

**THIRD  
EDITION**

# Applied Linear Regression Models

---

**John Neter**

*University of Georgia*

**Michael H. Kutner**

*The Cleveland Clinic Foundation*

**Christopher J. Nachtsheim**

*University of Minnesota*

**William Wasserman**

*Syracuse University*

**IRWIN**

Chicago • Bogotá • Boston • Buenos Aires • Caracas  
London • Madrid • Mexico City • Sydney • Toronto

# Diagnostics and Remedial Measures

When a regression model, such as the simple linear regression model (2.1), is considered for an application, we can usually not be certain in advance that the model is appropriate for that application. Any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this chapter, we discuss some simple graphic methods for studying the appropriateness of a model, as well as some formal statistical tests for doing so. We also consider some remedial techniques that can be helpful when the data are not in accordance with the conditions of regression model (2.1). We conclude the chapter with a case example that brings together the concepts and methods presented in this and the earlier chapters.

While the discussion in this chapter is in terms of the appropriateness of the simple linear regression model (2.1), the basic principles apply to all - statistical models discussed in this book. In later chapters, additional methods useful for examining the appropriateness of statistical models and other remedial measures will be presented, as well as methods for validating the statistical model.

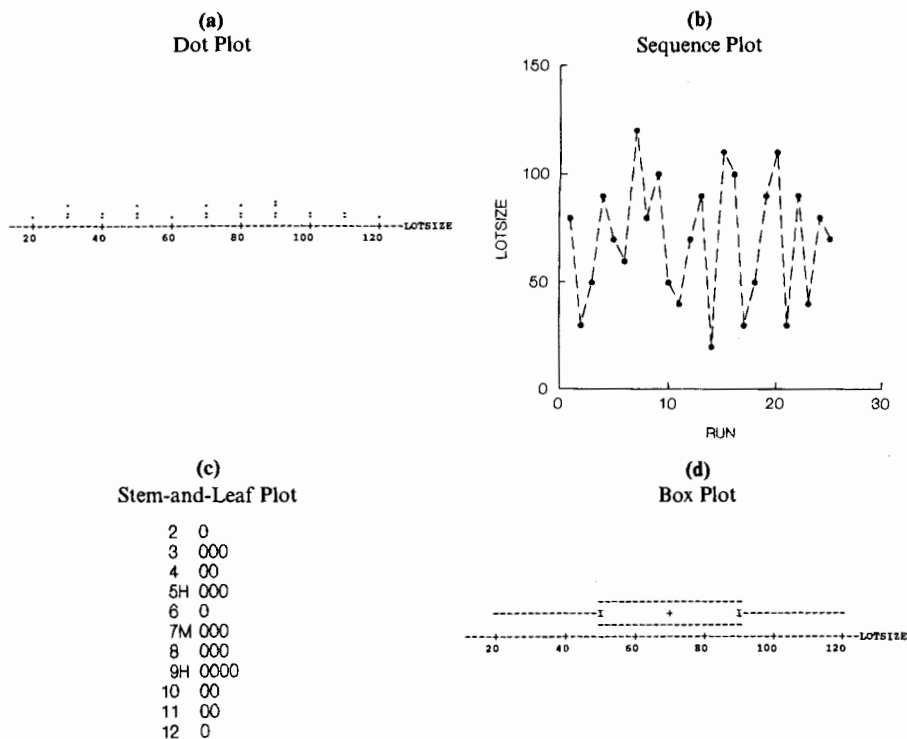
---

## 3.1 Diagnostics for Predictor Variable

We begin by considering some graphic diagnostics for the predictor variable. We need diagnostic information about the predictor variable to see if there are any outlying  $X$  values that could influence the appropriateness of the fitted regression function. We discuss the role of influential cases in detail in Chapter 9. Diagnostic information about the range and concentration of the  $X$  levels in the study is also useful for ascertaining the range of validity for the regression analysis.

Figure 3.1a contains a simple *dot plot* for the lot sizes in the Toluca Company example in Figure 1.10. A dot plot is helpful when the number of observations in the data set is not large. The dot plot in Figure 3.1a shows that the minimum and maximum lot sizes are 20 and 120, respectively, that the lot size levels are spread throughout this interval, and that there are no lot sizes that are far outlying. The dot plot also shows that in a number of cases several runs were made for the same lot size.

**FIGURE 3.1** MINITAB and SYGRAPH Diagnostic Plots for Predictor Variable—Toluca Company Example.



A second useful diagnostic for the predictor variable is a *sequence plot*. Figure 3.1b contains a time sequence plot of the lot sizes for the Toluca Company example. Lot size is here plotted against production run (i.e., against time sequence). The points in the plot are connected to show more effectively the time sequence. Sequence plots should be utilized whenever data are obtained in a sequence, such as over time or for adjacent geographic areas. The sequence plot in Figure 3.1b contains no special pattern. If, say, the plot had shown that smaller lot sizes had been utilized early on and larger lot sizes later on, this information could be very helpful for subsequent diagnostic studies of the aptness of the fitted regression model.

Figures 3.1c and 3.1d contain two other diagnostic plots that present information similar to the dot plot in Figure 3.1a. The *stem-and-leaf plot* in Figure 3.1c provides information similar to a frequency histogram. By displaying the last digits, this plot also indicates here that all lot sizes in the Toluca Company example were multiples of 10. The letter M in the SYGRAPH output denotes the stem where the median is located, and the letter H denotes the stems where the first and third quartiles (hinges) are located.

The *box plot* in Figure 3.1d shows the minimum and maximum lot sizes, the first and third quartiles, and the median lot size. We see that the middle half of the lot sizes range from 50 to 90, and that they are fairly symmetrically distributed because the median is located in the middle of the central box. A box plot is particularly helpful when there are many observations in the data set.

