7/14/99 5:33 PM Page 178

# Chapter Six

# Multiple Regression Analysis: Further Issues

This chapter brings together several issues in multiple regression analysis that we could not conveniently cover in earlier chapters. These topics are not as fundamental as the material in Chapters 3 and 4, but they are important for applying multiple regression to a broad range of empirical problems.

## 6.1 EFFECTS OF DATA SCALING ON OLS STATISTICS

In Chapter 2 on bivariate regression, we briefly discussed the effects of changing the units of measurement on the OLS intercept and slope estimates. We also showed that changing the units of measurement did not affect *R*-squared. We now return to the issue of data scaling and examine the effects of rescaling the dependent or independent variables on standard errors, *t* statistics, *F* statistics, and confidence intervals.

We will discover that everything we expect to happen, does happen. When variables are rescaled, the coefficients, standard errors, confidence intervals, *t* statistics, and *F* statistics change in ways that preserve all measured effects and testing outcomes. While this is no great surprise—in fact, we would be very worried if it were not the case—it is useful to see what occurs explicitly. Often, data scaling is used for cosmetic purposes, such as to reduce the number of zeros after a decimal point in an estimated coefficient. By judiciously choosing units of measurement, we can improve the appearance of an estimated equation while changing nothing that is essential.

We could treat this problem in a general way, but it is much better illustrated with examples. Likewise, there is little value here in introducing an abstract notation.

We begin with an equation relating infant birth weight to cigarette smoking and family income:

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc, \tag{6.1}$$

where *bwght* is child birth weight, in ounces, *cigs* is number of cigarettes smoked by the mother while pregnant, per day, and *faminc* is annual family income, in thousands of dollars. The estimates of this equation, obtained using the data in BWGHT.RAW, are given in the first column of Table 6.1. Standard errors are listed in parentheses. The estimate on *cigs* says that if a woman smoked 5 more cigarettes per day, birth weight is pre-

7/14/99 5:33 PM Page 179



**Table 6.1**

Effects of Data Scaling

| Dependent Variable | (1) *bwght* | (2) *bwghtlbs* | (3) *bwght* |
|---|---|---|---|
| **Independent Variables** | | | |
| *cigs* | −.4634<br>(.0916) | −.0289<br>(.0057) | ——— |
| *packs* | ——— | ——— | −9.268<br>(1.832) |
| *faminc* | .0927<br>(.0292) | .0058<br>(.0018) | .0927<br>(.0292) |
| *intercept* | 116.974<br>(1.049) | 7.3109<br>(0.0656) | 116.974<br>(1.049) |
| Observations: | 1,388 | 1,388 | 1,388 |
| *R*-squared: | .0298 | .0298 | .0298 |
| SSR: | 557,485.51 | 2,177.6778 | 557,485.51 |
| SER: | 20.063 | 1.2539 | 20.063 |

dicted to be about .4634(5) = 2.317 ounces less. The *t* statistic on *cigs* is −5.03, so the variable is very statistically significant.

Now, suppose that we decide to measure birth weight in pounds, rather than in ounces. Let *bwghtlbs* = *bwght*/16 be birth weight in pounds. What happens to our OLS statistics if we use this as the dependent variable in our equation? It is easy to find the effect on the coefficient estimates by simple manipulation of equation (6.1). Divide this entire equation by 16:

$$\widehat{bwght}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)faminc.$$

Since the left hand side is birth weight in pounds, it follows that each new coefficient will be the corresponding old coefficient divided by 16. To verify this, the regression of *bwghtlbs* on *cigs*, and *faminc* is reported in column (2) of Table 6.1. Up to four digits, the intercept and slopes in column (2) are just those in column (1) divided by 16. For example, the coefficient on *cigs* is now −.0289; this means that if *cigs* were higher by five, birth weight would be .0289(5) = .1445 *pounds* lower. In terms of ounces, we have

7/14/99 5:33 PM Page 180



.1445(16) = 2.312, which is slightly different from the 2.32 we obtained earlier due to rounding error. The point is, once the effects are transformed into the same units, we get exactly the same answer, regardless of how the dependent variable is measured.

What about statistical significance? As we expect, changing the dependent variable from ounces to pounds has no effect on how statistically important the independent variables are. The standard errors in column (2) are 16 times smaller than those in column (1). A few quick calculations show that the $t$ statistics in column (2) are indeed identical to the $t$ statistics in column (2). The endpoints for the confidence intervals in column (2) are just the endpoints in column (1) divided by 16. This is because the CIs change by the same factor as the standard errors. [Remember that the 95% CI here is $\hat{\beta}_j \pm 1.96 \text{ se}(\hat{\beta}_j)$.]

In terms of goodness-of-fit, the $R$-squareds from the two regressions are identical, as should be the case. Notice that the sum of squared residuals, SSR, and the standard error of the regression, SER, do differ across equations. These differences are easily explained. Let $\hat{u}_i$ denote the residual for observation $i$ in the original equation (6.1). Then the residual when *bwghtlbs* is the dependent variable is simply $\hat{u}_i/16$. Thus, the *squared* residual in the second equation is $(\hat{u}_i/16)^2 = \hat{u}_i^2/256$. This is why the sum of squared residuals in column (2) is equal to the SSR in column (1) divided by 256.

Since $\text{SER} = \hat{\sigma} = \sqrt{\text{SSR}/(n-k-1)} = \sqrt{\text{SSR}/1{,}385}$, the SER in column (2) is 16 times smaller than that in column (1). Another way to think about this is that the error in the equation with *bwghtlbs* as the dependent variable has a standard deviation 16 times smaller than the standard deviation of the original error. This does not mean that we have reduced the error by changing how birth weight is measured; the smaller SER simply reflects a difference in units of measurement.

