8 Nonlinear Regression Functions

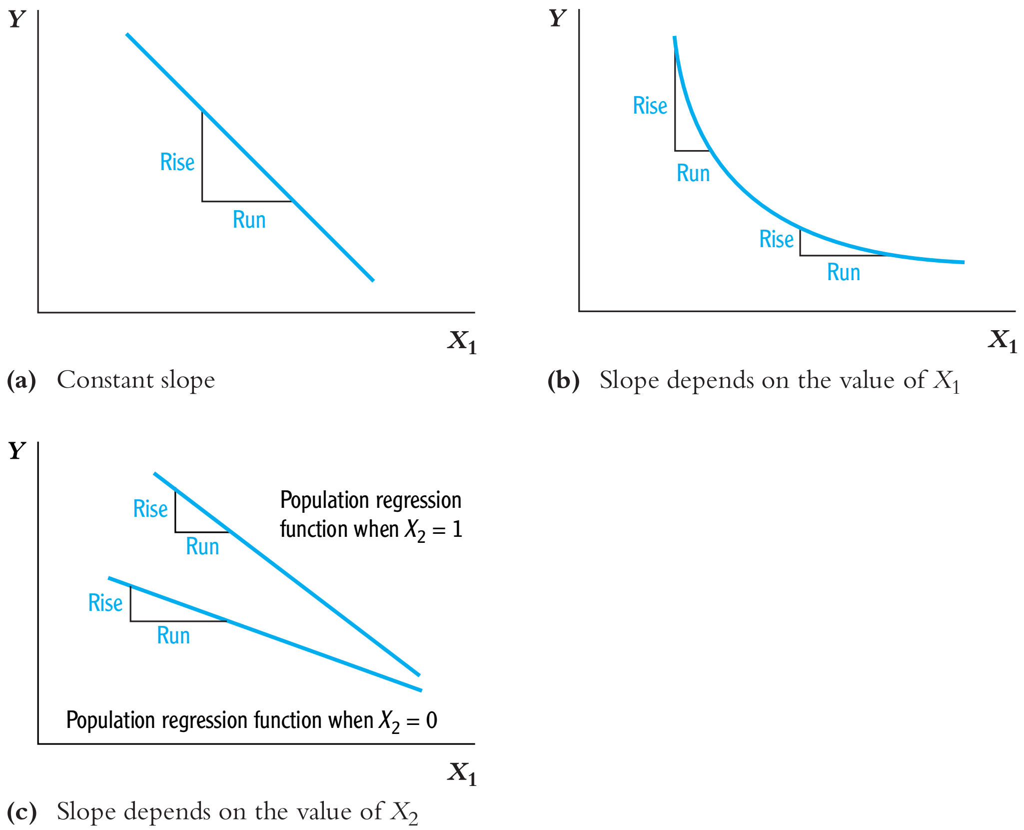
In Chapters 4 through 7, the population regression function was assumed to be linear. In other words, the slope of the population regression function was constant, so the effect on Y of a unit change in X does not itself depend on the value of X. But what if the effect on Y of a change in X does depend on the value of one or more of the independent variables? If so, the population regression function is nonlinear.

This chapter develops two groups of methods for detecting and modeling nonlinear population regression functions. The methods in the first group are useful when the effect on Y of a change in one independent variable, X_1, depends on the value of X_1 itself. For example, reducing class sizes by one student per teacher might have a greater effect if class sizes are already manageably small than if they are so large that the teacher can do little more than keep the class under control. If so, the test score (Y) is a nonlinear function of the student–teacher ratio (X_1), where this function is steeper when X_1 is small. An example of a nonlinear regression function with this feature is shown in Figure 8.1. Whereas the linear population regression function in Figure 8.1a has a constant slope, the nonlinear population regression function in Figure 8.1b has a steeper slope when X_1 is small than when it is large. This first group of methods is presented in Section 8.2.

The methods in the second group are useful when the effect on Y of a change in X_1 depends on the value of another independent variable, say X_2. For example, students still learning English might especially benefit from having more one-on-one attention; if so, the effect on test scores of reducing the student–teacher ratio will be greater in districts with many students still learning English than in districts with few English learners. In this example, the effect on test scores (Y) of a reduction in the student–teacher ratio (X_1) depends on the percentage of English learners in the district (X_2). As shown in Figure 8.1c, the slope of this type of population regression function depends on the value of X_2. This second group of methods is presented in Section 8.3.

In the models of Sections 8.2 and 8.3, the population regression function is a nonlinear function of the independent variables; that is, the conditional expectation $E(Y_i | X_{1i}, \ldots, X_{ki})$ is a nonlinear function of one or more of the X's. Although they are nonlinear in the X's, these models are linear functions of the unknown coefficients

FIGURE 8.1 Population Regression Functions with Different Slopes



In Figure 8.1a, the population regression function has a constant slope. In Figure 8.1b, the slope of the population regression function depends on the value of X_1. In Figure 8.1c, the slope of the population regression function depends on the value of X_2.

(or parameters) of the population regression model and thus are versions of the multiple regression model of Chapters 6 and 7. Therefore, the unknown parameters of these nonlinear regression functions can be estimated and tested using OLS and the methods of Chapters 6 and 7.

Sections 8.1 and 8.2 introduce nonlinear regression functions in the context of regression with a single independent variable, and Section 8.3 extends this to two independent variables. To keep things simple, additional control variables are omitted in the empirical examples of Sections 8.1 through 8.3. In practice, however, it is important to analyze nonlinear regression functions in models that control for omitted factors by including control variables as well. In Section 8.5, we combine nonlinear regression functions and additional control variables when we take a close look at possible nonlinearities in the relationship between test scores and the student–teacher ratio, holding student characteristics constant. In some applications, the regression function is a nonlinear function of the *X*'s *and* of the parameters. If so, the parameters cannot be estimated by OLS, but they can be estimated using nonlinear least squares. Appendix 8.1 provides examples of such functions and describes the nonlinear least squares estimator.

8.1 A General Strategy for Modeling Nonlinear Regression Functions

This section lays out a general strategy for modeling nonlinear population regression functions. In this strategy, the nonlinear models are extensions of the multiple regression model and therefore can be estimated and tested using the tools of Chapters 6 and 7. First, however, we return to the California test score data and consider the relationship between test scores and district income.

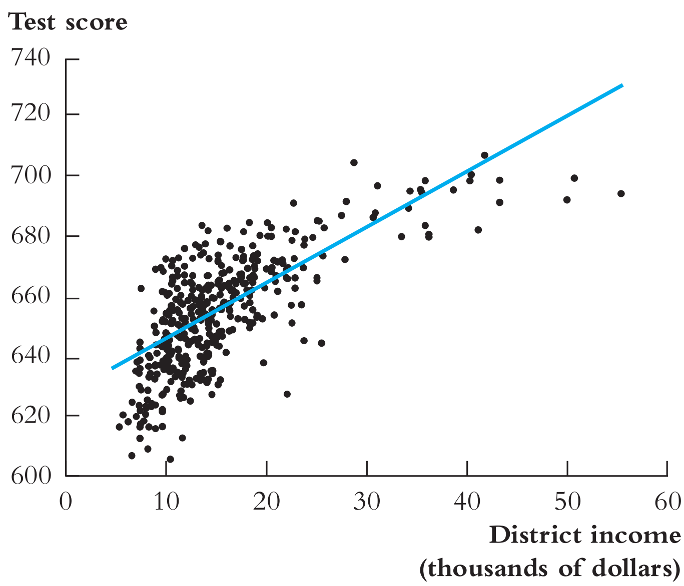
Test Scores and District Income

In Chapter 7, we found that the economic background of the students is an important factor in explaining performance on standardized tests. That analysis used two economic background variables (the percentage of students qualifying for a subsidized lunch and the percentage of district families qualifying for income assistance) to measure the fraction of students in the district coming from poor families. A different, broader measure of economic background is the average annual per capita income in the school district ("district income"). The California data set includes district income measured in thousands of 1998 dollars. The sample contains a wide range of income levels: For the 420 districts in our sample, the median district income is 13.7 (that is, $13,700 per person), and it ranges from 5.3 ($5300 per person) to 55.3 ($55,300 per person).

Figure 8.2 shows a scatterplot of fifth-grade test scores against district income for the California data set, along with the OLS regression line relating these two variables. Test scores and average income are strongly positively correlated, with a correlation coefficient of 0.71; students from affluent districts do better on the tests than students from poor districts. But this scatterplot has a peculiarity: Most of the points are below the OLS line when income is very low (under $10,000) or very high (over $40,000), but are above the line when income is between $15,000 and $30,000. There seems to be some curvature in the relationship between test scores and income that is not captured by the linear regression.

FIGURE 8.2 Scatterplot of Test Score vs. District Income with a Linear OLS Regression Function

There is a positive correlation between test scores and district income (correlation = 0.71), but the linear OLS regression line does not adequately describe the relationship between these variables.



In short, it seems that the relationship between district income and test scores is not a straight line. Rather, it is nonlinear. A nonlinear function is a function with a slope that is not constant: The function $f(X)$ is linear if the slope of $f(X)$ is the same for all values of X, but if the slope depends on the value of X, then $f(X)$ is nonlinear.

If a straight line is not an adequate description of the relationship between district income and test scores, what is? Imagine drawing a curve that fits the points in Figure 8.2. This curve would be steep for low values of district income and then would flatten out as district income gets higher. One way to approximate such a curve mathematically is to model the relationship as a quadratic function. That is, we could model test scores as a function of income *and* the square of income.

A quadratic population regression model relating test scores and income is written mathematically as

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i, \tag{8.1}$$

where β_0, β_1, and β_2 are coefficients, $Income_i$ is the income in the i^{th} district, $Income_i^2$ is the square of income in the i^{th} district, and u_i is an error term that, as usual, represents all the other factors that determine test scores. Equation (8.1) is called the **quadratic regression model** because the population regression function,

$E(TestScore_i|Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2$, is a quadratic function of the independent variable, *Income*.

If you knew the population coefficients β_0, β_1, and β_2 in Equation (8.1), you could predict the test score of a district based on its average income. But these population coefficients are unknown and therefore must be estimated using a sample of data.

At first, it might seem difficult to find the coefficients of the quadratic function that best fits the data in Figure 8.2. If you compare Equation (8.1) with the multiple regression model in Key Concept 6.2, however, you will see that Equation (8.1) is in fact a version of the multiple regression model with two regressors: The first regressor is *Income*, and the second regressor is $Income^2$. Mechanically, you can create this second regressor by generating a new variable that equals the square of *Income*, for example as an additional column in a spreadsheet. Thus, after defining the regressors as *Income* and $Income^2$, the nonlinear model in Equation (8.1) is simply a multiple regression model with two regressors!

Because the quadratic regression model is a variant of multiple regression, its unknown population coefficients can be estimated and tested using the OLS methods described in Chapters 6 and 7. Estimating the coefficients of Equation (8.1) using OLS for the 420 observations in Figure 8.2 yields

$$\underset{}{\widehat{TestScore}} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85}\,Income - \underset{(0.0048)}{0.0423}\,Income^2, \; \overline{R}^2 = 0.554, \qquad (8.2)$$

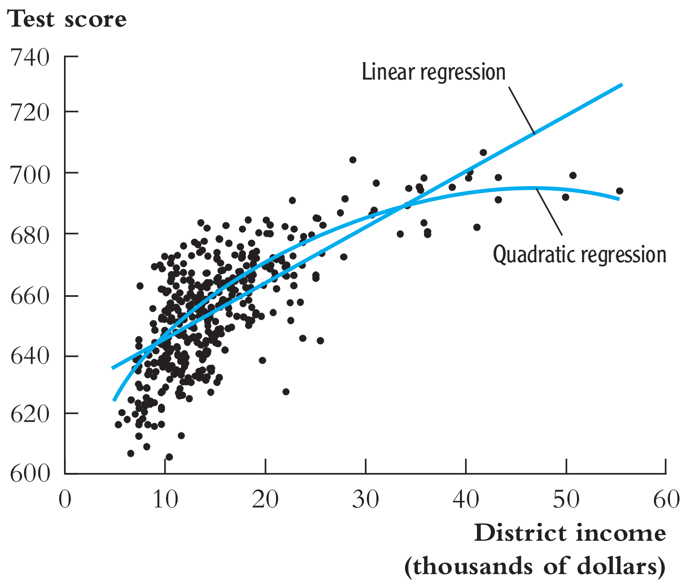
where (as usual) standard errors of the estimated coefficients are given in parentheses. The estimated regression function of Equation (8.2) is plotted in Figure 8.3, superimposed over the scatterplot of the data. The quadratic function captures the curvature in the scatterplot: It is steep for low values of district income but flattens out when district income is high. In short, the quadratic regression function seems to fit the data better than the linear one.

We can go one step beyond this visual comparison and formally test the hypothesis that the relationship between income and test scores is linear against the alternative that it is nonlinear. If the relationship is linear, then the regression function is correctly specified as Equation (8.1), except that the regressor $Income^2$ is absent; that is, if the relationship is linear, then Equation (8.1) holds with $\beta_2 = 0$. Thus we can test the null hypothesis that the population regression function is linear against the alternative that it is quadratic by testing the null hypothesis that $\beta_2 = 0$ against the alternative that $\beta_2 \neq 0$.

Because Equation (8.1) is just a variant of the multiple regression model, the null hypothesis that $\beta_2 = 0$ can be tested by constructing the t-statistic for this

FIGURE 8.3 Scatterplot of Test Score vs. District Income with Linear and Quadratic Regression Functions

The quadratic OLS regression function fits the data better than the linear OLS regression function.



hypothesis. This t-statistic is $t = (\hat{\beta}_2 - 0)/SE(\hat{\beta}_2)$, which from Equation (8.2) is $t = -0.0423/0.0048 = -8.81$. In absolute value, this exceeds the 5% critical value of this test (which is 1.96). Indeed the p-value for the t-statistic is less than 0.01%, so we can reject the hypothesis that $\beta_2 = 0$ at all conventional significance levels. Thus this formal hypothesis test supports our informal inspection of Figures 8.2 and 8.3: The quadratic model fits the data better than the linear model.

The Effect on Y of a Change in X in Nonlinear Specifications

Put aside the test score example for a moment and consider a general problem. You want to know how the dependent variable Y is expected to change when the independent variable X_1 changes by the amount ΔX_1, holding constant other independent variables $X_2, \ldots, X_k$. When the population regression function is linear, this effect is easy to calculate: As shown in Equation (6.4), the expected change in Y is $\Delta Y = \beta_1 \Delta X_1$, where β_1 is the population regression coefficient multiplying X_1. When the regression function is nonlinear, however, the expected change in Y is more complicated to calculate because it can depend on the values of the independent variables.

A general formula for a nonlinear population regression function.[1] The nonlinear population regression models considered in this chapter are of the form

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i, i = 1, \dots, n, \quad (8.3)$$

where $f(X_{1i}, X_{2i}, \dots, X_{ki})$ is the population **nonlinear regression function**, a possibly nonlinear function of the independent variables $X_{1i}, X_{2i}, \dots, X_{ki}$, and u_i is the error term. For example, in the quadratic regression model in Equation (8.1), only one independent variable is present, so X_1 is *Income* and the population regression function is $f(Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2$.

Because the population regression function is the conditional expectation of Y_i given $X_{1i}, X_{2i}, \dots, X_{ki}$, in Equation (8.3) we allow for the possibility that this conditional expectation is a nonlinear function of $X_{1i}, X_{2i}, \dots, X_{ki}$; that is, $E(Y_i | X_{1i}, X_{2i}, \dots, X_{ki}) = f(X_{1i}, X_{2i}, \dots, X_{ki})$, where f can be a nonlinear function. If the population regression function is linear, then $f(X_{1i}, X_{2i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$, and Equation (8.3) becomes the linear regression model in Key Concept 6.2. However, Equation (8.3) allows for nonlinear regression functions as well.

The effect on Y of a change in X_1. As discussed in Section 6.2, the effect on Y of a change in X_1, ΔX_1, holding $X_2, \dots, X_k$ constant, is the difference in the expected value of Y when the independent variables take on the values $X_1 + \Delta X_1, X_2, \dots, X_k$ and the expected value of Y when the independent variables take on the values $X_1, X_2, \dots, X_k$. The difference between these two expected values, say ΔY, is what happens to Y on average in the population when X_1 changes by an amount ΔX_1, holding constant the other variables $X_2, \dots, X_k$. In the nonlinear regression model of Equation (8.3), this effect on Y is $\Delta Y = f(X_1 + \Delta X_1, X_2, \dots, X_k) - f(X_1, X_2, \dots, X_k)$.

