originally presented in Fig. 4.2.5(F). These data were generated to have a nonlinear relationship between X and Y. The lowess curve for the original (raw) data indicated that the relationship is nonlinear. Suppose a researcher mistakenly specified a linear regression model to account for these data. The resulting regression equation, $\hat{Y} = 1.14X + 8.72$, appears to nicely account for these data: $R^2 = .56$; test of B_1, $t(98) = 11.2$, $p < .001$. Figure 4.4.2 plots the residuals from this regression equation against X. The lowess fit does not follow the 0-line. It clearly indicates that there is a relatively large and systematic nonlinear component in these data. By comparing Fig. 4.2.5(F) for the original data with Fig. 4.4.2 for the residuals, we see how the plot of residuals magnifies and more clearly depicts the nonlinear component of the X–Y relationship.

The graphical methods presented here are particularly powerful. The true relationship between the IVs and DVs may take many different mathematical forms. Graphical methods can detect a very wide range of types of misspecification of the form of the relationship. In contrast, statistical tests in polynomial regression presented in Section 6.2 are much more focused, contrasting only the fit of two different model specifications chosen by the analyst.



Note: The horizontal line corresponds to a value of 0 for the residuals (0-line). The lowess line is also shown. The clear curvilinear form of the lowess line indicates a nonlinear relationship between the residuals and X.

FIGURE 4.4.2 Scatterplot of residuals vs. X: nonlinear relationship.

4.4.2 Omitted Independent Variables

In specifying the regression model, we include all IVs specified by our hypothesis. Theory and prior empirical research will often provide strong guidelines for the variables that should be included. However, in some cases, the analyst will be unsure whether or not certain variables should be included in the regression equation. The analyst may believe that a theory has omitted important variables or that theory and prior empirical research may be unclear or contradictory about the importance of certain IVs. In such cases, the analyst can explore the effects of including additional variables in the regression equation.

The simplest method of approaching this issue is to construct a series of scatterplots. The analyst first runs a regression analysis in which the originally hypothesized model is specified and saves the residuals. Then, a series of scatterplots are constructed in which the value of the residuals is represented on the y axis and an additional candidate variable (omitted from the regression equation) is represented on the x axis. If the original regression model is correct and a lowess line is fitted to these data, the lowess line should ideally be very close to the 0-line (horizontal line where residuals = 0). In contrast, if the lowess curve suggests either a linear or nonlinear relationship, the omitted variable should receive further investigation.

An improvement over this basic scatterplot is the *added variable plot* (AVP, also known as the *partial regression leverage plot*). The AVP allows the analyst to directly visualize the effect of adding a candidate IV to the base regression model. To understand conceptually how the added variable plot is constructed, assume we have specified a base regression equation with three independent variables,

$$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_0. \tag{4.4.1}$$

We wish to investigate whether another candidate variable, X_4, should have also been included in the regression equation as an independent variable.

In constructing the added variable plot, the statistical package first estimates the regression equation predicting the candidate variable, X_4, from the other three IVs,

(4.4.2) $$\hat{X}_4 = B'_1X_1 + B'_2X_2 + B'_3X_3 + B'_0.$$

In Eq. (4.4.2) B'_0 through B'_3 are used to represent the unstandardized regression coefficients, as they will typically take on different values than those in Eq. (4.4.1). The residual $e'_i = X_4 - \hat{X}_4$ is calculated for each case. This residual represents the unique part of X_4 that remains after X_1, X_2, and X_3 have been accounted for. Using the predicted value from Eq. (4.4.1), the residual, $e_i = Y - \hat{Y}$, is computed. This residual represents the part of Y that is not accounted for by X_1, X_2, and X_3. Then the added variable plot is constructed; this is a scatterplot with the residuals e'_i from Eq. (4.4.2) on the x axis and the residuals e_i from Eq. (4.4.1) on the y axis. A straight line and lowess curve can be superimposed to the added variable plot to help elucidate the relationship of X_4 to Y, partialing out X_1, X_2, and X_3.

To illustrate, suppose in our example of faculty salaries that we had hypothesized a model with salary as the dependent variable and years since Ph.D. (X_1), number of publications (X_2), and female (X_3) as the independent variables. Using the 62 cases presented in Table 3.5.1, this regression equation is

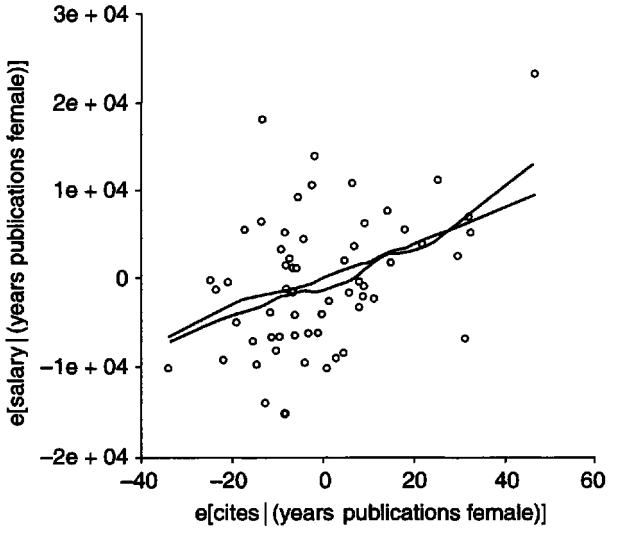
$$\hat{Y} = 1066 \text{ years} + 131 \text{ publications} - 1408 \text{ female} + 45{,}810.$$

The 62 residuals $Y - \hat{Y}$ are calculated using Eq. (4.4.1). Suppose some literature suggests that number of citations (X_4) is also an important IV, so we are concerned that we have incorrectly omitted X_4 from the regression model. A second regression equation with number of citations as the dependent variable and years since Ph.D., number of publications, and female as the independent variables is estimated,

$$\hat{X}_4 = 1.03 \text{ years} + 0.19 \text{ publications} - 2.43 \text{ female} + 30.81.$$

The 62 residuals $X_4 - \hat{X}_4$ are calculated. Plotting $Y - \hat{Y}$ on the y axis and $X_4 - \hat{X}_4$ on the x axis produces the added variable plot shown in Fig. 4.4.3. The positive slope of the straight line suggests there is a positive linear relationship between the candidate variable X_4 and Y, controlling for X_1, X_2, X_3. Indeed, the exact value of the slope of the straight line for the regression of $(Y - \hat{Y})$ on $(X_4 - \hat{X}_4)$ will equal the numerical value of B_4 in the regression equation including the candidate variable, $\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_0$. In the present case, the lowess fit does not deviate substantially or systematically from the straight line, suggesting that the relationship between X_4 and Y does not have a curvilinear component. Taken together, these results suggest that X_4 should be included in the specification of the linear regression equation as we did in our original analysis presented in Section 3.5.

Interpreting the results of added variable plots is straightforward. If the slope of the best fitting regression line produced by the added variable is 0, the independent variable has no unique relation to Y. If the slope is positive, the added variable will have a positive relationship to Y; if the slope is negative it will have a negative relationship to Y. If the lowess line indicates some form of systematic curvature, then the relationship of X and Y will be nonlinear. Added variable plots can be used to study the effects of adding a candidate independent variable to base regression equations involving one or more independent variables and with more complex base regression models involving interactions (Chapter 7) and nonlinear effects represented by power polynomials (Chapter 6). They can also be used to visualize and identify outlying data points that strongly influence the estimate of the regression coefficient associated with the



Note: The straight line depicts the linear relationship between number of citations and salary controlling for the independent years (X_1), publications (X_2), and female (X_3). The slope of this line = B_4 in the full regression model, $\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_0$. The curved line represent the lowess fit. Values on the y axis are residuals, $Y - \hat{Y}$ and values on the x axis are residuals, $X_4 - \hat{X}_4$. The value 0 occurs on each axis when the observed value of X_4 and Y equal their respective predicted values.

FIGURE 4.4.3 Added variable plot.

added variable.[11] However, the analyst must remember that the added variable plot provides information relative to the base regression model that has been specified. If the base model has not been properly specified (e.g., the relationship between X_1 and Y is in fact nonlinear), the added variable plot can give misleading results.

4.4.3 Measurement Error

Measurement error is easily detected with a measure of reliability. One common type of measure is a scale in which the participants' scores are based on the sum (or mean) of their responses to a set of items. In cross-sectional studies in which the measures are collected on a single occasion, the most commonly used measure of reliability (internal consistency) is *coefficient alpha* (Cronbach, 1951). Imagine that we have a 10-item scale. We can split the scale into two halves and correlate the subjects' scores on the two halves to get a measure of reliability. Coefficient alpha represents the mean of the correlations between all of the different possible splits of the scale into two halves. Another common form of reliability known as *test-retest reliability* is the correlation between subjects' scores on the scale measured at two different times. When two judges rate the subjects' score on the variable, *interrater reliability* is the

[11] If there is a single outlying point, the AVP corresponds to a visualization of $DFBETAS_{ij}$ presented in Chapter 10. However, the AVP also allows analysts to identify clumps of outliers that influence the regression coefficient for the candidate IV.

correlation between scores given by the two judges. With these correlation-based measures of reliability, scores close to 1.0 indicate high levels of reliability.