Next, let us return the dependent variable to its original units: *bwght* is measured in ounces. Instead, let us change the unit of measurement of one of the independent variables, *cigs*. Define *packs* to be the number of packs of cigarettes smoked per day. Thus, *packs* = *cigs*/20. What happens to the coefficients and other OLS statistics now? Well, we can write

$$\widehat{bwght} = \hat{\beta}_0 + (20\hat{\beta}_1)(cigs/20) + \hat{\beta}_2 faminc = \hat{\beta}_0 + (20\hat{\beta}_1)packs + \hat{\beta}_2 faminc.$$

Thus, the intercept and slope coefficient on *faminc* are unchanged, but the coefficient on *packs* is 20 times that on *cigs*. This is intuitively appealing. The results from the regression of *bwght* on *packs* and *faminc* are in column (3) of Table 6.1. Incidentally, remember that it would make no sense to include both *cigs* and *packs* in the same equation; this would induce perfect multicollinearity and would have no interesting meaning.

**QUESTION 6.1**

In the original birth weight equation (6.1), suppose that *faminc* is measured in dollars rather than in thousands of dollars. Thus, define the variable *fincdol* = 1,000·*faminc*. How will the OLS statistics change when *fincdol* is substituted for *faminc*? For the purposes of presenting the regression results, do you think it is better to measure income in dollars or in thousands of dollars?

Other than the coefficient on *packs*, there is one other statistic in column (3) that differs from that in column (1): the standard error on *packs* is 20 times larger than that on *cigs* in column (1). This means that the $t$ statistic for testing the significance of cigarette smoking is the same whether we measure smoking in terms of cigarettes or packs. This is only natural.

7/14/99 5:33 PM Page 181



The previous example spells out most of the possibilities that arise when the dependent and independent variables are rescaled. Rescaling is often done with dollar amounts in economics, especially when the dollar amounts are very large.

In Chapter 2 we argued that, if the dependent variable appears in logarithmic form, changing the units of measurement does not affect the slope coefficient. The same is true here: changing the units of measurement of the dependent variable, when it appears in logarithmic form, does not affect any of the slope estimates. This follows from the simple fact that $\log(c_1 y_i) = \log(c_1) + \log(y_i)$ for any constant $c_1 > 0$. The new intercept will be $\log(c_1) + \hat{\beta}_0$. Similarly, changing the units of measurement of any $x_j$, where $\log(x_j)$ appears in the regression, only affects the intercept. This corresponds to what we know about percentage changes and, in particular, elasticities: they are invariant to the units of measurement of either $y$ or the $x_j$. For example, if we had specified the dependent variable in (6.1) to be log(*bwght*), estimated the equation, and then reestimated it with log(*bwghtlbs*) as the dependent variable, the coefficients on *cigs* and *faminc* would be the same in both regressions; only the intercept would be different.

## Beta Coefficients

Sometimes in econometric applications, a key variable is measured on a scale that is difficult to interpret. Labor economists often include test scores in wage equations, and the scale on which these tests are scored is often arbitrary and not easy to interpret (at least for economists!). In almost all cases, we are interested in how a particular individual's score compares with the population. Thus, instead of asking about the effect on hourly wage if, say, a test score is 10 points higher, it makes more sense to ask what happens when the test score is one *standard deviation* higher.

Nothing prevents us from seeing what happens to the dependent variable when an independent variable in an estimated model increases by a certain number of standard deviations, assuming that we have obtained the sample standard deviation (which is easy in most regression packages). This is often a good idea. So, for example, when we look at the effect of a standardized test score, such as the SAT score, on college GPA, we can find the standard deviation of SAT and see what happens when the SAT score increases by one or two standard deviations.

Sometimes it is useful to obtain regression results when *all* variables involved, the dependent as well as all the independent variables, have been *standardized*. A variable is standardized in the sample by subtracting off its mean and dividing by its standard deviation (see Appendix C). This means that we compute the *z-score* for every variable in the sample. Then, we run a regression using the $z$-scores.

Why is standardization useful? It is easiest to start with the original OLS equation, with the variables in their original forms:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i. \qquad \textbf{(6.2)}$$

We have included the observation subscript $i$ to emphasize that our standardization is applied to all sample values. Now, if we average (6.2), use the fact that the $\hat{u}_i$ have a zero sample average, and subtract the result from (6.2), we get

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i.$$

7/14/99 5:33 PM Page 182



Now, let $\hat{\sigma}_y$ be the sample standard deviation for the dependent variable, let $\hat{\sigma}_1$ be the sample *sd* for $x_1$, let $\hat{\sigma}_2$ be the sample *sd* for $x_2$, and so on. Then, simple algebra gives the equation

$$(y_i - \bar{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \bar{x}_1)/\hat{\sigma}_1] + \ldots + (\hat{\sigma}_k/\hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \bar{x}_k)/\hat{\sigma}_k] + (\hat{u}_i/\hat{\sigma}_y). \quad \textbf{(6.3)}$$

Each variable in (6.3) has been standardized by replacing it with its $z$-score, and this has resulted in new slope coefficients. For example, the slope coefficient on $(x_{i1} - \bar{x}_1)/\hat{\sigma}_1$ is $(\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1$. This is simply the original coefficient, $\hat{\beta}_1$, multiplied by the ratio of the standard deviation of $x_1$ to the standard deviation of $y$. The intercept has dropped out altogether.

It is useful to rewrite (6.3), dropping the $i$ subscript as

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 \ldots + \hat{b}_k z_k + error, \quad \textbf{(6.4)}$$

where $z_y$ denotes the $z$-score of $y$, $z_1$ is the $z$-score of $x_1$, and so on. The new coefficients are

$$\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\hat{\beta}_j \text{ for } j = 1, \ldots, k. \quad \textbf{(6.5)}$$

These $\hat{b}_j$ are traditionally called **standardized coefficients** or **beta coefficients**. (The latter name is more common, which is unfortunate since we have been using beta hat to denote the *usual* OLS estimates.)