Because the regression function f is unknown, the population effect on Y of a change in X_1 is also unknown. To estimate the population effect, first estimate the population regression function. At a general level, denote this estimated function

[1]The term *nonlinear regression* applies to two conceptually different families of models. In the first family, the population regression function is a nonlinear function of the X's but is a linear function of the unknown parameters (the β's). In the second family, the population regression function is a nonlinear function of the unknown parameters and may or may not be a nonlinear function of the X's. The models in the body of this chapter are all in the first family. Appendix 8.1 takes up models from the second family.

The Expected Change on Y of a Change in X_1 in the Nonlinear Regression Model (8.3)

KEY CONCEPT 8.1

The expected change in Y, ΔY, associated with the change in X_1, ΔX_1, holding $X_2, \ldots, X_k$ constant, is the difference between the value of the population regression function before and after changing X_1, holding $X_2, \ldots, X_k$ constant. That is, the expected change in Y is the difference:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k). \tag{8.4}$$

The estimator of this unknown population difference is the difference between the predicted values for these two cases. Let $\hat{f}(X_1, X_2, \ldots, X_k)$ be the predicted value of Y based on the estimator $\hat{f}$ of the population regression function. Then the predicted change in Y is

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k). \tag{8.5}$$

by $\hat{f}$; an example of such an estimated function is the estimated quadratic regression function in Equation (8.2). The estimated effect on Y (denoted $\Delta\hat{Y}$) of the change in X_1 is the difference between the predicted value of Y when the independent variables take on the values $X_1 + \Delta X_1, X_2, \ldots, X_k$ and the predicted value of Y when they take on the values $X_1, X_2, \ldots, X_k$.

The method for calculating the expected effect on Y of a change in X_1 is summarized in Key Concept 8.1. The method in Key Concept 8.1 always works, whether ΔX_1 is large or small and whether the regressors are continuous or discrete. Appendix 8.2 shows how to evaluate the slope using calculus for the special case of a single continuous regressor when ΔX_1 small.

Application to test scores and income. What is the predicted change in test scores associated with a change in district income of \$1000, based on the estimated quadratic regression function in Equation (8.2)? Because that regression function is quadratic, this effect depends on the initial district income. We therefore consider two cases: an increase in district income from 10 to 11 (i.e., from \$10,000 per capita to \$11,000) and an increase in district income from 40 to 41.

To compute $\Delta\hat{Y}$ associated with the change in income from 10 to 11, we can apply the general formula in Equation (8.5) to the quadratic regression model. Doing so yields

$$\Delta\hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2), \quad (8.6)$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the OLS estimators.

The term in the first set of parentheses in Equation (8.6) is the predicted value of Y when $Income = 11$, and the term in the second set of parentheses is the predicted value of Y when $Income = 10$. These predicted values are calculated using the OLS estimates of the coefficients in Equation (8.2). Accordingly, when $Income = 10$, the predicted value of test scores is $607.3 + 3.85 \times 10 - 0.0423 \times 10^2 = 641.57$. When $Income = 11$, the predicted value is $607.3 + 3.85 \times 11 - 0.0423 \times 11^2 = 644.53$. The difference in these two predicted values is $\Delta\hat{Y} = 644.53 - 641.57 = 2.96$ points; that is, the predicted difference in test scores between a district with average income of \$11,000 and one with average income of \$10,000 is 2.96 points.

In the second case, when income changes from \$40,000 to \$41,000, the difference in the predicted values in Equation (8.6) is $\Delta\hat{Y} = (607.3 + 3.85 \times 41 - 0.0423 \times 41^2) - (607.3 + 3.85 \times 40 - 0.0423 \times 40^2) = 694.04 - 693.62 = 0.42$ points. Thus a change of income of \$1000 is associated with a larger change in predicted test scores if the initial income is \$10,000 than if it is \$40,000 (the predicted changes are 2.96 points versus 0.42 point). Said differently, the slope of the estimated quadratic regression function in Figure 8.3 is steeper at low values of income (like \$10,000) than at the higher values of income (like \$40,000).

Standard errors of estimated effects. The estimator of the effect on Y of changing X_1 depends on the estimator of the population regression function, $\hat{f}$, which varies from one sample to the next. Therefore, the estimated effect contains a sampling error. One way to quantify the sampling uncertainty associated with the estimated effect is to compute a confidence interval for the true population effect. To do so, we need to compute the standard error of $\Delta\hat{Y}$ in Equation (8.5).

It is easy to compute a standard error for $\Delta\hat{Y}$ when the regression function is linear. The estimated effect of a change in X_1 is $\hat{\beta}_1\Delta X_1$, so the standard error of $\Delta\hat{Y}$ is $SE(\Delta\hat{Y}) = SE(\hat{\beta}_1)\Delta X_1$ and a 95% confidence interval for the estimated change is $\hat{\beta}_1\Delta X_1 \pm 1.96SE(\hat{\beta}_1)\Delta X_1$.

In the nonlinear regression models of this chapter, the standard error of $\Delta\hat{Y}$ can be computed using the tools introduced in Section 7.3 for testing a single restriction involving multiple coefficients. To illustrate this method, consider the estimated change in test scores associated with a change in income from 10 to 11 in

Equation (8.6), which is $\Delta\hat{Y} = \hat{\beta}_1 \times (11 - 10) + \hat{\beta}_2 \times (11^2 - 10^2) = \hat{\beta}_1 + 21\hat{\beta}_2$. The standard error of the predicted change therefore is

$$SE(\Delta\hat{Y}) = SE(\hat{\beta}_1 + 21\hat{\beta}_2). \tag{8.7}$$

Thus, if we can compute the standard error of $\hat{\beta}_1 + 21\hat{\beta}_2$, then we have computed the standard error of $\Delta\hat{Y}$. There are two methods for doing this using standard regression software, which correspond to the two approaches in Section 7.3 for testing a single restriction on multiple coefficients.

The first method is to use Approach #1 of Section 7.3, which is to compute the F-statistic testing the hypothesis that $\beta_1 + 21\beta_2 = 0$. The standard error of $\Delta\hat{Y}$ is then given by[2]

$$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}. \tag{8.8}$$

When applied to the quadratic regression in Equation (8.2), the F-statistic testing the hypothesis that $\beta_1 + 21\beta_2 = 0$ is $F = 299.94$. Because $\Delta\hat{Y} = 2.96$, applying Equation (8.8) gives $SE(\Delta\hat{Y}) = 2.96/\sqrt{299.94} = 0.17$. Thus a 95% confidence interval for the change in the expected value of Y is $2.96 \pm 1.96 \times 0.17$ or (2.63, 3.29).

The second method is to use Approach #2 of Section 7.3, which entails transforming the regressors so that, in the transformed regression, one of the coefficients is $\beta_1 + 21\beta_2$. Doing this transformation is left as an exercise (Exercise 8.9).

A comment on interpreting coefficients in nonlinear specifications. In the multiple regression model of Chapters 6 and 7, the regression coefficients had a natural interpretation. For example, β_1 is the expected change in Y associated with a change in X_1, holding the other regressors constant. But, as we have seen, this is not generally the case in a nonlinear model. That is, it is not very helpful to think of β_1 in Equation (8.1) as being the effect of changing the district's income, holding the square of the district's income constant. In nonlinear models the regression function is best interpreted by graphing it and by calculating the predicted effect on Y of changing one or more of the independent variables.

[2]Equation (8.8) is derived by noting that the F-statistic is the square of the t-statistic testing this hypothesis—that is, $F = t^2 = [(\hat{\beta}_1 + 21\hat{\beta}_2)/SE(\hat{\beta}_1 + 21\hat{\beta}_1)]^2 = [\Delta\hat{Y}/SE(\Delta\hat{Y})]^2$—and solving for $SE(\Delta\hat{Y})$.

A General Approach to Modeling Nonlinearities Using Multiple Regression

The general approach to modeling nonlinear regression functions taken in this chapter has five elements:

1. ***Identify a possible nonlinear relationship.*** The best thing to do is to use economic theory and what you know about the application to suggest a possible nonlinear relationship. Before you even look at the data, ask yourself whether the slope of the regression function relating Y and X might reasonably depend on the value of X or on another independent variable. Why might such nonlinear dependence exist? What nonlinear shapes does this suggest? For example, thinking about classroom dynamics with 11-year-olds suggests that cutting class size from 18 students to 17 could have a greater effect than cutting it from 30 to 29.
2. ***Specify a nonlinear function and estimate its parameters by OLS.*** Sections 8.2 and 8.3 contain various nonlinear regression functions that can be estimated by OLS. After working through these sections you will understand the characteristics of each of these functions.
3. ***Determine whether the nonlinear model improves upon a linear model.*** Just because you think a regression function is nonlinear does not mean it really is! You must determine empirically whether your nonlinear model is appropriate. Most of the time you can use t-statistics and F-statistics to test the null hypothesis that the population regression function is linear against the alternative that it is nonlinear.
4. ***Plot the estimated nonlinear regression function.*** Does the estimated regression function describe the data well? Looking at Figures 8.2 and 8.3 suggested that the quadratic model fit the data better than the linear model.
5. ***Estimate the effect on* Y *of a change in* X.** The final step is to use the estimated regression to calculate the effect on Y of a change in one or more regressors X using the method in Key Concept 8.1.

8.2 Nonlinear Functions of a Single Independent Variable

This section provides two methods for modeling a nonlinear regression function. To keep things simple, we develop these methods for a nonlinear regression function that involves only one independent variable, X. As we see in Section 8.5, however, these models can be modified to include multiple independent variables.

The first method discussed in this section is polynomial regression, an extension of the quadratic regression used in the last section to model the relationship between test scores and income. The second method uses logarithms of X, of Y, or of both X and Y. Although these methods are presented separately, they can be used in combination.

Appendix 8.2 provides a calculus-based treatment of the models in this section.

Polynomials

One way to specify a nonlinear regression function is to use a polynomial in X. In general, let r denote the highest power of X that is included in the regression. The **polynomial regression model** of degree r is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i. \tag{8.9}$$

When $r = 2$, Equation (8.9) is the quadratic regression model discussed in Section 8.1. When $r = 3$ so that the highest power of X included is X^3, Equation (8.9) is called the **cubic regression model**.

The polynomial regression model is similar to the multiple regression model of Chapter 6 except that in Chapter 6 the regressors were distinct independent variables whereas here the regressors are powers of the same dependent variable, X; that is, the regressors are X, X^2, X^3, and so on. Thus the techniques for estimation and inference developed for multiple regression can be applied here. In particular, the unknown coefficients $\beta_0, \beta_1, \ldots, \beta_r$ in Equation (8.9) can be estimated by OLS regression of Y_i against $X_i, X_i^2, \ldots, X_i^r$.

Testing the null hypothesis that the population regression function is linear. If the population regression function is linear, then the quadratic and higher-degree terms do not enter the population regression function. Accordingly, the null hypothesis (H_0) that the regression is linear and the alternative (H_1) that it is a polynomial of degree r correspond to

$$H_0\colon \beta_2 = 0, \beta_3 = 0, \ldots, \beta_r = 0 \text{ vs. } H_1\colon \text{at least one } \beta_j \neq 0, j = 2, \ldots, r. \tag{8.10}$$

The null hypothesis that the population regression function is linear can be tested against the alternative that it is a polynomial of degree r by testing H_0 against H_1 in Equation (8.10). Because H_0 is a joint null hypothesis with $q = r - 1$ restrictions on the coefficients of the population polynomial regression model, it can be tested using the F-statistic as described in Section 7.2.

Which degree polynomial should I use? That is, how many powers of X should be included in a polynomial regression? The answer balances a trade-off between flexibility and statistical precision. Increasing the degree r introduces more flexibility into the regression function and allows it to match more shapes; a polynomial of degree r can have up to $r - 1$ bends (that is, inflection points) in its graph. But increasing r means adding more regressors, which can reduce the precision of the estimated coefficients.

Thus the answer to the question of how many terms to include is that you should include enough to model the nonlinear regression function adequately, but no more. Unfortunately, this answer is not very useful in practice!

A practical way to determine the degree of the polynomial is to ask whether the coefficients in Equation (8.9) associated with largest values of r are zero. If so, then these terms can be dropped from the regression. This procedure, which is called sequential hypothesis testing because individual hypotheses are tested sequentially, is summarized in the following steps:

1. Pick a maximum value of r and estimate the polynomial regression for that r.
2. Use the t-statistic to test the hypothesis that the coefficient on X^r [β_r in Equation (8.9)] is zero. If you reject this hypothesis, then X^r belongs in the regression, so use the polynomial of degree r.
3. If you do not reject $\beta_r = 0$ in step 2, eliminate X^r from the regression and estimate a polynomial regression of degree $r - 1$. Test whether the coefficient on X^{r-1} is zero. If you reject, use the polynomial of degree $r - 1$.
4. If you do not reject $\beta_{r-1} = 0$ in step 3, continue this procedure until the coefficient on the highest power in your polynomial is statistically significant.

This recipe has one missing ingredient: the initial degree r of the polynomial. In many applications involving economic data, the nonlinear functions are smooth, that is, they do not have sharp jumps, or "spikes." If so, then it is appropriate to choose a small maximum degree for the polynomial, such as 2, 3, or 4—that is, begin with $r = 2$ or 3 or 4 in step 1.

Application to district income and test scores. The estimated cubic regression function relating district income to test scores is

$$\begin{aligned}\widehat{TestScore} = \underset{(5.1)}{600.1} + \underset{(0.71)}{5.02}\,Income - \underset{(0.029)}{0.096}\,Income^2 + \underset{(0.00035)}{0.00069}\,Income^3,\\ \bar{R}^2 = 0.555. \qquad (8.11)\end{aligned}$$

The t-statistic on $Income^3$ is 1.97, so the null hypothesis that the regression function is a quadratic is rejected against the alternative that it is a cubic at the 5%

level. Moreover, the F-statistic testing the joint null hypothesis that the coefficients on $Income^2$ and $Income^3$ are both zero is 37.7, with a p-value less than 0.01%, so the null hypothesis that the regression function is linear is rejected against the alternative that it is either a quadratic or a cubic.

Interpretation of coefficients in polynomial regression models. The coefficients in polynomial regressions do not have a simple interpretation. The best way to interpret polynomial regressions is to plot the estimated regression function and calculate the estimated effect on Y associated with a change in X for one or more values of X.

Logarithms

Another way to specify a nonlinear regression function is to use the natural logarithm of Y and/or X. Logarithms convert changes in variables into percentage changes, and many relationships are naturally expressed in terms of percentages. Here are some examples:

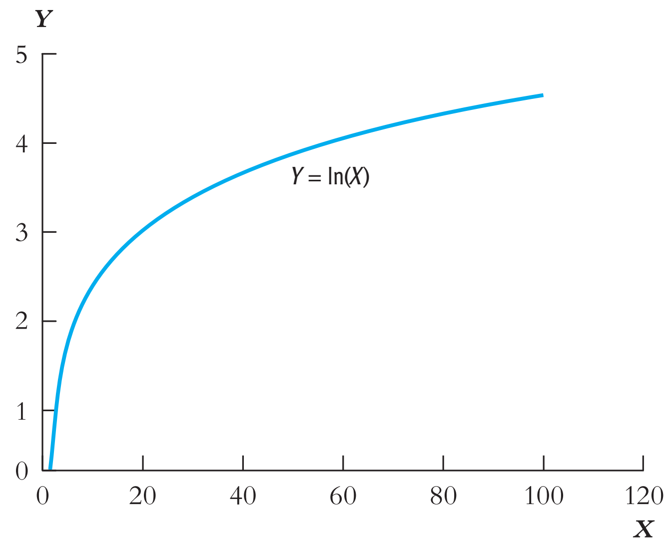
- A box in Chapter 3, "The Gender Gap of Earnings of College Graduates in the United States," examined the wage gap between male and female college graduates. In that discussion, the wage gap was measured in terms of dollars. However, it is easier to compare wage gaps across professions and over time when they are expressed in percentage terms.
- In Section 8.1, we found that district income and test scores were nonlinearly related. Would this relationship be linear using percentage changes? That is, might it be that a change in district income of 1%—rather than \$1000—is associated with a change in test scores that is approximately constant for different values of income?
- In the economic analysis of consumer demand, it is often assumed that a 1% increase in price leads to a certain *percentage* decrease in the quantity demanded. The percentage decrease in demand resulting from a 1% increase in price is called the price **elasticity**.

