12
Nonlinear Relationships

Assuming linearity between two variables when modeling their relationship often results in reasonably good models that are useful and easy to interpret. But sometimes we have reason to believe a relationship is not linear, or the evidence compels us to accept that it is not. In spite of its name, linear regression analysis can be used to model relationships that are better described with curves than with straight lines. In this chapter we discuss reasons you might choose to fit a curve to a relationship rather than a straight line, and we show how to detect nonlinearity visually as well as using polynomial regression. We also give a brief overview of spline regression, an interesting extension of regression analysis that allows for chaining of line or curve segments to capture complex forms of nonlinearity. We end with a discussion of transformations, often used to make nonlinear relationships approximate linear ones.

12.1 Linear Regression Can Model Nonlinear Relationships

Relationships between variables are sometimes better described with curves than with straight lines. A graph showing world population on the vertical axis against time on the horizontal axis would constantly curve upward, with the growth accelerating rapidly with time. Human height against age rises more slowly during childhood than in the early teen years but levels off later. A plot of "commitment to democracy" versus the extent to which a person identifies as politically conservative versus politically liberal might show greater commitment among those in the middle of the ideology continuum than among those on either the liberal or the conservative end of the spectrum. Desire to acquire more money might be especially high among people who have very little, slowly drop off as in-

come increases, and perhaps climb again among people who are already very wealthy.

It may come as a surprise that a statistical technique called *linear* regression analysis can be used to fit curves. It can. In this chapter, we show some of the ways this is done.

12.1.1 When Must Curves Be Fitted?

In a scatterplot of Y against X, sometimes you can see that a curve better describes a relationship than does a straight line. It may be that you could easily draw a curve freehand through the scatterplot that seems to fit better than any straight line that a regression program would generate. But there are times when you need to go beyond this informal means of representing curvilinearity. These include

- When you must estimate Y from X.
- When you want to test whether the relationship is curvilinear against the null hypothesis that it is linear.
- When you must estimate the value of X at which Y is maximized or minimized, such as the amplification volume at which a person's speech is perceived clearest, or the length of rest breaks that maximizes productivity.
- When you must correct for a nonlinear relationship between Y and a covariate when studying the relationship between Y and independent variable X.

Consider the data represented by the scatterplot in Figure 12.1. It is obvious that no straight line adequately characterizes the relationship between X and Y. The best-fitting regression line of the form $Y = b_0 + b_1X$ is superimposed on the scatterplot. The equation for this line is $\hat{Y} = 3.289 - 0.220X$. It is the best-fitting line by the least squares criterion. In this example, $R = 0.591$, $SS_{residual} = 6.003$, and we know that no equation of this form would result in a smaller $SS_{residual}$ or larger R.

But consider a *quadratic* equation of the form $Y = b_0 + b_1X + b_2X^2$. This is the equation for a parabola. The equation $\hat{Y} = 1.254 + 1.597X - 0.359X^2$ is superimposed on Figure 12.1, which is the best-fitting parabola for these data. Just looking at the plot, it obviously fits much better than the linear model. Statistics confirm the better fit, as $R = 0.905$ and $SS_{residual} = 1.666$ for this equation, which was found simply by regressing Y on X and X^2. R is

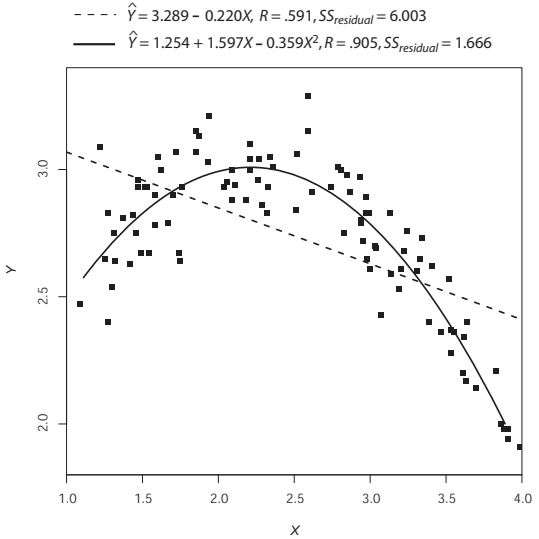

FIGURE 12.1. The best-fitting linear and quadratic model for these data.

much larger and $SS_{residual}$ is much smaller for the quadratic model than the linear model. So we have produced a better-fitting equation relating Y to X by adding the regressor X^2 to the model. Thus, linear regression analysis can be used to fit parabolas to data. Indeed, it can be used to fit other kinds of functions to data that are curves or semblances of curves.

This example illustrates each of the four points above. If your goal was to generate an estimate or prediction of Y from X, clearly you would do better using the model with X^2 than you would the model without it. You could also statistically compare the fit of the linear model to the nonlinear model to formally test whether the relationship is better described as curvilinear rather than linear. This would be the same as testing the null hypothesis that the regression coefficient for X^2 is equal to zero. Using calculus, you can derive that the estimated peak in Y occurs when $X = 2.224$; for those with a calculus background, the first derivative of the equation for $\hat{Y}$ with respect to X is $1.597 - 2 \times 0.359X$, which is equal to zero when $X = 2.224$. And suppose that X was a covariate. The procedures we described in Chapter 3 and elsewhere could result in improper control of X if you assumed that the relationship between Y and X was linear. But using

X^2 along with X as regressors in the model along with your independent variable of interest may reduce or eliminate this problem.

This latter point is worth developing further. Let the covariate be labeled C, and let X and Y be independent and dependent variables, respectively. Imagine that C has a mean near zero (either naturally or because you have made it so), meaning that C and C^2 are uncorrelated or nearly uncorrelated (we develop this point in section 12.2.4). Now suppose that Y is determined entirely by C^2 as $Y = C^2$. And further suppose that C also entirely determines X in the same way: $X = C^2$. Thus, $Y = X$, and both correlate zero or nearly so with C. If you failed to control for the curvilinear effect of C on Y, you would mistakenly conclude that X determines Y completely, when it actually has no effect at all, because Y is determined entirely by C.

The distortion in the apparent effect of X on Y occurs in this example because the relationship between X and C mirrors that between Y and C. But even in the absence of this, failure to control for curvilinear effects of covariates can distort results in the opposite direction by increasing $MS_{residual}$, which makes it harder to identify effects of X on Y that actually do exist, because all other things being equal, standard errors for regression coefficients are larger when $MS_{residual}$ is larger (recall equation 4.3).

12.1.2 The Graphical Display of Curvilinearity

When there are no covariates to complicate matters, a simple scatterplot depicting the relationship between two variables can be very useful both for seeing that a relationship is curvilinear and for discerning the nature of the curvilinearity. To take a few examples from the Roman alphabet, a scatterplot depicting nonlinearity between X and Y, with Y on the vertical axis and X on the horizonal axis, may look something like an L, with a sharp, rapid drop in Y as X increases, but a flattening of Y as X increases further. Or it could look like a U, with Y higher on the extremes of X than in the moderate values of X. The inverse of this would be a lowercase n, with Y lower in the extremes of X but higher in the middle of X. A J-shaped relationship would appear with Y relatively flat with increases in X with a sharp spike upward in Y once X reaches a certain value. Other forms of nonlinearity that are possible may not look like letters from the alphabet.

However, it is more difficult than you might think to depict or discern a nonlinear relationship between X and Y when there are covariates. If we have an independent variable X, a dependent variable Y, and one or more covariates C, and if there is a curvilinear relation between X and Y

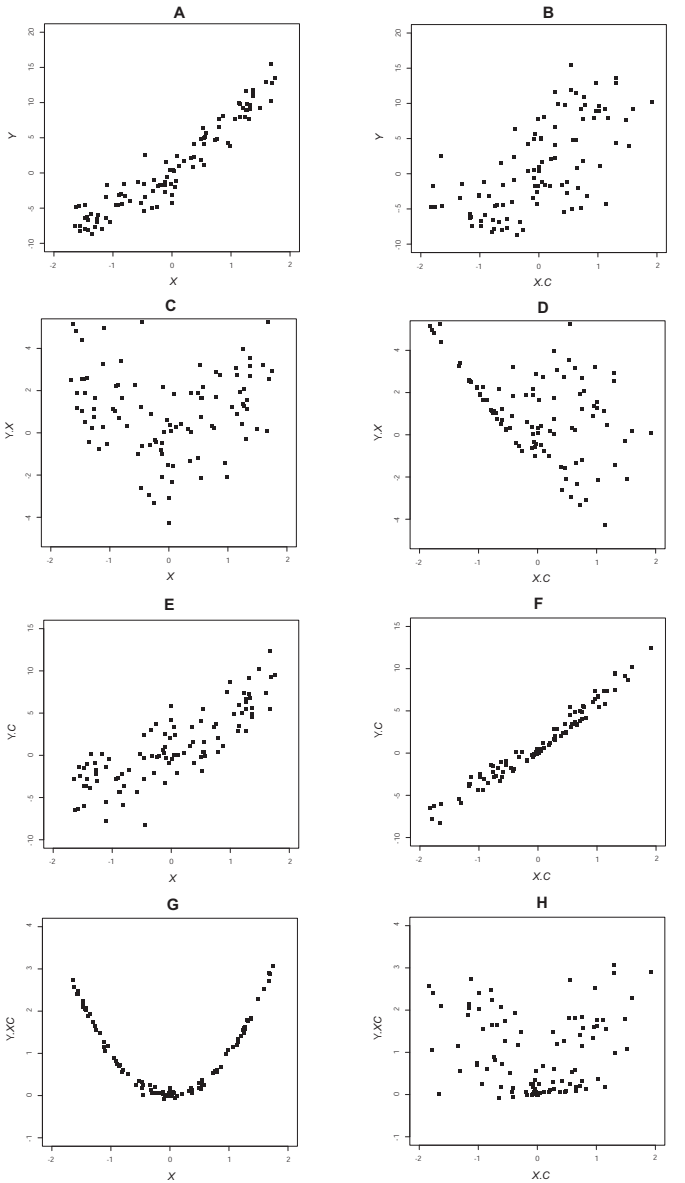

FIGURE 12.2. Eight possible scatterplots of *Y* against *X* with a covariate *C*. Plot G is the *residual scatterplot*.