Beta coefficients receive their interesting meaning from equation (6.4): If $x_1$ increases by one standard deviation, then $\hat{y}$ changes by $\hat{b}_1$ standard deviations. Thus, we are measuring effects not in terms of the original units of $y$ or the $x_j$, but in standard deviation units. Because it makes the scale of the regressors irrelevant, this equation puts the explanatory variables on equal footing. In a standard OLS equation, it is not possible to simply look at the size of different coefficients and conclude that the explanatory variable with the largest coefficient is "the most important." We just saw that the magnitudes of coefficients can be changed at will by changing the units of measurement of the $x_j$. But, when each $x_j$ has been standardized, comparing the magnitudes of the resulting beta coefficients is more compelling.

To obtain the beta coefficients, we can always standardize $y, x_1, \ldots, x_k$, and then run the OLS regression of the $z$-score of $y$ on the $z$-scores of $x_1, \ldots, x_k$—where it is not necessary to include an intercept, as it will be zero. This can be tedious with many independent variables. Some regression packages provide beta coefficients via a simple command. The following example illustrates the use of beta coefficients.

---

**EXAMPLE 6.1**
(Effects of Pollution on Housing Prices)

We use the data from Example 4.5 (in the file HPRICE2.RAW) to illustrate the use of beta coefficients. Recall that the key independent variable is *nox*, a measure of the nitrogen oxide in the air over each community. One way to understand the size of the pollution

7/14/99 5:33 PM Page 183



effect—without getting into the science underlying nitrogen oxide's effect on air quality—is to compute beta coefficients. (An alternative approach is contained in Example 4.5: we obtained a price elasticity with respect to *nox* by using *price* and *nox* in logarithmic form.)

The population equation is the level-level model

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u,$$

where all the variables except *crime* were defined in Example 4.5; *crime* is the number of reported crimes per capita. The beta coefficients are reported in the following equation (so each variable has been converted to its *z*-score):

$$\widehat{zprice} = -.340\ znox - .143\ zcrime + .514\ zrooms - .235\ zdist - .270\ zstratio.$$

This equation shows that a one standard deviation increase in *nox* decreases price by .34 standard deviations; a one standard deviation increase in *crime* reduces price by .14 standard deviation. Thus, the same relative movement of pollution in the population has a larger effect on housing prices than crime does. Size of the house, as measured by number of rooms (*rooms*), has the largest standardized effect. If we want to know the effects of each independent variable on the dollar value of median house price, we should use the unstandardized variables.

## 6.2 MORE ON FUNCTIONAL FORM

In several previous examples, we have encountered the most popular device in econometrics for allowing nonlinear relationships between the explained and explanatory variables: using logarithms for the dependent or independent variables. We have also seen models containing quadratics in some explanatory variables, but we have yet to provide a systematic treatment of them. In this section, we cover some variations and extensions on functional forms that often arise in applied work.

### More on Using Logarithmic Functional Forms

We begin by reviewing how to interpret the parameters in the model

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u, \qquad \textbf{(6.6)}$$

where these variables are taken from Example 4.5. Recall that throughout the text $\log(x)$ is the *natural* log of $x$. The coefficient $\beta_1$ is the elasticity of *price* with respect to *nox* (pollution). The coefficient $\beta_2$ is the change in $\log(price)$, when $\Delta rooms = 1$; as we have seen many times, when multiplied by 100, this is the approximate percentage change in price. Recall that $100 \cdot \beta_2$ is sometimes called the semi-elasticity of *price* with respect to *rooms*.

When estimated using the data in HPRICE2.RAW, we obtain

$$\begin{gathered} \widehat{\log(price)} = \underset{(0.19)}{9.23} - \underset{(.066)}{.718} \log(nox) + \underset{(.019)}{.306} rooms \\ n = 506,\ R^2 = .514. \end{gathered} \qquad \textbf{(6.7)}$$

7/14/99 5:33 PM Page 184



Thus, when *nox* increases by 1%, *price* falls by .718%, holding only *rooms* fixed. When *rooms* increases by one, *price* increases by approximately 100(.306) = 30.6%.

The estimate that one more room increases price by about 30.6% turns out to be somewhat inaccurate for this application. The approximation error occurs because, as the change in log($y$) becomes larger and larger, the approximation $\%\Delta y \approx 100\cdot\Delta\log(y)$ becomes more and more inaccurate. Fortunately, a simple calculation is available to compute the exact percentage change.

To describe the procedure, we consider the general estimated model

$$\hat{\log}(y) = \hat{\beta}_0 + \hat{\beta}_1\log(x_1) + \hat{\beta}_2 x_2.$$

(Adding additional independent variables does not change the procedure.) Now, fixing $x_1$, we have $\Delta\hat{\log}(y) = \hat{\beta}_2\Delta x_2$. Using simple algebraic properties of the exponential and logarithmic functions gives the *exact* percentage change in the predicted $y$ as

$$\%\hat{\Delta}y = 100\cdot[\exp(\hat{\beta}_2\Delta x_2) - 1], \tag{6.8}$$

where the multiplication by 100 turns the proportionate change into a percentage change. When $\Delta x_2 = 1$,

$$\%\hat{\Delta}y = 100\cdot[\exp(\hat{\beta}_2) - 1]. \tag{6.9}$$

Applied to the housing price example with $x_2$ = rooms and $\hat{\beta}_2 = .306$, $\%\Delta\hat{price} = 100[\exp(.306) - 1] = 35.8\%$, which is notably larger than the approximate percentage change, 30.6%, obtained directly from (6.7). {Incidentally, this is not an unbiased estimator because $\exp(\cdot)$ is a nonlinear function; it is, however, a consistent estimator of $100[\exp(\beta_2) - 1]$. This is because the probability limit passes through continuous functions, while the expected value operator does not. See Appendix C.}

The adjustment in equation (6.8) is not as crucial for small percentage changes. For example, when we include the student-teacher ratio in equation (6.7), its estimated coefficient is −.052, which means that if *stratio* increases by one, *price* decreases by approximately 5.2%. The exact proportionate change is $\exp(-.052) - 1 \approx -.051$, or −5.1%. On the other hand, if we increase *stratio* by five, then the approximate percentage change in price is −26%, while the exact change obtained from equation (6.8) is $100[\exp(-.26) - 1] \approx -22.9\%$.