## 3.2 Residuals

Direct diagnostic plots for the response variable  $Y$  are ordinarily not too useful in regression analysis because the values of the observations on the response variable are a function of the level of the predictor variable. Instead, diagnostics for the response variable are usually carried out indirectly through an examination of the residuals.

The residual  $e_i$ , as defined in (1.16), is the difference between the observed value  $Y_i$  and the fitted value  $\hat{Y}_i$ :

$$(3.1) \quad e_i = Y_i - \hat{Y}_i$$

The residual may be regarded as the observed error, in distinction to the unknown true error  $\varepsilon_i$  in the regression model:

$$(3.2) \quad \varepsilon_i = Y_i - E\{Y_i\}$$

For regression model (2.1), the error terms  $\varepsilon_i$  are assumed to be independent normal random variables, with mean 0 and constant variance  $\sigma^2$ . If the model is appropriate for the data at hand, the observed residuals  $e_i$  should then reflect the properties assumed for the  $\varepsilon_i$ . This is the basic idea underlying *residual analysis*, a highly useful means of examining the aptness of a statistical model.

### Properties of Residuals

**Mean.** The mean of the  $n$  residuals  $e_i$  for the simple linear regression model (2.1) is, by (1.17):

$$(3.3) \quad \bar{e} = \frac{\sum e_i}{n} = 0$$

where  $\bar{e}$  denotes the mean of the residuals. Thus, since  $\bar{e}$  is always 0, it provides no information as to whether the true errors  $\varepsilon_i$  have expected value  $E\{\varepsilon_i\} = 0$ .

**Variance.** The variance of the  $n$  residuals  $e_i$  is defined as follows for regression model (2.1):

$$(3.4) \quad \frac{\sum (e_i - \bar{e})^2}{n - 2} = \frac{\sum e_i^2}{n - 2} = \frac{SSE}{n - 2} = MSE$$

If the model is appropriate,  $MSE$  is, as noted earlier, an unbiased estimator of the variance of the error terms  $\sigma^2$ .

**Nonindependence.** The residuals  $e_i$  are not independent random variables because they involve the fitted values  $\hat{Y}_i$  which are based on the same fitted regression function. As a result, the residuals for regression model (2.1) are subject to two constraints. These are constraint (1.17)—that the sum of the  $e_i$  must be 0—and constraint (1.19)—that the products  $X_i e_i$  must sum to 0.

When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals  $e_i$  is relatively unimportant and can be ignored for most purposes.

### Semistudentized Residuals

At times, it is helpful to standardize the residuals for residual analysis. Since the standard deviation of the error terms  $\varepsilon_i$  is  $\sigma$ , which is estimated by  $\sqrt{MSE}$ , it is natural to consider the following form of standardization:

$$(3.5) \quad e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

If  $\sqrt{MSE}$  were an estimate of the standard deviation of the residual  $e_i$ , we would call  $e_i^*$  a studentized residual. However, the standard deviation of  $e_i$  is complex and varies for the different residuals  $e_i$ , and  $\sqrt{MSE}$  is only an approximation of the standard deviation of  $e_i$ . Hence, we call the statistic  $e_i^*$  in (3.5) a *semistudentized residual*. We shall take up studentized residuals in Chapter 9. Both semistudentized residuals and studentized residuals can be very helpful in identifying outlying observations.

### Departures from Model to Be Studied by Residuals

We shall consider the use of residuals for examining six important types of departures from the simple linear regression model (2.1) with normal errors:

1. The regression function is not linear.
2. The error terms do not have constant variance.
3. The error terms are not independent.
4. The model fits all but one or a few outlier observations.
5. The error terms are not normally distributed.
6. One or several important predictor variables have been omitted from the model.

