Similarly, scientists may prefer logarithmic transformations of both Y and X when studying the relation between radioactive decay (Y) of a substance and time (X) for a curvilinear relation of the type illustrated in Figure 3.15b because the slope of the regression line for the transformed variables then measures the decay rate.

2. After a transformation has been tentatively selected, residual plots and other analyses described earlier need to be employed to ascertain that the simple linear regression model (2.1) is appropriate for the transformed data.

3. When transformed models are employed, the estimators b_0 and b_1 obtained by least squares have the least squares properties with respect to the transformed observations, not the original ones.

4. The maximum likelihood estimate of λ with the Box-Cox procedure is subject to sampling variability. In addition, the error sum of squares *SSE* is often fairly stable in a neighborhood around the estimate. It is therefore often reasonable to use a nearby λ value for which the power transformation is easy to understand. For example, use of $\lambda = 0$ instead of the maximum likelihood estimate $\hat{\lambda} = .13$ or use of $\lambda = -.5$ instead of $\hat{\lambda} = -.79$ may facilitate understanding without sacrificing much in terms of the effectiveness of the transformation. To determine the reasonableness of using an easier-to-understand value of λ, one should examine the flatness of the likelihood function in the neighborhood of $\hat{\lambda}$, as we did in the plasma levels example. Alternatively, one may construct an approximate confidence interval for λ; the procedure for constructing such an interval is discussed in Reference 3.10.

5. When the Box-Cox procedure leads to a λ value near 1, no transformation of Y may be needed.

■

3.10 Exploration of Shape of Regression Function

Scatter plots often indicate readily the nature of the regression function. For instance, Figure 1.3 clearly shows the curvilinear nature of the regression relationship between steroid level and age. At other times, however, the scatter plot is complex and it becomes difficult to see the nature of the regression relationship, if any, from the plot. In these cases, it is helpful to explore the nature of the regression relationship by fitting a smoothed curve without any constraints on the regression function. These smoothed curves are also called *nonparametric regression curves*. They are useful not only for exploring regression relationships but also for confirming the nature of the regression function when the scatter plot visually suggests the nature of the regression relationship.

Many smoothing methods have been developed for obtaining smoothed curves for time series data, where the X_i denote time periods that are equally spaced apart. The *method of moving averages* uses the mean of the Y observations for adjacent time periods to obtain smoothed values. For example, the mean of the Y values for the first three time periods in the time series might constitute the first smoothed value corresponding to the middle of the three time periods, in other words, corresponding to time period 2. Then the mean of the Y values for the second, third, and fourth time periods would constitute the second smoothed value, corresponding to the middle of these three time periods, in other words, corresponding to time period 3, and so on. Special procedures are required for obtaining smoothed values at the two ends of the time series. The larger the successive neighborhoods used for obtaining the smoothed values, the smoother the curve will be.

The *method of running medians* is similar to the method of moving averages, except that the median is used as the average measure in order to reduce the influence of outlying

observations. With this method, as well as with the moving average method, successive smoothing of the smoothed values and other refinements may be undertaken to provide a suitable smoothed curve for the time series. Reference 3.11 provides a good introduction to the running median smoothing method.

Many smoothing methods have also been developed for regression data when the X values are not equally spaced apart. A simple smoothing method, *band regression,* divides the data set into a number of groups or "bands" consisting of adjacent cases according to their X levels. For each band, the median X value and the median Y value are calculated, and the points defined by the pairs of these median values are then connected by straight lines. For example, consider the following simple data set divided into three groups:

X	Y	Median X	Median Y
2.0	13.1	2.7	14.4
3.4	15.7		
3.7	14.9		
4.5	16.8	4.5	16.8
5.0	17.1		
5.2	16.9	5.55	17.35
5.9	17.8		

The three pairs of medians are then plotted on the scatter plot of the data and connected by straight lines as a simple smoothed nonparametric regression curve.