We have seen that using natural logs leads to coefficients with appealing interpretations, and we can be ignorant about the units of measurement of variables appearing in logarithmic form because the slope coefficients are invariant to rescalings. There are several other reasons logs are used so much in applied work. First, when $y > 0$, models using log($y$) as the dependent variable often satisfy the CLM assumptions more closely than models using the level of $y$. Strictly positive variables often have conditional distributions that are heteroskedastic or skewed; taking the log can mitigate, if not eliminate, both problems.

Moreover, taking logs usually narrows the range of the variable, in some cases by a considerable amount. This makes estimates less sensitive to outlying (or extreme) observations on the dependent or independent variables. We take up the issue of outlying observations in Chapter 9.

7/14/99 5:33 PM Page 185



There are some standard rules of thumb for taking logs, although none is written in stone. When a variable is a positive dollar amount, the log is often taken. We have seen this for variables such as wages, salaries, firm sales, and firm market value. Variables such as population, total number of employees, and school enrollment often appear in logarithmic form; these have the common feature of being large integer values.

Variables that are measured in years—such as education, experience, tenure, age, and so on—usually appear in their original form. A variable that is a proportion or a percent—such as the unemployment rate, the participation rate in a pension plan, the percentage of students passing a standardized exam, the arrest rate on reported crimes—can appear in either original or logarithmic form, although there is a tendency to use them in level forms. This is because any regression coefficients involving the *original* variable—whether it is the dependent or independent variable—will have a *percentage point* change interpretation. (See Appendix A for a review of the distinction between a percentage change and a percentage point change.) If we use, say, log(*unem*) in a regression, where *unem* is the percent of unemployed individuals, we must be very careful to distinguish between a percentage point change and a percentage change. Remember, if *unem* goes from 8 to 9, this is an increase of one percentage point, but a 12.5% increase from the initial unemployment level. Using the log means that we are looking at the percentage change in the unemployment rate: $\log(9) - \log(8) \approx .118$ or 11.8%, which is the logarithmic approximation to the actual 12.5% increase.

**QUESTION 6.2**

Suppose that the annual number of drunk driving arrests is determined by

$$\log(arrests) = \beta_0 + \beta_1 \log(pop) + \beta_2 age16\_25 + other\ factors,$$

where *age16_25* is the proportion of the population between 16 and 25 years of age. Show that $\beta_2$ has the following (ceteris paribus) interpretation: it is the percentage change in *arrests* when the percentage of the people aged 16 to 25 increases by one *percentage point*.

One limitation of the log is that it cannot be used if a variable takes on zero or negative values. In cases where a variable $y$ is nonnegative but can take on the value 0, $\log(1 + y)$ is sometimes used. The percentage change interpretations are often closely preserved, except for changes beginning at $y = 0$ (where the percentage change is not even defined). Generally, using $\log(1 + y)$ and then interpreting the estimates as if the variable were $\log(y)$ is acceptable when the data on $y$ are not dominated by zeros. An example might be where $y$ is hours of training per employee for the population of manufacturing firms, if a large fraction of firms provide training to at least one worker.

One drawback to using a dependent variable in logarithmic form is that it is more difficult to predict the original variable. The original model allows us to predict $\log(y)$, not $y$. Nevertheless, it is fairly easy to turn a prediction for $\log(y)$ into a prediction for $y$ (see Section 6.4). A related point is that it is *not* legitimate to compare $R$-squareds from models where $y$ is the dependent variable in one case and $\log(y)$ is the dependent variable in the other. These measures explained variations in different variables. We discuss how to compute comparable goodness-of-fit measures in Section 6.4.

## Models with Quadratics

**Quadratic functions** are also used quite often in applied economics to capture decreasing or increasing marginal effects. You may want to review properties of quadratic functions in Appendix A.

7/14/99 5:33 PM Page 186



In the simplest case, $y$ depends on a single observed factor $x$, but it does so in a quadratic fashion:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

For example, take $y = wage$ and $x = exper$. As we discussed in Chapter 3, this model falls outside of simple regression analysis but is easily handled with multiple regression.

It is important to remember that $\beta_1$ does not measure the change in $y$ with respect to $x$; it makes no sense to hold $x^2$ fixed while changing $x$. If we write the estimated equation as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \qquad \textbf{(6.10)}$$

then we have the approximation

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \text{ so } \Delta\hat{y}/\Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \qquad \textbf{(6.11)}$$

This says that the slope of the relationship between $x$ and $y$ depends on the value of $x$; the estimated slope is $\hat{\beta}_1 + 2\hat{\beta}_2 x$. If we plug in $x = 0$, we see that $\hat{\beta}_1$ can be interpreted as the approximate slope in going from $x = 0$ to $x = 1$. After that, the second term, $2\hat{\beta}_2 x$, must be accounted for.

If we are only interested in computing the predicted change in $y$ given a starting value for $x$ and a change in $x$, we could use (6.10) directly: there is no reason to use the calculus approximation at all. However, we are usually more interested in quickly summarizing the effect of $x$ on $y$, and the interpretation of $\hat{\beta}_1$ and $\hat{\beta}_2$ in equation (6.11) provides that summary. Typically, we might plug the average value of $x$ in the sample, or some other interesting values, such as the median or the lower and upper quartile values.

In many applications, $\hat{\beta}_1$ is positive, and $\hat{\beta}_2$ is negative. For example, using the wage data in WAGE1.RAW, we obtain

$$\widehat{wage} = \underset{}{3.73} + \underset{(0.35)}{.298}\, exper - \underset{(.041)}{.0061}\, exper^2 \qquad \textbf{(6.12)}$$
$$\underset{(.0009)}{}$$
$$n = 526,\ R^2 = .093.$$

This estimated equation implies that *exper* has a diminishing effect on *wage*. The first year of experience is worth roughly 30 cents per hour (.298 dollars). The second year of experience is worth less [about $.298 - 2(.0061)(1) \approx .286$, or 28.6 cents, according the approximation in (6.11) with $x = 1$]. In going from 10 to 11 years of experience, *wage* is predicted to increase by about $.298 - 2(.0061)(10) \approx .176$, or 17.6 cents. And so on.