Regression specifications that use natural logarithms allow regression models to estimate percentage relationships such as these. Before introducing those specifications, we review the exponential and natural logarithm functions.

The exponential function and the natural logarithm. The exponential function and its inverse, the natural logarithm, play an important role in modeling nonlinear regression functions. The **exponential function** of x is e^x (that is, e raised

FIGURE 8.4 The Logarithm Function, $Y = \ln(X)$

The logarithmic function $Y = \ln(X)$ is steeper for small than for large values of X, is only defined for $X > 0$, and has slope $1/X$.



to the power x), where e is the constant 2.71828 . . . ; the exponential function is also written as $\exp(x)$. The **natural logarithm** is the inverse of the exponential function; that is, the natural logarithm is the function for which $x = \ln(e^x)$ or, equivalently, $x = \ln[\exp(x)]$. The base of the natural logarithm is e. Although there are logarithms in other bases, such as base 10, in this book we consider only logarithms in base e—that is, the natural logarithm—so when we use the term *logarithm* we always mean "natural logarithm."

The logarithm function, $y = \ln(x)$, is graphed in Figure 8.4. Note that the logarithm function is defined only for positive values of x. The logarithm function has a slope that is steep at first and then flattens out (although the function continues to increase). The slope of the logarithm function $\ln(x)$ is $1/x$.

The logarithm function has the following useful properties:

$$\ln(1/x) = -\ln(x); \tag{8.12}$$

$$\ln(ax) = \ln(a) + \ln(x); \tag{8.13}$$

$$\ln(x/a) = \ln(x) - \ln(a);\ \text{and} \tag{8.14}$$

$$\ln(x^a) = a\ln(x). \tag{8.15}$$

Logarithms and percentages. The link between the logarithm and percentages relies on a key fact: When Δx is small, the difference between the logarithm of

$x + \Delta x$ and the logarithm of x is approximately $\Delta x/x$, the percentage change in x divided by 100. That is,

$$\ln(x + \Delta x) - \ln(x) \cong \frac{\Delta x}{x} \quad \left(\text{when } \frac{\Delta x}{x} \text{ is small}\right), \tag{8.16}$$

where "$\cong$" means "approximately equal to." The derivation of this approximation relies on calculus, but it is readily demonstrated by trying out some values of x and Δx. For example, when $x = 100$ and $\Delta x = 1$, then $\Delta x/x = 1/100 = 0.01$ (or 1%), while $\ln(x + \Delta x) - \ln(x) = \ln(101) - \ln(100) = 0.00995$ (or 0.995%). Thus $\Delta x/x$ (which is 0.01) is very close to $\ln(x + \Delta x) - \ln(x)$ (which is 0.00995). When $\Delta x = 5$, $\Delta x/x = 5/100 = 0.05$, while $\ln(x + \Delta x) - \ln(x) = \ln(105) - \ln(100) = 0.04879$.

The three logarithmic regression models. There are three different cases in which logarithms might be used: when X is transformed by taking its logarithm but Y is not; when Y is transformed to its logarithm but X is not; and when both Y and X are transformed to their logarithms. The interpretation of the regression coefficients is different in each case. We discuss these three cases in turn.

Case I: X is in logarithms, Y is not. In this case, the regression model is

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i, i = 1, \ldots, n. \tag{8.17}$$

Because Y is not in logarithms but X is, this is sometimes referred to as a **linear-log model**.

In the linear-log model, a 1% change in X is associated with a change in Y of $0.01\beta_1$. To see this, consider the difference between the population regression function at values of X that differ by ΔX: This is $[\beta_0 + \beta_1 \ln(X + \Delta X)] - [\beta_0 + \beta_1 \ln(X)] = \beta_1[\ln(X + \Delta X) - \ln(X)] \cong \beta_1(\Delta X/X)$, where the final step uses the approximation in Equation (8.16). If X changes by 1%, then $\Delta X/X = 0.01$; thus in this model a 1% change in X is associated with a change of Y of $0.01\beta_1$.

The only difference between the regression model in Equation (8.17) and the regression model of Chapter 4 with a single regressor is that the right-hand variable is now the logarithm of X rather than X itself. To estimate the coefficients β_0 and β_1 in Equation (8.17), first compute a new variable, $\ln(X)$, which is readily done using a spreadsheet or statistical software. Then β_0 and β_1 can be estimated by the OLS regression of Y_i on $\ln(X_i)$, hypotheses about β_1 can be tested using the t-statistic, and a 95% confidence interval for β_1 can be constructed as $\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)$.

As an example, return to the relationship between district income and test scores. Instead of the quadratic specification, we could use the linear-log specification in Equation (8.17). Estimating this regression by OLS yields

$$\widehat{TestScore} = \underset{(3.8)}{557.8} + \underset{(1.40)}{36.42}\ln(Income), \bar{R}^2 = 0.561. \tag{8.18}$$

According to Equation (8.18), a 1% increase in income is associated with an increase in test scores of $0.01 \times 36.42 = 0.36$ point.

To estimate the effect on Y of a change in X in its original units of thousands of dollars (not in logarithms), we can use the method in Key Concept 8.1. For example, what is the predicted difference in test scores for districts with average incomes of \$10,000 versus \$11,000? The estimated value of ΔY is the difference between the predicted values: $\Delta\hat{Y} = [557.8 + 36.42\ln(11)] - [557.8 + 36.42\ln(10)] = 36.42 \times [\ln(11) - \ln(10)] = 3.47$. Similarly, the predicted difference between a district with average income of \$40,000 and a district with average income of \$41,000 is $36.42 \times [\ln(41) - \ln(40)] = 0.90$. Thus, like the quadratic specification, this regression predicts that a \$1000 increase in income has a larger effect on test scores in poor districts than it does in affluent districts.

The estimated linear-log regression function in Equation (8.18) is plotted in Figure 8.5. Because the regressor in Equation (8.18) is the natural logarithm of income rather than income, the estimated regression function is not a straight line. Like the quadratic regression function in Figure 8.3, it is initially steep but then flattens out for higher levels of income.

Case II: Y is in logarithms, X is not. In this case, the regression model is

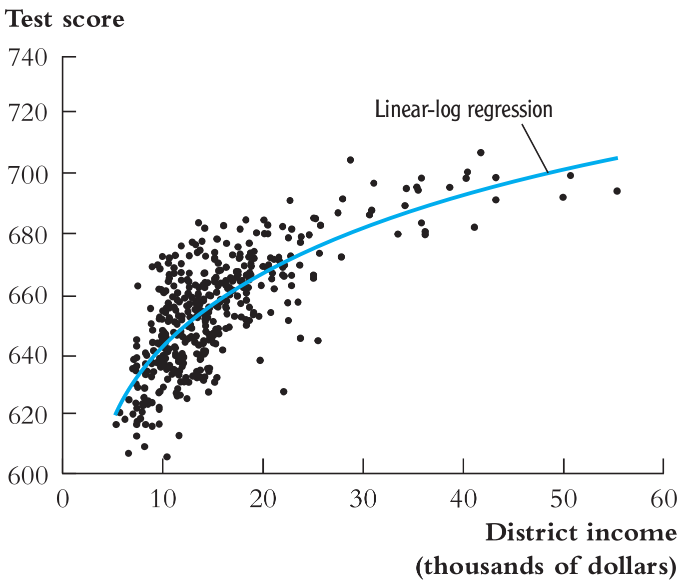
$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i. \tag{8.19}$$

Because Y is in logarithms but X is not, this is referred to as a **log-linear model**.

In the log-linear model, a one-unit change in X ($\Delta X = 1$) is associated with a $(100 \times \beta_1)\%$ change in Y. To see this, compare the expected values of $\ln(Y)$ for values of X that differ by ΔX. The expected value of $\ln(Y)$ given X is $\ln(Y) = \beta_0 + \beta_1 X$. When X is $X + \Delta X$, the expected value is given by $\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X)$. Thus the difference between these expected values is $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1(X + \Delta X)] - [\beta_0 + \beta_1 X] = \beta_1 \Delta X$. From the approximation in Equation (8.16), however, if $\beta_1 \Delta X$ is small, then $\ln(Y + \Delta Y) - \ln(Y) \cong \Delta Y/Y$. Thus $\Delta Y/Y \cong \beta_1 \Delta X$. If $\Delta X = 1$ so that X changes by one unit, then $\Delta Y/Y$ changes

FIGURE 8.5 The Linear-Log Regression Function

The estimated linear-log regression function $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$ captures much of the nonlinear relation between test scores and district income.



by β_1. Translated into percentages, a one-unit change in X is associated with a $(100 \times \beta_1)\%$ change in Y.

As an illustration, we return to the empirical example of Section 3.7, the relationship between age and earnings of college graduates. Many employment contracts specify that, for each additional year of service, a worker gets a certain percentage increase in his or her wage. This percentage relationship suggests estimating the log-linear specification in Equation (8.19) so that each additional year of age (X) is, on average in the population, associated with some constant percentage increase in earnings (Y). By first computing the new dependent variable, $\ln(Earnings_i)$, the unknown coefficients β_0 and β_1 can be estimated by the OLS regression of $\ln(Earnings_i)$ against Age_i. When estimated using the 14,752 observations on college graduates in the March 2013 Current Population Survey (the data are described in Appendix 3.1), this relationship is

$$\widehat{\ln(Earnings)} = \underset{(0.018)}{2.811} + \underset{(0.0004)}{0.0096} Age, \ \bar{R}^2 = 0.034. \tag{8.20}$$

According to this regression, earnings are predicted to increase by 0.96% $[(100 \times 0.0096)\%]$ for each additional year of age.

Case III: Both X and Y are in logarithms. In this case, the regression model is

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i. \tag{8.21}$$

Because both Y and X are specified in logarithms, this is referred to as a **log-log model**.

In the log-log model, a 1% change in X is associated with a β_1% change in Y. Thus in this specification β_1 is the elasticity of Y with respect to X. To see this, again apply Key Concept 8.1; thus $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1 \ln(X + \Delta X)] - [\beta_0 + \beta_1 \ln(X)] = \beta_1[\ln(X + \Delta X) - \ln(X)]$. Application of the approximation in Equation (8.16) to both sides of this equation yields

$$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X} \text{ or}$$

$$\beta_1 = \frac{\Delta Y/Y}{\Delta X/X} = \frac{100 \times (\Delta Y/Y)}{100 \times (\Delta X/X)} = \frac{\text{percentage change in } Y}{\text{percentage change in } X}. \tag{8.22}$$

Thus in the log-log specification β_1 is the ratio of the percentage change in Y associated with the percentage change in X. If the percentage change in X is 1% (that is, if $\Delta X = 0.01X$), then β_1 is the percentage change in Y associated with a 1% change in X. That is, β_1 is the elasticity of Y with respect to X.

As an illustration, return to the relationship between income and test scores. When this relationship is specified in this form, the unknown coefficients are estimated by a regression of the logarithm of test scores against the logarithm of income. The resulting estimated equation is

$$\widehat{\ln(TestScore)} = \underset{(0.006)}{6.336} + \underset{(0.0021)}{0.0554} \ln(Income), \bar{R}^2 = 0.557. \tag{8.23}$$

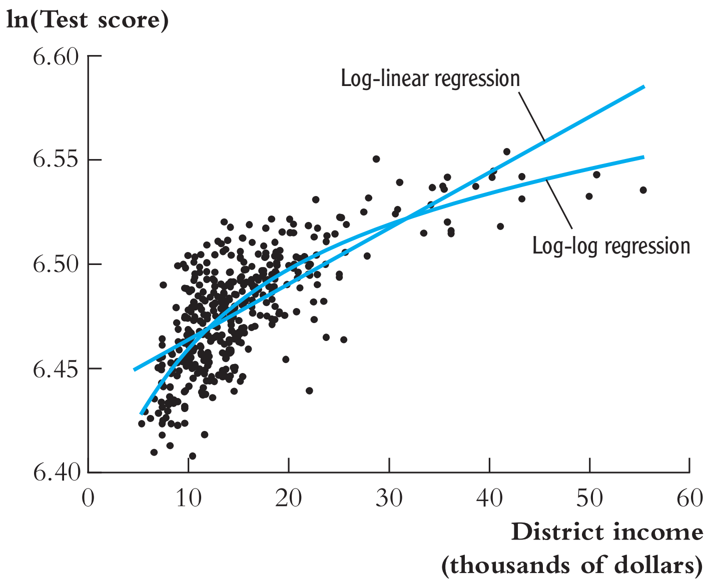
According to this estimated regression function, a 1% increase in income is estimated to correspond to a 0.0554% increase in test scores.

The estimated log-log regression function in Equation (8.23) is plotted in Figure 8.6. Because Y is in logarithms, the vertical axis in Figure 8.6 is the logarithm of the test score and the scatterplot is the logarithm of test scores versus district income. For comparison purposes, Figure 8.6 also shows the estimated regression function for a log-linear specification, which is

$$\widehat{\ln(TestScore)} = \underset{(0.003)}{6.439} + \underset{(0.00018)}{0.00284}\, Income, \bar{R}^2 = 0.497. \tag{8.24}$$

FIGURE 8.6 The Log-Linear and Log-Log Regression Functions

In the log-linear regression function, ln(Y) is a linear function of X. In the log-log regression function, ln(Y) is a linear function of ln(X).



Because the vertical axis is in logarithms, the regression function in Equation (8.24) is the straight line in Figure 8.6.

As you can see in Figure 8.6, the log-log specification fits better than the log-linear specification. This is consistent with the higher $\overline{R}^2$ for the log-log regression (0.557) than for the log-linear regression (0.497). Even so, the log-log specification does not fit the data especially well: At the lower values of income most of the observations fall below the log-log curve, while in the middle income range most of the observations fall above the estimated regression function.

The three logarithmic regression models are summarized in Key Concept 8.2.

A difficulty with comparing logarithmic specifications. Which of the log regression models best fits the data? As we saw in the discussion of Equations (8.23) and (8.24), the $\overline{R}^2$ can be used to compare the log-linear and log-log models; as it happened, the log-log model had the higher $\overline{R}^2$. Similarly, the $\overline{R}^2$ can be used to compare the linear-log regression in Equation (8.18) and the linear regression of Y against X. In the test score and income regression, the linear-log regression has an $\overline{R}^2$ of 0.561 while the linear regression has an $\overline{R}^2$ of 0.508, so the linear-log model fits the data better.

How can we compare the linear-log model and the log-log model? Unfortunately, the $\overline{R}^2$ *cannot* be used to compare these two regressions because their dependent variables are different [one is Y, the other is ln(Y)]. Recall that the $\overline{R}^2$

KEY CONCEPT 8.2

Logarithms in Regression: Three Cases

Logarithms can be used to transform the dependent variable Y, an independent variable X, or both (but the variable being transformed must be positive). The following table summarizes these three cases and the interpretation of the regression coefficient β_1. In each case, β_1 can be estimated by applying OLS after taking the logarithm of the dependent and/or independent variable.

Case	Regression Specification	Interpretation of β_1
I	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	A 1% change in X is associated with a change in Y of $0.01\beta_1$.
II	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	A change in X by one unit ($\Delta X = 1$) is associated with a $100\beta_1\%$ change in Y.
III	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	A 1% change in X is associated with a $\beta_1\%$ change in Y, so β_1 is the elasticity of Y with respect to X.

measures the fraction of the variance of the dependent variable explained by the regressors. Because the dependent variables in the log-log and linear-log models are different, it does not make sense to compare their $\overline{R}^2$'s.

Because of this problem, the best thing to do in a particular application is to decide, using economic theory and either your or other experts' knowledge of the problem, whether it makes sense to specify Y in logarithms. For example, labor economists typically model earnings using logarithms because wage comparisons, contract wage increases, and so forth are often most naturally discussed in percentage terms. In modeling test scores, it seems (to us, anyway) natural to discuss test results in terms of points on the test rather than percentage increases in the test scores, so we focus on models in which the dependent variable is the test score rather than its logarithm.

Computing predicted values of Y when Y is in logarithms.[3] If the dependent variable Y has been transformed by taking logarithms, the estimated regression can be used to compute directly the predicted value of $\ln(Y)$. However, it is a bit trickier to compute the predicted value of Y itself.

[3] This material is more advanced and can be skipped without loss of continuity.