Measurement is a large and important area of study in its own right (see Crocker & Algina, 1986; Nunnally & Bernstein, 1993) so we cannot provide a full treatment here. Readers should keep in mind three points when they encounter measures of reliability with which they are unfamiliar. First, not all indices of reliability are based on correlations, so different criteria for judging the reliability of the measure will be needed for some indices. Values near 1.0 are not expected for all measures of reliability. Second, the best measure of reliability will depend on the design of the study and the nature of the construct being studied. For example, some constructs, such as adult personality traits, are expected to be very stable over time whereas other constructs, such as moods, are expected to change very quickly. Measures of test-retest reliability are unlikely to be appropriate for mood measures. Finally, some newer approaches to measurement such as item response theory do not yield traditional measures of reliability. However, some of the newer methods can help researchers attain interval level measurement scales that are nearly free of measurement error (Embretson & Reise, 2000).

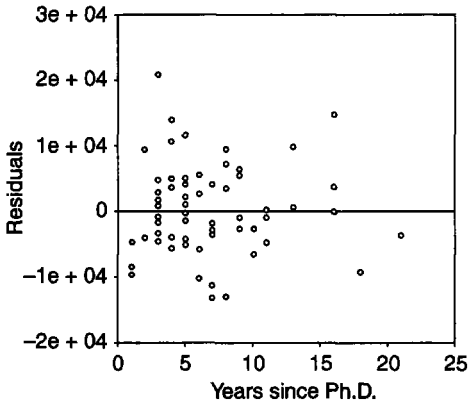
4.4.4 Homoscedasticity of Residuals

If the variance of the residuals $sd^2_{Y-\hat{Y}}$ is not constant, but is related to any of the IVs or to $\hat{Y}$, the standard methods of developing confidence intervals and conducting significance tests presented in Chapters 2 and 3 may potentially become compromised. A simple graphical method of detecting this problem is to construct a set of scatterplots, plotting the residuals in turn against each of the independent variables $X_1, X_2, \ldots, X_k$ and the predicted value, $\hat{Y}$. Figure 4.4.4(A) plots the residuals against years since Ph.D. for the faculty salary data ($n = 62$) using the full regression equation, $\hat{Y} = 857$ years + 93 publications − 918 female + 202 citations + 39,587, originally presented in Section 3.5. Figure 4.4.4(B) plots the residuals against the predicted values, $\hat{Y}$. These plots do not suggest that there is a relationship between the variability of the residuals and either years since Ph.D. or the predicted value.

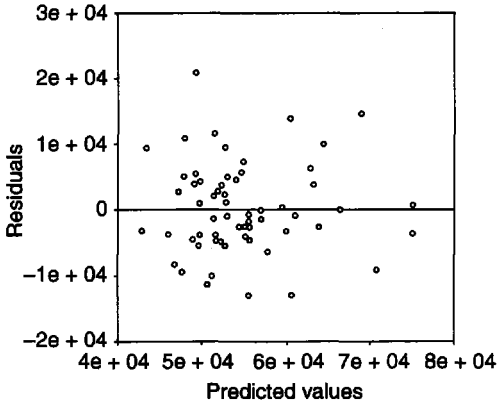
Some statistical packages allow the analyst to plot lowess fit lines at the mean of the residuals (0-line), 1 standard deviation above the mean, and 1 standard deviation below the mean of the residuals. Figure 4.4.4(C) replots the residuals against years since Ph.D. adding these lowess lines. The middle line corresponds to the lowess line described in Section 4.2.2. The other two lines are created using the lowess procedure to estimate values 1 standard deviation above and 1 standard deviation below the lowess line. In the present case, the two lines remain roughly parallel to the lowess line, consistent with the interpretation that the variance of the residuals does not change as a function of X. Examination of plots of the residuals against number of publications and against number of citations (not depicted) also do not suggest any relationship.

What do these scatterplots look like when there is a relationship between the variance of the residuals and X or $\hat{Y}$? Figure 4.4.5 displays three relatively common patterns using data sets with 400 cases. Figure 4.4.5(A) again shows the relationship when the data are homoscedastic. In Fig. 4.4.5(B), the variance in the residuals increases in magnitude as the value of X increases, often termed a right-opening megaphone. For example, such a pattern can occur in experimental psychology in such tasks as people's judgments of distances or the number of identical objects in a standard container—low values are judged more accurately. In contrast, Fig. 4.4.5(C) shows a pattern in which the residuals are highest for middle values of X and become smaller as X becomes smaller or larger. Such patterns are found for some

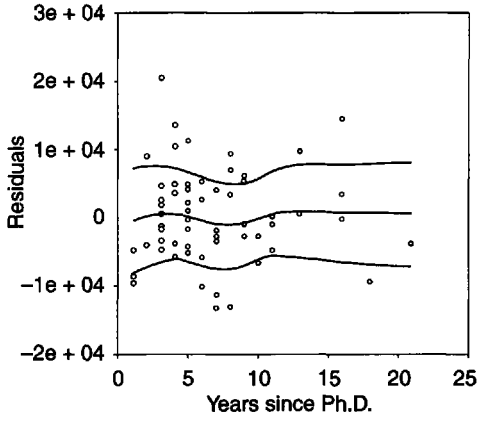
(A) Residuals vs. years since Ph.D.



(B) Residuals vs. predicted values ($\hat{Y}$).



(C) Residuals vs. years since Ph.D. (lowess fit added).



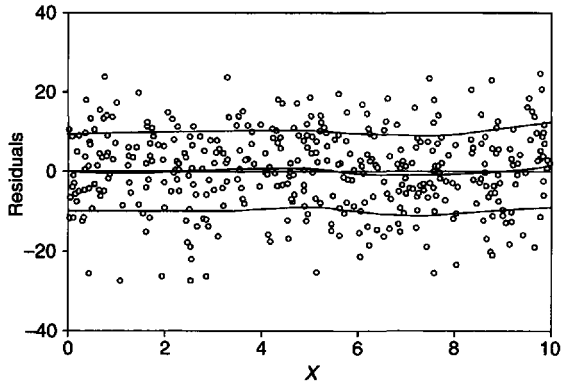
Note: Lowess fit values have been added. The middle line is the lowess fit. The upper line is the lowess fit + 1 *sd*. The lower line is the lowess fit − 1 *sd*. The set of lowess lines do not suggest any evidence of a substantial departure from linearity or heteroscedasticity.

FIGURE 4.4.4 Plots of residuals.

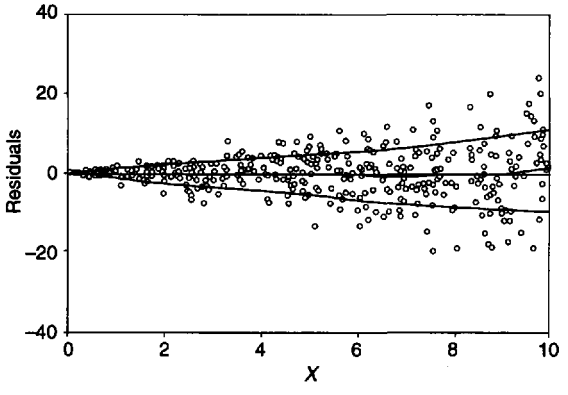
personality tests where people who are extremely high or low on the personality trait are more accurately measured. In each panel lowess fit lines have been added at the mean, 1 standard deviation above the mean, and 1 standard deviation below the mean of the residuals so that these patterns can be more easily discerned.

Formal statistical tests of nonconstant variance have also been developed. Most of these tests focus on detecting a specific pattern of heteroscedasticity. Interested readers can find an introduction to some of these tests in Box 4.4.1.

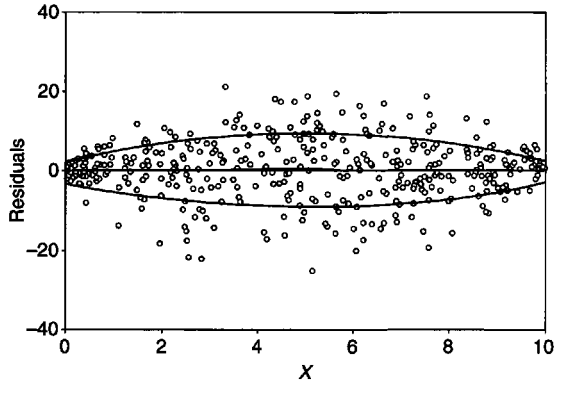
(A) Homoscedasticity: constant variance across values of X.



(B) Heteroscedasticity: variance increases with X (right-opening megaphone).



(C) Heteroscedasticity: curvilinear relationship between X and variance of residuals.



Note: $n = 400$ cases. The 0-line, the lowess line for the mean, the lowess lines for the mean + 1 *sd*, and the lowess line for the mean − 1 *sd* are superimposed.

FIGURE 4.4.5 Plots of residuals versus X. Illustrations of homoscedasticity and heteroscedasticity.