when covariates are controlled, then there are no fewer than eight different scatterplots that we might think would display this curvilinearity. This is because there are four "forms" of Y we might consider: Y itself, the portion of Y independent of X ($Y.X$), the portion of Y independent of C ($Y.C$), and the portion of Y independent of X and C ($Y.XC$). In Chapter 3 we discussed that these portions of Y are residuals from a regression (e.g., $Y.C$ is the residual from a regression estimating Y from C). There are also two forms of X we might consider: X itself, and the portion of X independent of C ($X.C$). By combining the four forms of Y with the two forms of X, we can generate eight different scatterplots that we might imagine would display any curvilinearity between X and Y. And indeed any of these eight will work if X is independent of C and neither X nor C has any linear effect on Y. But abandoning any one of these three conditions can make the curvilinearity invisible, or nearly so, in four of these eight scatterplots, abandoning a second condition makes it invisible, or nearly so, in two more, and abandoning a third makes it invisible, or nearly so, in one more. The only scatterplot that is impervious to violations of all three conditions is the *residual scatterplot*, which is the plot of $Y.XC$ against X.

This point is illustrated in Figure 12.2, which shows these eight scatterplots for a sample with two regressors. This artificial data set is fairly typical, except that Y was defined as an exact nonlinear function of X and C to make any nonlinearity as visible as possible. The exact definition of Y used was $Y = 5X + 1X^2 + 10C$, meaning X is nonlinearity related to Y when C is controlled. Curvilinearity is clearly visible only in the residual scatterplot, which is plot G in the lower left corner ($Y.XC$ against X). The semipartial scatterplot (Y against $X.C$) described in section 3.3.1 is plot B, and the partial scatterplot ($Y.C$ against $X.C$) described in section 3.3.2 is plot F. They can hide even substantial nonlinearity. Curvilinearity is barely visible in plots C ($Y.X$ against X) and H ($Y.XC$ against $X.C$) but is crystal clear in the residual scatterplot.

Residual scatterplots provide the best graphical method for detecting nonlinearity and discovering its nature, but they have a major limitation that creates the need for nongraphical methods. One limitation of any graphical approach is the inefficiency of the human eye in detecting nonlinearity. This is illustrated in Figure 12.3. If you didn't know otherwise, you would probably think that the relationship between X and Y depicted there is linear. Yet in these data, nonlinearity is statistically significant at the .01 level and can easily be detected by polynomial regression introduced in section 12.2, even though that nonlinearity is essentially invisible

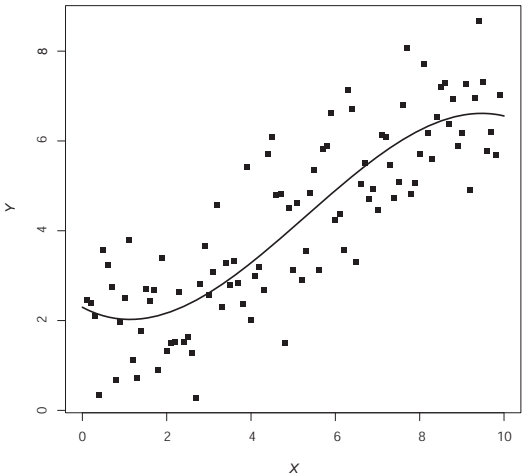

FIGURE 12.3. Real nonlinearity is sometimes hard to see in a scatterplot.

to the eye. This can be particularly problematic when there are nonlinear relations among regressors. In that situation, nonlinearity between one regressor and Y may be totally invisible even in a residual scatterplot.

An alternative problem is the tendency for the human mind to see patterns among even a random dispersion of dots. That is, you might think you see nonlinearity, but that nonlinearity is not actually present when formally tested. But whether it is failing to see real nonlinearity, or interpreting linearity as if it were nonlinearity, nongraphical methods are a good addition to and typically even better than graphical methods that rely on the subjective assessments of the perceiver. We cover some nongraphical methods in the next two sections.

12.2 Polynomial Regression

12.2.1 Basic Principles

Polynomial regression fits curves to data by using regressors that are successive powers, such as X, X^2, X^3, and so forth. The "order" of the polynomial is defined by the largest power in the polynomial. Figure 12.4 graphically depicts four equations relating Y to X. The linear equation is the one we have focused on throughout most of this book, where Y changes by the

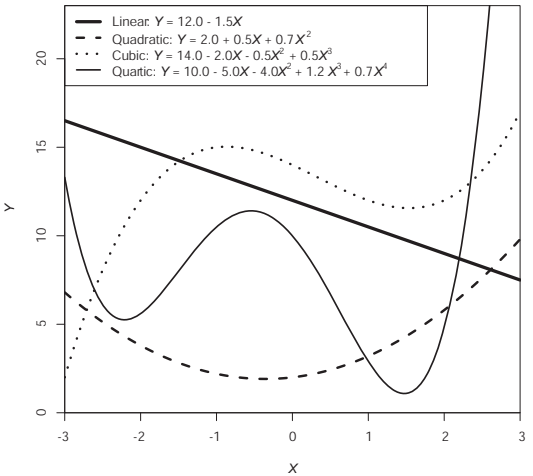

FIGURE 12.4. Some example polynomial models of the relationship between *X* and *Y*.

same amount as X increases by a fixed amount. A quadratic polynomial would take the form $Y = b_0 + b_1X + b_2X^2$ and is thus of "second order," because two is the largest power of X. A quadratic polynomial allows only a single "bend" in the relationship between X and Y, as in Figure 12.4. Adding a third power of X (thus yielding a "third-order" polynomial) results in a *cubic model*: $Y = b_0 + b_1X + b_2X^2 + b_3X^3$. This model allows for two bends in the curve, as can be seen in Figure 12.4. It would be exceedingly rare when using polynomial regression to add more than a third power of a variable to a model, but it is possible. Figure 12.4 depicts a quartic model, which by definition has a fourth power and thus is of the form $Y = b_0 + b_1X + b_2X^2 + b_3X^3 + b_4X^4$. This function allows three bends in the curve.

As Figure 12.4 depicts, the higher the order of the polynomial for X, the more complex the curve relating X to Y can be. The shape of the curve is also determined by the regression coefficients given to each of the powers of the variable. A characteristic of a polynomial of second order or higher is that the amount Y changes as X changes by a fixed unit depends on the starting point of X. So adding one unit to X will have a different effect on the amount Y changes depending on the value of X at which you start. Indeed, this is an informal definition of a curvilinear relationship between X and Y.

Some people criticize polynomial regression as excessively mechanical. Such critics argue that one should choose a curve whose shape makes scientific sense, which a polynomial may not. This is certainly good practice when possible. But polynomial regression can do a decent job representing curvilinear relationships that may not conform exactly to other kinds of functions (e.g., a logarithmic function; see section 12.4). Polynomial regression is also very versatile, because the shape a polynomial takes can be modified substantially by the amount of weight each power receives in the generation of Y, and your regression program will figure out how to weight each power in order to minimize $SS_{residual}$ and thus maximize the correlation between Y and $\hat{Y}$. Although it may be true that very few nonlinear relationships are truly parabolic, taking a U or inverted U shape, some nonlinear relationships between X and Y can be well described with a quadratic function within the domain of measurement of X.

Polynomials can also be nice ways of dealing with nonlinearity in covariates. Even if the relationship between an independent variable X and a dependent variable Y is linear, when those variables relate nonlinearly to a covariate C, it is important to allow for that nonlinearity in order to properly visualize and estimate the partial association between X and Y. We wouldn't typically care if the polynomial is a substantively or theoretically meaningful representation of the nonlinear relationship between a covariate and independent and dependent variables if it does a good job at capturing that nonlinearity and thereby affords a better adjustment for constructing measures of partial association between key variables in your analysis.