To see this, consider the log-linear regression model in Equation (8.19) and rewrite it so that it is specified in terms of Y rather than $\ln(Y)$. To do so, take the exponential function of both sides of the Equation (8.19); the result is

$$Y_i = \exp(\beta_0 + \beta_1 X_i + u_i) = e^{\beta_0 + \beta_1 X_i} e^{u_i}. \tag{8.25}$$

The expected value of Y_i given X_i is $E(Y_i | X_i) = E(e^{\beta_0 + \beta_1 X_i} e^{u_i} | X_i) = e^{\beta_0 + \beta_1 X_i} E(e^{u_i} | X_i)$. The problem is that even if $E(u_i | X_i) = 0$, $E(e^{u_i} | X_i) \neq 1$. Thus the appropriate predicted value of Y_i is not simply obtained by taking the exponential function of $\hat{\beta}_0 + \hat{\beta}_1 X_i$, that is, by setting $\hat{Y}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}$: This predicted value is biased because of the missing factor $E(e^{u_i} | X_i)$.

One solution to this problem is to estimate the factor $E(e^{u_i} | X_i)$ and use this estimate when computing the predicted value of Y. Exercise 17.12 works through several ways to estimate $E(e^{u_i} | X_i)$, but this gets complicated, particularly if u_i is heteroskedastic, and we do not pursue it further.

Another solution, which is the approach used in this book, is to compute predicted values of the logarithm of Y but not transform them to their original units. In practice, this is often acceptable because when the dependent variable is specified as a logarithm, it is often most natural just to use the logarithmic specification (and the associated percentage interpretations) throughout the analysis.

Polynomial and Logarithmic Models of Test Scores and District Income

In practice, economic theory or expert judgment might suggest a functional form to use, but in the end the true form of the population regression function is unknown. In practice, fitting a nonlinear function therefore entails deciding which method or combination of methods works best. As an illustration, we compare logarithmic and polynomial models of the relationship between district income and test scores.

Polynomial specifications. We considered two polynomial specifications specified using powers of *Income*, quadratic [Equation (8.2)] and cubic [Equation (8.11)]. Because the coefficient on $Income^3$ in Equation (8.11) was significant at the 5% level, the cubic specification provided an improvement over the quadratic, so we select the cubic model as the preferred polynomial specification.

Logarithmic specifications. The logarithmic specification in Equation (8.18) seemed to provide a good fit to these data, but we did not test this formally. One way to do so is to augment it with higher powers of the logarithm of income. If these additional terms are not statistically different from zero, then we can conclude that the specification in Equation (8.18) is adequate in the sense that it cannot be rejected against a polynomial function of the logarithm. Accordingly, the estimated cubic regression (specified in powers of the logarithm of income) is

$$\begin{aligned}\widehat{TestScore} = &\underset{(79.4)}{486.1} + \underset{(87.9)}{113.4}\ln(Income) - \underset{(31.7)}{26.9}[\ln(Income)^2] \\ &+ \underset{(3.74)}{3.06}[\ln(Income)]^3, \bar{R}^2 = 0.560.\end{aligned} \tag{8.26}$$

The t-statistic on the coefficient on the cubic term is 0.818, so the null hypothesis that the true coefficient is zero is not rejected at the 10% level. The F-statistic testing the joint hypothesis that the true coefficients on the quadratic and cubic term are both zero is 0.44, with a p-value of 0.64, so this joint null hypothesis is not rejected at the 10% level. Thus the cubic logarithmic model in Equation (8.26) does not provide a statistically significant improvement over the model in Equation (8.18), which is linear in the logarithm of income.

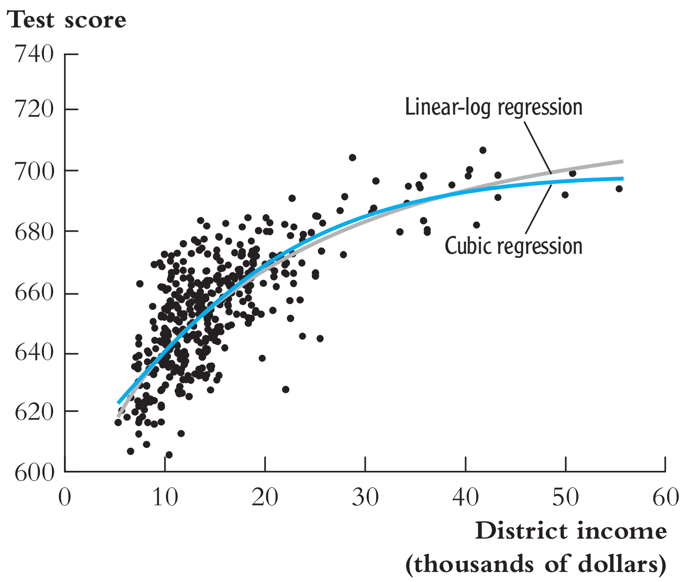
Comparing the cubic and linear-log specifications. Figure 8.7 plots the estimated regression functions from the cubic specification in Equation (8.11) and the linear-log specification in Equation (8.18). The two estimated regression functions are quite similar. One statistical tool for comparing these specifications is the $\bar{R}^2$. The $\bar{R}^2$ of the logarithmic regression is 0.561, and for the cubic regression it is 0.555. Because the logarithmic specification has a slight edge in terms of the $\bar{R}^2$ and because this specification does not need higher-degree polynomials in the logarithm of income to fit these data, we adopt the logarithmic specification in Equation (8.18).

8.3 Interactions Between Independent Variables

In the introduction to this chapter we wondered whether reducing the student–teacher ratio might have a bigger effect on test scores in districts where many students are still learning English than in those with few still learning English. This could arise, for example, if students who are still learning English benefit differentially from one-on-one or small-group instruction. If so, the presence of many English learners in a district would interact with the student–teacher ratio in such a way that the effect on test scores of a change in the student–teacher ratio would depend on the fraction of English learners.

FIGURE 8.7 The Linear-Log and Cubic Regression Functions

The estimated cubic regression function [Equation (8.11)] and the estimated linear-log regression function [Equation (8.18)] are nearly identical in this sample.



This section explains how to incorporate such interactions between two independent variables into the multiple regression model. The possible interaction between the student–teacher ratio and the fraction of English learners is an example of the more general situation in which the effect on Y of a change in one independent variable depends on the value of another independent variable. We consider three cases: when both independent variables are binary, when one is binary and the other is continuous, and when both are continuous.

Interactions Between Two Binary Variables

Consider the population regression of log earnings [Y_i, where $Y_i = \ln(Earnings_i)$] against two binary variables: whether a worker has a college degree (D_{1i}, where $D_{1i} = 1$ if the i^{th} person graduated from college) and the worker's gender (D_{2i}, where $D_{2i} = 1$ if the i^{th} person is female). The population linear regression of Y_i on these two binary variables is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i. \tag{8.27}$$

In this regression model, β_1 is the effect on log earnings of having a college degree, holding gender constant, and β_2 is the effect of being female, holding schooling constant.

The specification in Equation (8.27) has an important limitation: The effect of having a college degree in this specification, holding constant gender, is the same for men and women. There is, however, no reason that this must be so. Phrased mathematically, the effect on Y_i of D_{1i}, holding D_{2i} constant, could depend on the value of D_{2i}. In other words, there could be an interaction between having a college degree and gender so that the value in the job market of a degree is different for men and women.

Although the specification in Equation (8.27) does not allow for this interaction between having a college degree and gender, it is easy to modify the specification so that it does by introducing another regressor, the product of the two binary variables, $D_{1i} \times D_{2i}$. The resulting regression is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3(D_{1i} \times D_{2i}) + u_i. \tag{8.28}$$

The new regressor, the product $D_{1i} \times D_{2i}$, is called an **interaction term** or an **interacted regressor**, and the population regression model in Equation (8.28) is called a binary variable **interaction regression model**.

The interaction term in Equation (8.28) allows the population effect on log earnings (Y_i) of having a college degree (changing D_{1i} from $D_{1i} = 0$ to $D_{1i} = 1$) to depend on gender (D_{2i}). To show this mathematically, calculate the population effect of a change in D_{1i} using the general method laid out in Key Concept 8.1. The first step is to compute the conditional expectation of Y_i for $D_{1i} = 0$, given a value of D_{2i}; this is $E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_0 + \beta_1 \times 0 + \beta_2 \times d_2 + \beta_3 \times (0 \times d_2) = \beta_0 + \beta_2 d_2$, where we use the conditional mean zero assumption, $E(u_i | D_{1i}, D_{2i}) = 0$. The next step is to compute the conditional expectation of Y_i after the change—that is, for $D_{1i} = 1$—given the same value of D_{2i}; this is $E(Y_i | D_{1i} = 1, D_{2i} = d_2) = \beta_0 + \beta_1 \times 1 + \beta_2 \times d_2 + \beta_3 \times (1 \times d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2$. The effect of this change is the difference of expected values [that is, the difference in Equation (8.4)], which is

$$E(Y_i | D_{1i} = 1, D_{2i} = d_2) - E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_1 + \beta_3 d_2. \tag{8.29}$$

Thus, in the binary variable interaction specification in Equation (8.28), the effect of acquiring a college degree (a unit change in D_{1i}) depends on the person's gender [the value of D_{2i}, which is d_2 in Equation (8.29)]. If the person is male ($d_2 = 0$), the effect of acquiring a college degree is β_1, but if the person is female ($d_2 = 1$), the effect is $\beta_1 + \beta_3$. The coefficient β_3 on the interaction term is the difference in the effect of acquiring a college degree for women versus men.

A Method for Interpreting Coefficients in Regressions with Binary Variables

KEY CONCEPT 8.3

First compute the expected values of Y for each possible case described by the set of binary variables. Next compare these expected values. Each coefficient can then be expressed either as an expected value or as the difference between two or more expected values.

Although this example was phrased using log earnings, having a college degree, and gender, the point is a general one. The binary variable interaction regression allows the effect of changing one of the binary independent variables to depend on the value of the other binary variable.

The method we used here to interpret the coefficients was, in effect, to work through each possible combination of the binary variables. This method, which applies to all regressions with binary variables, is summarized in Key Concept 8.3.

Application to the student–teacher ratio and the percentage of English learners. Let $HiSTR_i$ be a binary variable that equals 1 if the student–teacher ratio is 20 or more and equals 0 otherwise, and let $HiEL_i$ be a binary variable that equals 1 if the percentage of English learners is 10% or more and equals 0 otherwise. The interacted regression of test scores against $HiSTR_i$ and $HiEL_i$ is

$$\begin{aligned}\widehat{TestScore} &= \underset{(1.4)}{664.1} - \underset{(1.9)}{1.9}\,HiSTR - \underset{(2.3)}{18.2}\,HiEL - \underset{(3.1)}{3.5}(HiSTR \times HiEL),\\ \bar{R}^2 &= 0.290. \end{aligned} \tag{8.30}$$

The predicted effect of moving from a district with a low student–teacher ratio to one with a high student–teacher ratio, holding constant whether the percentage of English learners is high or low, is given by Equation (8.29), with estimated coefficients replacing the population coefficients. According to the estimates in Equation (8.30), this effect thus is $-1.9 - 3.5HiEL$. That is, if the fraction of English learners is low ($HiEL = 0$), then the effect on test scores of moving from $HiSTR = 0$ to $HiSTR = 1$ is for test scores to decline by 1.9 points. If the fraction of English learners is high, then test scores are estimated to decline by $1.9 + 3.5 = 5.4$ points.

The estimated regression in Equation (8.30) also can be used to estimate the mean test scores for each of the four possible combinations of the binary variables.

This is done using the procedure in Key Concept 8.3. Accordingly, the sample average test score for districts with low student–teacher ratios ($HiSTR_i = 0$) and low fractions of English learners ($HiEL_i = 0$) is 664.1. For districts with $HiSTR_i = 1$ (high student–teacher ratios) and $HiEL_i = 0$ (low fractions of English learners), the sample average is 662.2 (= 664.1 − 1.9). When $HiSTR_i = 0$ and $HiEL_i = 1$, the sample average is 645.9 (= 664.1 − 18.2), and when $HiSTR_i = 1$ and $HiEL_i = 1$, the sample average is 640.5 (= 664.1 − 1.9 − 18.2 − 3.5).

Interactions Between a Continuous and a Binary Variable

Next consider the population regression of log earnings [$Y_i = \ln(Earnings_i)$] against one continuous variable, the individual's years of work experience (X_i), and one binary variable, whether the worker has a college degree (D_i, where $D_i = 1$ if the i^{th} person is a college graduate). As shown in Figure 8.8, the population regression line relating Y and the continuous variable X can depend on the binary variable D in three different ways.

In Figure 8.8a, the two regression lines differ only in their intercept. The corresponding population regression model is

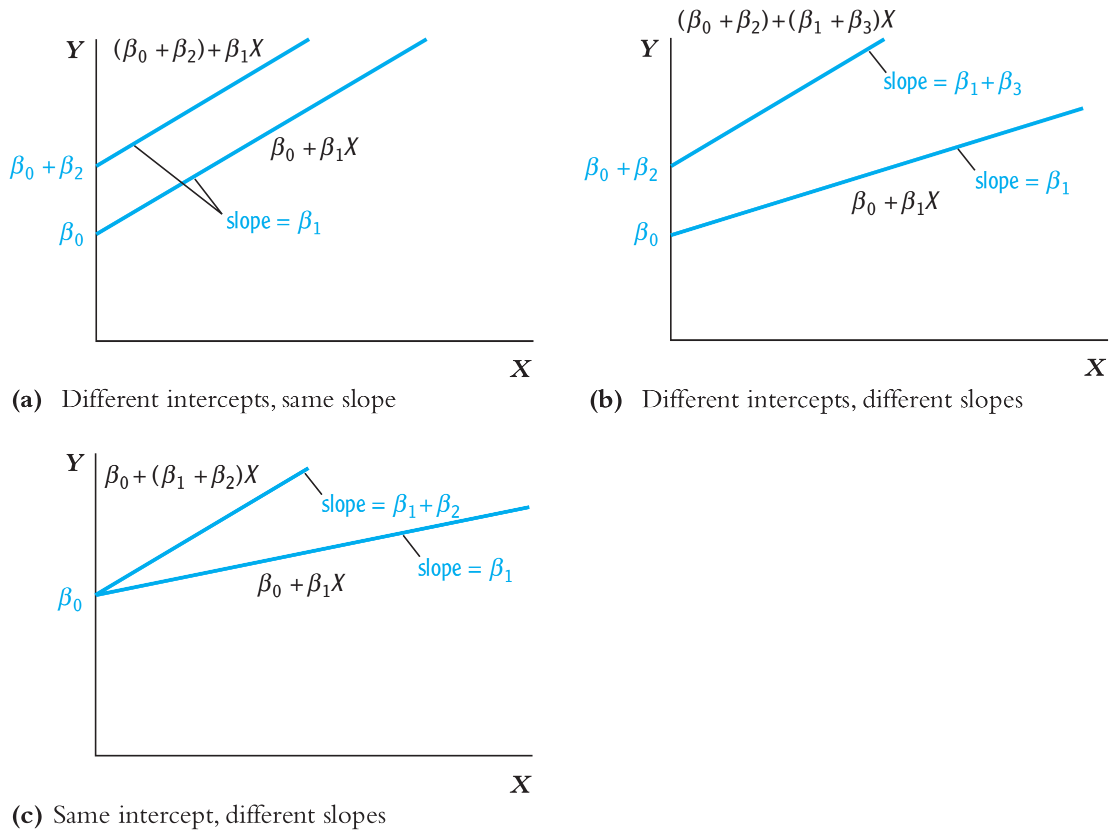
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i. \tag{8.31}$$

This is the familiar multiple regression model with a population regression function that is linear in X_i and D_i. When $D_i = 0$, the population regression function is $\beta_0 + \beta_1 X_i$, so the intercept is β_0 and the slope is β_1. When $D_i = 1$, the population regression function is $\beta_0 + \beta_1 X_i + \beta_2$, so the slope remains β_1 but the intercept is $\beta_0 + \beta_2$. Thus β_2 is the difference between the intercepts of the two regression lines, as shown in Figure 8.8a. Stated in terms of the earnings example, β_1 is the effect on log earnings of an additional year of work experience, holding college degree status constant, and β_2 is the effect of a college degree on log earnings, holding years of experience constant. In this specification, the effect of an additional year of work experience is the same for college graduates and nongraduates; that is, the two lines in Figure 8.8a have the same slope.