BOX 4.4.1
Statistical Tests of Nonconstant Variance

The *modified Levene test* provides a formal statistical test of the pattern depicted in Fig. 4.4.5(B) in which the variance of the residuals appears to increase (or decrease) as a function of the IV. The residuals are initially divided into two groups, one containing the cases that are high and one containing the cases that are low relative to a threshold value on the independent variable that is chosen by the analyst. For example, in our salary example ($n = 62$), an analyst might choose to examine whether the residuals are related to years since Ph.D. Using the 62 cases presented in Table 3.5.1, she would pick a threshold value—for example a score of 6 years, which is near the median of the distribution—and classify the $n_L = 30$ scores < 6 as low and $n_H = 32$ scores ≥ 6 as high. She would then calculate the median value for the residuals in each group: In the low group, the 30 residuals have a median value of -408.49; in the high group, the 32 residuals have a median value of -1634. She would then calculate the absolute deviation (ignoring sign) of the residuals in each group from the corresponding group median.

$$d_{\text{low}} = |e_{i_{\text{low}}} - mdn_{e_{\text{low}}}| \qquad d_{\text{high}} = |e_{i_{\text{high}}} - mdn_{e_{\text{high}}}|$$

The variance of the absolute deviations is

$$s^2 = \frac{\sum_1^{n_{\text{low}}} (d_{i_{\text{low}}} - M_{d_{\text{low}}})^2 + \sum_1^{n_{\text{high}}} (d_{i_{\text{high}}} - M_{d_{\text{high}}})^2}{n_{\text{low}} + n_{\text{high}} - 2}$$

where $M_{d_{\text{low}}}$ is the mean of the absolute residuals corresponding to the low values on and $M_{d_{\text{high}}}$ is the mean of the absolute residuals corresponding to the high values on X. Finally, she would calculate Levene's t^*

$$\text{Levene's } t^* = \frac{M_{d_{\text{low}}} - M_{d_{\text{high}}}}{\sqrt{s^2\left(\frac{1}{n_{\text{low}}} + \frac{1}{n_{\text{high}}}\right)}}.$$

The result of the Levene's t^* test is compared to the critical values of the t distribution from Appendix Table A with $df = n_{\text{low}} + n_{\text{high}} - 2$. Failure to reject the null hypothesis is nearly always the desired outcome—it is consistent with the use of standard OLS regression models, which consider the residuals to be homoscedastic.

Several other tests have also been proposed to test for various forms of homoscedasticity. R. D. Cook and Weisberg (1983) and Breusch and Pagan (1979) independently developed an alternative test that detects increases or decreases in the variance of the residuals. This test performs very well in large samples when the residuals have a normal distribution. This test can also be modified to test for other specified relationships (e.g., quadratic) between X or $\hat{Y}$ and the variance of the residuals (see Weisberg, 1985). White (1980) has developed a general test that can potentially detect all forms of hetereoscedasticity. This test requires large sample size, has relatively low statistical power, and can yield misleading results if there are other problems in the regression model (e.g., the regression model is misspecified). A discussion of advantages and disadvantages of several of the tests of heterogeneity of variance of residuals can be found in Greene (1997).

4.4.5 Nonindependence of Residuals

Multiple regression assumes that the residuals are independent. Index plots (also termed case-wise plots) provide a simple method for exploring whether the residuals are related to some systematic feature of the manner in which the data were collected. In the present context, index plots are simply scatterplots in which the value of the residual is presented on the y axis and an ordered numerical value is presented on the x axis. Statistics and graphical displays that are more sensitive to specific forms of nonindependence can also be used.

One form of dependency in the residuals occurs when there is a systematic change over time in the nature of the participants or in the research procedures. To illustrate, patients with more severe diagnoses may be recruited in the later phase of a clinical study, or less conscientious students in introductory psychology classes may delay their participation in experiments until late in the semester, or the delivery of an experimental treatment may improve over time, yielding greater improvement on the dependent variable. In such cases, plots of the residuals against the order of participation (i.e., first $= 1$, second $= 2, \ldots$) can potentially show systematic relationships. Adding a lowess line to the plot can be useful in revealing the form of the relationship. Joiner (1981) presents several illustrative examples in which the use of plots of residuals against variables related to order of participation have helped uncover what he terms "lurking variables" in the data.

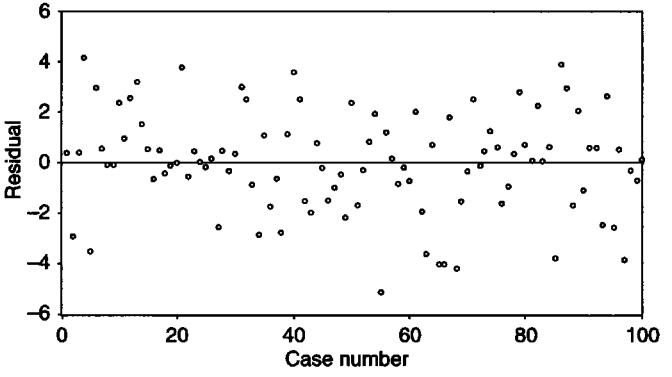
A second form of dependence of residuals, known as clustering, occurs when the data are collected in groups or other clusters. In this case the residuals may be more similar within clusters than between clusters. Figure 4.4.6(A) illustrates 100 residuals from a random sample in which the residuals are independent. Figure 4.4.6(B) and Fig. 4.4.6(C) present two different ways of depicting residuals from a data set in which observations are collected from each of 10 clusters—observations 1–10 are from cluster 1, 11–20 are from cluster 2, and so on. Figure 4.4.6(B) presents an index plot of the residuals using different plotting symbols to represent each cluster. Note that the residuals within clusters tend to bunch together more than the residuals in Fig. 4.4.6(A). For example, in Fig. 4.4.6(B) residuals in cluster 3 (case numbers 21–30, represented by the symbol ×) tend to bunch together below the 0-line, whereas residuals in cluster 7 (case numbers 61–70, represented by the symbol o) tend to bunch together above the 0-line. Figure 4.4.6(C) presents a series of 10 side-by-side boxplots of the same data with each boxplot in turn representing a different cluster. The median value, depicted by the horizontal line in each box, suggests that there is substantial variability in the typical (median) value of each cluster.

A more precise statistical estimate of the amount of clustering will be presented in Section 14.1.2 when we consider the intraclass correlation coefficient. Briefly the intraclass correlation can theoretically range[12] from 0 to 1. To the extent the intraclass correlation exceeds 0, the standard errors of the regression coefficients will be too small. This problem is further exacerbated as the number of cases in each cluster increases (Barcikowski, 1981). Significance tests of regression coefficients will be too liberal, meaning that the null hypothesis will be rejected when it is true at rates far exceeding the stated value (e.g., $\alpha = .05$). The width of confidence intervals will typically be smaller than the true value.

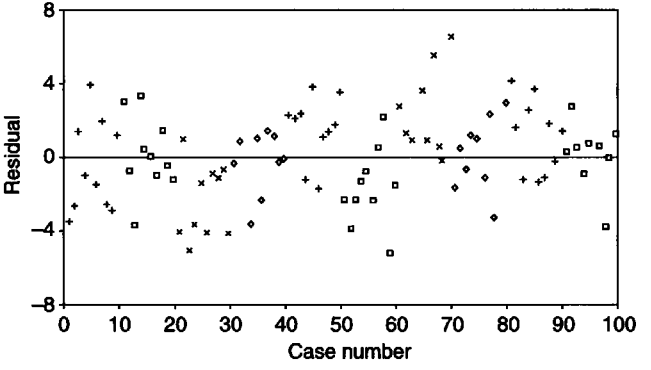
Finally, if the data are repeatedly collected from a single individual or the same sample of individuals over time, then the residuals will often show *serial dependency*. Figure 4.4.6(D) presents an illustration of residuals that exhibit serial dependency. Note that temporally adjacent observations tend to have more similar values than in Fig. 4.4.6(A). For example, the last 15 residuals of the series all have positive values. A more precise statistical measure of serial

[12] In practice, the intraclass correlation coefficient can take on small negative values due to sampling error. In such cases, it is assigned a value of 0.

(A) Independent residuals from a random sample.

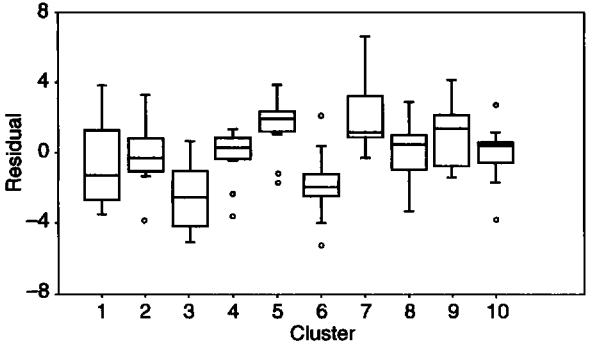


(B) Residuals from clustered data (10 cases per cluster).



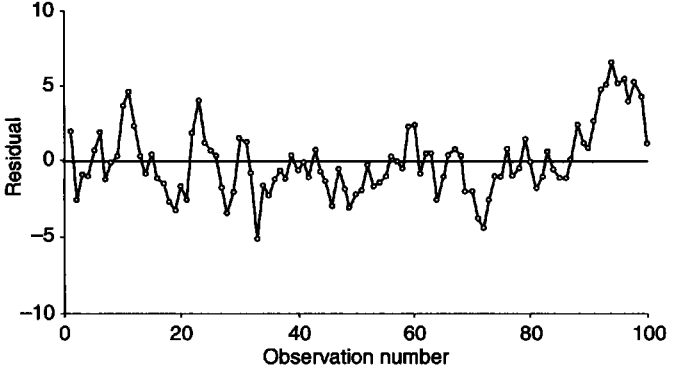
Note: Each cluster of 10 cases (1–10; 11–20; . . .; 91–100) is represented by a different symbol.