Polynomial regression is often used as a means of testing for nonlinearity in the relationship between X and Y. Because polynomials can describe such a wide range of curves, a test of nonlinearity can be conducted by determining if adding successive powers or sets of powers of X improves the fit of the model to a statistically significant degree. The test described in section 5.3.3 can be used for this purpose. We will see an example of this in section 12.2.2.

When a variable X is included as a regressor along with various powers of that variable, we usually think of that set of variables as a compound variable representing X. So, for example, if you think that age is nonlinearly related to something like attitudes toward gun control, you could use age as well as age^2 and perhaps even age^3 as regressors in the model. Any test involving age would involve all three of these. For instance, you could test whether gun control is related to age while controlling for sex

and income by adding age, age^2, and age^3 to a model of gun control that already contains income and sex. An improvement in fit as indexed by a statistically significant increase in R is evidence of a partial relationship between age and gun control, without imposing the assumption that this relationship is linear. But ordinarily, you would start with age and then decide whether adding powers of age improves the fit of the model, because it is easier to interpret linear relationships, and we wouldn't want to add an unnecessary complexity to a model unless the data (or relevant theory or past literature) suggested it was necessary to do so.

You would almost never include higher powers of a regressor in a model without including all of the lower powers as regressors as well. Consider, for example, the equation $Y = 2 + 3X^2$. This equation contains the second power of X but not the first. As a result, the line for this equation must pass through the point $X = 0, Y = 2$. This is very restrictive and not likely to be consistent with your data. When you include X, the function is no longer so restricted. Notice that $Y = 2 + 3X^2$ could be written in equivalent form as $Y = 2 + 0X + X^2$. Leaving X out of the model but including X^2 is like forcing the regression coefficient for X to be zero, and this is not likely to fit the data as well as if you let X's regression coefficient be something else. It is better to let your regression program figure out how to weight X in tandem with X^2 rather than imposing this constraint on the estimation process.

12.2.2 An Example

We illustrate polynomial regression using the POLITICS data file, which comes from a nationally representative survey of people living in the United States at the time of data collection. The dependent variable Y is score on a test of political knowledge (*pknow*), and we will estimate political knowledge from frequency of use of traditional news sources (X), named *news* in the data file. Participants in the study were asked three questions about how many days (0 through 7) during the typical week they read the newspaper, watch the national network news broadcast, and watch their local televised news broadcast. Responses to these three questions were averaged to produce the measure of traditional news use. We will look for evidence of nonlinearity between news use and political knowledge, while holding constant the respondent's age (C_1), sex (C_2), and SES (C_3, labeled *ses* in the data, defined as the average of the person's standardized education level and income).

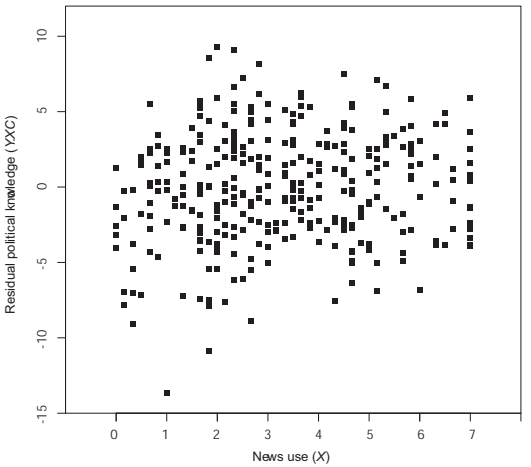

FIGURE 12.5. A residual scatterplot depicting the association between political knowledge and news use. The residuals are departures from estimated knowledge from a model that includes sex, age, SES, and news use.

Regressing political knowledge on news use X and the covariates, but without any higher powers of news use (without X^2, X^3, etc.), yields $R = 0.566$. The regression coefficient for news use is 0.265 and statistically significant, $t(335) = 2.214, p = .028$, indicating that people who use the news more frequently know more about politics. More specifically, two people who differ by 1 day in their typical news use but are equal on the covariates are estimated to differ by 0.265 units in their knowledge, with the more frequent news user being more political knowledgeable. But the meaningfulness of this depends on the partial relationship being linear.

Figure 12.5 is a residual scatterplot depicting the relationship between covariate-adjusted political knowledge and unadjusted news use. You can probably see some evidence of nonlinearity, as the residuals appear to be larger (more positive) in the center of the X distribution than in the extremes of X. But we should do a formal test.

When the square of news use is added to the model, the resulting model is

$$\hat{Y} = 7.168 + 1.372X - 0.156X^2 + 0.022C_1 + 1.720C_2 + 2.472C_3 \qquad (12.1)$$

The regression coefficient for the square of news use is statistically significant, $t(334) = -2.807, p = .005$. This test is equivalent to the change in the fit of the model when the square of news use is added to the model. Without X^2, $R^2 = 0.320$, but with X^2, $R^2 = 0.336$. This is a statistically significant increase, $F(1, 334) = 7.879, p = .005$. The increase in R^2 of .016 is the proportion of the variability in political knowledge uniquely attributable to the *square* of news use. If we wanted the proportion attributable uniquely to news use, we'd have to look at difference in the squared multiple correlations between a model that excludes news use *and* the square of news use, because news use is a compound variable in this model. Doing so, along with a test of significance as described in section 5.3.3, yields a difference of 0.026 in the two model R^2s, $F(2, 334) = 6.441, p = .002$. So news use uniquely accounts for about 2.6% of the variance in political knowledge.

Figure 12.6 visually depicts equation 12.1. This figure was generated by setting C_1, C_2, and C_3 to their sample means[1] and plotting estimated political knowledge for many values of news use (X and therefore X^2). As can be seen, holding age, SES, and sex constant, political knowledge is estimated as higher among those moderate in their news use, with more extreme users (less or more) estimated as lower in political knowledge. As you can see, the curvilinear effect is quite large even though it was barely visible in the partial scatterplot of Figure 12.5.

Just to make sure more complex curvilinearity is not missed, the cube of news use (X^3) was added to the model that includes news use and its square. The cubed term was not significant, meaning that adding it to the quadratic model does not improve the fit of the model to a statistically significant degree.

12.2.3 The Meaning of the Regression Coefficients for Lower-Order Regressors

We define a *global* property of a model as a property of the entire model, while a *local* property applies to only part of the model. For instance, a straight line relating X to Y has the same slope at all points, so the slope, estimated by the regression coefficient for X, is a global property of the model. But a curve defined by quadratic model that includes X and X^2 as regressors has different slopes at different points and may even slope downward in some sections but upward in others. Thus, the slope of

[1]It is legitimate to use the sample mean of a dichotomous variable when generating a plot such as this, even if the mean has no inherent meaning. In this case, sex is coded 0 for females and 1 for males, so the mean is the proportion of the sample that is male.

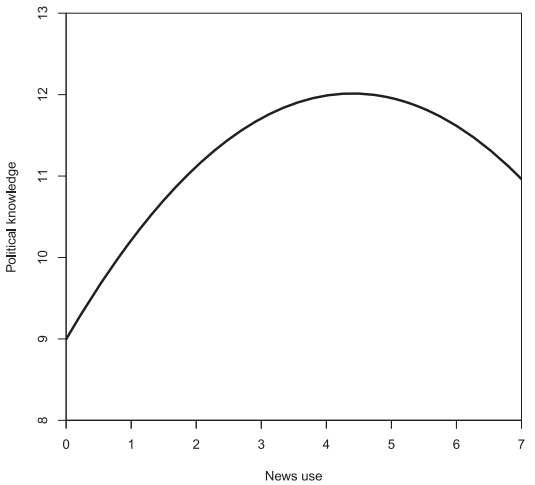

FIGURE 12.6. A quadratic polynomial model of political knowledge from news use frequency.

a curve defined by a quadratic model is a local property of the model. But a quadratic model is either concave, with a slope that becomes more positive as X increases, or convex, meaning that the slope is becoming more negative as X increases. So the concavity or convexity of a quadratic model is a global property of the model.

In a quadratic equation, it can be shown that if the regression coefficient for X^2 is positive, then the function is concave, but if this coefficient is negative, then the function is convex. It can also be shown that this regression coefficient measures the curvature of the relationship between X and Y, defined as the difference between the $\hat{Y}$ value of at any X point and the average of the two $\hat{Y}$ values corresponding to the values of X one unit to the left and one unit to the right. For instance, if $\hat{Y} = 2X^2$ and we arbitrarily use $X = 5$, then $\hat{Y} = 32$, 50, and 72 when X is 4, 5, and 6, respectively. We then have $(32 + 72)/2 - 50 = 2$, which is the coefficient for X^2. We would find the same value of 2 if we chose any other X value besides 5. Thus, the coefficient for X^2 measures a global property of the model, and we shall call X^2 a *global term* in the regression.