When the coefficient on $x$ is positive, and the coefficient on $x^2$ is negative, the quadratic has a parabolic shape. There is always a positive value of $x$, where the effect of $x$ on $y$ is zero; before this point, $x$ has a positive effect on $y$; after this point, $x$ has a negative effect on $y$. In practice, it can be important to know where this turning point is.

7/14/99 5:33 PM Page 187



In the estimated equation (6.10) with $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$, the turning point (or maximum of the function) is always achieved at the coefficient on $x$ over *twice* the absolute value of the coefficient on $x^2$:

$$x^* = |\hat{\beta}_1/(2\hat{\beta}_2)|. \quad \textbf{(6.13)}$$

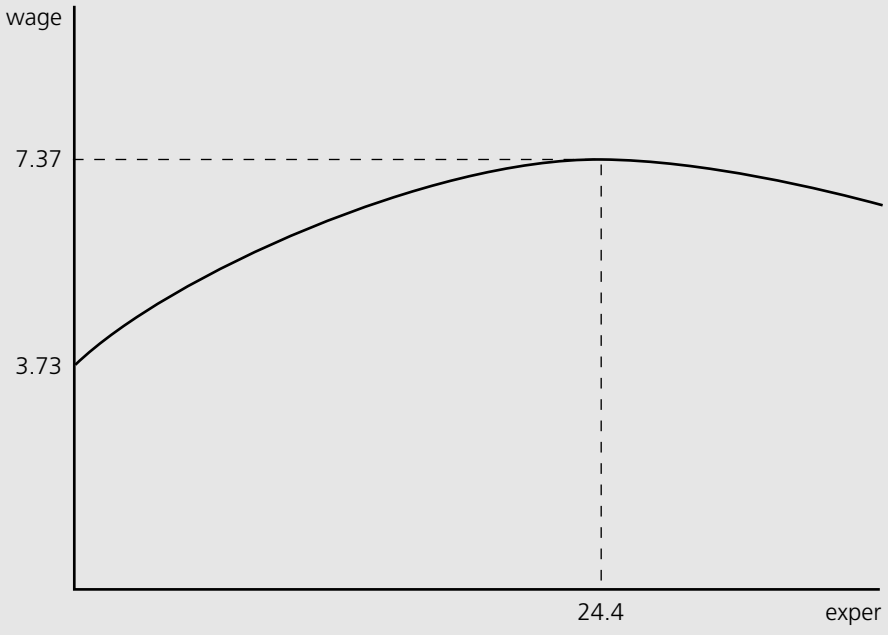
In the wage example, $x^* = exper^*$ is $.298/[2(.0061)] \approx 24.4$. (Note how we just drop the minus sign on $-.0061$ in doing this calculation.) This quadratic relationship is illustrated in Figure 6.1.

In the wage equation (6.12), the return to experience becomes zero at about 24.4 years. What should we make of this? There are at least three possible explanations. First, it may be that few people in the sample have more than 24 years of experience, and so the part of the curve to the right of 24 can be ignored. The cost of using a quadratic to capture diminishing effects is that the quadratic must eventually turn around. If this point is beyond all but a small percentage of the people in the sample, then this is not of much concern. But in the data set WAGE1.RAW, about 28% of the people in the sample have more than 24 years of experience; this is too high a percentage to ignore.

It is possible that the return to *exper* really become negative at some point, but it is hard to believe that this happens at 24 years of experience. A more likely possibil-

**Figure 6.1**

Quadratic relationship between *wage* and *exper*.



7/14/99 5:33 PM Page 188



ity is that the estimated effect of *exper* on *wage* is biased, because we have controlled for no other factors, or because the functional relationship between *wage* and *exper* in equation (6.12) is not entirely correct. Problem 6.9 asks you to explore this possibility by controlling for education, in addition to using log(*wage*) as the dependent variable.

When a model has a dependent variable in logarthmic form and an explanatory variable entering as a quadratic, some care is needed in making a useful interpretation. The following example also shows the quadratic can have a U-shape, rather than a parabolic shape. A U-shape arises in the equation (6.10) when $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive; this captures an increasing effect of $x$ on $y$.

---

**EXAMPLE 6.2**
(Effects of Pollution on Housing Prices)

We modify the housing price model from Example 4.5 to include a quadratic term in *rooms*:

$$\log(price) = \beta_0 + \beta_1\log(nox) + \beta_2\log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + u. \quad \textbf{(6.14)}$$

The model estimated using the data in HPRICE2.RAW is

$$\widehat{\log(price)} = \underset{(0.57)}{13.39} - \underset{(.115)}{.902}\log(nox) - \underset{(.043)}{.087}\log(dist) - \underset{(.165)}{.545}\,rooms + \underset{(.013)}{.062}\,rooms^2 - \underset{(.006)}{.048}\,stratio$$
$$n = 506,\ R^2 = .603.$$

The quadratic term $rooms^2$ has a $t$ statistic of about 4.77, and so it is very statistically significant. But what about interpreting the effect of *rooms* on log(*price*)? Initially, the effect appears to be strange. Since the coefficient on *rooms* is negative and the coefficient on $rooms^2$ is positive, this equation literally implies that, at low values of *rooms*, an additional room has a *negative* effect on log(*price*). At some point, the effect becomes positive, and the quadratic shape means that the semi-elasticity of *price* with respect to *rooms* is increasing as *rooms* increases. This situation is shown in Figure 6.2.