(C) Side by side boxplots of the 10 clusters.



Note: Each boxplot represents a different cluster of 10 cases. The horizontal line in each box represents the median of the cluster. The medians of the 10 clusters show more variation than would be expected by chance.

(D) Autocorrelated residuals ($\rho_1 = .7$).



Note: Observations are equally spaced over time. Temporally adjacent observations are connected by straight lines.

FIGURE 4.4.6 Index plots of residuals.

dependency is provided by a measure known as *autocorrelation*. Autocorrelation assesses the correlation of the temporal series with itself after the series has been shifted forward a specified number of time periods. The number of time periods the series is shifted forward is termed the lag. In most applications of multiple regression using temporal data, only the lag 1 autocorrelation is investigated. Like Pearson *r*s, autocorrelations may theoretically range in value from -1 to $+1$. An autocorrelation of 0 indicates that there is no relationship between the original and shifted series. A positive lag 1 autocorrelation indicates that the residual at time $t-1$ is positively related to the residual at time t. For example, yesterday's mood level will tend to be positively related to today's mood. A negative lag 1 autocorrelation indicates that the residual at time $t-1$ is negatively related to the residual at time t. Negative lag 1 autocorrelations can arise when there is a homeostatic process. For example, regular cigarette smokers are likely to smoke at lower than normal rates the hour after they have consumed a larger than normal number of cigarettes. Both positive and negative autocorrelations can lead to incorrect standard errors and consequently incorrect hypothesis tests and confidence intervals. Interested readers will find an illustration of the calculation and test of significance of lag 1 autocorrelation in Box 4.4.2.

BOX 4.4.2
Calculating Autocorrelation and the Durbin-Watson Test

We illustrate here the calculation of the lag 1 autocorrelation. Imagine we have recorded the number of cigarettes smoked for eight consecutive 2-hour periods. After we fit a regression equation, we have the eight residuals listed here and wish to calculate the lag 1 autocorrelation. In this example, the residuals are presented in order by time period in row 2, but are shifted 1 period forward in time in row 3.

Time period (t)	1	2	3	4	5	6	7	8	
Residual	4	−3	−2	3	−1	3	2	−1	
Residual (shifted)		4	−3	−2	3	−1	3	2	−1

The lag 1 autocorrelation can then be calculated between the residual series and the shifted residual series using Eq. (4.4.3),

$$r_1 = \frac{\sum_{t=2}^{T}(e_t)(e_{t-1})/(T-1)}{\sum_{t=1}^{T}(e_t)^2/(T)}, \tag{4.4.3}$$

where e_t is the value of the residual at time t, and e_{t-1} is the value of the residual at time $t-1$, T is the number of equally spaced observations in the original temporal series (here, $T=8$), and r_1 is the value of the autocorrelation at lag 1. To form the product in the numerator of Eq. (4.4.3), we start with second residual in the original series and multiply it by the shifted residual immediately below, continuing to do this until we get to the final original residual, here $t=8$. Note that there are now $T-1$ pairs, here 7, in the residual and shifted residual series. In this example,

$$r_1 = \frac{[(-3)(4)+(-2)(-3)+(3)(-2)+\cdots+(-1)(2)]/7}{[(4)^2+(-3)^2+(-2)^2+\cdots+(-1)^2]/8} = -0.30$$

Standard statistical packages will calculate the lag 1 autocorrelation. In long time series with say 100 observations, autocorrelations can also be calculated at lags 2, 3, 4, etc. using the time series routines within the statistical packages; however, in most applications serial dependency is investigated by only examining the autocorrelation at lag 1.

The Durbin-Watson test is used to test the null hypothesis that the lag 1 autocorrelation is 0 in the population. As a focused test, the Durbin-Watson D test does not address autocorrelation of lag 2 or higher. The expression for the Durbin-Watson D is shown in Eq. (4.4.4),

$$D = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}. \tag{4.4.4}$$

Given that $D \approx 2(1 - r_1)$, values close to $D = 2$ will lead to retention of the null hypothesis of no lag 1 autocorrelation in the population. The exact critical value of the Durbin-Watson statistic is difficult to calculate. Consequently, the value of D is compared with both upper bound (D_U) and lower bound (D_L) critical values. The null hypothesis that the lag 1 autocorrelation in the population $\rho_1 = 0$ is rejected if $D < D_L$ for positive autocorrelations or $D > 4 - D_L$ for negative autocorrelations. The null hypothesis that $\rho_1 = 0$ in the population is retained if $D > D_U$ for positive autocorrelations or $D < 4 - D_U$ for negative autocorrelations. When D falls between D_L and D_U or between $4 - D_L$ and $4 - D_U$, most analysts consider the results of the Durbin-Watson test to be inconclusive since the exact critical value of the Durbin-Watson statistic is unknown. The regression modules of SAS, SPSS, and SYSTAT all calculate the Durbin-Watson statistic.

4.4.6 Normality of Residuals

Two different graphical methods can provide an indication of whether the residuals follow a normal distribution. In the first, more straightforward, but less accurate method, the analyst plots a histogram of the residuals and then overlays a normal curve with the same mean and standard deviation as the data. If the distribution is normal, then the histogram and the normal curve should be similar. Figure 4.4.7(A) depicts a histogram of the residuals and a normal curve for the 62 residuals from the salary example with four independent variables: $\hat{Y} = 857$ years $+ 93$ publications $- 918$ female $+ 202$ citations $+ 39{,}587$. The histogram of the residuals does not appear to be obviously discrepant from the normal curve overlay, although these judgments are often very difficult in small samples. Most standard statistical packages will now generate normal curve distribution overlays for histograms.

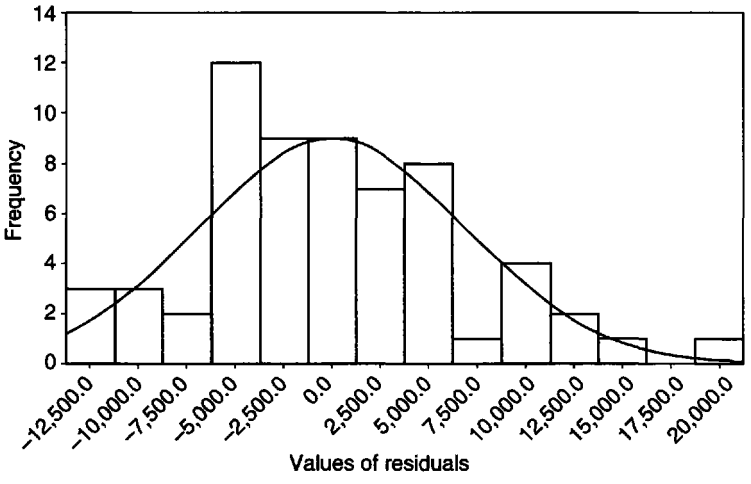
The second method known as a *normal q-q plot* takes advantage of the great accuracy of humans in the perception of straight lines. Many standard statistical packages[13] including SAS and SYSTAT construct the normal q-q plot, and the analyst has only to judge whether the plot approximates a straight line. This judgment task is far easier than with the normal curve overlay.

Figure 4.4.7(B) displays a normal q-q plot for the 62 residuals from the salary data set. As can be seen, the residuals do appear to be close to the straight line which is superimposed. Figure 4.4.7(C) presents the same plot but overlays an approximate 95% confidence interval[14] around the values expected from the normal curve. Nearly all of the residuals from the actual sample fall inside the approximate confidence interval, supporting the interpretation that the residuals have close to a normal distribution. As illustrated in Fig. 4.4.7(C), the 95% confidence

[13]Both SAS and SYSTAT can produce normal q-q plots.

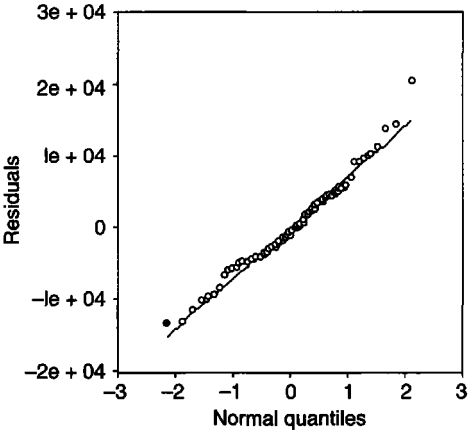
[14]A computer intensive method (Atkinson, 1985) is used to construct this confidence interval. The computer program draws a large number (e.g., 1000) of samples from a normally distributed population. All samples are the same size as the sample being studied, here $n = 62$. These simulated data are then used to construct the empirical distributions for upper 2.5% and the lower 2.5% of the residuals.

(A) Histogram of residuals with normal curve overlay.



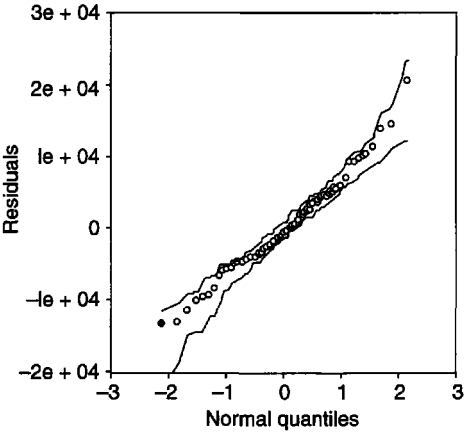
Note: The mean ($M = 0$) and standard deviation of the normal curve are set equal to the mean and standard deviation of the residuals ($n = 62$).