In Figure 8.8b, the two lines have different slopes and intercepts. The different slopes permit the effect of an additional year of work to differ for college graduates and nongraduates. To allow for different slopes, add an interaction term to Equation (8.31):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i, \tag{8.32}$$

FIGURE 8.8 Regression Functions Using Binary and Continuous Variables



Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1 X + \beta_2 D$ allows for different intercepts but has the same slope, (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$ allows for different intercepts and different slopes, and (c) $\beta_0 + \beta_1 X + \beta_2 (X \times D)$ has the same intercept but allows for different slopes.

where $X_i \times D_i$ is a new variable, the product of X_i and D_i. To interpret the coefficients of this regression, apply the procedure in Key Concept 8.3. Doing so shows that, if $D_i = 0$, the population regression function is $\beta_0 + \beta_1 X_i$, whereas if $D_i = 1$, the population regression function is $(\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_i$. Thus this specification allows for two different population regression functions relating Y_i and X_i, depending on the value of D_i, as is shown in Figure 8.8b. The difference between the two intercepts is β_2, and the difference between the two slopes is β_3. In the earnings example, β_1 is the effect of an additional year of work experience for nongraduates ($D_i = 0$) and $\beta_1 + \beta_3$ is this effect for graduates, so β_3 is the *difference* in the effect of an additional year of work experience for college graduates versus nongraduates.

KEY CONCEPT 8.4

Interactions Between Binary and Continuous Variables

Through the use of the interaction term $X_i \times D_i$, the population regression line relating Y_i and the continuous variable X_i can have a slope that depends on the binary variable D_i. There are three possibilities:

1. Different intercept, same slope (Figure 8.8a):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i;$$

2. Different intercept and slope (Figure 8.8b):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i;$$

3. Same intercept, different slope (Figure 8.8c):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i.$$

A third possibility, shown in Figure 8.8c, is that the two lines have different slopes but the same intercept. The interacted regression model for this case is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i. \tag{8.33}$$

The coefficients of this specification also can be interpreted using Key Concept 8.3. In terms of the earnings example, this specification allows for different effects of experience on log earnings between college graduates and nongraduates, but requires that expected log earnings be the same for both groups when they have no prior experience. Said differently, this specification corresponds to the population mean entry-level wage being the same for college graduates and nongraduates. This does not make much sense in this application, and in practice this specification is used less frequently than Equation (8.32), which allows for different intercepts and slopes.

All three specifications—Equations (8.31), (8.32), and (8.33)—are versions of the multiple regression model of Chapter 6, and once the new variable $X_i \times D_i$ is created, the coefficients of all three can be estimated by OLS.

The three regression models with a binary and a continuous independent variable are summarized in Key Concept 8.4.

Application to the student–teacher ratio and the percentage of English learners. Does the effect on test scores of cutting the student–teacher ratio depend on whether the percentage of students still learning English is high or low? One way to answer this question is to use a specification that allows for two different regression lines, depending on whether there is a high or a low percentage of English learners. This is achieved using the different intercept/different slope specification:

$$\begin{aligned}\widehat{TestScore} &= \underset{(11.9)}{682.2} - \underset{(0.59)}{0.97}STR + \underset{(19.5)}{5.6}HiEL - \underset{(0.97)}{1.28}(STR \times HiEL),\\ \overline{R}^2 &= 0.305,\end{aligned} \tag{8.34}$$

where the binary variable $HiEL_i$ equals 1 if the percentage of students still learning English in the district is greater than 10% and equals 0 otherwise.

For districts with a low fraction of English learners ($HiEL_i = 0$), the estimated regression line is $682.2 - 0.97STR_i$. For districts with a high fraction of English learners ($HiEL_i = 1$), the estimated regression line is $682.2 + 5.6 - 0.97STR_i - 1.28STR_i = 687.8 - 2.25STR_i$. According to these estimates, reducing the student–teacher ratio by 1 is predicted to increase test scores by 0.97 point in districts with low fractions of English learners but by 2.25 points in districts with high fractions of English learners. The difference between these two effects, 1.28 points, is the coefficient on the interaction term in Equation (8.34).

The interaction regression model in Equation (8.34) allows us to estimate the effect of more nuanced policy interventions than the across-the-board class size reduction considered so far. For example, suppose that the state considered a policy to reduce the student–teacher ratio by 2 in districts with a high fraction of English learners ($HiEL_i = 1$) but to leave class size unchanged in other districts. Applying the method of Key Concept 8.1 to Equations (8.32) and (8.34) shows that the estimated effect of this reduction for the districts for which $HiEL = 1$ is $-2(\hat{\beta}_1 + \hat{\beta}_3) = 4.50$. The standard error of this estimated effect is $SE(-2\hat{\beta}_1 - 2\hat{\beta}_3) = 1.53$, which can be computed using Equation (8.8) and the methods of Section 7.3.

The OLS regression in Equation (8.34) can be used to test several hypotheses about the population regression line. First, the hypothesis that the two lines are in fact the same can be tested by computing the F-statistic testing the joint hypothesis that the coefficient on $HiEL_i$ and the coefficient on the interaction term $STR_i \times HiEL_i$ are both zero. This F-statistic is 89.9, which is significant at the 1% level.

Second, the hypothesis that two lines have the same slope can be tested by testing whether the coefficient on the interaction term is zero. The t-statistic, $-1.28/0.97 = -1.32$, is less than 1.64 in absolute value, so the null hypothesis that the two lines have the same slope cannot be rejected using a two-sided test at the 10% significance level.

Third, the hypothesis that the lines have the same intercept corresponds to the restriction that the population coefficient on $HiEL$ is zero. The t-statistic testing this restriction is $t = 5.6/19.5 = 0.29$, so the hypothesis that the lines have the same intercept cannot be rejected at the 5% level.

These three tests produce seemingly contradictory results: The joint test using the F-statistic rejects the joint hypothesis that the slope and the intercept are the same, but the tests of the individual hypotheses using the t-statistic fail to reject. The reason is that the regressors, $HiEL$ and $STR \times HiEL$, are highly correlated. This results in large standard errors on the individual coefficients. Even though it is impossible to tell which of the coefficients is nonzero, there is strong evidence against the hypothesis that *both* are zero.

Finally, the hypothesis that the student–teacher ratio does not enter this specification can be tested by computing the F-statistic for the joint hypothesis that the coefficients on STR and on the interaction term are both zero. This F-statistic is 5.64, which has a p-value of 0.004. Thus the coefficients on the student–teacher ratio are statistically significant at the 1% significance level.

Interactions Between Two Continuous Variables

Now suppose that both independent variables (X_{1i} and X_{2i}) are continuous. An example is when Y_i is log earnings of the i^{th} worker, X_{1i} is his or her years of work experience, and X_{2i} is the number of years he or she went to school. If the population regression function is linear, the effect on wages of an additional year of experience does not depend on the number of years of education, or, equivalently, the effect of an additional year of education does not depend on the number of years of work experience. In reality, however, there might be an interaction between these two variables so that the effect on wages of an additional year of experience depends on the number of years of education. This interaction can be modeled by augmenting the linear regression model with an interaction term that is the product of X_{1i} and X_{2i}:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i. \tag{8.35}$$

The interaction term allows the effect of a unit change in X_1 to depend on X_2. To see this, apply the general method for computing effects in nonlinear regression models in Key Concept 8.1. The difference in Equation (8.4), computed for the interacted regression function in Equation (8.35), is $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1$ [Exercise 8.10(a)]. Thus the effect on Y of a change in X_1, holding X_2 constant, is

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2, \tag{8.36}$$

which depends on X_2. For example, in the earnings example, if β_3 is positive, then the effect on log earnings of an additional year of experience is greater, by the amount β_3, for each additional year of education the worker has.

The Return to Education and the Gender Gap

In addition to its intellectual pleasures, education has economic rewards. As the boxes in Chapters 3 and 5 show, workers with more education tend to earn more than their counterparts with less education. The analysis in those boxes was incomplete, however, for at least three reasons. First, it failed to control for other determinants of earnings that might be correlated with educational achievement, so the OLS estimator of the coefficient on education could have omitted variable bias. Second, the functional form used in Chapter 5—a simple linear relation—implies that earnings change by a constant dollar amount for each additional year of education, whereas one might suspect that the dollar change in earnings is actually larger at higher levels of education. Third, the box in Chapter 5 ignores the gender differences in earnings highlighted in the box in Chapter 3.

All these limitations can be addressed by a multiple regression analysis that controls for determinants of earnings that, if omitted, could cause omitted variable bias and that uses a nonlinear functional form relating education and earnings. Table 8.1 summarizes regressions estimated using data on full-time workers, ages 30 through 64, from the Current Population Survey (the CPS data are described in Appendix 3.1). The dependent variable is the logarithm of hourly earnings, so another year of education is associated with a constant percentage increase (not dollar increase) in earnings.

Table 8.1 has four salient results. First, the omission of gender in regression (1) does not result in substantial omitted variable bias: Even though gender enters regression (2) significantly and with a large coefficient, gender and years of education are uncorrelated; that is, on average men and women have nearly the same levels of education. Second, the returns to education are economically and statistically significantly different for men and women: In regression (3), the *t*-statistic testing the hypothesis that they are the same is 4.55 (= 0.008/0.0018). Third, regression (4) controls for the region of the country in which the individual lives, thereby addressing potential omitted variable bias that might arise if years of education differ systematically by region. Controlling for region makes a small difference to the estimated coefficients on the education terms, relative to those reported in regression (3). Fourth, regression (4) controls for the potential experience of the worker, as measured by years since completion of schooling. The estimated coefficients imply a declining marginal value for each year of potential experience.

The estimated economic return to education in regression (4) is 11.26% for each year of education for men and 12.25% (= 0.1126 + 0.0099, in percent) for women. Because the regression functions for men and women have different slopes, the gender gap depends on the years of education. For 12 years of education, the gender gap is estimated to be 27.3% (= 0.0099 × 12 − 0.392, in percent); for 16 years of education, the gender gap is less in percentage terms, 23.4%.

These estimates of the return to education and the gender gap still have limitations, including the possibility of other omitted variables, notably the native ability of the worker, and potential problems associated with the way variables are measured in the CPS. Nevertheless, the estimates in Table 8.1

continued on next page

TABLE 8.1 The Return to Education and the Gender Gap: Regression Results for the United States in 2012

Dependent variable: logarithm of *Hourly Earnings*.

Regressor	(1)	(2)	(3)	(4)
Years of education	0.1082** (0.0009)	0.1111** (0.0009)	0.1078** (0.0012)	0.1126** (0.0012)
Female		−0.251** (0.005)	−0.367** (0.026)	−0.392** (0.025)
Female × *Years of education*			0.0081** (0.0018)	0.0099** (0.0018)
Potential experience				0.0186** (0.0012)
Potential experience^2				−0.000263** (0.000024)
Midwest				−0.080** (0.007)
South				−0.083** (0.007)
West				−0.018** (0.007)
Intercept	1.515** (0.013)	1.585** (0.013)	1.632** (0.016)	1.335** (0.024)
$\overline{R}^2$	0.221	0.263	0.264	0.276

The data are from the March 2013 Current Population Survey (see Appendix 3.1). The sample size is $n = 50{,}174$ observations for each regression. *Female* is an indicator variable that equals 1 for women and 0 for men. *Midwest*, *South*, and *West* are indicator variables denoting the region of the United States in which the worker lives: For example, *Midwest* equals 1 if the worker lives in the Midwest and equals 0 otherwise (the omitted region is *Northeast*). Standard errors are reported in parentheses below the estimated coefficients. Individual coefficients are statistically significant at the *5% or **1% significance level.

are consistent with those obtained by economists who carefully address these limitations. A survey by the econometrician David Card (1999) of dozens of empirical studies concludes that labor economists' best estimates of the return to education generally fall between 8% and 11%, and that the return depends on the quality of the education. If you are interested in learning more about the economic return to education, see Card (1999).

Interactions in Multiple Regression

KEY CONCEPT 8.5

The interaction term between the two independent variables X_1 and X_2 is their product $X_1 \times X_2$. Including this interaction term allows the effect on Y of a change in X_1 to depend on the value of X_2 and, conversely, allows the effect of a change in X_2 to depend on the value of X_1.

The coefficient on $X_1 \times X_2$ is the effect of a one-unit increase in X_1 *and* X_2, above and beyond the sum of the individual effects of a unit increase in X_1 alone and a unit increase in X_2 alone. This is true whether X_1 and/or X_2 are continuous or binary.

A similar calculation shows that the effect on Y of a change ΔX_2 in X_2, holding X_1 constant, is $\Delta Y/\Delta X_2 = (\beta_2 + \beta_3 X_1)$.

Putting these two effects together shows that the coefficient β_3 on the interaction term is the effect of a unit increase in X_1 and X_2, above and beyond the sum of the effects of a unit increase in X_1 alone and a unit increase in X_2 alone. That is, if X_1 changes by ΔX_1 and X_2 changes by ΔX_2, then the expected change in Y is $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$ [Exercise 8.10(c)]. The first term is the effect from changing X_1 holding X_2 constant; the second term is the effect from changing X_2 holding X_1 constant; and the final term, $\beta_3 \Delta X_1 \Delta X_2$, is the extra effect from changing both X_1 and X_2.

Interactions between two variables are summarized as Key Concept 8.5.

When interactions are combined with logarithmic transformations, they can be used to estimate price elasticities when the price elasticity depends on the characteristics of the good (see the box "The Demand for Economics Journals" on page 290 for an example).

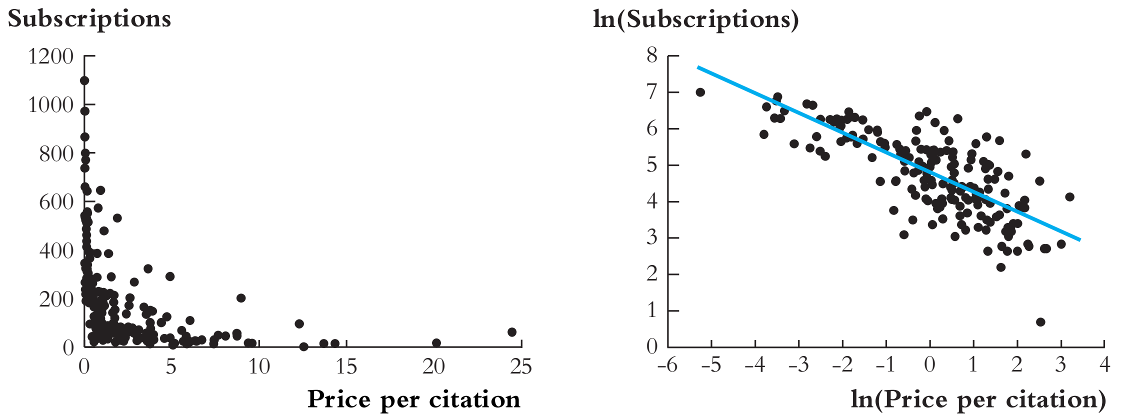
Application to the student–teacher ratio and the percentage of English learners. The previous examples considered interactions between the student–teacher ratio and a binary variable indicating whether the percentage of English learners is large or small. A different way to study this interaction is to examine the interaction between the student–teacher ratio and the continuous variable,

The Demand for Economics Journals

Professional economists follow the most recent research in their areas of specialization. Most research in economics first appears in economics journals, so economists—or their libraries—subscribe to economics journals.

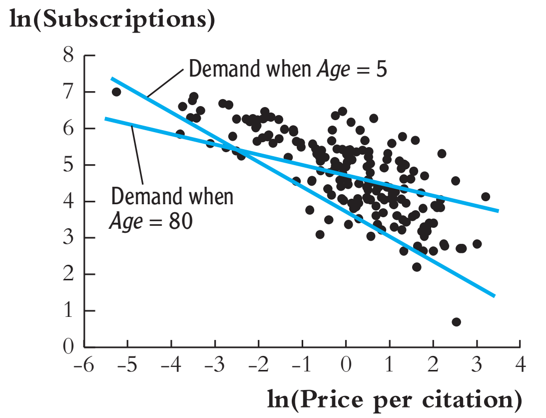
How elastic is the demand by libraries for economics journals? To find out, we analyzed the relationship between the number of subscriptions to a journal at U.S. libraries (Y_i) and the journal's library subscription price using data for the year 2000 for 180 economics journals. Because the product of a journal is not the paper on which it is printed but rather the ideas it contains, its price is logically measured not in dollars per year or dollars per page but instead in dollars per idea. Although we cannot measure "ideas" directly, a good indirect measure is the number of times that articles in a journal are subsequently cited by other researchers. Accordingly, we measure price

FIGURE 8.9 Library Subscriptions and Prices of Economics Journals



(a) Subscriptions and Price per citation

(b) ln(Subscriptions) and ln(Price per citation)



(c) ln(Subscriptions) and ln(Price per citation)

There is a nonlinear inverse relation between the number of U.S. library subscriptions (quantity) and the library price per citation (price), as shown in Figure 8.9a for 180 economics journals in 2000. But as seen in Figure 8.9b, the relation between log quantity and log price appears to be approximately linear. Figure 8.9c shows that demand is more elastic for young journals (Age $=$ 5) than for old journals (Age $=$ 80).

continued on next page

as the "price per citation" in the journal. The price range is enormous, from $\frac{1}{2}$¢ per citation (the *American Economic Review*) to 20¢ per citation or more. Some journals are expensive per citation because they have few citations, others because their library subscription price per year is very high. In 2014, a library print subscription to the *Journal of Econometrics* cost \$4089, compared to only \$455 for a bundled subscription to all seven journals published by the American Economics Association, including the *American Economic Review*!