We obtain the turnaround value of *rooms* using equation (6.13) (even though $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive). The absolute value of the coefficient on *rooms*, .545, divided by twice the coefficient on $rooms^2$, .062, gives $rooms^* = .545/[2(.062)] \approx 4.4$; this point is labeled in Figure 6.2.

Do we really believe that starting at three rooms and increasing to four rooms actually reduces a house's expected value? Probably not. It turns out that only five of the 506 communities in the sample have houses averaging 4.4 rooms or less, about 1% of the sample. This is so small that the quadratic to the left of 4.4 can, for practical purposes, be ignored. To the right of 4.4, we see that adding another room has an increasing effect on the percentage change in price:

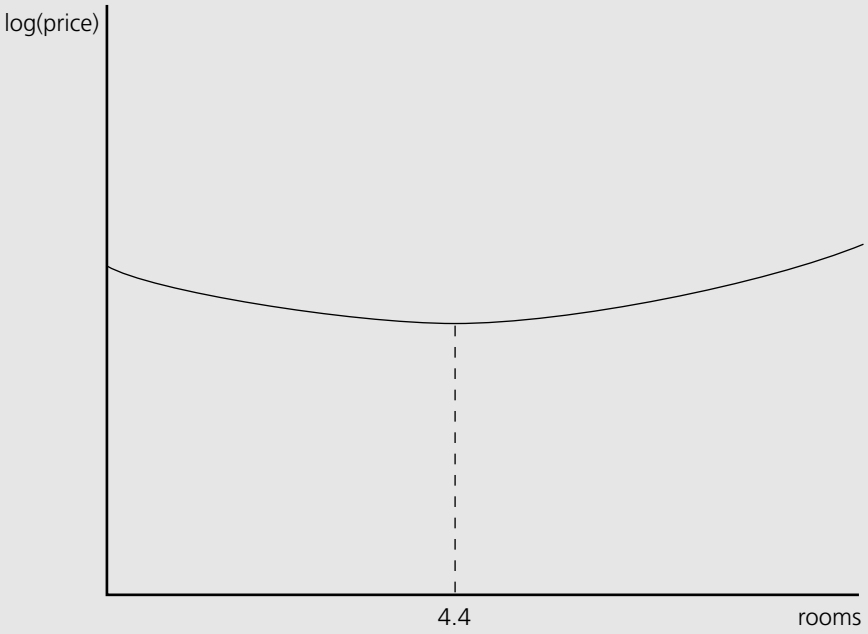
$$\Delta\widehat{\log(price)} \approx \{[-.545 + 2(.062)]rooms\}\Delta rooms$$

7/14/99 5:33 PM Page 189



**Figure 6.2**

log(*price*) as a quadratic function of *rooms*.




and so

$$\%\Delta\hat{price} \approx 100\{[-.545 + 2(.062)]rooms\}\Delta rooms$$
$$= (-54.5 + 12.4\ rooms)\Delta rooms.$$

Thus, an increase in *rooms* from, say, five to six increases price by about $-54.5 + 12.4(5) = 7.5\%$; the increase from six to seven increases price by roughly $-54.5 + 12.4(6) = 19.9\%$. This is a very strong increasing effect.

There are many other possibilities for using quadratics along with logarithms. For example, an extension of (6.14) that allows a nonconstant elasticity between *price* and *nox* is

$$\log(price) = \beta_0 + \beta_1\log(nox) + \beta_2[\log(nox)]^2 + \beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u. \quad \textbf{(6.15)}$$

7/14/99 5:33 PM Page 190



If $\beta_2 = 0$, then $\beta_1$ is the elasticity of *price* with respect to *nox*. Otherwise, this elasticity depends on the level of *nox*. To see this, we can combine the arguments for the partial effects in the quadratic and logarithmic models to show that

$$\%\Delta price \approx [\beta_1 + 2\beta_2 \log(nox)]\%\Delta nox, \tag{6.16}$$

and therefore the elasticity of *price* with respect to *nox* is $\beta_1 + 2\beta_2 \log(nox)$, so that it depends on $\log(nox)$.

Finally, other polynomial terms can be included in regression models. Certainly the quadratic is seen most often, but a cubic and even a quartic term appear now and then. An often reasonable functional form for a total cost function is

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u.$$

Estimating such a model causes no complications. Interpreting the parameters is more involved (though straightforward using calculus); we do not study these models further.

## Models with Interaction Terms

Sometimes it is natural for the partial effect, elasticity, or semi-elasticity of the dependent variable with respect to an explanatory variable to depend on the magnitude of yet *another* explanatory variable. For example, in the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \cdot bdrms + \beta_4 bthrms + u,$$

the partial effect of *bdrms* on *price* (holding all other variables fixed) is

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft. \tag{6.17}$$

If $\beta_3 > 0$, then (6.17) implies that an additional bedroom yields a higher increase in housing price for larger houses. In other words, there is an **interaction effect** between square footage and number of bedrooms. In summarizing the effect of *bdrms* on *price*, we must evaluate (6.17) at interesting values of *sqrft*, such as the mean value, or the lower and upper quartiles in the sample. Whether or not $\beta_3$ is zero is something we can easily test.

---

**EXAMPLE 6.3**
(Effects of Attendance on Final Exam Performance)

A model to explain the standardized outcome on a final exam (*stndfnl*) in terms of percentage of classes attended, prior college grade point average, and ACT score is

$$\begin{aligned} stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 \\ + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u. \end{aligned} \tag{6.18}$$

(We use the standardized exam score for the reasons discussed in Section 6.1: it is easier to interpret a student's performance relative to the rest of the class.) In addition to quadratics

7/14/99 5:33 PM Page 191



in *priGPA* and *ACT*, this model includes an interaction between *priGPA* and the attendance rate. The idea is that class attendance might have a different effect for students who have performed differently in the past, as measured by *priGPA*. We are interested in the effects of attendance on final exam score: $\Delta stndfnl/\Delta atndrte = \beta_1 + \beta_6 priGPA$.