(B) Normal q-q plot of residuals with superimposed straight line.



Note: The points do not exhibit substantial discrepancy from the superimposed straight line. Thus, the residuals appear to be approximately normally distributed. The darkened point is the most negative residual (discussed in the text).

(C) Normal q-q plot of residuals with an approximate 95% confidence interval.



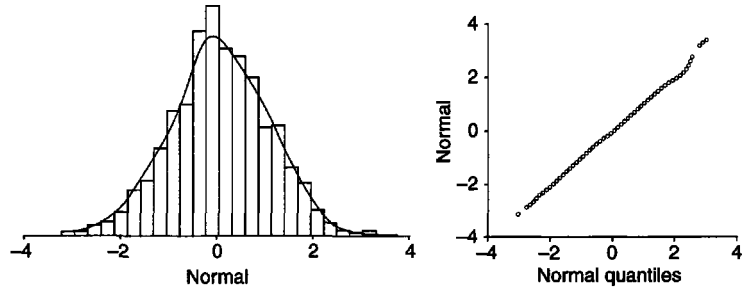
Note: The lines represent an approximate 95% confidence interval for a normal distribution. The points represent the actual values of the residuals. The residuals appear to follow a normal distribution.

FIGURE 4.4.7 Plots to assess normality of the residuals.

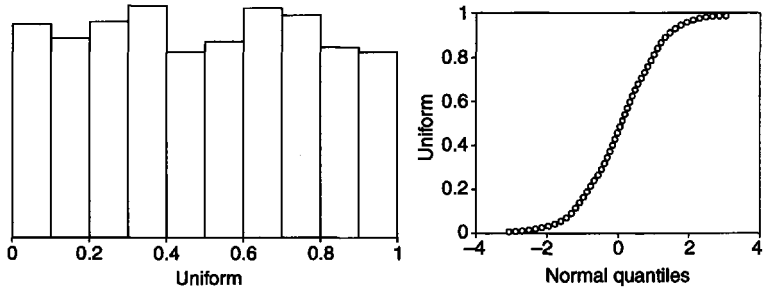
interval for the residuals is narrower (more precise) near the center than near the ends of the distribution of residuals.

Although the normal q-q plot provides an excellent method of determining whether the data follow a normal distribution, interpreting the meaning of a substantial deviation from normality in a q-q plot can be difficult at first. Many analysts supplement normal q-q plots with histograms or kernel density plots of the residuals to provide a more familiar display of

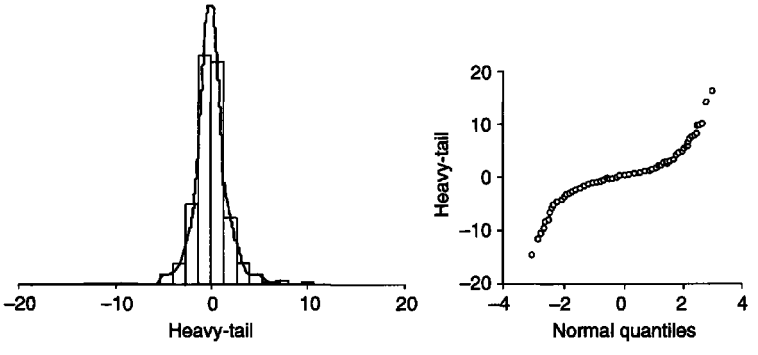
(A) Normal.



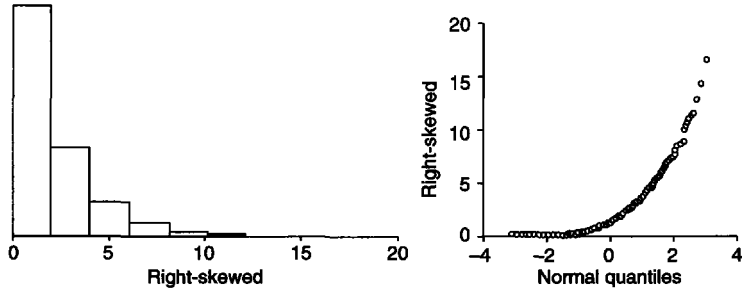
(B) Uniform or rectangular distribution



(C) Heavy or long tailed distribution



(D) Right skewed distribution.



Note: The histogram are on the left and corresponding q-q plots are on the right in each panel. Kernel density estimates are superimposed on the histograms in (A) and (C). Data sets represent random samples of $n = 1000$ from the following population distributions: (A) normal, (B) uniform, (C) t-distribution, $df = 2$, and (D) chi-square distribution, $df = 2$.

FIGURE 4.4.8 Histograms and q-q plots illustrating some common distributions.

BOX 4.4.3
Inside the Black Box: How Normal q-q Plots Are Constructed

The construction of the normal q-q plot starts by putting the cases in rank order $(1, 2, 3, \ldots, n)$ according to the values of their residuals. The smallest (most negative) residual receives rank 1 and the case with the largest (most positive) residual receives rank n. For a case with rank i, we first calculate $f(i)$, the approximate fraction of the data that falls at or below this rank:

$$f(i) = \frac{(i - .5)}{n}.$$

For our salary example with $n = 62$, $f(i = 1)$ for the lowest case would correspond to $(1 - 0.5)/62 = .0089$ and $f(2)$ for the second lowest case would correspond to $(2 - 0.5)/62 = .0242$. These values represent the proportion of the area under the normal curve that falls at or below rank i. Looking in the normal curve table (Appendix Table C gives an abbreviated normal curve table), we find the z score that most closely corresponds to $f(i)$ for each of the cases in the data set. For $f(1) = .0089$, we look in the column labeled P and find the closest value, .009, which corresponds to a z score of -2.35. For $f(2) = .0242$, $z = -2.00$. z scores corresponding to the normal distribution are calculated in this manner for each of the cases in the data set from 1 to n. These z scores are termed normal quantiles. The original values of the residuals are then plotted on the y axis and the quantiles (z scores) corresponding to the normal curve are plotted on the x axis. For example, in Fig. 4.4.7(B), the darkened point at the lower left represents case 28, the most negative residual. The value of this residual $= -13{,}376.8$ is represented on the y axis; the quantile $= -2.35$ corresponding to its rank $i = 1$ of 62 cases is plotted on the x axis. The data are scaled so that the physical length of the y axis is equal to the physical length of the x axis (e.g., 3 inches by 3 inches). The result of this scaling is that the q-q plot of the residuals against the quantiles will result in a straight line at a 45° angle if the residuals precisely follow a normal distribution. The task of the analyst then is simply to judge the extent to which the actual q-q plot matches a straight line.

the distribution of residuals. Some analysts also compare the obtained q-q plot to examples depicting common alternative distributions. For example, Fig. 4.4.8(A) to (D) presents side by side plots of several common distributions. In each panel a histogram of the distribution is plotted on the left (in some cases with a kernel density plot overlayed) and a q-q plot of the same distribution against the normal distribution is presented on the right. Figure 4.4.8(A) portrays a normal distribution, (B) a uniform or rectangular distribution, (C) a heavy or long-tailed distribution, and (D) a right skewed distribution. Only in the case of the normal distribution does the q-q plot follow a straight line. Box 4.4.3 presents the details of the calculation of q-q plots for interested readers.


CH04EX02

Several formal statistical tests of normality have also been proposed. For example, using a method similar to the idea underlying the normal q-q plot, Looney and Gulledge (1985; see also Shapiro & Wilk, 1965) compute the correlation between the value of each residual in order from lowest to highest and the value of the residual that would expected based on a normal distribution. The obtained correlation is then tested against a population value of 1. Another

approach (D'Agostino, 1986) performs a joint test of whether skewness and excess kurtosis[15] of the residuals both equal 0 in the population. Recall, however, that the primary value of examining the normality of the residuals in multiple regression is to help identify problems in the specification of the regression model. Consequently, we have stressed the graphical examination of the distribution of the residuals. Such graphical examination helps reveal the magnitude and nature of any non-normality in the residuals, information that is nearly always far more useful than the significance or nonsignificance of a formal statistical test.

4.5 REMEDIES: ALTERNATIVE APPROACHES WHEN PROBLEMS ARE DETECTED

The diagnostic procedures discussed in Section 4.4 are useful in identifying a variety of potential problems that result from violations of the assumptions of regression analysis. As is the case with routine medical tests, these diagnostic procedures will often indicate that no important problems have occurred. When no problems are detected, standard OLS regression with linear relations specified between the independent and dependent variables will provide the best approach to the analysis. However, researchers cannot know that this will be the case without examining their own data. Problems stemming from violations of the assumptions of linear regression do occur with some regularity in analyses in all areas of the behavioral sciences. These problems raise questions about the conclusions that are reached based on standard linear regression analyses. When problems are diagnosed, then potential remedial actions should be explored. In this section, we identify several common problems that may occur in regression analysis and outline potential remedies. Several of these remedies will be developed in greater detail in later chapters in this book.