On the other hand, it can be shown that the coefficient of X in a quadratic equation measures the slope of the curve at the single point where $X = 0$. Readers who know calculus can see why this is so; if $\hat{Y} = b_0 + b_1X + b_2X^2$, then the first derivative of this function is $d\hat{Y}/dX = b_1 + 2b_2X$, which equals b_1 when $X = 0$. In the political knowledge example, $b_1 = 1.372$, and you can see by inspecting Figure 12.6 that this is about the slope of the parabola where it meets the Y-axis, when $X = 0$. Therefore, we call X a *local term*, since its regression coefficient measures a local property of the model.

This logic applies to higher-order polynomials, though understanding it requires knowledge of some calculus. For example, in a cubic model, the regression coefficient for X^3 is a global property of the model, but the regression coefficients for X and X^2 are local properties. In calculus terms, the first derivative of a cubic model $\hat{Y} = b_0 + b_1X + b_2X^2 + b_3X^3$ is $d\hat{Y}/dX = b_1 + 2b_2X + 3b_3X^2$. The first derivative is the slope of the curve at given point X, and you can see that if you set X to 0 in the equation for the first derivative, then you get b_1. Thus, b_1 is the slope of the curve when $X = 0$; thus, it is a local property of a cubic regression model.

The second derivative of a cubic model is $2b_2 + 6b_3X$. The second derivative quantifies how *quickly* and in what *direction* the slope is *changing* at a point X. This is sometimes called the acceleration of the function. If the second derivative is positive, that means that the slope is increasing as X is increasing in value. But if the second derivative is negative, that means that the slope is decreasing in value as X is increasing. The larger the second derivative ignoring sign, the faster the slope is changing. In this case, if you set X to 0 in the equation for the second derivative, you get $2b_2$. So b_2 is one-half of the speed at which the slope is changing at the point $X = 0$. This makes b_2 a local property in a cubic regression model.

12.2.4 Centering Variables in Polynomial Regression

A variable is *mean-centered* by subtracting its mean from all measurements, creating a new variable with a mean of zero. A variable can be mean-centered relative to its sample mean, or relative to its population mean if that happens to be known. There are two reasons why you might choose to mean-center X in a polynomial regression involving powers of X.

First, if $\overline{X}$ is high relative to s_X, then the successive powers X, X^2, X^3, and so on, might correlate with each other so highly that rounding error is produced, or you will reach the lower limit on the tolerance for a regressor that your regression program allows. For instance, in a sample of size $N = 5$ containing the values of X equal to 1,000, 1,001, 1,002, 1,003,

and 1,004, the correlation between X and X^2 is 0.99999983, which may be large enough to start introducing nontrivial rounding error into some regression computations. This can be corrected by centering X around its mean before computing the powers of X. If we subtract 1,002 from these five measurements (which is their mean), they become −2, −1, 0, 1, and 2, and now the correlation between X and X^2 is exactly zero. This can reduce computational problems and allow your regression program to estimate the model.

The second reason for mean-centering X before computing powers of X is that the regression coefficient for X is then the effect of X on Y at the mean of X, instead of when $X = 0$. This is likely to be more interpretable. A proof of this point was given in section 12.2.3.

Mean-centering a variable has no effect on regression coefficients for regressors or correlations when only first-order terms are used (e.g., X itself). But the situation with polynomial regression is more complex. Measures of simple relationship, such as correlations or simple regression coefficients, are affected for all but the first-order terms. For instance, if five measurements on X are 1, 2, 3, 4, and 5, then the five values of X^2 are 1, 4, 9, 16, and 25. But if we subtract 5 points from X before computing X^2, the new X values are −4, −3, −2, −1, and 0, and the new values of X^2 are 16, 9, 4, 1, and 0. Thus, the cases having the highest values on X^2 originally now have the lowest values. This, of course, will change the correlation between X^2 and other variables.

Measures of unique contribution for X or one of its powers, such as b_j, pr_j, sr_j, and the values of t or F that test their significance, are affected by centering for all but the highest power term. This is illustrated in Figure 12.7. Consider curve A. Its equation is $Y = 11.75 - 5.50X + 0.75X^2$. The slope of this curve is negative at $X = 0$ because Y is decreasing as X increases past zero. Thus, the regression coefficient for X is negative (see section 12.2.3). If we subtract 6 from X, then the curve shifts and becomes curve B in Figure 12.7, which has the same shape as curve A but is shifted horizontally in space. The equation for this curve is $Y = 5.75 + 3.50X + 0.75X^2$. Its slope at $X = 0$ is the coefficient for X, which is now positive because Y is increasing as X increases past $X = 0$. So centering X has changed the value of the regression coefficient for X, but the regression coefficient for X^2 is unaffected.

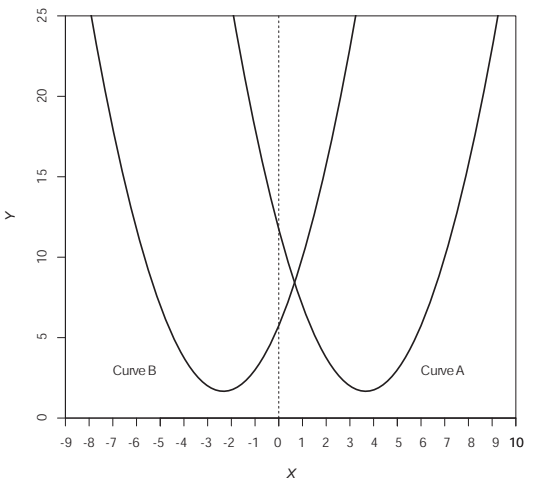

FIGURE 12.7. The effect of centering X on the regression coefficient for X in a quadratic model.

12.2.5 Finding a Parabola's Maximum or Minimum

Suppose you have estimated a model of the form $Y = b_0 + b_1X + b_2X^2$. The model could contain additional regressors as well, without changing the discussion that follows. However, the model should not include any regressors formed as the product of X and some other variable. The reasons for using a regressor that is a product of variables is discussed starting in Chapter 13.

In such a model, the value of X that either maximizes or minimizes Y (when all other variables are held constant, if the model contains additional regressors) is

$$X = \frac{-0.5b_1}{b_2} \tag{12.2}$$

Readers familiar with calculus will recognize this as the value at which the first derivative of Y with respect to X is equal to zero. The first derivative of a function of X with respect to X quantifies the amount Y is changing as X changes at a particular value of X. In a parabola, there is a point at which Y stops increasing or decreasing with changes in X and then "reverses course," such that if it was increasing with X, it now begins to decrease, or if it were decreasing with X, it now begins to increase. This point is either the minimum or maximum value. If the sign of b_2 is positive, then this

value is a minimum. If the sign of b_2 is negative, then this is a maximum value. But keep in mind that this point may not be within the range of the observed data.

To illustrate, in the political knowledge example in section 12.2.2 we had $b_1 = 1.372$ and $b_2 = -0.156$. Applying equation 12.2 gives $X = -0.5(1.372)/-0.156 = 4.397$. So we can say that holding constant education, age, sex, and SES, political knowledge is at its peak among those who use traditional news sources a bit over 4 days per week. We know it is a maximum and not a minimum because b_2 is negative, and we can also tell this from Figure 12.6.

12.3 Spline Regression

The scatterplot in Figure 12.8 depicts the association between two variables X and Y. As can be seen, the relationship is complex, with Y increasing with increasing X in some ranges of X, but decreasing Y with increasing X in other ranges. After reading section 12.2, you might think a quartic function would fit these data well. This would involving estimating Y from X, X^2, X^3, and X^4. Doing so results in $\hat{Y} = 5.161 + 6.217X - 0.845X^2 + 0.037X^3 - 0.001X^4$, and $R = 0.883$. This function is depicted in Figure 12.8 with the curve running through the scatterplot. It is apparent that even though R is fairly large, there is quite a bit of room for improvement. Observe that the vast majority of residuals are positive when X is between about 6 and 13, most are negative when X is between about 16 and 23, most are again positive between 23 and 26, and then again mostly negative beyond 26. This model is consistently underestimating Y in some ranges of X but overestimating Y in other ranges.

Spline regression is an alternative to polynomial regression. *Segmented regression* might be a better term, as the methods we discuss here all focus on fitting a set of models to various segments of the relationship between X and Y. But we will stick with the traditional term *spline regression*. Spline regression can model complex curves and do many other things, such as fitting lines with different slopes in different ranges of X. It can also be used when Y is expected to abruptly jump up or down at a specific value of X.

In this section, we introduce the fundamentals of spline regression, focusing first on *linear* spline models, which approximate a complex curve with a set of straight lines that are connected at joints. After describing these fundamentals, we discuss polynomial spline regression, which connects polynomials at joints. Polynomial splines are more versatile and therefore

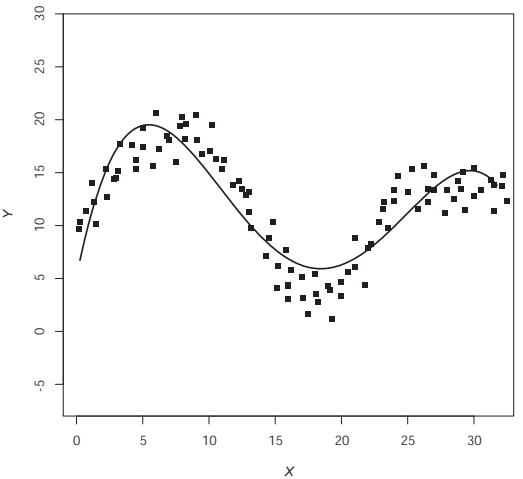

FIGURE 12.8. A scatterplot depicting a complex relationship and a quartic model superimposed.

more useful than linear splines, but you may find occasion to use linear splines, and it is easier to understand polynomial spline models by first learning how linear spline models work.