Lowess Method

The *lowess method,* developed by Cleveland (Ref. 3.12), is a more refined nonparametric method than band regression. It obtains a smoothed curve by fitting successive linear regression functions in local neighborhoods. The name lowess stands for *locally weighted regression scatter plot smoothing.* The method is similar to the moving average and running median methods in that it uses a neighborhood around each X value to obtain a smoothed Y value corresponding to that X value. It obtains the smoothed Y value at a given X by fitting a linear regression to the data in the neighborhood of the X value and then using the fitted value at X as the smoothed value. To illustrate this concretely, let (X_1, Y_1) denote the sample case with the smallest X value, (X_2, Y_2) denote the sample case with the second smallest X value, and so on. If neighborhoods of three X values are used with the lowess method, then a linear regression would be fitted to the data:

$$(X_1, Y_1) \quad (X_2, Y_2) \quad (X_3, Y_3)$$

The fitted value at X_2 would constitute the smoothed value corresponding to X_2. Another linear regression would be fitted to the data:

$$(X_2, Y_2) \quad (X_3, Y_3) \quad (X_4, Y_4)$$

and the fitted value at X_3 would constitute the smoothed value corresponding to X_3. Smoothed values at each end of the X range are also obtained by the lowess procedure.

The lowess method uses a number of refinements in obtaining the final smoothed values to improve the smoothing and to make the procedure robust to outlying observations.

1. The linear regression is weighted to give cases further from the middle X level in each neighborhood smaller weights.
2. To make the procedure robust to outlying observations, the linear regression fitting is repeated, with the weights revised so that cases that had large residuals in the first fitting receive smaller weights in the second fitting.
3. To improve the robustness of the procedure further, step 2 is repeated one or more times by revising the weights according to the size of the residuals in the latest fitting.

To implement the lowess procedure, one must choose the size of the successive neighborhoods to be used when fitting each linear regression. One must also choose the weight function that gives less weight to neighborhood cases with X values far from each center X level and another weight function that gives less weight to cases with large residuals. Finally, the number of iterations to make the procedure robust must be chosen.

In practice, two iterations appear to be sufficient to provide robustness. Also, the weight functions suggested by Cleveland appear to be adequate for many circumstances. Hence, the primary choice to be made for a particular application is the size of the successive neighborhoods. The larger the size, the smoother the function but the greater the danger that the smoothing will lose essential features of the regression relationship. It may require some experimentation with different neighborhood sizes in order to find the size that best brings out the regression relationship. We explain the lowess method in detail in Chapter 11 in the context of multiple regression. Specific choices of weight functions and neighborhood sizes are discussed there.

Example

Figure 3.18a contains a scatter plot based on a study of research quality at 24 research laboratories. The response variable is a measure of the quality of the research done at the laboratory, and the explanatory variable is a measure of the volume of research performed at the laboratory. Note that it is very difficult to tell from this scatter plot whether or not a relationship exists between research quality and quantity. Figure 3.18b repeats the scatter plot and also shows the lowess smoothed curve. The curve suggests that there might be somewhat higher research quality for medium-sized laboratories. However, the scatter is great so that this suggested relationship should be considered only as a possibility. Also, because any particular measures of research quality and quantity are so limited, other measures should be considered to see if these corroborate the relationship suggested in Figure 3.18b.

Use of Smoothed Curves to Confirm Fitted Regression Function

Smoothed curves are useful not only in the exploratory stages when a regression model is selected but they are also helpful in confirming the regression function chosen. The procedure for confirmation is simple: The smoothed curve is plotted together with the confidence band for the fitted regression function. If the smoothed curve falls within the confidence band, we have supporting evidence of the appropriateness of the fitted regression function.