4.5.1 Form of the Relationship

When the form of the relationship between X and Y is not properly specified, the estimate of the regression coefficient and its standard error will both be biased. Chapters 2 and 3 have focused on linear relationships between independent and dependent variables. However, a variety of forms of nonlinear relationships may exist between X and Y. In some cases these relationships may be specified by theory or prior research. In other cases, a linear relationship may be initially specified, but the fit of the lowess curve will strongly indicate that the relationship between X and Y is nonlinear. When nonlinear relationships are specified or detected, an alternative approach that accounts for the nonlinear relationship will need to be taken. Chapter 6 considers these approaches in detail.

Four questions should be posed as a basis for choosing among methods for restructuring the regression equation to properly capture the form of the relationship.

1. Is there a theory that predicts a specific form of nonlinear relationship? To cite two examples, the Yerkes-Dodson law predicts there will be an inverted-U (quadratic) relationship between motivation (X) and performance (Y) in which moderate levels of motivation lead to the highest levels of performance. Many learning theories predict that performance on a new task will follow an exponential form in which it initially increases rapidly and then increases more slowly up to a maximum level of performance (asymptote).

2. What is the observed relationship between each pair of IVs? OLS regression only controls for linear relationships among the IVs. If a scatterplot shows a strong nonlinear relationship between two IVs, then these IVs should be re-expressed to make their relationship more linear.

[15]Excess kurtosis = kurtosis − 3. This index rescales the value so that it will be 0 when the distribution is normal.

3. What is the shape of the lowess curve of the original data? The lowess curve portrays the best estimate of the X–Y relationship in the sample. What is the shape of the curve? Does it need to bend in one direction or have multiple bends? Does the curve appear to have an asymptote in which it approaches a maximum or minimum level? Identifying the general shape of the lowess curve is very helpful in trying to identify a mathematical function that can represent the data.

4. Does the variance of the residuals remain constant or does it increase or decrease systematically? This question can be addressed through plots of the residuals against each IV and $\hat{Y}$, particularly when enhanced with lines 1 *sd* above and below the lowess line (see Fig. 4.4.5).

The diagnostic tools we have considered in this chapter help provide answers to questions 2, 3, and 4. The answers help guide the analyst toward the approach to representing nonlinearity that is most likely to prove fruitful. We provide brief guidelines here, and Chapter 6 discusses approaches to nonlinearity in detail.

Theoretically Predicted Relationship

The analyst should specify a regression equation that conforms to theoretically specified mathematical relationship. The test of the Yerkes-Dobson law involves specifying a quadratic relationship between X and Y; The test of the learning theory prediction involves specifying an exponential relationship between $X$and Y.

Nonlinear Relationship Between Independent Variables and Nonlinear Relationship of Independent Variables to the Dependent Variable

The analyst should consider transforming the IVs involved in the nonlinear relationships. In transformation, the original variable is replaced by a mathematical function of the original variable. As one example, X_1 and X_2 might be replaced by their logs so that the new regression equation would be $\hat{Y} = B_1 \log X_1 + B_2 \log X_2 + B_0$. Ideally, the proper choice of transformation will yield a linear relationship between the pair of IVs and between each of the IVs and the DV. Residuals that are homoscedastic in lowess plots remain homoscedastic following transformation of the IVs. Section 6.4 presents rules for choosing transformations.

Nonlinear Relationship Between Independent Variables and the Dependent Variable and Homoscedasticity

The most common approach to this situation is to include power polynomial terms in the regression equation. Power polynomials are power functions of the original IV such as X_1^2 and X_1^3. For example, the regression equation $\hat{Y} = B_1X_1 + B_2X_1^2 + B_0$ can represent any form of quadratic relationship including U-shaped, inverted U-shaped, and relationships in which the strength of the relationship between X and Y increases or decreases as X gets larger. Residuals that are homoscedastic in lowess plots will also remain homoscedastic if the proper polynomial regression model is specified. In some cases simple polynomial functions may not be adequate, so more complicated nonparametric functions may be needed to represent the relationship between the IVs and DVs. Sections 6.2 and 6.3 give a full presentation of multiple regression including power polynomials; Section 6.6 gives an introduction to nonparametric regression.

Nonlinear Relationship Between Independent Variables and the Dependent Variable and Heteroscedasticity

Proper choice of a transformations of Y can potentially simultaneously address problems of nonlinearity, heteroscedasticity, and non-normal residuals in the original linear regression model.[16] Here, the dependent variable is replaced by a nonlinear mathematical function of Y such as $\log Y$ or $\sqrt{Y}$ and the Xs are not changed. As one example, we could replace Y with $\log Y$ so that the new regression equation would be $\log Y = B_1X_1 + B_2X_2 + B_0 + e$. In this transformed regression equation, the residuals are now equal to the observed value of $\log Y$ minus the predicted value of $\log Y$. This means that the relationship between the variance of the residuals and the IVs will be different than was observed with the original data. Ideally, heteroscedasticity can be minimized with the proper choice of transformation of Y. Section 6.4 provides a thorough discussion of methods of transformation of Y.

4.5.2 Inclusion of All Relevant Independent Variables

When a theory or prior research states that a set of Xs (e.g., X_1, X_2, X_3, X_4) should be included in the regression model, omission of any of the independent variables (e.g., X_4) leads to *potential* bias in the estimates of the remaining regression coefficients and their standard errors. Similarly, when there is theoretical reason to believe that a candidate independent variable should be added to the model and the added variable plot supports its inclusion, then it is nearly always a good idea to estimate a new regression model with this variable included. These are the clear-cut cases.

In contrast, other cases may be less clear-cut. We can gain insight into these more difficult situations by considering the various possible relationships between the candidate IV, the other IVs, in the equation, and the DV.

Consider a situation in which X_1 and X_2 are important independent variables, but the candidate IV X_3 has no relevance to the actual process generating the dependent variable in the population. If the irrelevant variable X_3 is not related to either the other independent variables or the dependent variable, then the regression coefficients for the other variables, B_1 and B_2, will be unbiased estimates of the true values in the population. The primary cost of including X_3 in the regression model is a small increase in the standard errors of the B_1 and B_2 coefficients. This increase in the standard errors implies that confidence intervals will be slightly too large and the statistical power of the tests of B_1 and B_2 will be slightly reduced.

Now consider a situation in which X_3 is related to the DV but is unrelated to the IVs. This situation commonly occurs in randomized experiments in which the two IVs are experimental treatments and X_3 is an individual difference characteristic (e.g., IQ, an attitude, or a personality trait). Once again, the regression coefficients for X_1 and X_2 will be unbiased estimates of the true values in the population. There are two potential costs of failing to include X_3 in the regression equation. First, to the extent that the researchers are interested in individual differences, they have omitted an important predictor of behavior, particularly if there is a good theoretical rationale for inclusion of the candidate variable. Second, the standard errors of B_1 and B_2 will be larger if X_3 is not included in the regression model so that the confidence intervals for B_1 and B_2 will be larger and the corresponding significance tests will be lower in power. We will further consider this second issue in Section 8.7 in our discussion of analysis of covariance.

---

[16]Researchers who predict specific forms of nonlinear relationships between X and Y or interactions between IVs should be very cautious in the use of transformations of either X or Y. Transformations change the form of the relationship between the new variables potentially eliminating the predicted X–Y relationship when it exists in the original data.

Finally, the most difficult situation occurs when X_3 is related to X_1 and X_2 as well as Y. Here, the correct answer depends on knowing the actual process that determines Y in the population. If the analyst omits X_3 from the regression equation when it should be included, the estimates of B_1 and B_2 will be biased. If the analyst includes X_3 when it should be omitted, the estimates of B_1 and B_2 will be biased. In practice, the analyst can never know for sure which regression model is correct. Careful thought about candidate variables often provides the best solution to the omitted variable problem. Analysts should never simply "throw independent variables into a regression equation." Careful consideration of whether the addition of the candidate variable makes conceptual sense in the substantive area of research is required.[17] In Chapter 5 we will consider two statistical approaches to this dilemma. Hierarchical regression adds a set of candidate variables to the regression equation to determine how much the set of candidate variables adds to the prediction of Y over and above the contribution of the previously included independent variables (Sections 5.3–5.5). Sensitivity analysis involves estimating regression models that include and do not include the candidate variables and comparing the results. If the effect of an independent variable (e.g., X_1) that is included in all of the analyses does not change appreciably, then the analyst can claim that the effect of this variable appears to be robust regardless of which model is estimated. Other more advanced statistical methods of probing the effects of omitted variables can be found in Maddala (1988) and Mauro (1990).

Beginning analysts are advised to be very clear about which variables are included in their originally hypothesized regression model. The originally hypothesized model has a special status in statistical inference because it has been developed independently of the current data. When model modifications are made based on the current data set, then all findings become exploratory. Analysts cannot know for certain whether they have discovered an important new relationship that holds in general or whether they have detected a relationship that is peculiar to this one particular sample. Exploratory findings are potentially important and should not be ignored. However, they should be clearly labeled as exploratory in any report or publication, and their interpretation should be very tentative. Exploratory findings should be replicated in a fresh data set before any strong conclusions are reached (see Diaconis, 1985, for a fuller discussion of inferential issues in exploratory data analysis).