Because we are interested in estimating elasticities, we use a log-log specification (Key Concept 8.2). The scatterplots in Figures 8.9a and 8.9b provide empirical support for this transformation. Because some of the oldest and most prestigious journals are the cheapest per citation, a regression of log quantity against log price could have omitted variable bias.

TABLE 8.2 Estimates of the Demand for Economics Journals

Dependent variable: logarithm of subscriptions at U.S. libraries in the Year 2000; 180 observations.

Regressor	(1)	(2)	(3)	(4)
ln(*Price per citation*)	−0.533** (0.034)	−0.408** (0.044)	−0.961** (0.160)	−0.899** (0.145)
[ln(*Price per citation*)]^2			0.017 (0.025)	
[ln(*Price per citation*)]^3			0.0037 (0.0055)	
ln(*Age*)		0.424** (0.119)	0.373** (0.118)	0.374** (0.118)
ln(*Age*) × ln(*Price per citation*)			0.156** (0.052)	0.141** (0.040)
ln(*Characters* ÷ 1,000,000)		0.206* (0.098)	0.235* (0.098)	0.229* (0.096)
Intercept	4.77** (0.055)	3.21** (0.38)	3.41** (0.38)	3.43** (0.38)
***F* Statistics and Summary Statistics**				
F-statistic testing coefficients on quadratic and cubic terms (*p*-value)			0.25 (0.779)	
SER	0.750	0.705	0.691	0.688
$\bar{R}^2$	0.555	0.607	0.622	0.626

The *F*-statistic tests the hypothesis that the coefficients on [ln(*Price per citation*)]^2 and [ln(*Price per citation*)]^3 are both zero. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% level or **1% level.

continued on next page

Our regressions therefore include two control variables: the logarithm of age and the logarithm of the number of characters per year in the journal.

The regression results are summarized in Table 8.2. Those results yield the following conclusions (see if you can find the basis for these conclusions in the table!):

1. Demand is less elastic for older than for newer journals.
2. The evidence supports a linear, rather than a cubic, function of log price.
3. Demand is greater for journals with more characters, holding price and age constant.

So what is the elasticity of demand for economics journals? It depends on the age of the journal. Demand curves for an 80-year-old journal and a 5-year-old upstart are superimposed on the scatterplot in Figure 8.9c; the older journal's demand elasticity is -0.28 ($SE = 0.06$), while the younger journal's is -0.67($SE = 0.08$).

This demand is very inelastic: Demand is very insensitive to price, especially for older journals. For libraries, having the most recent research on hand is a necessity, not a luxury. By way of comparison, experts estimate the demand elasticity for cigarettes to be in the range of -0.3 to -0.5. Economics journals are, it seems, as addictive as cigarettes, but a lot better for your health![1]

[1]These data were graciously provided by Professor Theodore Bergstrom of the Department of Economics at the University of California, Santa Barbara. If you are interested in learning more about the economics of economics journals, see Bergstrom (2001).

the percentage of English learners ($PctEL$). The estimated interaction regression is

$$\begin{aligned}\widehat{TestScore} = \underset{(11.8)}{686.3} - \underset{(0.59)}{1.12}STR - \underset{(0.37)}{0.67}PctEL + \underset{(0.019)}{0.0012}(STR \times PctEL),\\ \overline{R}^2 = 0.422. \qquad (8.37)\end{aligned}$$

When the percentage of English learners is at the median ($PctEL = 8.85$), the slope of the line relating test scores and the student–teacher ratio is estimated to be -1.11 ($= -1.12 + 0.0012 \times 8.85$). When the percentage of English learners is at the 75th percentile ($PctEL = 23.0$), this line is estimated to be flatter, with a slope of -1.09 ($= -1.12 + 0.0012 \times 23.0$). That is, for a district with 8.85% English learners, the estimated effect of a one-unit reduction in the student–teacher ratio is to increase test scores by 1.11 points, but for a district with 23.0% English learners, reducing the student–teacher ratio by one unit is predicted to increase test scores by only 1.09 points. The difference between these estimated effects is not statistically significant, however: The t-statistic testing whether the coefficient on the interaction term is zero is $t = 0.0012/0.019 = 0.06$, which is not significant at the 10% level.

To keep the discussion focused on nonlinear models, the specifications in Sections 8.1 through 8.3 exclude additional control variables such as the students' economic background. Consequently, these results arguably are subject to omitted variable bias. To draw substantive conclusions about the effect on test scores of reducing the student–teacher ratio, these nonlinear specifications must be augmented with control variables, and it is to such an exercise that we now turn.

8.4 Nonlinear Effects on Test Scores of the Student–Teacher Ratio

This section addresses three specific questions about test scores and the student–teacher ratio. First, after controlling for differences in economic characteristics of different districts, does the effect on test scores of reducing the student–teacher ratio depend on the fraction of English learners? Second, does this effect depend on the value of the student–teacher ratio? Third, and most important, after taking economic factors and nonlinearities into account, what is the estimated effect on test scores of reducing the student–teacher ratio by two students per teacher, as our superintendent from Chapter 4 proposes to do?

We answer these questions by considering nonlinear regression specifications of the type discussed in Sections 8.2 and 8.3, extended to include two measures of the economic background of the students: the percentage of students eligible for a subsidized lunch and the logarithm of average district income. The logarithm of income is used because the empirical analysis of Section 8.2 suggests that this specification captures the nonlinear relationship between test scores and income. As in Section 7.6, we do not include expenditures per pupil as a regressor and in so doing we are considering the effect of decreasing the student–teacher ratio, allowing expenditures per pupil to increase (that is, we are not holding expenditures per pupil constant).

Discussion of Regression Results

The OLS regression results are summarized in Table 8.3. The columns labeled (1) through (7) each report separate regressions. The entries in the table are the coefficients, standard errors, certain F-statistics and their p-values, and summary statistics, as indicated by the description in each row.

The first column of regression results, labeled regression (1) in the table, is regression (3) in Table 7.1 repeated here for convenience. This regression does not

TABLE 8.3 Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (*STR*)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English learners ≥ 10%? (Binary, *HiEL*)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.2** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)
***F*-Statistics and *p*-Values on Joint Hypotheses**							
(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR, HiEL \times STR^2,$ $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\bar{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

control for income, so the first thing we do is check whether the results change substantially when log income is included as an additional economic control variable. The results are given in regression (2) in Table 8.3. The log of income is statistically significant at the 1% level and the coefficient on the student–teacher ratio becomes somewhat closer to zero, falling from -1.00 to -0.73, although it remains statistically significant at the 1% level. The change in the coefficient on STR is large enough between regressions (1) and (2) to warrant including the logarithm of income in the remaining regressions as a deterrent to omitted variable bias.

Regression (3) in Table 8.3 is the interacted regression in Equation (8.34) with the binary variable for a high or low percentage of English learners, but with no economic control variables. When the economic control variables (percentage eligible for subsidized lunch and log income) are added [regression (4) in the table], the coefficients change, but in neither case is the coefficient on the interaction term significant at the 5% level. Based on the evidence in regression (4), the hypothesis that the effect of STR is the same for districts with low and high percentages of English learners cannot be rejected at the 5% level (the t-statistic is $t = -0.58/0.50 = -1.16$).

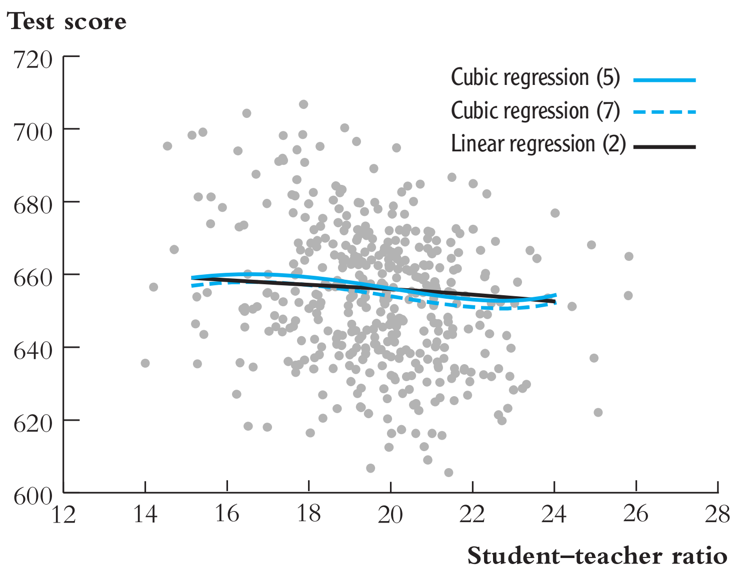
Regression (5) examines whether the effect of changing the student–teacher ratio depends on the value of the student–teacher ratio by including a cubic specification in STR in addition to the other control variables in regression (4) [the interaction term, $HiEL \times STR$, was dropped because it was not significant in regression (4) at the 10% level]. The estimates in regression (5) are consistent with the student–teacher ratio having a nonlinear effect. The null hypothesis that the relationship is linear is rejected at the 1% significance level against the alternative that it is cubic (the F-statistic testing the hypothesis that the true coefficients on STR^2 and STR^3 are zero is 6.17, with a p-value of < 0.001).

Regression (6) further examines whether the effect of the student–teacher ratio depends not just on the value of the student–teacher ratio but also on the fraction of English learners. By including interactions between $HiEL$ and STR, STR^2, and STR^3, we can check whether the (possibly cubic) population regressions functions relating test scores and STR are different for low and high percentages of English learners. To do so, we test the restriction that the coefficients on the three interaction terms are zero. The resulting F-statistic is 2.69, which has a p-value of 0.046 and thus is significant at the 5% but not the 1% significance level. This provides some evidence that the regression functions are different for districts with high and low percentages of English learners; however, comparing regressions (6) and (4) makes it clear that these differences are associated with the quadratic and cubic terms.

Regression (7) is a modification of regression (5), in which the continuous variable $PctEL$ is used instead of the binary variable $HiEL$ to control for the percentage of English learners in the district. The coefficients on the other regressors

FIGURE 8.10 Three Regression Functions Relating Test Scores and Student–Teacher Ratio

The cubic regressions from columns (5) and (7) of Table 8.3 are nearly identical. They indicate a small amount of nonlinearity in the relation between test scores and student–teacher ratio.



do not change substantially when this modification is made, indicating that the results in regression (5) are not sensitive to what measure of the percentage of English learners is actually used in the regression.

In all the specifications, the hypothesis that the student–teacher ratio does not enter the regressions is rejected at the 1% level.

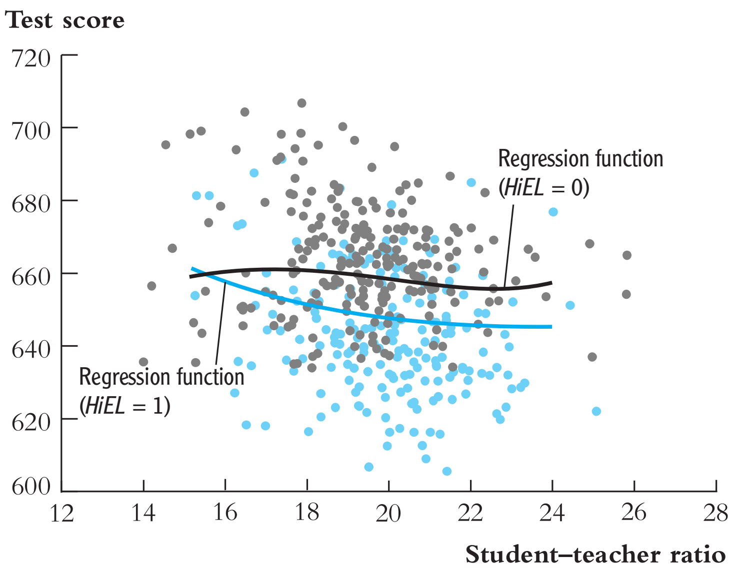
The nonlinear specifications in Table 8.3 are most easily interpreted graphically. Figure 8.10 graphs the estimated regression functions relating test scores and the student–teacher ratio for the linear specification (2) and the cubic specifications (5) and (7), along with a scatterplot of the data.[4] These estimated regression functions show the predicted value of test scores as a function of the student–teacher ratio, holding fixed other values of the independent variables in the regression. The estimated regression functions are all close to one another, although the cubic regressions flatten out for large values of the student–teacher ratio.

Regression (6) indicates a statistically significant difference in the cubic regression functions relating test scores and *STR*, depending on whether the percentage of English learners in the district is large or small. Figure 8.11 graphs these two estimated regression functions so that we can see whether this difference, in addition

[4]For each curve, the predicted value was computed by setting each independent variable, other than *STR*, to its sample average value and computing the predicted value by multiplying these fixed values of the independent variables by the respective estimated coefficients from Table 8.3. This was done for various values of *STR*, and the graph of the resulting adjusted predicted values is the estimated regression function relating test scores and the *STR*, holding the other variables constant at their sample averages.

FIGURE 8.11 Regression Functions for Districts with High and Low Percentages of English Learners

Districts with low percentages of English learners ($HiEL = 0$) are shown by gray dots, and districts with $HiEL = 1$ are shown by colored dots. The cubic regression function for $HiEL = 1$ from regression (6) in Table 8.3 is approximately 10 points below the cubic regression function for $HiEL = 0$ for $17 \leq STR \leq 23$, but otherwise the two functions have similar shapes and slopes in this range. The slopes of the regression functions differ most for very large and small values of STR, for which there are few observations.



to being statistically significant, is of practical importance. As Figure 8.11 shows, for student–teacher ratios between 17 and 23—a range that includes 88% of the observations—the two functions are separated by approximately 10 points but otherwise are very similar; that is, for STR between 17 and 23, districts with a lower percentage of English learners do better, holding constant the student–teacher ratio, but the effect of a change in the student–teacher ratio is essentially the same for the two groups. The two regression functions are different for student–teacher ratios below 16.5, but we must be careful not to read more into this than is justified. The districts with $STR < 16.5$ constitute only 6% of the observations, so the differences between the nonlinear regression functions are reflecting differences in these very few districts with very low student–teacher ratios. Thus, based on Figure 8.11, we conclude that the effect on test scores of a change in the student–teacher ratio does not depend on the percentage of English learners for the range of student–teacher ratios for which we have the most data.

Summary of Findings

These results let us answer the three questions raised at the start of this section.

First, after controlling for economic background, whether there are many or few English learners in the district does not have a substantial influence on the effect on test scores of a change in the student–teacher ratio. In the linear specifications, there is no statistically significant evidence of such a difference. The cubic specification in regression (6) provides statistically significant evidence (at the 5% level) that the regression functions are different for districts with high and low percentages of English learners; as shown in Figure 8.11, however, the estimated regression functions have similar slopes in the range of student–teacher ratios containing most of our data.

Second, after controlling for economic background, there is evidence of a nonlinear effect on test scores of the student–teacher ratio. This effect is statistically significant at the 1% level (the coefficients on STR^2 and STR^3 are always significant at the 1% level).

Third, we now can return to the superintendent's problem that opened Chapter 4. She wants to know the effect on test scores of reducing the student–teacher ratio by two students per teacher. In the linear specification (2), this effect does not depend on the student–teacher ratio itself, and the estimated effect of this reduction is to improve test scores by 1.46 ($= -0.73 \times -2$) points. In the nonlinear specifications, this effect depends on the value of the student–teacher ratio. If her district currently has a student–teacher ratio of 20 and she is considering cutting it to 18, then based on regression (5) the estimated effect of this reduction is to improve test scores by 3.00 points, while based on regression (7) this estimate is 2.93. If her district currently has a student–teacher ratio of 22 and she is considering cutting it to 20, then based on regression (5) the estimated effect of this reduction is to improve test scores by 1.93 points, while based on regression (7) this estimate is 1.90. The estimates from the nonlinear specifications suggest that cutting the student–teacher ratio has a greater effect if this ratio is already small.