---

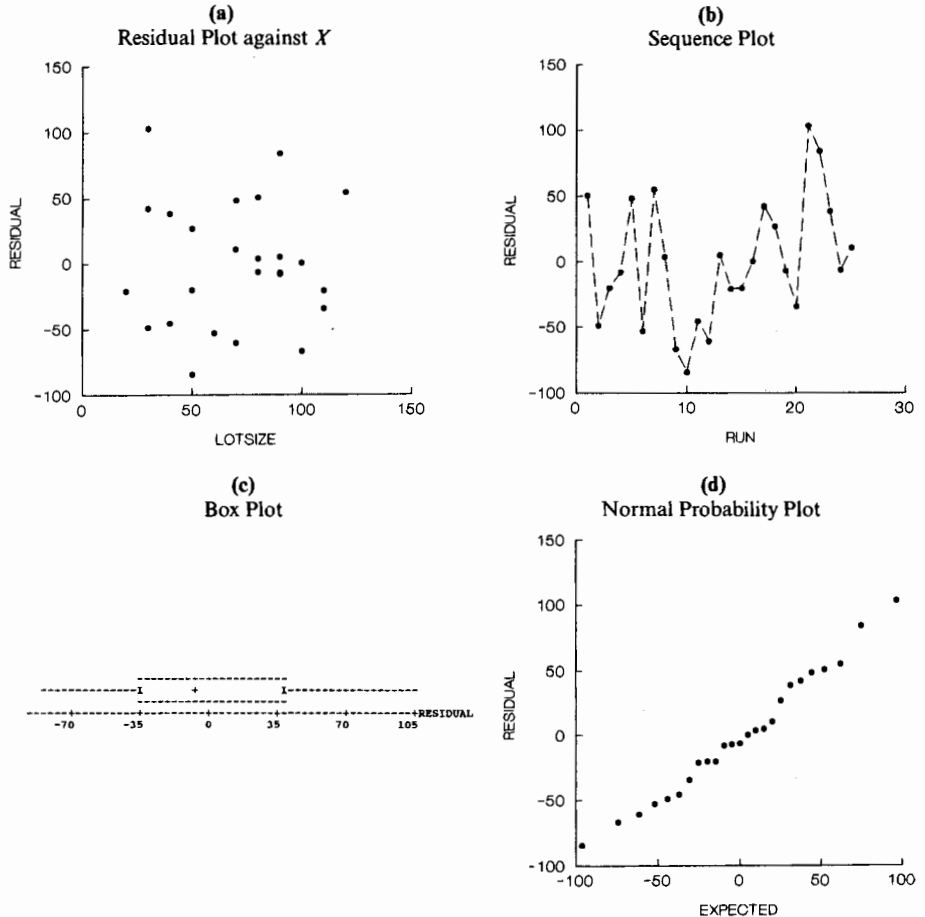
## 3.3 Diagnostics for Residuals

We take up now some informal diagnostic plots of residuals to provide information on whether any of the six types of departures from the simple linear regression model (2.1) just mentioned are present. The following plots of residuals (or semistudentized residuals) will be utilized here for this purpose:

1. Plot of residuals against predictor variable
2. Plot of absolute or squared residuals against predictor variable
3. Plot of residuals against fitted values
4. Plot of residuals against time or other sequence
5. Plots of residuals against omitted predictor variables
6. Box plot of residuals
7. Normal probability plot of residuals

Figure 3.2 contains, for the Toluca Company example, MINITAB and SYGRAPH plots of the residuals in Table 1.2 against the predictor variable and against time, a box plot, and

**FIGURE 3.2** MINITAB and SYGRAPH Diagnostic Residual Plots—Toluca Company Example.

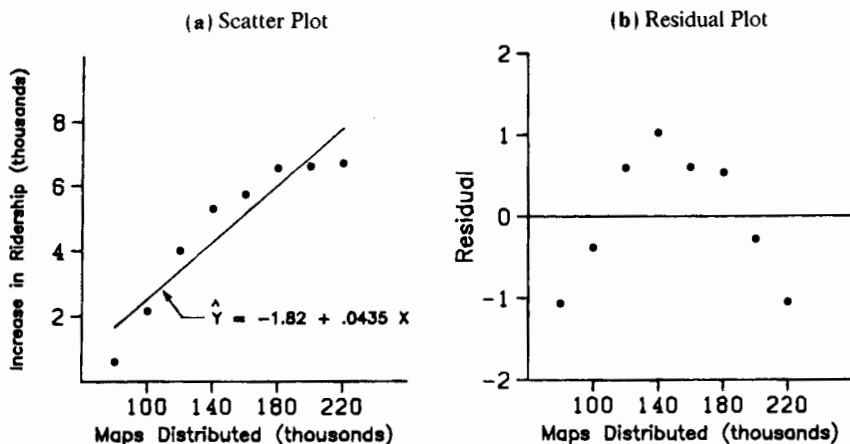


a normal probability plot. All of these plots, as we shall see, support the appropriateness of regression model (2.1) for the data.

We turn now to consider how residual analysis can be helpful in studying each of the six departures from regression model (2.1).

### **Nonlinearity of Regression Function**

Whether a linear regression function is appropriate for the data being analyzed can be studied from a *residual plot against the predictor variable* or, equivalently, from a *residual plot against the fitted values*. Nonlinearity of the regression function can also be studied from a *scatter plot*, but this plot is not always as effective as a residual plot. Figure 3.3a contains a scatter plot of the data and the fitted regression line for a study of the relation between maps distributed and bus ridership in eight test cities. Here,  $X$  is the number of bus transit maps distributed free to residents of the city at the beginning of the test period and  $Y$  is the increase during the test period in average daily bus rid-

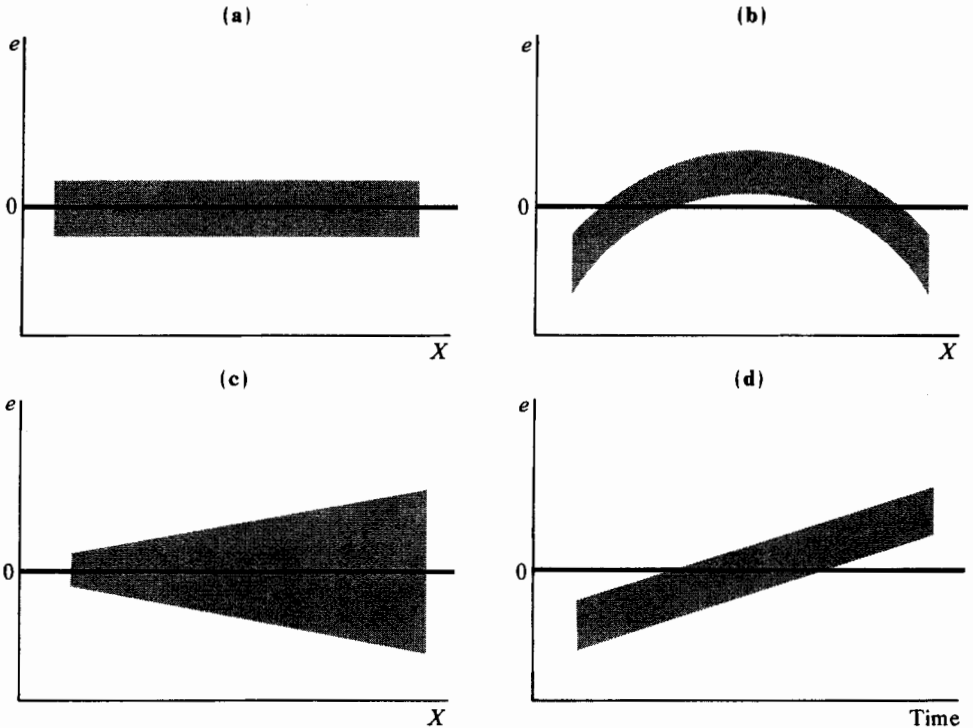
**FIGURE 3.3** Scatter Plot and Residual Plot Illustrating Nonlinear Regression Function—Transit Example.**TABLE 3.1** Number of Maps Distributed and Increase in Ridership—Transit Example.