4.5.3 Measurement Error in the Independent Variables

When one or more independent variables has been measured with less than perfect reliability, the estimates of the regression coefficients and their standard errors for each independent variable will be biased. Methods of correcting for measurement error in the one-IV and two-IV cases were introduced in Section 4.5. A general strategy with more than two IVs is to correct each separate correlation in the full correlation matrix including all independent variables and the dependent variable for measurement error. A simple method is to apply Eq. (2.10.5) (reproduced here) to each pair of variables in the correlation matrix:

(2.10.5)
$$r_{X_t Y_t} = \frac{r_{XY}}{\sqrt{r_{XX} r_{YY}}}.$$

The corrected correlation matrix can then be used as input to standard statistical packages such as SAS, SPSS, and SYSTAT, and the desired measures of partial relationship can be

[17]Recall that the addition of IVs also changes the meaning of the regression coefficients. In the equation $\hat{Y} = B_1 X_1 + B_0, B_1$ represents the linear relationship between X_1 and Y. In the equation, $\hat{Y} = B_1 X_1 + B_2 X_2 + B_0, B_1$ represents the conditional linear relationship between X_1 and Y given that X_2 is held constant.

computed. A particular value of this method is that it produces estimates of the unstandardized and standardized regression coefficients that are corrected for measurement error. The analyst can compare these results to the corresponding results based on the uncorrected data to get an idea of the extent to which the direction and magnitude of the relationships may be affected by measurement error.

The central drawback of this simple approach to correction for unreliability is that the standard errors of the corrected regression coefficients, and hence significance tests and confidence intervals, will not be accurate. Proper estimation of the standard errors of corrected regression coefficients can be obtained from structural equation modeling programs (see, e.g., Bryne, 1998, for details of reliability correction using the LISREL program). Chapter 12 introduces the basic concepts of path analysis and structural equation modeling.

Correction for measurement error may involve difficult issues. The correction procedures assume that we have a very good estimate of the reliability for each variable. And *all* independent variables in the regression equation must be corrected for unreliability or the estimates of the regression coefficients will be biased (Won, 1982). Precise estimates of reliability are typically only available for tests that report reliabilities for large standardization samples or when the researchers' own work with the measures has involved very large samples. Estimates of reliability based on small samples will often be too high or too low, and may produce estimates of correlations between true scores that are grossly inaccurate.

Another problem that may occur when inaccurate estimates of reliability are used is that the corrected correlation matrix may no longer have the standard mathematical properties that define correlation matrices.[18] For example, correlations between true scores are sometimes found that are greater than 1.0 in magnitude, or the correlation between two variables may be higher or lower than is mathematically possible given their pattern of correlations with other variables. In such cases, more advanced techniques of correcting for measurement error may be needed (Fuller, 1987).

Although these correction methods can lead to improved estimates of the regression coefficients, the best strategy is to confront measurement error before the study is designed. Choosing the most reliable available measure of each construct will minimize the bias due to measurement error. Using multiple measures of each construct and analyzing the data using multiple indicator, structural equation models (see Chapter 12; Bollen, 1989) also leads to regression coefficients and standard errors that are corrected for measurement error. Implementing one of these procedures at the design stage helps avoid the potential problems with the reliability correction methods that occur when one or more of the reliabilities is inaccurate or if other conditions for the application of the correction procedure are not met.

4.5.4 Nonconstant Variance

We now consider situations in which our graphical examination of the residuals suggests that the form of the regression model was properly specified, but that the variance of the residuals is not constant (heteroscedasticy). Recall from Section 4.4 that estimates of the regression coefficients are unbiased in this situation, but that the standard errors may be inaccurate. Section 4.4 also presented graphical displays and statistical tests of the residuals against each IV and $\hat{Y}$. These approaches are useful in detecting whether nonconstant variance exists. However, in deciding whether or not corrective action is needed, it is more important to get an

[18]For each pair of variables, no observed correlation r_{12} can exceed the product of the square root of the reliabilities, $\sqrt{r_{11}r_{22}}$, of X_1 and X_2. For each triplet of variables, r_{12} has mathematical upper and lower limits of $r_{13}r_{23} \pm \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}$. Correlation matrices that do not have these properties are described as ill conditioned. Mathematically, correlation and covariance matrices are ill conditioned if they have at least one negative eigenvalue.

estimate of the magnitude of the nonconstant variance problem. Recall that heteroscedasticity does not have a material effect on the results of the regression analysis until the magnitude of the problem becomes "large."

One simple method of determining the magnitude of the nonconstant variance problem is to order the cases from lowest to highest according to their values on X and then to divide them into a number of sets with approximately an equal number of cases in each set. Each of these sets is termed a *slice*. The number of slices chosen reflects a compromise between having a relatively stable estimate of the variance in each slice versus having the ability to examine different portions of the data.

For example, with 200 cases, the analyst might divide the data into five sets by putting the 40 cases with the smallest values of X in set 1, the next lowest 40 cases in set 2, and so on, yielding 5 sets. The analyst calculates the variance of the residuals around the regression line within each slice separately,

$$sd^2_{Y-\hat{Y}|\text{slice}} = \sum_{i=1}^{n_{\text{slice}}} \frac{(Y-\hat{Y})^2}{(n_{\text{slice}}-2)}.$$

In this equation, n_{slice} is the number of cases in the slice and $sd^2_{Y-\hat{Y}|\text{slice}}$ represents the conditional variance of the residuals within the slice. If the ratio of the largest to the smallest conditional variance for the slices exceeds 10 or the conditional variance changes in a regular and systematic way as the IV increases in value, the analyst may wish to consider the remedial procedure of weighted least squares regression.

Weighted least squares (WLS) regression is the most commonly used remedial procedure for heteroscedasticity. Recall from Section 4.3 that in OLS estimation each case is given the same weight ($w_i = 1$) in calculating the regression coefficients. The values of B_0 and B_1 are chosen so as to minimize the value of sum of the squared residuals. With one independent variable, this expression is

$$\min\left(\sum e_i^2\right) = \min \sum (Y_i - B_1X_i - B_0)^2.$$

In contrast, in WLS each case is given a weight, w_i, depending on the precision of the observation of Y for that case. For observations for which the variance of the residuals around the regression line is high, the case is given a low weight. For observations for which the variance of the residuals around the regression line is low, the case is given a high weight. In the regression equation the values of B_0 and B_1 are chosen so as to minimize the sum of the weighted squared residuals,

(4.5.1) $$\min\left(\sum w_i e_i^2\right) = \min \sum w_i(Y_i - B_1X_i - B_0)^2.$$

When there is heteroscedasticity, WLS produces regression coefficients with the smallest possible standard errors when w_i is the inverse of the conditional variance of the residuals in the population corresponding to the specified value of X,

$$w_i = \frac{1}{\sigma^2_{Y-\hat{Y}|X}}.$$

The notation $\sigma^2_{Y-\hat{Y}|X}$ represents the variance of the residuals in the population conditional on the specified value of X. In practice, $\sigma^2_{Y-\hat{Y}|X}$ will not usually be known and must be estimated from the data. Interested readers will find an illustration of how weights are estimated in Box 4.5.1.

BOX 4.5.1
An Example of Estimating Weights for Weighted Least Squares

WLS regression is often used when the variance of the residuals has an increasing or decreasing linear relationship with X. In this case, the two-step process may be used to estimate the weights. We illustrate this process in the one predictor case.

1. We estimate the usual OLS regression equation, $Y = B_1X + B_0 + e$. The residuals are saved for each case in the sample.
2. The residual for each case is squared. The squared residuals are regressed on X in a second regression equation,

$$\hat{e}_i^2 = B_1'X + B_0'.$$

In this equation, $\hat{e}_i^2$ is the predicted value of the squared residual for case i, B_0' is the intercept, and B_1' is the slope. Note that B_0 and B_1 in the regression equation from step 1 will not typically equal B_0' and B_1' in the regression equation from step 2. To estimate w_i for case i, the value of X for case i is substituted in Eq. (4.5.1), and the weight is the inverse of the predicted value of the squared residual, $w_i = 1/\hat{e}_i^2$. Once the weights are calculated, they are typically added to the data set as a new variable with a value for each case.[19] Standard statistical packages will then perform weighted least squares regression.

Two observations about WLS regression are in order. First, the primary difficulty in using WLS regression is choosing the proper value of the weight for each case. To the degree that the weights are not accurately estimated, WLS will not give optimal performance. Consequently WLS will show the best performance when there is a large sample size or when there are multiple cases (replicates) for each fixed value of X, as occurs in a designed experiment. If the WLS weights have been properly estimated in a sample, the regression coefficients from OLS, in which each case receives an equal weight, and WLS, in which each case is weighted according to the precision of the observation, should be very similar. Because of the imprecision in estimating the weights, OLS regression will often perform nearly as well as (or sometimes even better than) WLS regression when the sample size is small, the degree to which the variance of the residuals varies as a function of X is not large, or both. Second, WLS regression involves one important cost relative to OLS regression. In WLS the measures of standardized effect size, such as R^2 and pr^2, do not have a straightforward meaning as they do in OLS. Although standard computer programs include these measures in their output, they should not normally be presented when WLS regression is used. Taken together, these two observations suggest that OLS regression will be preferable to WLS regression except in cases where the sample size is large or there is a very serious problem of nonconstant variance.

4.5.5 Nonindependence of Residuals

Nonindependence of the residuals arises from two different sources, clustering and serial dependency. We consider the remedies for each of these problems in turn.