8.5 Conclusion

This chapter presented several ways to model nonlinear regression functions. Because these models are variants of the multiple regression model, the unknown coefficients can be estimated by OLS, and hypotheses about their values can be tested using t- and F-statistics as described in Chapter 7. In these models, the expected effect on Y of a change in one of the independent variables, X_1, holding the other independent variables $X_2, \ldots, X_k$ constant in general depends on the values of $X_1, X_2, \ldots, X_k$.

There are many different models in this chapter, and you could not be blamed for being a bit bewildered about which to use in a given application. How should you analyze possible nonlinearities in practice? Section 8.1 laid out a general approach for such an analysis, but this approach requires you to make decisions and exercise judgment along the way. It would be convenient if there were a single recipe you could follow that would always work in every application, but in practice data analysis is rarely that simple.

The single most important step in specifying nonlinear regression functions is to "use your head." Before you look at the data, can you think of a reason, based on economic theory or expert judgment, why the slope of the population regression function might depend on the value of that, or another, independent variable? If so, what sort of dependence might you expect? And, most important, which nonlinearities (if any) could have major implications for the substantive issues addressed by your study? Answering these questions carefully will focus your analysis. In the test score application, for example, such reasoning led us to investigate whether hiring more teachers might have a greater effect in districts with a large percentage of students still learning English, perhaps because those students would differentially benefit from more personal attention. By making the question precise, we were able to find a precise answer: After controlling for the economic background of the students, we found no statistically significant evidence of such an interaction.

Summary

1. In a nonlinear regression, the slope of the population regression function depends on the value of one or more of the independent variables.
2. The effect on Y of a change in the independent variable(s) can be computed by evaluating the regression function at two values of the independent variable(s). The procedure is summarized in Key Concept 8.1.
3. A polynomial regression includes powers of X as regressors. A quadratic regression includes X and X^2, and a cubic regression includes X, X^2, and X^3.
4. Small changes in logarithms can be interpreted as proportional or percentage changes in a variable. Regressions involving logarithms are used to estimate proportional changes and elasticities.
5. The product of two variables is called an interaction term. When interaction terms are included as regressors, they allow the regression slope of one variable to depend on the value of another variable.

Key Terms

quadratic regression model (259)
nonlinear regression function (262)
polynomial regression model (267)
cubic regression model (267)
elasticity (269)
exponential function (269)
natural logarithm (270)
linear-log model (271)
log-linear model (272)
log-log model (274)
interaction term (280)
interacted regressor (280)
interaction regression model (280)
nonlinear least squares (311)
nonlinear least squares estimators (311)

Review the Concepts

8.1 Sketch a regression function that is increasing (has a positive slope) and is steep for small values of X but less steep for large values of X. Explain how you would specify a nonlinear regression to model this shape. Can you think of an economic relationship with a shape like this?

8.2 A "Cobb–Douglas" production function relates production (Q) to factors of production, capital (K), labor (L), and raw materials (M), and an error term u using the equation $Q = \lambda K^{\beta_1} L^{\beta_2} M^{\beta_3} e^{u}$, where λ, β_1, β_2, and β_3 are production parameters. Suppose that you have data on production and the factors of production from a random sample of firms with the same Cobb–Douglas production function. How would you use regression analysis to estimate the production parameters?

8.3 Can you use $\overline{R}^2$ to compare the fit of a log-log and log-linear regression? Why? Can you use $\overline{R}^2$ to compare the fit of a log-log and linear-log regression? Why?

8.4 Suppose the regression in Equation (8.30) is estimated using $LoSTR$ and $LoEL$ in place of $HiSTR$ and $HiEL$, where $LoSTR = 1 - HiSTR$ is an indicator for a low-class-size district and $LoEL = 1 - HiEL$ is an indicator for a district with a low percentage of English learners. What are the values of the estimated regression coefficients?

8.5 Suppose that in Exercise 8.2 you thought that the value of β_2 was not constant but rather increased when K increased. How could you use an interaction term to capture this effect?

8.6 You have estimated a linear regression model relating Y to X. Your professor says, "I think that the relationship between Y and X is nonlinear." Explain how you would test the adequacy of your linear regression.

Exercises

8.1 Sales in a company are \$196 million in 2013 and increase to \$198 million in 2014.

a. Compute the percentage increase in sales, using the usual formula $100 \times \frac{(Sales_{2014} - Sales_{2013})}{Sales_{2013}}$. Compare this value to the approximation $100 \times [\ln(Sales_{2014}) - \ln(Sales_{2013})]$.

b. Repeat (a), assuming that $Sales_{2014} = 205$, $Sales_{2014} = 250$, and $Sales_{2014} = 500$.

c. How good is the approximation when the change is small? Does the quality of the approximation deteriorate as the percentage change increases?

8.2 Suppose that a researcher collects data on houses that have sold in a particular neighborhood over the past year and obtains the regression results in the table shown below.

a. Using the results in column (1), what is the expected change in price of building a 500-square-foot addition to a house? Construct a 95% confidence interval for the percentage change in price.

b. Comparing columns (1) and (2), is it better to use *Size* or ln(*Size*) to explain house prices?

c. Using column (2), what is the estimated effect of pool on price? (Make sure you get the units right.) Construct a 95% confidence interval for this effect.

d. The regression in column (3) adds the number of bedrooms to the regression. How large is the estimated effect of an additional bedroom? Is the effect statistically significant? Why do you think the estimated effect is so small? (*Hint:* Which other variables are being held constant?)

e. Is the quadratic term $\ln(Size)^2$ important?

f. Use the regression in column (5) to compute the expected change in price when a pool is added to a house that doesn't have a view. Repeat the exercise for a house that has a view. Is there a large difference? Is the difference statistically significant?

Regression Results for Exercise 8.2

Dependent variable: ln(*Price*)

Regressor	(1)	(2)	(3)	(4)	(5)
Size	0.00042 (0.000038)				
ln(*Size*)		0.69 (0.054)	0.68 (0.087)	0.57 (2.03)	0.69 (0.055)
ln(*Size*)^2				0.0078 (0.14)	
Bedrooms			0.0036 (0.037)		
Pool	0.082 (0.032)	0.071 (0.034)	0.071 (0.034)	0.071 (0.036)	0.071 (0.035)
View	0.037 (0.029)	0.027 (0.028)	0.026 (0.026)	0.027 (0.029)	0.027 (0.030)
Pool × *View*					0.0022 (0.10)
Condition	0.13 (0.045)	0.12 (0.035)	0.12 (0.035)	0.12 (0.036)	0.12 (0.035)
Intercept	10.97 (0.069)	6.60 (0.39)	6.63 (0.53)	7.02 (7.50)	6.60 (0.40)
Summary Statistics					
SER	0.102	0.098	0.099	0.099	0.099
$\bar{R}^2$	0.72	0.74	0.73	0.73	0.73

Variable definitions: *Price* = sale price ($); *Size* = house size (in square feet); *Bedrooms* = number of bedrooms; *Pool* = binary variable (1 if house has a swimming pool, 0 otherwise); *View* = binary variable (1 if house has a nice view, 0 otherwise); *Condition* = binary variable (1 if real estate agent reports house is in excellent condition, 0 otherwise).

8.3 After reading this chapter's analysis of test scores and class size, an educator comments, "In my experience, student performance depends on class size, but not in the way your regressions say. Rather, students do well when class size is less than 20 students and do very poorly when class size is greater than 25. There are no gains from reducing class size below 20 students, the relationship is constant in the intermediate region between 20 and 25 students, and there is no loss to increasing class size when it is already greater than 25." The educator is describing a "threshold effect" in which performance is constant for class sizes less than 20, then jumps and is constant for class sizes between 20 and 25, and then jumps again for class sizes greater than 25. To model these threshold effects, define the binary variables

$$STRsmall = 1 \text{ if } STR < 20, \text{ and } STRsmall = 0 \text{ otherwise;}$$

$$STRmoderate = 1 \text{ if } 20 \leq STR \leq 25, \text{ and } STRmoderate = 0 \text{ otherwise; and}$$

$$STRlarge = 1 \text{ if } STR > 25, \text{ and } STRlarge = 0 \text{ otherwise.}$$

a. Consider the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRlarge_i + u_i$. Sketch the regression function relating *TestScore* to *STR* for hypothetical values of the regression coefficients that are consistent with the educator's statement.

b. A researcher tries to estimate the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRmoderate_i + \beta_3 STRlarge_i + u_i$ and finds that the software gives an error message. Why?

8.4 Read the box "The Return to Education and the Gender Gap" in Section 8.3.

a. Consider a man with 16 years of education and 2 years of experience who is from a western state. Use the results from column (4) of Table 8.1 and the method in Key Concept 8.1 to estimate the expected change in the logarithm of average hourly earnings (AHE) associated with an additional year of experience.

b. Repeat (a), assuming 10 years of experience.

c. Explain why the answers to (a) and (b) are different.

d. Is the difference in the answers to (a) and (b) statistically significant at the 5% level? Explain.

e. Would your answers to (a) through (d) change if the person were a woman? If the person were from the South? Explain.

f. How would you change the regression if you suspected that the effect of experience on earnings was different for men than for women?

8.5 Read the box "The Demand for Economics Journals" in Section 8.3.

a. The box reaches three conclusions. Looking at the results in the table, what is the basis for each of these conclusions?

b. Using the results in regression (4), the box reports that the elasticity of demand for an 80-year-old journal is −0.28.

i. How was this value determined from the estimated regression?

ii. The box reports that the standard error for the estimated elasticity is 0.06. How would you calculate this standard error?

(*Hint:* See the discussion "Standard errors of estimated effects" on page 264.)

c. Suppose that the variable *Characters* had been divided by 1000 instead of 1,000,000. How would the results in column (4) change?

8.6 Refer to Table 8.3.

a. A researcher suspects that the effect of *%Eligible for subsidized lunch* has a nonlinear effect on test scores. In particular, he conjectures that increases in this variable from 10% to 20% have little effect on test scores but that changes from 50% to 60% have a much larger effect.

i. Describe a nonlinear specification that can be used to model this form of nonlinearity.

ii. How would you test whether the researcher's conjecture was better than the linear specification in column (7) of Table 8.3?

b. A researcher suspects that the effect of income on test scores is different in districts with small classes than in districts with large classes.

i. Describe a nonlinear specification that can be used to model this form of nonlinearity.

ii. How would you test whether the researcher's conjecture was better than the linear specification in column (7) of Table 8.3?

8.7 This problem is inspired by a study of the "gender gap" in earnings in top corporate jobs [Bertrand and Hallock (2001)]. The study compares total compensation among top executives in a large set of U.S. public corporations in the 1990s. (Each year these publicly traded corporations must report total compensation levels for their top five executives.)

a. Let *Female* be an indicator variable that is equal to 1 for females and 0 for males. A regression of the logarithm of earnings onto *Female* yields

$$\widehat{\ln(Earnings)} = \underset{(0.01)}{6.48} - \underset{(0.05)}{0.44}\,Female,\ SER = 2.65.$$

i. The estimated coefficient on *Female* is −0.44. Explain what this value means.

ii. The *SER* is 2.65. Explain what this value means.

iii. Does this regression suggest that female top executives earn less than top male executives? Explain.

iv. Does this regression suggest that there is gender discrimination? Explain.

b. Two new variables, the market value of the firm (a measure of firm size, in millions of dollars) and stock return (a measure of firm performance, in percentage points), are added to the regression:

$$\widehat{\ln(Earnings)} = \underset{(0.03)}{3.86} - \underset{(0.04)}{0.28}\,Female + \underset{(0.004)}{0.37}\ln(MarketValue) + \underset{(0.003)}{0.004}\,Return,$$
$$n = 46{,}670,\ \overline{R}^2 = 0.345.$$

i. The coefficient on ln(*MarketValue*) is 0.37. Explain what this value means.

ii. The coefficient on *Female* is now −0.28. Explain why it has changed from the regression in (a).

c. Are large firms more likely than small firms to have female top executives? Explain.

8.8 X is a continuous variable that takes on values between 5 and 100. Z is a binary variable. Sketch the following regression functions (with values of X between 5 and 100 on the horizontal axis and values of $\hat{Y}$ on the vertical axis):

a. $\hat{Y} = 2.0 + 3.0 \times \ln(X)$.

b. $\hat{Y} = 2.0 - 3.0 \times \ln(X)$.

c. i. $\hat{Y} = 2.0 + 3.0 \times \ln(X) + 4.0Z$, with $Z = 1$.

ii. Same as (i), but with $Z = 0$.

d. i. $\hat{Y} = 2.0 + 3.0 \times \ln(X) + 4.0Z - 1.0 \times Z \times \ln(X)$, with $Z = 1$.

ii. Same as (i), but with $Z = 0$.

e. $\hat{Y} = 1.0 + 125.0X - 0.01X^2$.

8.9 Explain how you would use Approach #2 from Section 7.3 to calculate the confidence interval discussed below Equation (8.8). [*Hint:* This requires estimating a new regression using a different definition of the regressors and the dependent variable. See Exercise (7.9).]

8.10 Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$. Use Key Concept 8.1 to show:

a. $\Delta Y/\Delta X_1 = \beta_1 + \beta_3 X_2$ (effect of change in X_1, holding X_2 constant).

b. $\Delta Y/\Delta X_2 = \beta_2 + \beta_3 X_1$ (effect of change in X_2, holding X_1 constant).

c. If X_1 changes by ΔX_1 and X_2 changes by ΔX_2, then $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$.

8.11 Derive the expressions for the elasticities given in Appendix 8.2 for the linear and log-log models. (*Hint:* For the log-log model, assume that u

and X are independent, as is done in Appendix 8.2 for the log-linear model.)

8.12 The discussion following Equation (8.28) interprets the coefficient on interacted binary variables using the conditional mean zero assumption. This exercise shows that interpretation also applies under conditional mean independence. Consider the hypothetical experiment in Exercise 7.11.

a. Suppose that you estimate the regression $Y_i = \gamma_0 + \gamma_1 X_{1i} + u_i$ using only the data on returning students. Show that γ_1 is the class size effect for returning students—that is, that $\gamma_1 = E(Y_i | X_{1i} = 1, X_{2i} = 0) - E(Y_i | X_{1i} = 0, X_{2i} = 0)$. Explain why $\hat{\gamma}_1$ is an unbiased estimator of γ_1.

b. Suppose that you estimate the regression $Y_i = \delta_0 + \delta_1 X_{1i} + u_i$ using only the data on new students. Show that δ_1 is the class size effect for new students—that is, that $\delta_1 = E(Y_i | X_{1i} = 1, X_{2i} = 1) - E(Y_i | X_{1i} = 0, X_{2i} = 1)$. Explain why $\hat{\delta}_1$ is an unbiased estimator of δ_1.

c. Consider the regression for both returning and new students, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$. Use the conditional mean independence assumption $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$ to show that $\beta_1 = \gamma_1$, $\beta_1 + \beta_3 = \delta_1$, and $\beta_3 = \delta_1 - \gamma_1$ (the difference in the class size effects).

d. Suppose that you estimate the interaction regression in (c) using the combined data and that $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$. Show that $\hat{\beta}_1$ and $\hat{\beta}_3$ are unbiased but that $\hat{\beta}_2$ is in general biased.

Empirical Exercises

(Only two empirical exercises for this chapter are given in the text, but you can find more on the text website **http://www.pearsonhighered.com/stock_watson/**.)

E8.1 Lead is toxic, particularly for young children, and for this reason government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leached into drinking water. In this exercise you will investigate the effect of these lead water pipes on infant mortality. On the text website **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **Lead_Mortality**, which contains data on infant mortality, type of water pipes (lead or non-lead), water acidity (pH), and several demographic variables for 172 U.S. cities in 1900.[5] A detailed description is given in **Lead_Mortality_Description**, also available on the website.

a. Compute the average infant mortality rate (*Inf*) for cities with lead pipes and for cities with non-lead pipes. Is there a statistically significant difference in the averages?

b. The amount of lead leached from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water (that is, the lower its pH), the more lead is leached. Run a regression of *Inf* on *Lead*, *pH*, and the interaction term $Lead \times pH$.

i. The regression includes four coefficients (the intercept and the three coefficients multiplying the regressors). Explain what each coefficient measures.

ii. Plot the estimated regression function relating *Inf* to *pH* for $Lead = 0$ and for $Lead = 1$. Describe the differences in the regression functions and relate these differences to the coefficients you discussed in (i).

iii. Does *Lead* have a statistically significant effect on infant mortality? Explain.

iv. Does the effect of *Lead* on infant mortality depend on *pH*? Is this dependence statistically significant?

v. What is the average value of *pH* in the sample? At this *pH* level, what is the estimated effect of *Lead* on infant mortality? What is the standard deviation of *pH*? Suppose that the *pH* level is one standard deviation lower than the average level of *pH* in the sample; what is the estimated effect of *Lead* on infant mortality? What if *pH* is one standard deviation higher than the average value?

vi. Construct a 95% confidence interval for the effect of *Lead* on infant mortality when $pH = 6.5$.

c. The analysis in (b) may suffer from omitted variable bias because it neglects factors that affect infant mortality and that might potentially be correlated with *Lead* and *pH*. Investigate this concern, using the other variables in the data set.