	(1)	(2)	(3)	(4)
	Increase in Ridership (thousands)	Maps Distributed (thousands)	Fitted Value	Residual
City <i>i</i>	$Y_i$	$X_i$	$\hat{Y}_i$	$Y_i - \hat{Y}_i = e_i$
1	.60	80	1.66	-1.06
2	6.70	220	7.75	-1.05
3	5.30	140	4.27	1.03
4	4.00	120	3.40	.60
5	6.55	180	6.01	.54
6	2.15	100	2.53	-.38
7	6.60	200	6.88	-.28
8	5.75	160	5.14	.61

$$\hat{Y} = -1.82 + .0435X$$

ership during nonpeak hours. The original data and fitted values are given in Table 3.1, columns 1, 2, and 3. The plot suggests strongly that a linear regression function is not appropriate.

Figure 3.3b presents a plot of the residuals, shown in Table 3.1, column 4, against the predictor variable  $X$ . The lack of fit of the linear regression function is even more strongly suggested by the residual plot against  $X$  in Figure 3.3b than by the scatter plot. Note that the residuals depart from 0 in a systematic fashion; they are negative for smaller  $X$  values, positive for medium-size  $X$  values, and negative again for large  $X$  values.

**FIGURE 3.4** Prototype Residual Plots.

In this case, both Figures 3.3a and 3.3b point out the lack of linearity of the regression function. In general, however, the residual plot is to be preferred, because it has some important advantages over the scatter plot. First, the residual plot can easily be used for examining other facets of the aptness of the model. Second, there are occasions when the scaling of the scatter plot places the  $Y_i$  observations close to the fitted values  $\hat{Y}_i$ , for instance, when there is a steep slope. It then becomes more difficult to study the appropriateness of a linear regression function from the scatter plot. A residual plot, on the other hand, can clearly show any systematic pattern in the deviations around the fitted regression line under these conditions.

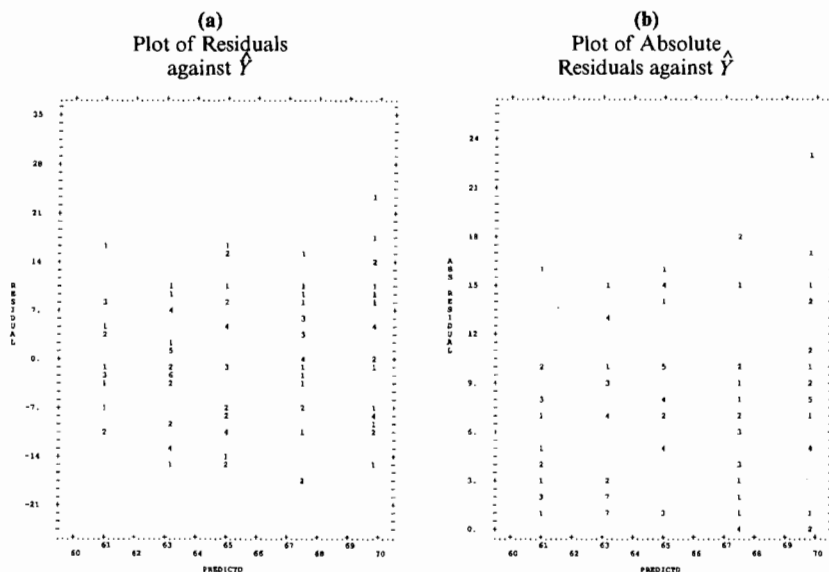
Figure 3.4a shows a prototype situation of the residual plot against  $X$  when a linear regression model is appropriate. The residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative. This is the case in Figure 3.2a for the Toluca Company example.

Figure 3.4b shows a prototype situation of a departure from the linear regression model that indicates the need for a curvilinear regression function. Here the residuals tend to vary in a systematic fashion between being positive and negative. This is the case in Figure 3.3b for the transit example. A different type of departure from linearity would, of course, lead to a picture different from the prototype pattern in Figure 3.4b.

### Note

A plot of residuals against the fitted values  $\hat{Y}$  provides equivalent information as a plot of residuals against  $X$  for the simple linear regression model, and thus is not needed in addition

**FIGURE 3.5 BMDP Residual Plots Illustrating Nonconstant Error Variance—Blood Pressure Example.**



to the residual plot against  $X$ . The two plots provide the same information because the fitted values  $\hat{Y}_i$  are a linear function of the values  $X_i$  for the predictor variable. Thus, only the  $X$  scale values, not the basic pattern of the plotted points, are affected by whether the residual plot is against the  $X_i$  or the  $\hat{Y}_i$ . For curvilinear regression and multiple regression, on the other hand, separate plots of the residuals against the fitted values and against the predictor variable(s) are usually helpful. ■

### Nonconstancy of Error Variance

Plots of the residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant. Figure 3.5a shows a BMDP residual plot against the fitted values  $\hat{Y}$  for a study of the relation between diastolic blood pressure of female children ( $Y$ ) and their age ( $X$ ). Note that the horizontal axis is labeled PREDICTD, which stands for “predicted,” an alternative term often used for “fitted” value. The numerical values shown in the graph indicate the number of residuals falling on or near a point. The plot suggests that the larger the fitted value  $\hat{Y}$  is, the more spread out the residuals are. Since the relation between blood pressure and age is positive, this suggests that the error variance is larger for older children than for younger ones.

The prototype plot in Figure 3.4a exemplifies residual plots when the error term variance is constant. The residual plot in Figure 3.2a for the Toluca Company example is of this type, suggesting that the error terms have constant variance here.