As a category of methods rather than a single method, spline regression includes more complex variants than we describe here. For a discussion of some of these more complex variants and their applications, see Ahlberg, Nilson, and Walsh (1967), Greville (1969), and Marsh and Cormier (2002). We focus only on methods that can be applied with an ordinary regression program.

12.3.1 Linear Spline Regression

In its simplest form, linear spline regression is a method for fitting to data a jagged line, like the solid line in Figure 12.9. Observe that this "curve" is formed by the four line segments that are joined together. By increasing the number of line segments, even extremely complex shapes can be fitted. The user of linear spline regression chooses the values of the regressor X but not the Y values that define the "joints" in a spline model. In Figure 12.9, these are marked J_1, J_2, and J_3. These could be chosen after examining a scatterplot, as we did in this example, or they could be chosen before

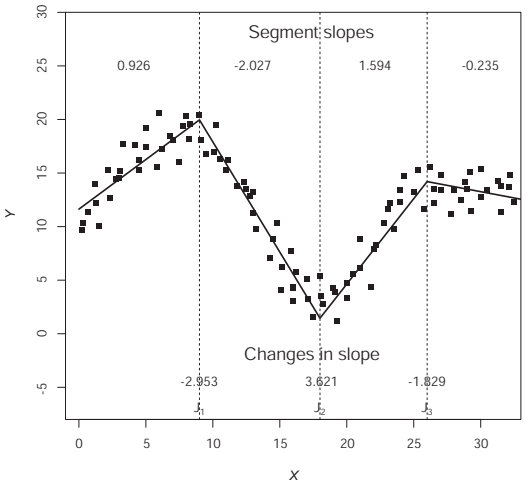

FIGURE 12.9. A linear spline regression model with three joints.

examining the data if you had an a priori basis for expecting a change in the relationship between X and Y at some values of X.

Spline regression using linear splines essentially estimates the slope of each line segment relating X to Y by computing the slope of the first segment and then the *change* in the slope at each joint. In Figure 12.9, the slopes of the four line segments displayed are 0.926, −2.027, 1.594, and −0.235, so a spline regression would estimate the changes in slope at J_1, J_2, and J_3 as −2.953, 3.621, and −1.829, respectively. These changes in slopes will be manifested in the regression solution as the regression weights for artificial variables created based on values of X.

To see how this is achieved, consider Figure 12.10. Line segment A, which applies when $X \leq 4$, is defined by the equation $Y = 1.00 + 1.00X$. Line segment *B* applies when $X > 4$, and it is defined by $Y = 13.00 - 2.00X$. That is,

$$Y = 1.00 + 1.00X \text{ when } X \leq 4 \text{ (segment A)}$$
$$Y = 13.00 - 2.00X \text{ when } X > 4 \text{ (segment B)}$$

Suppose we used the formula for line segment A to estimate all the Y values, regardless of whether or not X was greater than 4. As can be seen in Figure 12.10, doing so fits the Y values of the first four points perfectly,

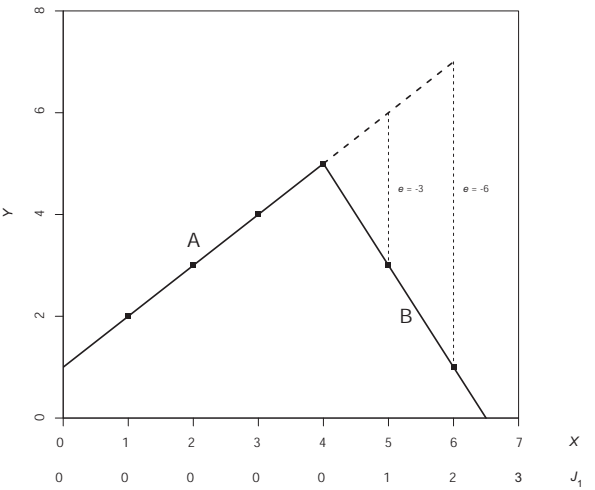

FIGURE 12.10. Why spline regression works.

but extending it beyond $X = 4$ (the dotted section of the line representing a continuation of line segment A) overestimates the next two Y values by 3 and 6 units, respectively. What we want to do is find a way of integrating the equations above into one equation that applies regardless of X.

Here is how we do it. Suppose we create a variable J_1 set to 0 when X is 4 or less, but set to $X - 4$ when $X > 4$. These values of J_1 can be found on the horizontal axis in Figure 12.10 below the values of X. Let e be defined as the errors in the estimation of Y from the equation for line segment A: $1.00 + 1.00X$. Notice that $e = 0$ when $J_1 = 0$, $e = -3$ when $J_1 = 1$, and $e = -6$ when $J_1 = 2$. In other words, $e = -3 \times J_1$. So an equation that perfectly fits the Y data would be

$$\hat{Y} = 1.00 + 1.00X - 3.00J_1 \tag{12.3}$$

This is the equation for the jagged line AB, and it integrates the equations for segments A and B into one equation. Observe that the line segment A has a slope of 1.00, and line segment B has a slope of −2.00, so the difference between these slopes is −3.00, which is the coefficient for J_1 in equation 12.3. So by creating the variable J_1 in this fashion, we were able to model the amount the slope of the line relating X to Y changes once X is higher than the location defining the joint. This example is atypical in that we do not

normally achieve a perfect fit. But once the X values of the joints have been selected, the regression program will fit the jagged line that minimizes the sum of the squared residuals and maximizes R.

Returning to the more complex example in Figure 12.9, using this approach we can fit a series of line segments with different slopes for different ranges of X but connected to each other at the joints. In Figure 12.9 there are three joints at X values of 9, 18, and 26. So we construct three new variables defined as X minus the joint location, but conditioned on X exceeding that joint value. If X does not exceed the joint value, then that variable is set to zero. In this example,

$$\begin{aligned}
&\text{if } X > 9, \text{ then } J_1 = X - 9 \text{ else } J_1 = 0\\
&\text{if } X > 18, \text{ then } J_2 = X - 18 \text{ else } J_2 = 0\\
&\text{if } X > 26, \text{ then } J_3 = X - 26 \text{ else } J_3 = 0
\end{aligned}$$

Once J_1, J_2, and J_3 are created, then regressing Y on X, J_1, J_2, and J_3 yields the equation

$$\hat{Y} = 11.628 + 0.926X - 2.953J_1 + 3.621J_2 - 1.829J_3 \tag{12.4}$$

The regression weight for X is the slope of the first line segment, and the values of b_j for J_1, J_2, and J_3 equal the changes in slope at joints J_1, J_2, and J_3. As in other forms of regression, regression programs routinely provide a test of significance for each b_j. When b_j represents a change in slope, we are testing the null hypothesis of no change in slope. In this example, these changes in slope at each joint are all statistically significant. Joints with nonsignificant changes may be deleted, given that the results in such a case suggest that there is no change in the size or direction of the association at that joint.

For this model, $R = 0.948, F(4, 95) = 209.101, p < .001$. This is a decent improvement from the quartic model (recall that in that model, $R = 0.883$), and as can be seen by comparing Figure 12.8 and Figure 12.9, the linear spline model does a better job estimating Y across the range of X. As discussed in section 4.3.2, the F-ratio for this model tests the null hypothesis that ${}_TR = 0$. This can be interpreted as a test of the null hypothesis of no relationship between X and Y, where X is a compound variable consisting of X itself as well as J_1, J_2, and J_3. We can test whether the relationship between X and Y is linear against the null hypothesis that it is nonlinear using the method in section 5.3.3. First estimate a model of Y from X alone.

Then add the J variables to the model. A statistically significant increase in R means that the linear spline model fits better than the ordinary linear model. In this case, the model with just X has $R = 0.263$, whereas the model with X and the three J variables has $R = 0.948$. This is a statistically significant increase, $F(3, 95) = 257.351, p < .001$. The linear spline model fits better than the simple linear model.

To better understand how this test works, consider that the model with just X as a regressor is equivalent to the spline model but with the constraint that all the regression weights for the J variables are equal to zero, meaning no change in slope at the joints. If the spline model fits better, then allowing for at least one joint with a change in slope produces a better-fitting model.

But we cannot use this test to compare the fit of this spline model to the quartic model. This test works only when the model with more variables (the spline model) contains all the same variables as the model with fewer variables (the quartic model), plus at least one extra variable. The J variables are not the same as the X^2, X^3, and X^4 variables, so we can't formally test the significance of the difference in fit of these two models.