Using the 680 observations in ATTEND.RAW, for students in microeconomic principles, the estimated equation is

$$\begin{aligned} \widehat{stndfnl} = \underset{}{2.05} &- \underset{(1.36)}{.0067}\, atndrte - \underset{(0.48)}{1.63}\, priGPA - \underset{(.098)}{.128}\, ACT \\ &+ \underset{(.101)}{.296}\, priGPA^2 + \underset{(.0022)}{.0045}\, ACT^2 + \underset{(.0043)}{.0056}\, priGPA{\cdot}atndrte \\ n = 680,\ R^2 = .229,\ &\bar{R}^2 = .222. \end{aligned} \tag{6.19}$$

We must interpret this equation with extreme care. If we simply look at the coefficient on *atndrte*, we will incorrectly conclude that attendance has a *negative* effect on final exam score. But this coefficient supposedly measures the effect when *priGPA* = 0, which is not interesting (in this sample, the smallest prior GPA is about .86). We must also take care not to look separately at the estimates of $\beta_1$ and $\beta_6$ and conclude that, because each *t* statistic is insignificant, we cannot reject $H_0$: $\beta_1 = 0$, $\beta_6 = 0$. In fact, the *p*-value for the *F* test of this joint hypothesis is .014, so we certainly reject $H_0$ at the 5% level. This is a good example of where looking at separate *t* statistics when testing a joint hypothesis can lead one far astray.

How should we estimate the partial effect of *atndrte* on *stndfnl*? We must plug in interesting values of *priGPA* to obtain the partial effect. The mean value of *priGPA* in the sample is 2.59, so at the mean *priGPA*, the effect of *atndrte* on *stndfnl* is $-.0067 + .0056(2.59) \approx .0078$. What does this mean? Because *atndrte* is measured as a percent, it means that a 10 percentage point increase in *atndrte* increases $\widehat{stndfnl}$ by .078 standard deviations from the mean final exam score.

How can we tell whether the estimate .0078 is statistically different from zero? We need to rerun the regression, where we replace *priGPA*·*atndrte* with (*priGPA* − 2.59)·*atndrte*. This gives, as the new coefficient on *atndrte*, the estimated effect at *priGPA* = 2.59, along with its standard error; nothing else in the regression changes. (We described this device in Section 4.4.) Running this new regression gives the standard error of $\hat{\beta}_1 + \hat{\beta}_6(2.59) = .0078$ as .0026, which yields $t = .0078/.0026 = 3$. Therefore, at the average *priGPA*, we conclude that attendance has a statistically significant positive effect on final exam score.

**QUESTION 6.3**

If we add the term $\beta_7 ACT{\cdot}atndrte$ to equation (6.18), what is the partial effect of *atndrte* on *stndfnl*?

Things are even more complicated for finding the effect of *priGPA* on *stndfnl* because of the quadratic term $priGPA^2$. To find the effect at the mean value of *priGPA* and the mean attendance rate, .82, we would replace $priGPA^2$ with $(priGPA - 2.59)^2$ and *priGPA*·*atndrte* with *priGPA*·(*atndrte* − .82). The coefficient on *priGPA* becomes the partial effect at the mean values, and we would have its standard error. (See Problem 6.14.)

7/14/99 5:33 PM Page 192



## 6.3 MORE ON GOODNESS-OF-FIT AND SELECTION OF REGRESSORS

Until now, we have not focused much on the size of $R^2$ in evaluating our regression models, because beginning students tend to put too much weight on *R*-squared. As we will see now, choosing a set of explanatory variables based on the size of the *R*-squared can lead to nonsensical models. In Chapter 10, we will discover that *R*-squareds obtained from time series regressions can be artificially high and can result in misleading conclusions.

Nothing about the classical linear model assumptions requires that $R^2$ be above any particular value; $R^2$ is simply an estimate of how much variation in $y$ is explained by $x_1$, $x_2,\ldots,x_k$ in the population. We have seen several regressions that have had pretty small *R*-squareds. While this means that we have not accounted for several factors that affect $y$, this does not mean that the factors in $u$ are correlated with the independent variables. The zero conditional mean assumption MLR.3 is what determines whether we get unbiased estimators of the ceteris paribus effects of the independent variables, and the size of the *R*-squared has no direct bearing on this.

Remember, though, that the relative *change* in the *R*-squared, when variables are added to an equation, is very useful: the *F* statistic in (4.41) for testing the joint significance crucially depends on the difference in *R*-squareds between the unrestricted and restricted models.

### Adjusted *R*-Squared

Most regression packages will report, along with the *R*-squared, a statistic called the **adjusted *R*-squared**. Since the adjusted *R*-squared is reported in much applied work, and since it has some useful features, we cover it in this subsection.

To see how the usual *R*-squared might be adjusted, it is usefully written as

$$R^2 = 1 - (\text{SSR}/n)/(\text{SST}/n), \tag{6.20}$$

where SSR is the sum of squared residuals and SST is the total sum of squares; compared with equation (3.28), all we have done is divide both SSR and SST by $n$. This expression reveals what $R^2$ is actually estimating. Define $\sigma_y^2$ as the population variance of $y$ and let $\sigma_u^2$ denote the population variance of the error term, $u$. (Until now, we have used $\sigma^2$ to denote $\sigma_u^2$, but it is helpful to be more specific here.) The **population *R*-squared** is defined as $1 - \sigma_u^2/\sigma_y^2$; this is the proportion of the variation in $y$ in the population explained by the independent variables. This is what $R^2$ is supposed to be estimating.