Figure 3.4c shows a prototype picture of residual plots when the error variance increases with  $X$ . In many business, social science, and biological science applications, departures from constancy of the error variance tend to be of the “megaphone” type shown in

Figure 3.4c, as in the blood pressure example in Figure 3.5a. One can also encounter error variances decreasing with increasing levels of the predictor variable and occasionally varying in some more complex fashion.

Plots of the absolute values of the residuals or of the squared residuals against the predictor variable  $X$  or against the fitted values  $\hat{Y}$  are also useful for diagnosing nonconstancy of the error variance since the signs of the residuals are not meaningful for examining the constancy of the error variance. These plots are especially useful when there are not many cases in the data set because plotting of either the absolute or squared residuals places all of the information on changing magnitudes of the residuals above the horizontal zero line so that one can more readily see whether the magnitude of the residuals (irrespective of sign) is changing with the level of  $X$  or  $\hat{Y}$ .

Figure 3.5b contains a BMDP plot of the absolute residuals against the fitted values for the blood pressure example. This plot shows more clearly that the residuals tend to be larger in absolute magnitude for larger fitted values.

### Presence of Outliers

Outliers are extreme observations. Residual outliers can be identified from *residual plots against  $X$  or  $\hat{Y}$* , as well as from *box plots*, *stem-and-leaf plots*, and *dot plots* of the residuals. Plotting of semistudentized residuals is particularly helpful for distinguishing outlying observations, since it then becomes easy to identify residuals that lie many standard deviations from zero. A rough rule of thumb when the number of cases is large is to consider semistudentized residuals with absolute value of four or more to be outliers. We shall take up more refined procedures for identifying outliers in Chapter 9.

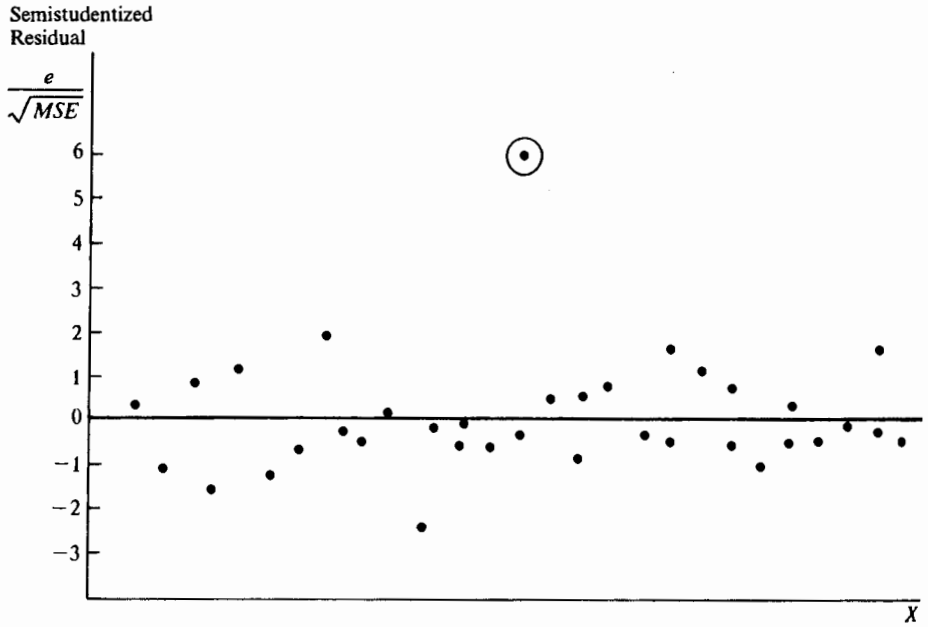
The residual plot in Figure 3.6 presents semistudentized residuals and contains one outlier, which is circled. Note that this residual represents an observation almost six standard deviations from the fitted value.

Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect, and hence should be discarded. A major reason for discarding it is that under the least squares method, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the *squared* deviations is minimized. This could cause a misleading fit if indeed the outlying observation resulted from a mistake or other extraneous cause. On the other hand, outliers may convey significant information, as when an outlier occurs because of an interaction with another predictor variable omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

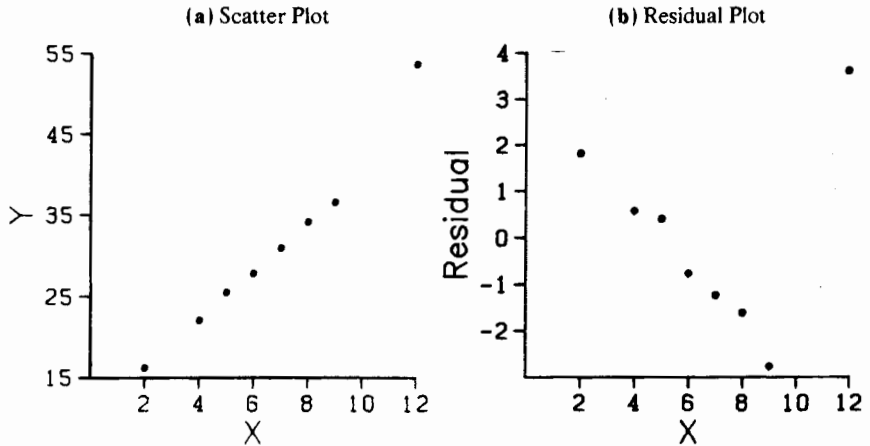
#### Note

When a linear regression model is fitted to a data set with a small number of cases and an outlier is present, the fitted regression can be so distorted by the outlier that the residual plot may improperly suggest a lack of fit of the linear regression model, in addition to flagging the outlier. Figure 3.7 illustrates this situation. The scatter plot in Figure 3.7a presents a situation where all observations except the outlier fall around a straight-line statistical relationship. When a linear regression function is fitted to these data, the outlier causes such a shift in the fitted regression line as to lead to a systematic pattern of deviations from the fitted line for the other observations, suggesting a lack of fit of the linear regression function. This is shown by the residual plot in Figure 3.7b. ■