[19] When there are a few unusually large positive or negative values of the raw residuals, e_i, that are highly discrepant from the rest of the residuals (i.e., outlying values, see Section 10.2), the weights will be very poorly estimated. When there are outlying values, regressing $|e_i|$ on X and then using $w_i = 1/(|\hat{e}_i|)^2$ will yield improved estimates of the weights (Davidian & Carroll, 1987). $(|\hat{e}_i|)^2$ is the square of the predicted value of the absolute value (ignoring sign) of the residual.

Clustering

When data are collected from groups or other clusters, the estimates of the regression coefficients are unbiased, but the standard errors will typically be too small. Although historically a number of approaches were proposed to remedy this problem, two different general approaches appear to be the most promising.

In the first approach, a set of dichotomous variables known as dummy variables is used to indicate group membership. Coding schemes for representing group membership are initially presented in Chapter 8 and then are applied to the specific problem of clustering in Section 14.2.1 (disaggregated analysis with dummy coded groups). This approach removes the effects of mean differences among the groups on the results of the regression analysis. The results are easy to interpret, and the analysis does not require assumptions beyond those of OLS regression. However, the dummy-variable approach does not permit generalization of the results beyond the specific set of g groups that have been studied.[20] This approach also excludes the study of many interesting research questions that involve group-level variables such as the influence of group-level variables (e.g., total amount of interaction among family members) on individual-level variables (e.g., individual happiness).

The second approach is known variously by the terms *multilevel models* or hierarchical linear models. A form of regression analysis know as *random coefficient regression* is utilized. Conceptually, this approach may be thought of as specifying two levels of regression equations. At level 1, a separate regression equation is used to specify the intercept and slope of relationships within each of the groups. At level 2, regression equations specify the relationships between group-level variables and the slope and intercept obtained within each group from the level 1 analyses. For example, an educational researcher might study whether the income of the neighborhood in which schools are located is related to the overall level (intercept) of math achievement in each school and also to the relationship (slope) between student IQ and math achievement within each school.

Random coefficient regression models represent an important extension of multiple regression analysis that have opened up important new lines of inquiry in several substantive areas. At the same time, these models use more complex estimation procedures than OLS. These estimation procedures have several statistical advantages over OLS, but they come at a cost of requiring more stringent assumptions than OLS regression. Chapter 14 provides an extensive introduction to both approaches to the analysis of clustered data. Raudenbush and Bryk (2002), Kreft and de Leeuw (1998), and Snijders and Bosker (1999) provide more advanced treatments.

Serial Dependency

Analyses of temporal data that include substantial serial dependency lead to regression coefficients that are unbiased, but that have standard errors that are incorrect. A data transformation procedure is used to remedy the problem, here with the goal of removing the serial dependency. Successful transformation yields transformed values of each observation that are independent. Regression analyses may then be performed on the transformed values. For interested readers Box 4.5.2 provides an illustration of the transformation procedure when the residuals show a lag 1 autocorrelation.

---

[20]In some applications generalization is not an issue because the entire population of clusters is included in the sample. For example, a political scientist studying samples of voters selected from each of the 50 states in the United States would not wish to generalize beyond this set of states. However, in most research contexts, researchers wish to make inferences about a population of groups rather than a specific set of groups.

BOX 4.5.2
Transformation of Data with Lag 1 Autocorrelation

The transformation strategy illustrated here involves separately removing the part of Y that relates to the previous observation in the series:

(4.5.2) $$Y_t^* = Y_t - r_1 Y_{t-1}.$$

In Eq. 4.5.2, Y_t^* is the transformed value of Y at time t, Y_t is the observed value of Y at time t, and Y_{t-1} is the observed value of Y at time $t-1$. r_1 is the lag 1 autocorrelation. The regression analysis is then conducted on the transformed data,

(4.5.3) $$\hat{Y}_t^* = B_1^* X_t + B_0^*$$

where $\hat{Y}_t^*$ is the predicted value of transformed Y, B_1^* is the slope for the transformed data, and B_0^* is the intercept for the transformed data.

How do the results of the regression analysis performed on the transformed Y_t^* data using Eq. 4.5.3 compare with the results of the regression analysis, $\hat{Y}_t = B_1 X_t + B_0$, performed on the original data? The transformation has no effect on the slope: $B_1^* = B_1$. We can use the results of the analysis of the transformed data directly and report the estimate of B_1, its standard error, and the significance test and confidence interval. However, $B_0^* = (1 - r_1)B_0$. To recover the original (untransformed) intercept B_0 with its corrected standard error SE_{B_0}, an adjustment of the results of the analysis of the transformed data is necessary. The adjusted values of B_0, SE_{B_0}, and the t test of B_0 are calculated as follows:

$$\text{Adjusted } B_0 = \frac{B_0^*}{1 - r_1}$$

$$\text{Adjusted } SE_{B_0} = \frac{SE_{B_0^*}}{1 - r_1}$$

$$t = \frac{\text{Adjusted } B_0}{\text{Adjusted } SE_{B_0}}$$

The transformation procedure described here assumes that r_1 is a very good estimate of the value of the lag 1 autocorrelation ρ_1 in the population. Econometric texts (e.g., Greene, 1997) present more advanced analysis procedures that simultaneously estimate the values of the regression coefficients and the autocorrelation parameter. When more complicated forms of serial dependency than lag 1 autocorrelation are detected or there are several independent variables, statistical procedures known as time series analysis are used. These procedures include specialized methods for detecting complex forms of serial dependency and for transforming each series so that unbiased estimates of the regression coefficients and their standard errors can be obtained. Section 15.8 presents an introduction to time series analysis and other approaches to temporal data. McCleary and Hay (1980) provide a comprehensive introduction to time series analysis for behavioral science researchers. Box, Jenkins, and Reinsel (1994) and Chatfield (1996) provide more advanced treatments.

4.6 SUMMARY

Chapter 4 considers the full variety of problems that may arise in multiple regression analysis and offers remedies for those problems. A key feature of good statistical analysis is becoming very familiar with one's data. We present a variety of graphical tools that can quickly provide this familiarity and help detect a number of potential problems in multiple regression analysis.

Univariate displays include the frequency histogram, stem and leaf displays, kernel density estimates, and boxplots. Scatterplots are useful in seeing both linear and nonlinear relationships between two variables, particularly when enhanced with superimposed straight lines or lowess lines. Scatterplot matrices make it possible to examine all possible pairwise relationships between the IVs and the DV (Section 4.2).

The assumptions underlying multiple regression and ordinary least squares estimation are then considered. Violations of some of the assumptions can lead to biased estimates of regression coefficients and incorrect standard errors (Section 4.3). Violations of other assumptions lead to incorrect standard errors. Serious violations of the assumptions potentially lead to incorrect significance tests and confidence intervals. Graphical and statistical methods of detecting violations of several of the assumptions including incorrect specification of the form of the regression model, omitted variables, heteroscedasticity of residuals, clustering and serial dependency, and non-normality of residuals are then presented (Section 4.5). A variety of remedies that are useful when the assumptions of regression analysis are violated are then introduced. Some of the remedies address various issues in the specification of the regression model including the form of the IV-DV relationship, omitted IVs, and measurement error in the IVs. Other remedies address nonconstant variance of the residuals, clustering, and serial dependency. A fuller presentation of some of the remedies is deferred to later chapters, where the problems receive a more in-depth treatment.

5

Data-Analytic Strategies Using Multiple Regression/Correlation

5.1 RESEARCH QUESTIONS ANSWERED BY CORRELATIONS AND THEIR SQUARES

Until this point we have presented regression/correlation analysis as if the typical investigation proceeded by selecting a single set of IVs and producing a single regression equation that is then used to summarize the findings. Life, however, is seldom so simple for the researcher. The coefficient or set of coefficients that provide the answers depend critically on the questions being asked. There is a wealth of information about the interrelationships among the variables not extractable from a single equation. It is, perhaps, the skill with which other pertinent information can be ferreted out that distinguishes the expert data analyst from the novice. In this chapter we address five major issues of strategy that should be considered in using MRC analysis. The first examines the fit between the research questions and the coefficients that answer them. The second examines some options and considerations for making regression coefficients more substantively interpretable. The third strategic consideration is the use of sequential or hierarchical analysis to wrest the best available answers from the data. The fourth is the employment of sets of independent variables in hierarchical analyses. The final section discusses strategies for controlling and balancing Type I and Type II errors of inference in MRC.

It is often the case that regression coefficients provide the most informative answers to scientific questions. However, there are a number of questions that are best answered by correlation coefficients and their comparisons. Indeed, it is sometimes hard to avoid the suspicion that correlation coefficients and squared correlations of various kinds are not reported or not focused on, even when most relevant, because they are typically so small. There is something rather discouraging about a major effort to study a variable that turns out to account uniquely for 1 or 2 percent of the dependent variable variance. We have tried to indicate that such a small value may still represent a material effect (see Section 2.10.1), but there is no getting around the more customary disparagement of effects of this magnitude.

Different questions are answered by different coefficients and comparisons among them. Standard statistical programs in MRC produce both regression and correlation coefficients for the use of scientists in interpreting their findings. All coefficients, but especially correlation coefficients, need a definable population to which to generalize, of which one has a random, or at least representative or unbiased sample. Without a population framework some coefficients may