However, there is an alternative approach that can be used to assess the relative value of the polynomial (X^2, X^3, and X^4) and spline terms (the J regressors). Combining these two models as

$$\hat{Y} = b_0 + b_1X + b_2J_1 + b_3J_2 + b_4J_3 + b_5X^2 + b_6X^3 + b_7X^4 \tag{12.5}$$

yields $R = .952$ when applied to these data. We can ask how much the polynomial terms add to fit by removing them from equation 12.5 and seeing if fit is significantly worse (which is the same as asking whether adding the polynomial terms to the linear spline model significantly improves fit). We already know that the linear spline model has $R = 0.948$. When the polynomial terms are added to the model, the test from section 5.3.3, which is appropriate here, does not quite achieve statistical significance, $F(3, 92) = 2.564, p = .059$. But when only the linear spline terms are removed from equation 12.5, the result is the quartic model, and we know that for this model $R = 0.883$. This reduction in fit relative to the combined model is statistically significant using this same test, $F(3, 92) = 41.034, p < .001$. That is, the inclusion of the linear spline terms significantly improves the fit of the model relative to when Y is modeled as a quartic function of X.

12.3.2 Implementation in Statistical Software

Although spline regression is not built into any commonly used statistical software packages of which we are aware, it can be implemented with any regression program. Assuming X is in your data and named as such, the SPSS code below constructs the three J variables in the four-segment linear spline model described in section 12.3.1 and then estimates the model.

```
compute j1=0.
compute j2=0.
compute j3=0.
if (x>9) j1=x-9.
if (x>18) j2=x-18.
if (x>26) j3=x-26.
regression/dep=y/method=enter x j1 j2 j3.
```

Assuming the data reside in a file named SPLINE, the comparable SAS code is

```
data spline;set spline;j1=0;j2=0;j3=0;
    if (x>9) then j1=x-9;if (x>18) then j2=x-18;if (x>26) then j3=x-26;
run;
proc reg data=spline;
    model y=x j1 j2 j3;
run;
```

and in STATA, use

```
gen j1=0
gen j2=0
gen j3=0
replace j1=x-9 if x>9
replace j2=x-18 if x>18
replace j3=x-26 if x>26
regress y x j1 j2 j3
```

The RLM macro documented in Appendix A has an option for linear spline regression. The user specifies the location of the joints, and RLM constructs all of the necessary J variables and then estimates the model. For instance, the SPSS RLM command below is comparable to the SPSS code above.

```
rlm y=y/x=x/spline=9,18,26.
```

See Appendix A for information on the use of the **spline** option in RLM.

12.3.3 Polynomial Spline Regression

Linear spline regression works as means of modeling nonlinearity, because any curve can be approximated by a set of line segments tied together at joints. The more joints you include, the better the approximation to the curvilinearity, in the same way that an octagon approximates a circle better than does a pentagon. But one restriction of linear spline regression is that between joints, the relationship between X and Y is fixed to be linear. As a result, the curve ends up jagged, with "elbows" at the joints and potentially very abrupt shifts in slope at the joints. A polynomial model doesn't have this problem, but a polynomial may not fit the relationship between X and Y as well, as in this example.

Polynomial spline regression combines the strengths of both polynomial and linear spline regression while eliminating the largest weakness of each. This procedure fits a polynomial rather than a straight line within each segment of the regressor. In principle, one could model the relationship between joints with a polynomial of any order, but we focus only on parabolic models (i.e., involving X^2) between joints, because this is usually sufficient. This will allow for different models of the relationship between X and Y in the segments, but will produce a smooth curve (rather than a line) between joints, with sets of smooth curves tied together at the joint points and no jaggedness at the joints.

When we fitted straight lines between joints, we constructed new variables defined as a set of one or more new variables quantifying whether and by how much X exceeded a particular joint value. To fit polynomials between the joints, we follow a similar procedure except that the new variables are higher powers of X conditioned on X exceeding the joint value. So if you want to fit a parabola between joint values, then the new variable will be set to 0 if X is less than or equal to the joint value, but if X exceeds the joint value, then set the new variable to the square of how much X exceeds that joint value. For instance, if X ranged between 0 and 20 and you placed a joint at 10, then J_1 would be set to 0 unless $X > 10$. If $X > 10$, then J_1 would be set to $(X - 10)^2$. You could include a higher additional power if desired, such as $(X - 10)^3$, if you wanted to fit a cubic function, although in practice, squares will usually suffice. You would typically also

include these same powers of X in the model to allow a polynomial of the same order for the first segment.

The scatterplot in Figure 12.11 is the same as the scatterplot in Figures 12.8 and 12.9, with the quartic model superimposed as a dashed line. This is a nice smooth curve, but as discussed already, its fit leaves something to be desired. The solid line depicts a polynomial spline model, the splines defined by second powers of X. Clearly, this does a better job describing the relationship between X and Y. We now describe how this model was constructed and estimated.

Examination of the scatterplot suggests that an inverted parabola may characterize the relationship between X and Y for values of X below 11. Between 11 and 16, the relationship appears linear or nearly so. Between 16 and 22, we can see what appears to be an upright parabola, but the left side of an inverted parabola appears to describe the relationship between X and Y between the values of 22 and 25. Finally, for X higher than 25, the relationship between X and Y looks linear or nearly so. So we define the five segments of the range of X with X values of 11, 16, 22, and 25. These are depicted in Figure 12.11.

With these five segments defined, we then create four J variables set to zero unless X exceeds the joint value. If X exceeds the joint value, then the J variable is set to the square of the amount X exceeds that joint. The algorithm for constructing these four J variables is

$$\begin{aligned}
&\text{if } X > 11, \text{ then } J_1 = (X - 11)^2 \text{ else } J_1 = 0\\
&\text{if } X > 16, \text{ then } J_2 = (X - 16)^2 \text{ else } J_2 = 0\\
&\text{if } X > 22, \text{ then } J_3 = (X - 22)^2 \text{ else } J_3 = 0\\
&\text{if } X > 25, \text{ then } J_4 = (X - 25)^2 \text{ else } J_4 = 0
\end{aligned}$$

We then regress Y on X and X^2 (which fits a parabola to the first segment), as well as J_1, J_2, J_3, and J_4. The resulting model is

$$\hat{Y} = 8.688 + 2.884X - 0.207X^2 + 0.148J_1 + 0.513J_2 - 0.972J_3 + 0.507J_4 \quad (12.6)$$

with $R = 0.956$. It is represented by the solid line in Figure 12.11. Observe it is a smooth curve, with no jaggedness at the joints as occurs when using linear splines. The fit of this model is clearly superior to the quartic model, and it is not obvious in looking at the scatterplot how this model could be changed to improve it further. All of the regression coefficients in this model are statistically significant, with p-values below .0001.

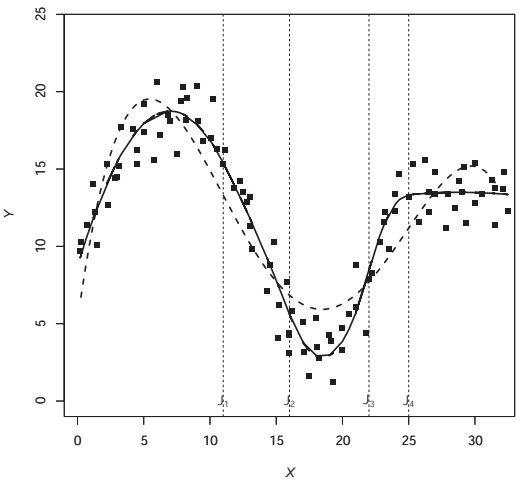

FIGURE 12.11. A quartic model (dashed line) and a polynomial spine model with four joints (solid line).

The code in section 12.3.2 can easily be modified to produce the *J* variables based on the algorithm above. For example, in SPSS, you can use

```
compute j1=0.
compute j2=0.
compute j3=0.
compute j4=0.
compute xsq=x*x.
if (x>11) j1=(x-11)*(x-11).
if (x>16) j2=(x-16)*(x-16).
if (x>22) j3=(x-22)*(x-22).
if (x>25) j4=(x-25)*(x-25).
regression/dep=y/method=enter x xsq j1 j2 j3 j4.
```

You may find with some statistics programs that the correlation between the regressors is sufficiently large that the model won't estimate, or the program may remove one or more of the regressors to deal with the near singularity (see section 17.3.3). If this occurs, center *X* around the mean of *X* (i.e., subtract $\overline{X}$ from all *X* values) and set up the joints and *J* variables using this transformed *X*. This likely will raise the tolerances of the regressors to more acceptable levels and may allow your program to estimate the model.