**FIGURE 3.6** Residual Plot with Outlier.

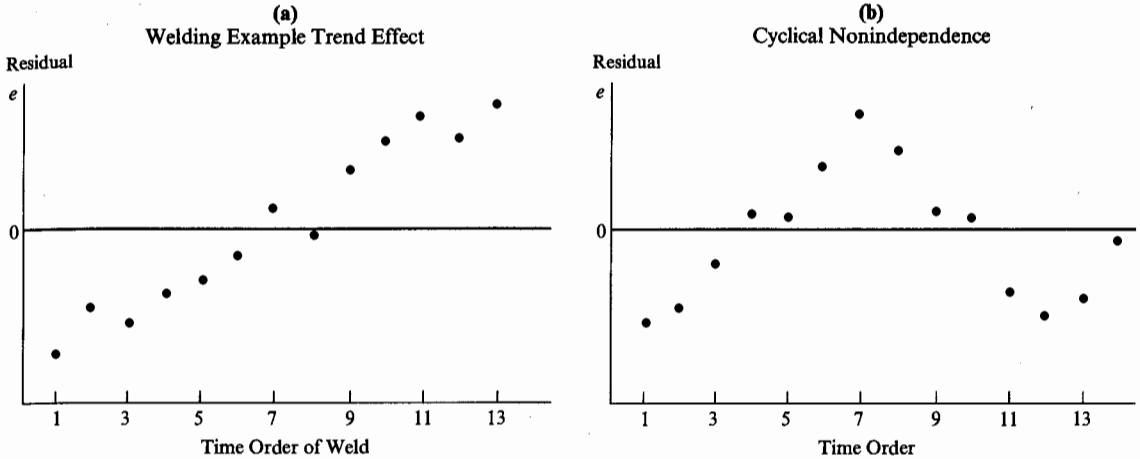


**FIGURE 3.7** Distorting Effect on Residuals Caused by an Outlier When Remaining Data Follow Linear Regression.



**Nonindependence of Error Terms**

Whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographic areas, it is a good idea to prepare a *sequence plot of the residuals*. The purpose of plotting the residuals against time or in some other type of sequence is to

**FIGURE 3.8 Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.**

see if there is any correlation between error terms that are near each other in the sequence. Figure 3.8a contains a time sequence plot of the residuals in an experiment to study the relation between the diameter of a weld ( $X$ ) and the shear strength of the weld ( $Y$ ). An evident correlation between the error terms stands out. Negative residuals are associated mainly with the early trials, and positive residuals with the later trials. Apparently, some effect connected with time was present, such as learning by the welder or a gradual change in the welding equipment, so the shear strength tended to be greater in the later welds because of this effect.

A prototype residual plot showing a time-related trend effect is presented in Figure 3.4d, which portrays a linear time-related trend effect, as in the welding example. It is sometimes useful to view the problem of nonindependence of the error terms as one in which an important variable (in this case, time) has been omitted from the model. We shall discuss this type of problem shortly.

Another type of nonindependence of the error terms is illustrated in Figure 3.8b. Here the adjacent error terms are also related, but the resulting pattern is a cyclical one with no trend effect present.

When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the base line 0, such as the scattering shown in Figure 3.2b for the Toluca Company example. Lack of randomness can take the form of too much or too little alternation of points around the zero line. In practice, there is little concern with the former because it does not arise frequently. Too little alternation, in contrast, frequently occurs, as in the welding example in Figure 3.8a.

### Note

When the residuals are plotted against  $X$ , as in Figure 3.3b for the transit example, the scatter may not appear to be random. For this plot, however, the basic problem is probably not lack of independence of the error terms but a poorly fitting regression function. This, indeed, is the situation portrayed in the scatter plot in Figure 3.3a. ■

**TABLE 3.2** Residuals and Expected Values under Normality—Toluca Company Example.

Run $i$	(1) Residual $e_i$	(2) Rank $k$	(3) Expected Value under Normality
1	51.02	22	51.95
2	-48.47	5	-44.10
3	-19.88	10	-14.76
...	...	...	...
23	38.83	19	31.05
24	-5.98	13	0
25	10.72	17	19.93

### Nonnormality of Error Terms

As we noted earlier, small departures from normality do not create any serious problems. Major departures, on the other hand, should be of concern. The normality of the error terms can be studied informally by examining the residuals in a variety of graphic ways.

**Distribution Plots.** A *box plot* of the residuals is helpful for obtaining summary information about the symmetry of the residuals and about possible outliers. Figure 3.2c contains a box plot of the residuals in the Toluca Company example. No serious departures from symmetry are suggested by this plot. A *histogram*, *dot plot*, or *stem-and-leaf plot* of the residuals can also be helpful for detecting gross departures from normality. However, the number of cases in the regression study must be reasonably large for any of these plots to convey reliable information about the shape of the distribution of the error terms.

**Comparison of Frequencies.** Another possibility when the number of cases is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 68 percent of the residuals  $e_i$  fall between  $\pm\sqrt{MSE}$  or about 90 percent fall between  $\pm 1.645\sqrt{MSE}$ . When the sample size is moderately large, corresponding  $t$  values may be used for the comparison.

To illustrate this procedure, we again consider the Toluca Company example of Chapter 1. Table 3.2, column 1, repeats the residuals from Table 1.2. We see from Figure 2.2 that  $\sqrt{MSE} = 48.82$ . Using the  $t$  distribution, we expect under normality about 90 percent of the residuals to fall between  $\pm t(.95; 23)\sqrt{MSE} = \pm 1.714(48.82)$ , or between -83.68 and 83.68. Actually, 22 residuals, or 88 percent, fall within these limits. Similarly, under normality, we expect about 60 percent of the residuals to fall between -41.89 and 41.89. The actual percentage here is 52 percent. Thus, the actual frequencies here are reasonably consistent with those expected under normality.