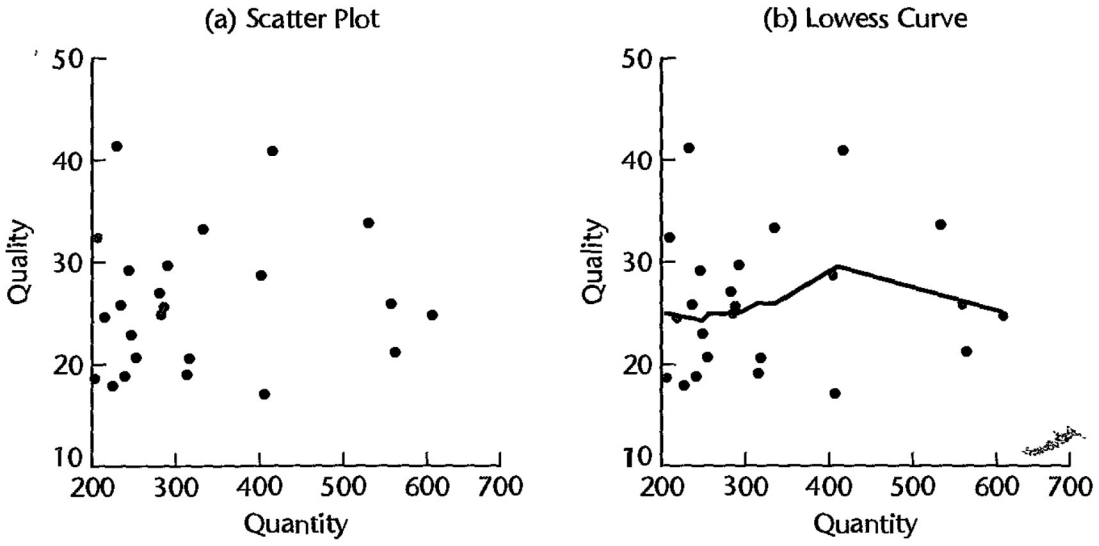

FIGURE 3.18
MINITAB Scatter Plot and Lowess Smoothed Curve—Research Laboratories Example.

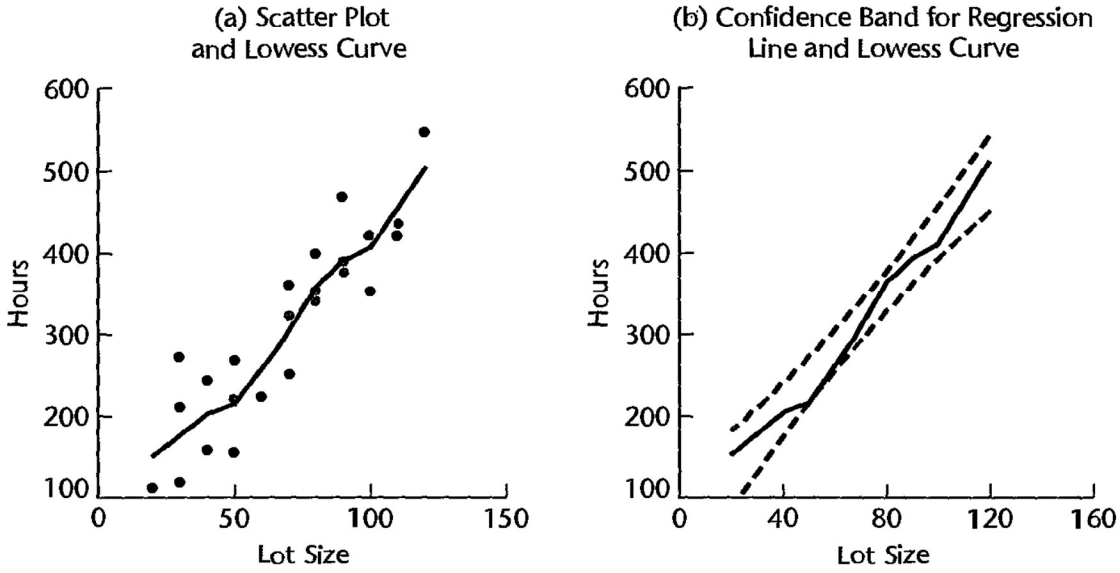

FIGURE 3.19
MINITAB Lowess Curve and Confidence Band for Regression Line—Toluca Company Example.

Example

Figure 3.19a repeats the scatter plot for the Toluca Company example from Figure 1.10a and shows the lowess smoothed curve. It appears that the regression relation is linear or possibly slightly curved. Figure 3.19b repeats the confidence band for the regression line from Figure 2.6 and shows the lowess smoothed curve. We see that the smoothed curve falls within the confidence band for the regression line and thereby supports the appropriateness of a linear regression function.

Comments

1. Smoothed curves, such as the lowess curve, do not provide an analytical expression for the functional form of the regression relationship. They only suggest the shape of the regression curve.

2. The lowess procedure is not restricted to fitting linear regression functions in each neighborhood. Higher-degree polynomials can also be utilized with this method.