In section 12.3.1, we saw that the regression coefficients for the J variables quantify the difference in the slope relating X to Y between adjacent segments. In this quadratic spline regression model, the regression coefficients for the J variables quantify the change in curvilinearity of the relationship between X and Y between adjacent segments. Mathematically, this corresponds to the change in the regression coefficient for the squared term between adjacent segments. To see how this works, consider that for all values of $X \leq 11$, $J_1 = J_2 = J_3 = J_4 = 0$. So equation 12.6 reduces to $\hat{Y} = 8.688 + 2.884X - 0.207X^2$. This is the model relating X to Y when $X \leq 11$. The regression coefficient for X^2 is -0.207.

For the next segment defined as $11 < X \leq 15$, $J_1 = (X - 11)^2$ and $J_2 = J_3 = J_4 = 0$, so equation 12.6 simplifies to $\hat{Y} = 8.688 + 2.884X - 0.207X^2 + 0.148(X - 11)^2$. A little algebra results in

$$\begin{aligned}\hat{Y} &= 8.688 + 2.884X - 0.207X^2 + 0.148(X - 11)^2 \\ &= 8.688 + 2.884X - 0.207X^2 + 0.148(X^2 - 22X + 121) \\ &= 26.744 - 0.372X - 0.059X^2\end{aligned}$$

Thus, when $11 < X \leq 16$, the model relating Y to X is $\hat{Y} = 26.744 - 0.372X - 0.059X^2$. The regression coefficient for X^2 is -0.059, which is a change of 0.148 relative to the regression coefficient for X in the segment defined by $X \leq 11$. Notice that the regression coefficient for J_1 is 0.148. It is statistically significant from zero. If it were not statistically significant, then J_1 could be excluded from the model, because this would mean that allowing for a shift in the curvilinearity of the relationship between X and Y at this joint does not improve the fit of the model to a statistically significant degree.

Using this same logic and algebra for each segment produces a quadratic model for each segment, with the regression coefficient for successive J variables quantifying the difference in the regression coefficient for X^2 in a given segment relative to the prior segment. These changes add up cumulatively, so you can derive the weight for X^2 for any segment by starting with the regression coefficient for X^2 and then adding up the regression coefficients for each successive J variable, stopping once you reach the desired segment. For example, the regression coefficient for X^2 for the fourth segment ($X > 22 \leq 25$) is $0.207 + 0.148 + 0.513 - 0.972 = -0.518$. You can see in Figure 12.11 that, indeed, the model in this segment looks like the left half of a downward pointing parabola, consistent with a negative coefficient for X^2. And for the segment defined as $X > 25$, the weight for X^2

is $0.207 + 0.148 + 0.513 - 0.972 + 0.507 = -0.011$. This too is consistent with Figure 12.11. Observe that the regression line is nearly straight in the last segment, as you would expect for a polynomial model with such a small weight for the squared term.

12.3.4 Covariates, Weak Curvilinearity, and Choosing Joints

In the examples of spline regression we have described, there were no covariates, and we chose where to locate the joints by eyeballing a scatterplot. Covariates are easily added to a spline regression model simply by including them as regressors, and no modification to the procedure is needed. But we saw in section 12.1.2 that nonlinearity in the partial association between X and Y may be hard to see unless you construct the right scatterplot. And in real data, nonlinearity in the simple or partial association between X and Y may be so weak that it can't be detected with the eye even in the proper scatterplot. In such cases, you may not be able to eyeball a scatterplot and figure out where to locate the joints.

We don't have any silver bullet solutions to this problem, but is important to acknowledge the problem exists. If your sample size is sufficiently large, one option is to use a large number of joints equally dispersed across the range of X and then estimate a linear or polynomial spline model as discussed here. As you know, the p-values for the regression coefficients for the J variables can be used to decide whether a change in slope or curvilinearity is needed at specific joints. If not, those joints can be deleted. You can iteratively apply this procedure, adding or removing joints until you settle on a model that is satisfying to you. There are more advanced versions of spline regression that don't require the joints to be specified by the analyst but, rather, are derived mathematically from the data. You can read about some of these methods in the literature on spline regression, including the references we provided earlier in this chapter.

When you choose joints by eyeballing a scatterplot or using an exploratory method such as that just described, the concern is overfitting the data. Choosing joints by examining the data will tend to increase the variance explained by X. When X is a covariate, this produces a conservative bias into tests on independent variables. But if X is an independent variable, then the bias is toward exaggerating the importance of X in explaining variation in Y, and the nonlinearity captured by your spline model may not replicate in another sample.

But joint values need not always be chosen arbitrarily or by exploring the data and looking at scatterplots for visual evidence of transitions in the

relationship. You may have some a priori basis for choosing certain joint values. For example, if X were time and Y were something like a stock price, you might know that at a certain point in time (perhaps even a point in time of your choosing), some event happened that you think would change the trajectory for Y, making it increase or decrease in a particular manner that is different from what it was before that point in time. Or perhaps X is score on some kind of psychological test, such as a test of depression. If you assume, believe, or hypothesize that the relationship between depression and some dependent variable of interest is different for people who are below a certain score on the test relative to those who are above it, then that score would be natural choice for a joint in a spline regression model.

12.4 Transformations of Dependent Variables or Regressors

The natural relationship between two variables may be nonlinear, but sometimes nonlinear relationships can be made linear or nearly so by some kind of *transformation* of one of the variables. There are many kinds of transformations, but we focus on *monotonic* transformations here. A transformation is monotonic if the original and transformed values have the same rank order, such that the highest value on the original variable is the highest after transformation, the second highest original value is the second highest transformed value, and so forth. Technically, we should distinguish between positive and negative monotonic transformations. What we have described just now is positive monotonic. A negative monotonic transformation exactly reverses the ranks, so that the highest original value is the lowest transformed value, the second highest original is the second lowest transformed value, and so forth. Unless we say otherwise, when we say *monotonic* assume we are talking about positive monotonic.

Monotonic transformations of a variable may produce as many as three benefits at once. The first we have already discussed: Two variables may be nonlinearly related in their original form, but linear if one or both is transformed monotonically. This often simplifies interpretation of regression results. The second benefit is that a transformation may improve the prediction of one variable from another. Third, they can make residuals more normally distributed. Normality of the errors in estimation (manifested as residuals in a specific analysis) is an assumption of linear regression analysis.

12.4.1 Logarithmic Transformation

A logarithmic transformation can be used when the importance of the difference between two values is judged to be proportional to their ratio rather than their absolute difference. For instance, if we are studying the effect of an animal's size on some feature of its behavior or structure, we might consider the difference between body weights of 100 and 200 kilograms to be no more important than a difference between 1 and 2 kilograms. In absolute terms, a difference of 100 kilograms is 100 times larger than a difference of 1 kilogram, but both ratios are 2:1. Whereas weight may be nonlinearly related to many things (e.g., brain size), a logarithmic transformation may make the relationship linear. Or if the difference between incomes of $50,000 and $100,000 has the same average effect on attitudes toward wealth as the difference between $10,000 and $20,000, then income will have a nonlinear relationship with attitude, but a logarithmic transformation can make the relationship linear.

Only positive numbers have logarithms, but there are many kinds of logarithms. The most commonly used logarithms are the common logarithm, also called a base 10 log and often denoted *log*, and the natural logarithm or base *e* log, most often denoted as *ln*. The common logarithm of a number *X* is the power of 10, which equals *X*. For instance, the common logarithms of 10, 100, and 1,000 are, respectively, 1, 2, and 3, because $10^1 = 10$, $10^2 = 100$, and $10^3 = 1{,}000$.

Whereas a common logarithm is a power of 10, a natural logarithm is a power of *e*, where *e* is approximately 2.718828. Like the number pi, *e* cannot be written exactly. The natural logs of 10, 100, and 1,000 are, respectively, 2.30259, 4.60517, and 9.21034, because $e^{2.30259} = 10$, $e^{4.60516} = 100$, and $e^{9.21034} = 1{,}000$. Natural logarithms are proportional to common logarithms; for any number *X*, the natural logarithm of *X* equals approximately 2.302589 times the common logarithm of *X*.

An interesting property of natural logarithms is that when two numbers *A* and *B* are nearly equal, the difference between their natural logarithms approximately equals the proportional difference between them. For instance, 63 is 5% larger than 60, and their natural logarithms are 4.1431 and 4.0943, which differ by .0488, which is close to 0.05. Thus, if the weights of two animals differed by 0.05 on a natural logarithm scale, you would know without calculation that one was about 5% heavier than the other. As two numbers approach equality, this relationship approaches exactness. For instance, the natural logarithms of 1,000 and 1,001 differ by .0009995, which to four significant digits is .001, or 1/1,000.