**Normal Probability Plot.** Still another possibility is to prepare a *normal probability plot of the residuals*. Here each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

Table 3.2, column 1, contains the residuals for the Toluca Company example. To find the expected values of the ordered residuals under normality, we utilize the facts that (1) the expected value of the error terms for regression model (2.1) is zero, and (2) the standard deviation of the error terms is estimated by  $\sqrt{MSE}$ . Statistical theory has shown that for a normal random variable with mean 0 and estimated standard deviation  $\sqrt{MSE}$ , a good approximation of the expected value of the  $k$ th smallest observation in a random sample of  $n$  is:

$$(3.6) \quad \sqrt{MSE} \left[ z \left( \frac{k - .375}{n + .25} \right) \right]$$

where  $z(A)$  as usual denotes the  $(A)100$  percentile of the standard normal distribution.

Using this approximation, let us calculate the expected values of the residuals under normality for the Toluca Company example. Column 2 of Table 3.2 shows the ranks of the residuals, with the smallest residual being assigned rank 1. We see that the rank of the residual for run 1,  $e_1 = 51.02$ , is 22, which indicates that this residual is the 22nd smallest among the 25 residuals. Hence, for this residual  $k = 22$ . We found earlier (Table 2.1) that  $MSE = 2,384$ . Hence:

$$\frac{k - .375}{n + .25} = \frac{22 - .375}{25 + .25} = \frac{21.625}{25.25} = .8564$$

so that the expected value of this residual under normality is:

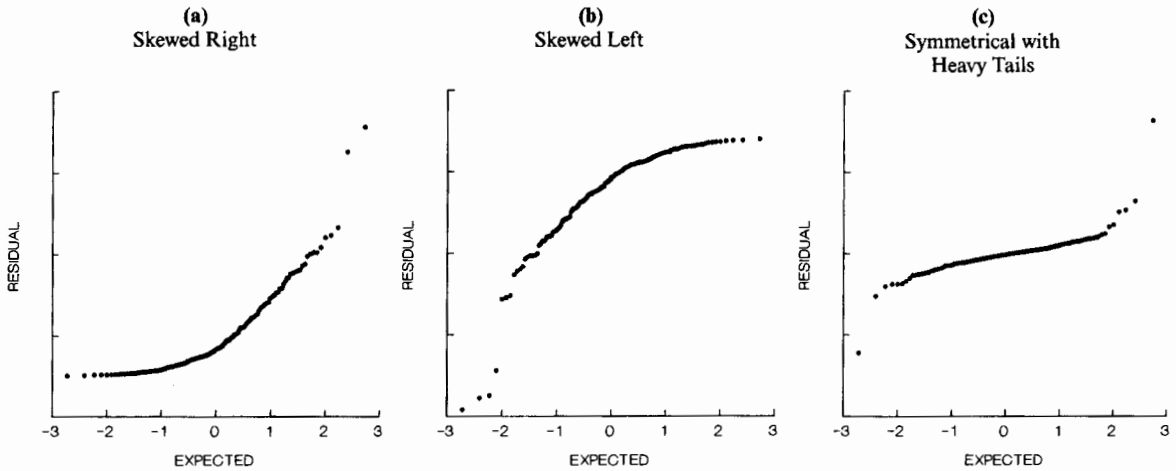
$$\sqrt{2,384}[z(.8564)] = \sqrt{2,384}(1.064) = 51.95$$

Similarly, the expected value of the residual for run 2,  $e_2 = -48.47$ , is obtained by noting that the rank of this residual is  $k = 5$ ; in other words, this residual is the fifth smallest one among the 25 residuals. Hence, we require  $(k - .375)/(n + .25) = (5 - .375)/(25 + .25) = .1832$ , so that the expected value of this residual under normality is:

$$\sqrt{2,384}[z(.1832)] = \sqrt{2,384}(-.9032) = -44.10$$

Table 3.2, column 3, contains the expected values under the assumption of normality for a portion of the 25 residuals. Figure 3.2d presents a plot of the residuals against their expected values under normality. Note that the points in Figure 3.2d fall reasonably close to a straight line, suggesting that the distribution of the error terms does not depart substantially from a normal distribution.

Figure 3.9 shows three normal probability plots when the distribution of the error terms departs substantially from normality. Figure 3.9a shows a normal probability plot when the error term distribution is highly skewed to the right. Note the concave-upward shape of the plot. Figure 3.9b shows a normal probability plot when the error term distribution is highly skewed to the left. Here, the pattern is concave downward. Finally, Figure 3.9c shows a normal probability plot when the distribution of the error terms is symmetrical but has heavy tails; in other words, the distribution has higher probabilities in the tails than a normal distribution. Note the concave-downward curvature in the plot at the left end, corresponding to the plot for a left-skewed distribution, and the concave-upward plot at the right end, corresponding to a right-skewed distribution.