[5]These data were provided by Professor Karen Clay of Carnegie Mellon University and were used in her paper with Werner Troesken and Michael Haines, "Lead and Mortality," *The Review of Economics and Statistics,* 2014, 96(3).

E8.2 On the text website **http://www.pearsonhighered.com/stock_watson/** you will find a data file **CPS12**, which contains data for full-time, full-year workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS12_Description**, also available on the website. (These are the same data as in **CPS92_12**, used in Empirical Exercise 3.1, but are limited to the year 2012.) In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and higher earnings.)

a. Run a regression of average hourly earnings (*AHE*) on age (*Age*), gender (*Female*), and education (*Bachelor*). If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

b. Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on *Age*, *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

c. Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on ln(*Age*), *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

d. Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on *Age*, Age^2, *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

e. Do you prefer the regression in (c) to the regression in (b)? Explain.

f. Do you prefer the regression in (d) to the regression in (b)? Explain.

g. Do you prefer the regression in (d) to the regression in (c)? Explain.

h. Plot the regression relation between *Age* and ln(*AHE*) from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

i. Run a regression of ln(*AHE*) on *Age*, Age^2, *Female*, *Bachelor*, and the interaction term *Female* × *Bachelor*. What does the coefficient on the interaction term measure? Alexis is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of $\ln(AHE)$? Jane is a 30-year-old female with a high school degree. What does the regression predict for her value of $\ln(AHE)$? What is the predicted difference between Alexis's and Jane's earnings? Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of $\ln(AHE)$? Jim is a 30-year-old male with a high school degree. What does the regression predict for his value of $\ln(AHE)$? What is the predicted difference between Bob's and Jim's earnings?

j. Is the effect of *Age* on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.

k. Is the effect of *Age* on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.

l. After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.

APPENDIX

8.1 Regression Functions That Are Nonlinear in the Parameters

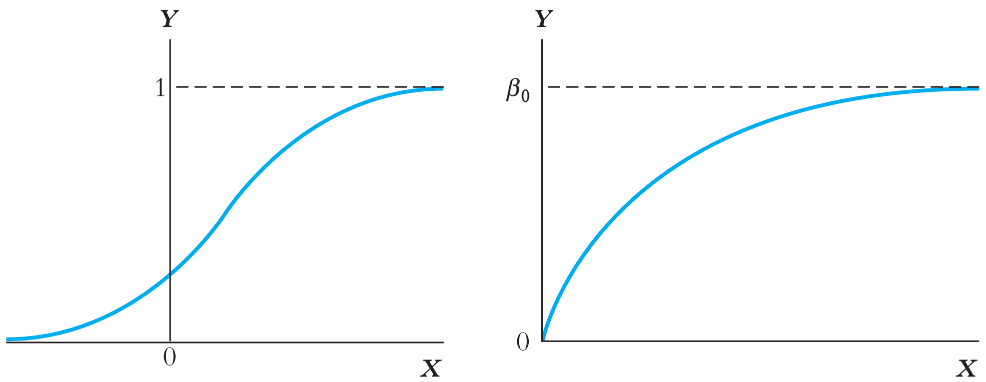
The nonlinear regression functions considered in Sections 8.2 and 8.3 are nonlinear functions of the X's but are linear functions of the unknown parameters. Because they are linear in the unknown parameters, those parameters can be estimated by OLS after defining new regressors that are nonlinear transformations of the original X's. This family of nonlinear regression functions is both rich and convenient to use. In some applications, however, economic reasoning leads to regression functions that are not linear in the parameters. Although such regression functions cannot be estimated by OLS, they can be estimated using an extension of OLS called nonlinear least squares.

Functions That Are Nonlinear in the Parameters

We begin with two examples of functions that are nonlinear in the parameters. We then provide a general formulation.

Logistic curve. Suppose that you are studying the market penetration of a technology, such as the adoption of database management software in different industries. The dependent variable is the fraction of firms in the industry that have adopted the software, a single

FIGURE 8.12 Two Functions That Are Nonlinear in Their Parameters



(a) A logistic curve **(b)** A negative exponential growth curve

Part (a) plots the logistic function of Equation (8.38), which has predicted values that lie between 0 and 1. Part (b) plots the negative exponential growth function of Equation (8.39), which has a slope that is always positive and decreases as X increases, and an asymptote at β_0 as X tends to infinity.

independent variable X describes an industry characteristic, and you have data on n industries. The dependent variable is between 0 (no adopters) and 1 (100% adoption). Because a linear regression model could produce predicted values less than 0 or greater than 1, it makes sense to use instead a function that produces predicted values between 0 and 1.

The logistic function smoothly increases from a minimum of 0 to a maximum of 1. The logistic regression model with a single X is

$$Y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} + u_i. \tag{8.38}$$

The logistic function with a single X is graphed in Figure 8.12a. As can be seen in the graph, the logistic function has an elongated "S" shape. For small values of X, the value of the function is nearly 0 and the slope is flat; the curve is steeper for moderate values of X; and for large values of X, the function approaches 1 and the slope is flat again.

Negative exponential growth. The functions used in Section 8.2 to model the relation between test scores and income have some deficiencies. For example, the polynomial models can produce a negative slope for some values of income, which is implausible. The logarithmic specification has a positive slope for all values of income; however, as income gets very large, the predicted values increase without bound, so for some incomes the predicted value for a district will exceed the maximum possible score on the test.

The negative exponential growth model provides a nonlinear specification that has a positive slope for all values of income, has a slope that is greatest at low values of income and decreases as income rises, and has an upper bound (that is, an asymptote as income increases to infinity). The negative exponential growth regression model is

$$Y_i = \beta_0[1 - e^{-\beta_1(X_i - \beta_2)}] + u_i. \tag{8.39}$$

The negative exponential growth function is graphed in Figure 8.12b. The slope is steep for low values of X, but as X increases, it reaches an asymptote of β_0.

General functions that are nonlinear in the parameters. The logistic and negative exponential growth regression models are special cases of the general nonlinear regression model

$$Y_i = f(X_{1i}, \ldots, X_{ki}; \beta_0, \ldots, \beta_m) + u_i, \tag{8.40}$$

in which there are k independent variables and $m + 1$ parameters, $\beta_0, \ldots, \beta_m$. In the models of Sections 8.2 and 8.3, the X's entered this function nonlinearly, but the parameters entered linearly. In the examples of this appendix, the parameters enter nonlinearly as well. If the parameters are known, then predicted effects can be computed using the method described in Section 8.1. In applications, however, the parameters are unknown and must be estimated from the data. Parameters that enter nonlinearly cannot be estimated by OLS, but they can be estimated by nonlinear least squares.

Nonlinear Least Squares Estimation

Nonlinear least squares is a general method for estimating the unknown parameters of a regression function when those parameters enter the population regression function nonlinearly.

Recall the discussion in Section 5.3 of the OLS estimator of the coefficients of the linear multiple regression model. The OLS estimator minimizes the sum of squared prediction mistakes in Equation (5.8), $\sum_{i=1}^{n}[Y_i - (b_0 + b_1X_{1i} + \cdots + b_kX_{ki})]^2$. In principle, the OLS estimator can be computed by checking many trial values of $b_0, \ldots, b_k$ and settling on the values that minimize the sum of squared mistakes.

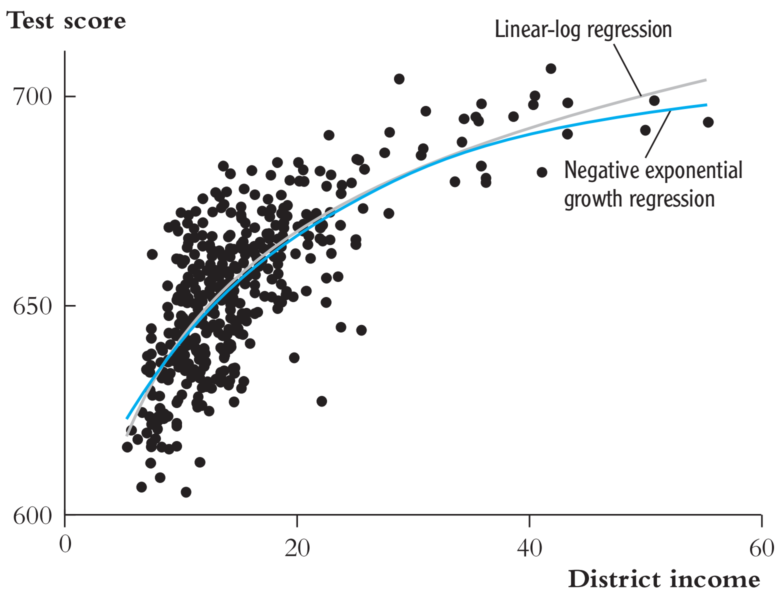
This same approach can be used to estimate the parameters of the general nonlinear regression model in Equation (8.40). Because the regression function is nonlinear in the coefficients, this method is called **nonlinear least squares**. For a set of trial parameter values $b_0, b_1, \ldots, b_m$ construct the sum of squared prediction mistakes:

$$\sum_{i=1}^{n}[Y_i - f(X_{1i}, \ldots, X_{ki}, b_1, \ldots, b_m)]^2. \tag{8.41}$$

The **nonlinear least squares estimators** of $\beta_0, \beta_1, \ldots, \beta_m$ are the values of $b_0, b_1, \ldots, b_m$ that minimize the sum of squared prediction mistakes in Equation (8.41).

FIGURE 8.13 The Negative Exponential Growth and Linear-Log Regression Functions

The negative exponential growth regression function [Equation (8.42)] and the linear-log regression function [Equation (8.18)] both capture the nonlinear relation between test scores and district income. One difference between the two functions is that the negative exponential growth model has an asymptote as *Income* increases to infinity, but the linear-log regression function does not.



In linear regression, a relatively simple formula expresses the OLS estimator as a function of the data. Unfortunately, no such general formula exists for nonlinear least squares, so the nonlinear least squares estimator must be found numerically using a computer. Regression software incorporates algorithms for solving the nonlinear least squares minimization problem, which simplifies the task of computing the nonlinear least squares estimator in practice.

Under general conditions on the function f and the X's, the nonlinear least squares estimator shares two key properties with the OLS estimator in the linear regression model: It is consistent, and it is normally distributed in large samples. In regression software that supports nonlinear least squares estimation, the output typically reports standard errors for the estimated parameters. As a consequence, inference concerning the parameters can proceed as usual; in particular, t-statistics can be constructed using the general approach in Key Concept 5.1, and a 95% confidence interval can be constructed as the estimated coefficient, plus or minus 1.96 standard errors. Just as in linear regression, the error term in the nonlinear regression model can be heteroskedastic, so heteroskedasticity-robust standard errors should be used.

Application to the Test Score–Income Relation

A negative exponential growth model, fit to district income (X) and test scores (Y), has the desirable features of a slope that is always positive [if β_1 in Equation (8.39) is positive] and an asymptote of β_0 as income increases to infinity. The result of estimating β_0, β_1, and β_2 in

Equation (8.39) using the California test score data yields $\hat{\beta}_0 = 703.2$ (heteroskedasticity-robust standard error $= 4.44$), $\hat{\beta}_1 = 0.0552$ ($SE = 0.0068$), and $\hat{\beta}_2 = -34.0$ ($SE = 4.48$). Thus the estimated nonlinear regression function (with standard errors reported below the parameter estimates) is

$$\underset{}{\widehat{TestScore}} = \underset{(4.44)}{703.2}[1 - e^{-\underset{(0.0068)}{0.0552}(Income + \underset{(4.48)}{34.0})}]. \tag{8.42}$$

This estimated regression function is plotted in Figure 8.13, along with the logarithmic regression function and a scatterplot of the data. The two specifications are, in this case, quite similar. One difference is that the negative exponential growth curve flattens out at the highest levels of income, consistent with having an asymptote.

APPENDIX

8.2 Slopes and Elasticities for Nonlinear Regression Functions

This appendix uses calculus to evaluate slopes and elasticities of nonlinear regression functions with continuous regressors. We focus on the case of Section 8.2, in which there is a single X. This approach extends to multiple X's, using partial derivatives.

Consider the nonlinear regression model, $Y_i = f(X_i) + u_i$, with $E(u_i|X_i) = 0$. The slope of the population regression function, $f(X)$, evaluated at the point $X = x$, is the derivative of f, that is, $df(X)/dX|_{X=x}$. For the polynomial regression function in Equation (8.9), $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_r X^r$ and $dX^a/dX = aX^{a-1}$ for any constant a, so $df(X)/dX|_{X=x} = \beta_1 + 2\beta_2 x + \cdots + r\beta_r x^{r-1}$. The estimated slope at x is $d\hat{f}(X)/dX|_{X=x} = \hat{\beta}_1 + 2\hat{\beta}_2 x + \cdots + r\hat{\beta}_r x^{r-1}$. The standard error of the estimated slope is $SE(\hat{\beta}_1 + 2\hat{\beta}_2 x + \cdots + r\hat{\beta}_r x^{r-1})$; for a given value of x, this is the standard error of a weighted sum of regression coefficients, which can be computed using the methods of Section 7.3 and Equation (8.8).

The elasticity of Y with respect to X is the percentage change in Y for a given percentage change in X. Formally, this definition applies in the limit that the percentage change in X goes to zero, so the slope appearing in the definition in Equation (8.22) is replaced by the derivative and the elasticity is

$$\text{elasticity of } Y \text{ with respect to } X = \frac{dY}{dX} \times \frac{X}{Y} = \frac{d \ln Y}{d \ln X}.$$

In a regression model, Y depends both on X and on the error term u. Because u is random, it is conventional to evaluate the elasticity as the percentage change not of Y but of the predicted component of Y—that is, the percentage change in $E(Y|X)$. Accordingly, the elasticity of $E(Y|X)$ with respect to X is

$$\frac{dE(Y|X)}{dX} \times \frac{X}{E(Y|X)} = \frac{d\ \ln E(Y|X)}{d \ln X}.$$

The elasticities for the linear model and for the three logarithmic models summarized in Key Concept 8.2 are given in the table below.

Case	Population Regression Model	Elasticity of $E(Y\|X)$ with Respect to X
linear	$Y = \beta_0 + \beta_1 X + u$	$\frac{\beta_1 X}{\beta_0 + \beta_1 X}$
linear-log	$Y = \beta_0 + \beta_1 \ln(X) + u$	$\frac{\beta_1}{\beta_0 + \beta_1 \ln(X)}$
log-linear	$\ln(Y) = \beta_0 + \beta_1 X + u$	$\beta_1 X$
log-log	$\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$	β_1

The log-log specification has a constant elasticity, but in the other three specifications, the elasticity depends on X.

We now derive the expressions for the linear-log and log-linear models. For the linear-log model, $E(Y|X) = \beta_0 + \beta_1 \ln(X)$. Because $d\ln(X)/dX = 1/X$, applying the chain rule yields $dE(Y|X)/dX = \beta_1/X$. Thus the elasticity is $dE(Y|X)/dX \times X/E(Y|X) = (\beta_1/X) \times X/[\beta_0 + \beta_1\ln(X)] = \beta_1/[\beta_0 + \beta_1\ln(X)]$, as is given in the table. For the log-linear model, it is conventional to make the additional assumption that u and X are independently distributed, so the expression for $E(Y|X)$ given following Equation (8.25) becomes $E(Y|X) = ce^{\beta_0+\beta_1 X}$, where $c = E(e^u)$ is a constant that does not depend on X because of the additional assumption that u and X are independent. Thus $dE(Y|X)/dX = ce^{\beta_0+\beta_1 X}\beta_1$ and the elasticity is $dE(Y|X)/dX \times X/E(Y|X) = ce^{\beta_0+\beta_1 X}\beta_1 \times X/(ce^{\beta_0+\beta_1 X}) = \beta_1 X$. The derivations for the linear and log-log models are left as Exercise 8.11.