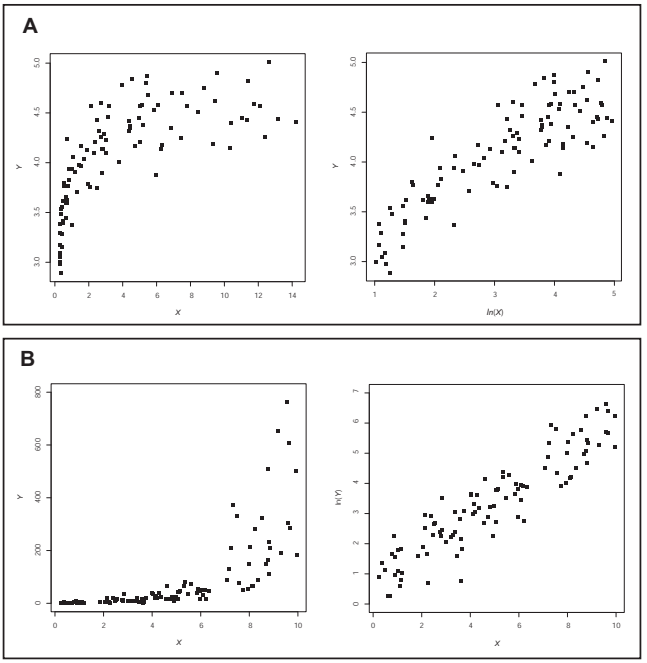

FIGURE 12.12. A log tranformation of *X* and *Y* can turn a nonlinear relationship into a linear relationship.

When a scatterplot depicting the association between two variables appears nonlinear, for some forms of nonlinearity a logarithmic transformation of *X* or *Y* may make the relationship more linear. If small changes in *X* result in large positive changes in *Y* at first but then the size of the change in *Y* levels off as *X* increases, as in Figure 12.12, panel A, on the left, then a logarithmic transformation of *X* may reduce or eliminate the nonlinearity. The scatterplot on the right of Figure 12.12, panel A, depicts the association between *X* and *Y* after a natural log transformation of *X*. As you can see, the relationship appears more linear after transformation than before.

But if small changes in *X* result in little changes in *Y* at first, but the change in *Y* with a change in *X* accelerates rapidly, as in Figure 12.12, panel B, on the left, then a logarithmic transformation of *Y* rather than *X* may reduce the nonlinearity. A scatterplot of the natural log of *Y* against *X* can be seen in the scatterplot on the right side of Figure 12.12, panel B; the

relationship between X and Y following the transformation now appears to be linear rather than nonlinear.

When using a logarithmic transformation, we often don't have to make distinctions between the different forms. If the common logarithms of X are linearly related to another variable Y, then the natural logarithms will be also. Thus, if we say that a logarithmic transformation makes a relationship linear, we need not specify which type of logarithm. But when reporting the results of an analysis that uses a transformation, it is a good idea to be explicit about what transformation was employed.

12.4.2 The Box–Cox Transformation

Box and Cox (1964) describe a family of transformations that includes logarithmic transformations as special cases. In this approach, one chooses a constant m, which may be any positive or negative real number (i.e., not zero). Then one transforms the original variable X to a transformed variable X_T by the equation

$$X_T = \frac{X^m - 1}{m} \tag{12.7}$$

In practice, you can try different values of m and see which one is best by some criterion of interest, such as making some extreme scores less extreme, improving linearity, or eliminating the need for an interaction (a concept introduced in Chapter 13). Although we use X in equation 12.7, the transformation can be applied to dependent variables, independent variables, or covariates.

Figure 12.13 displays the results of the transformation for $0 < X \leq 5$ for different values of m. The dashed line corresponding to $m = 1$ reflects no transformation (actually, when $m = 1$, $X_T = X - 1$). $X = 1$ is a pivoting point in the transformation, and what happens to the relative sizes of X after transformation depends on the distance from 1 and the value of m.

Define *measurement expansion* as making differences between values of X larger after the transformation, and define *measurement compression* as making differences between values of X smaller after transformation. Given these definitions, setting $m > 1$ results in measurement expansion when $X > 1$, with the expansion larger with higher values of m. But when $X < 1$, measurement compression is the result. But when $m < 1$, the transformation has the opposite effect on X. When $X > 1$, measurements are compressed, with greater compression occurring with smaller values of m. But when $X < 1$, measurement expansion occurs.

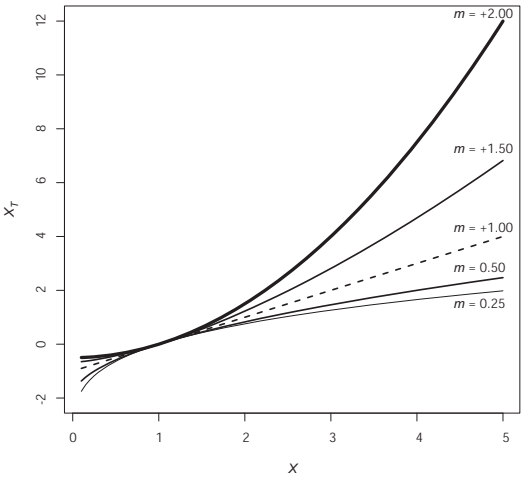

FIGURE 12.13. The Box–Cox tranformation as a function of *m*.

By making m arbitrarily close to zero (either positive or negative), we can make the Box–Cox transformation approach arbitrarily close to a logarithmic transformation. Thus, we can think of a logarithmic transformation as the special case of the Box–Cox transformation in which $m = 0$.

A Box–Cox transformation requires all measurements on the original variable to be positive since a negative number cannot be raised to a noninteger power. But if all measurements are negative we lose no information by replacing the original measurements with their absolute values before making the transformation. Thus, the requirement really is that all measurements have the same sign. This usually means that all measurements in the *population* must have the same sign, not just the measurements themselves, because inferences to the population have no meaning if some measurements in the population cannot be transformed.

Could one add points to a variable to make all its measurements have the same sign? Theoretically, a Box–Cox transformation is scientifically meaningful only if the original scale is a ratio scale—a scale with a meaningful zero point, so that it is meaningful to talk abut the ratios of two measurements. Thus, for instance, height and weight are ratio scales, but an attitude scale running from 1 to 9, with 1 denoting "very negative" and 9 denoting "very positive," is not. But in practice this restriction is not very important when using Box–Cox, because the effect of changing m is

often very similar to adding a constant to X before transformation. For instance, consider five cases scoring 1, 2, 3, 4, and 5 on a variable X. If we set $m = 0.5$, these five values transform to 0, 0.8284, 1.4641, 2.0000, and 2.4721, respectively. Thus, the second, third, and fourth transformed values are, respectively, 33.5, 59.2, and 80.9% of the distance from the first transformed value to the last. But if we add 7.841 to each of the original scores, then apply a Box–Cox transformation using $m = -1$, the percentages are nearly identical to the previous ones, now being 32.6, 59.2, and 81.3%. Thus, using $m = 0.5$ is nearly equivalent to adding 7.841 to each score and then using $m = -1$. Since trying different values of m is often very similar to adding different positive and negative constants to the original scores before making the transformation, the original zero point does not seem particularly sacrosanct.

12.5 Chapter Summary

Linear regression analysis can be used to model relationships between variables even when those relationships are not linear. It is always worth checking for nonlinearity by constructing a scatterplot, but it is important to construct the right scatterplot. The residual scatterplot is the best choice for detecting nonlinearity between X and Y when a model contains covariates. In a residual scatterplot, the residuals in the estimation of Y from X and the covariates are plotted against X. But even with the help of a residual scatterplot, the human eye is not very good at detecting relationships, so such eyeballing should be accompanied by some kind of formal analysis of nonlinearity.

Polynomial regression analysis is a versatile approach to testing for nonlinearity between X and Y, as well as modeling nonlinear relationships. This method involves estimating Y from X and successive powers of X, such as X^2 and, if desired, X^3 and (rarely) X^4. A statistically significant regression coefficient for one of the higher powers of X implies nonlinearity, as does an incremental increase in the fit of the model when one or more powers of X is added. Interpretation of the regression coefficients is complex and aided with an understanding of calculus. Most important is that in a model with a power of X higher than 1, the regression coefficient for X is a local term of the model and quantifies the relationship between X and Y when $X = 0$. Higher-order terms are interpreted in terms of changes in rates of changes of Y as X is changing.

Spline regression can be used to fit a jagged line to data. Any curve can be approximated by a set of jagged lines, and sometimes a spline model will fit better than a polynomial, because spline models better capture abrupt shifts in the relationship between X and Y as X increases or decreases. Spline and polynomial regression can be combined into polynomial spline regression. This involves estimating and tying together polynomials at various points in the distribution of X, thereby increasing the complexity of the kinds of curves that can be estimated.

Some nonlinear relationships can be made linear or nearly so through the use of a transformation, and transformations can sometimes help in meeting the other assumptions of regression. Logarithmic transformations of X and Y can be used in different circumstances, depending on the form of nonlinearity. A logarithmic transformation is a special form of the more general Box–Cox transformation. Using this transformation, the analyst selects an exponent in the function that produces the most appealing transformation, as defined by how well it makes a nonlinear relationship linear or removes skew or heteroscedasticity in the errors in estimation, for instance.

One could define a nonlinear relationship as one in which the relationship between X and Y depends on X. This could be thought of as a special kind of *moderation*, the topic of the next two chapters. With nonlinearity, X moderates its own effect on Y. In the following chapter we introduce how to build flexibility into a regression analysis by allowing the effect of a regressor X on Y to vary linearly with another regressor in the model.