**FIGURE 3.9 Normal Probability Plots when Error Term Distribution Is Not Normal.**

### Comments

1. Many computer packages will prepare normal probability plots, either automatically or at the option of the user. Some of these plots utilize semistudentized residuals, others omit the factor  $\sqrt{MSE}$  in (3.6), but neither of these variations affect the nature of the plot.
2. For continuous data, ties among the residuals should occur only rarely. If two residuals do have the same value, a simple procedure is to use the average rank for the tied residuals for calculating the corresponding expected values. ■

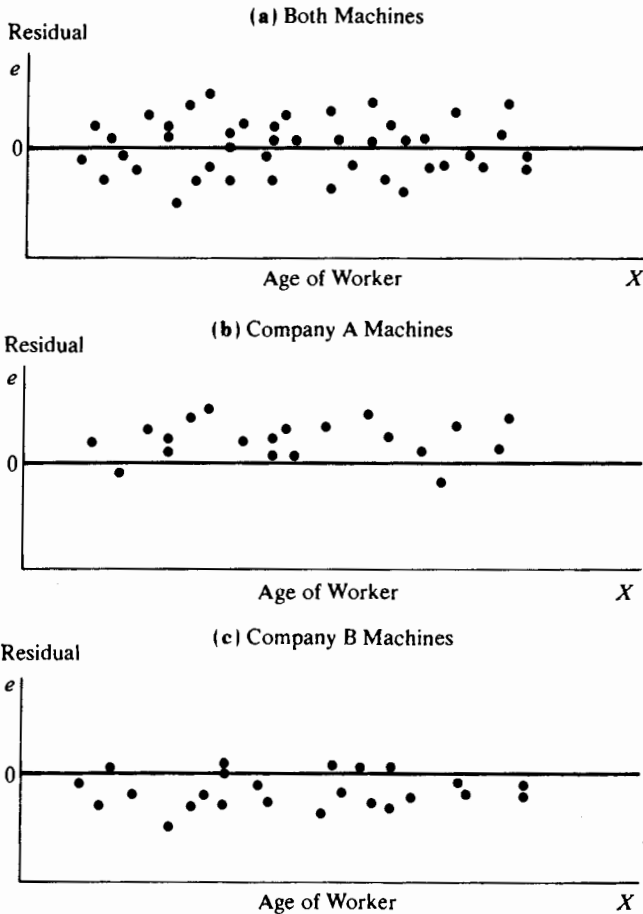
**Difficulties in Assessing Normality.** The analysis for model departures with respect to normality is, in many respects, more difficult than that for other types of departures. In the first place, random variation can be particularly mischievous when studying the nature of a probability distribution unless the sample size is quite large. Even worse, other types of departures can and do affect the distribution of the residuals. For instance, residuals may appear to be not normally distributed because an inappropriate regression function is used or because the error variance is not constant. Hence, it is usually a good strategy to investigate these other types of departures first, before concerning oneself with the normality of the error terms.

### Omission of Important Predictor Variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response. The time variable cited earlier in the welding example is an illustration. The purpose of this additional analysis is to determine whether there are any other key variables that could provide important additional descriptive and predictive power to the model.

As another example, in a study to predict output by piece-rate workers in an assembling operation, the relation between output ( $Y$ ) and age ( $X$ ) of worker was studied for a sample of employees. The plot of the residuals against  $X$ , shown in Figure 3.10a, indicates no ground for suspecting the appropriateness of the linearity of the regression function or the constancy

**FIGURE 3.10** Residual Plots for Possible Omission of Important Predictor Variable—Productivity Example.



of the error variance. Since machines produced by two companies (A and B) are used in the assembling operation and could have an effect on output, residual plots against  $X$  by type of machine were undertaken and are shown in Figures 3.10b and 3.10c. Note that the residuals for Company A machines tend to be positive, while those for Company B machines tend to be negative. Thus, type of machine appears to have a definite effect on productivity, and output predictions may turn out to be far superior when this variable is added to the model.

While this second example dealt with a qualitative variable (type of machine), the residual analysis for an additional quantitative variable is analogous. The residuals are plotted against the additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable.

### Note

We do not say that the original model is “wrong” when it can be improved materially by adding one or more predictor variables. Only a few of the factors operating on any response variable

$Y$  in real-world situations can be included explicitly in a regression model. The chief purpose of residual analysis in identifying other important predictor variables is therefore to test the adequacy of the model and see whether it could be improved materially by adding one or more predictor variables. ■

### **Some Final Comments**

1. We discussed model departures one at a time. In actuality, several types of departures may occur together. For instance, a linear regression function may be a poor fit and the variance of the error terms may not be constant. In these cases, the prototype patterns of Figure 3.4 can still be useful, but they would need to be combined into composite patterns.
2. Although graphic analysis of residuals is only an informal method of analysis, in many cases it suffices for examining the aptness of a model.
3. The basic approach to residual analysis explained here applies not only to simple linear regression but also to more complex regression and other types of statistical models.
4. Most of the routine work in residual analysis can be handled on computers. Almost all regression programs supply the fitted values and residuals, and routines are generally available whereby the various types of residual plots can be obtained. ■

---

## **3.4 Overview of Tests Involving Residuals**

Graphic analysis of residuals is inherently subjective. Nevertheless, subjective analysis of a variety of interrelated residual plots will frequently reveal difficulties with the model more clearly than particular formal tests. There are occasions, however, when one wishes to put specific questions to a test. We now briefly review some of the relevant tests.

Most statistical tests require independent observations. As we have seen, however, the residuals are dependent. Fortunately, the dependencies become quite small for large samples, so that one can usually then ignore them.

### **Tests for Randomness**

A runs test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specifically designed for lack of randomness in least squares residuals, is the Durbin-Watson test. This test is discussed in Chapter 12.

### **Tests for Constancy of Variance**

When a residual plot gives the impression that the variance may be increasing or decreasing in a systematic manner related to  $X$  or  $E\{Y\}$ , a simple test is based on the rank correlation between the absolute values of the residuals and the corresponding values of the predictor variable. Two other simple tests for constancy of the error variance—the modified Levene test and the Breusch-Pagan test—are discussed in Section 3.6.

### **Tests for Outliers**

A simple test for identifying an outlier observation involves fitting a new regression line to the other  $n - 1$  observations. The suspect observation, which was not used in fitting the new line, can now be regarded as a new observation. One can calculate the probability that in  $n$