$R^2$ estimates $\sigma_u^2$ by SSR/$n$, which we know to be biased. So why not replace SSR/$n$ with SSR/$(n - k - 1)$? Also, we can use SST/$(n - 1)$ in place of SST/$n$, as the former is the unbiased estimator of $\sigma_y^2$. Using these estimators, we arrive at the adjusted *R*-squared:

$$\begin{aligned}\bar{R}^2 &\equiv 1 - [\text{SSR}/(n - k - 1)]/[\text{SST}/(n - 1)] \\ &= 1 - \hat{\sigma}^2/[\text{SST}/(n - 1)],\end{aligned} \tag{6.21}$$

7/14/99 5:33 PM Page 193



since $\hat{\sigma}^2 = \text{SSR}/(n - k - 1)$. Because of the notation used to denote the adjusted *R*-squared, it is sometimes called *R-bar squared*.

The adjusted *R*-squared is sometimes called the *corrected R-squared*, but this is not a good name because it implies that $\bar{R}^2$ is somehow better than $R^2$ as an estimator of the population *R*-squared. Unfortunately, $\bar{R}^2$ is *not* generally known to be a better estimator. It is tempting to think that $\bar{R}^2$ corrects the bias in $R^2$ for estimating the population *R*-squared, but it does not: the ratio of two unbiased estimators is not an unbiased estimator.

The primary attractiveness of $\bar{R}^2$ is that it imposes a penalty for adding additional independent variables to a model. We know that $R^2$ can never fall when a new independent variable is added to a regression equation: this is because SSR never goes up (and usually falls) as more independent variables are added. But the formula for $\bar{R}^2$ shows that it depends explicitly on $k$, the number of independent variables. If an independent variable is added to a regression, SSR falls, but so does the *df* in the regression, $n - k - 1$. $\text{SSR}/(n - k - 1)$ can go up or down when a new independent variable is added to a regression.

An interesting algebraic fact is the following: if we add a new independent variable to a regression equation, $\bar{R}^2$ increases if, and only if, the *t* statistic on the new variable is greater than one in absolute value. (An extension of this is that $\bar{R}^2$ increases when a group of variables is added to a regression if, and only if, the *F* statistic for joint significance of the new variables is greater than unity.) Thus, we see immediately that using $\bar{R}^2$ to decide whether a certain independent variable (or set of variables) belongs in a model gives us a different answer than standard *t* or *F* testing (since a *t* or *F* statistic of unity is not statistically significant at traditional significance levels).

It is sometimes useful to have a formula for $\bar{R}^2$ in terms of $R^2$. Simple algebra gives

$$\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1). \qquad \textbf{(6.22)}$$

For example, if $R^2 = .30$, $n = 51$, and $k = 10$, then $\bar{R}^2 = 1 - .70(50)/40 = .125$. Thus, for small $n$ and large $k$, $\bar{R}^2$ can be substantially below $R^2$. In fact, if the usual *R*-squared is small, and $n - k - 1$ is small, $\bar{R}^2$ can actually be negative! For example, you can plug in $R^2 = .10$, $n = 51$, and $k = 10$ to verify that $\bar{R}^2 = -.125$. A negative $\bar{R}^2$ indicates a very poor model fit relative to the number of degrees of freedom.

The adjusted *R*-squared is sometimes reported along with the usual *R*-squared in regressions, and sometimes $\bar{R}^2$ is reported in place of $R^2$. It is important to remember that it is $R^2$, not $\bar{R}^2$, that appears in the *F* statistic in (4.41). The same formula with $\bar{R}^2_r$ and $\bar{R}^2_{ur}$ is *not* valid.

## Using Adjusted *R*-Squared to Choose Between Nonnested Models

In Section 4.5, we learned how to compute an *F* statistic for testing the joint significance of a group of variables; this allows us to decide, at a particular significance level, whether at least one variable in the group affects the dependent variable. This test does not allow us to decide *which* of the variables has an effect. In some cases, we want to

7/14/99 5:33 PM Page 194



choose a model without redundant independent variables, and the adjusted *R*-squared can help with this.

In the major league baseball salary example in Section 4.4, we saw that neither *hrunsyr* nor *rbisyr* was individually significant. These two variables are highly correlated, so we might want to choose between the models

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + u$$

and

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 rbisyr + u.$$

These two examples are **nonnested models**, because neither equation is a special case of the other. The *F* statistics we studied in Chapter 4 only allow us to test *nested* models: one model (the restricted model) is a special case of the other model (the unrestricted model). See equations (4.32) and (4.28) for examples of restricted and unrestricted models. One possibility is to create a composite model that contains *all* explanatory variables from the original models and then to test each model against the general model using the *F* test. The problem with this process is that either both models might be rejected, or neither model might be rejected (as happens with the major league baseball salary example in Section 4.4). Thus, it does not always provide a way to distinguish between models with nonnested regressors.

In the baseball player salary regression, $\bar{R}^2$ for the regression containing *hrunsyr* is .6211, and $\bar{R}^2$ for the regression containing *rbisyr* is .6226. Thus, based on the adjusted *R*-squared, there is a very slight preference for the model with *rbisyr*. But the difference is practically very small, and we might obtain a different answer by controlling for some of the variables in Problem 4.16. (Because both nonnested models contain five parameters, the usual *R*-squared can be used to draw the same conclusion.)

Comparing $\bar{R}^2$ to choose among different nonnested sets of independent variables can be valuable when these variables represent different functional forms. Consider two models relating R&D intensity to firm sales:

$$rdintens = \beta_0 + \beta_1 \log(sales) + u. \qquad \textbf{(6.23)}$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u. \qquad \textbf{(6.24)}$$

The first model captures a diminishing return by including *sales* in logarithmic form; the second model does this by using a quadratic. Thus, the second model contains one more parameter than the first.

When equation (6.23) is estimated using the 32 observations on chemical firms in RDCHEM.RAW, $R^2$ is .061, and $R^2$ for equation (6.24) is .148. Therefore, it appears that the quadratic fits much better. But a comparison of the usual *R*-squareds is unfair to the first model because it contains one less parameter than (6.24). That is, (6.23) is a more parsimonious model than (6.24).

Everything else being equal, simpler models are better. Since the usual *R*-squared does not penalize more complicated models, it is better to use $\bar{R}^2$. $\bar{R}^2$ for (6.23) is .030, while $\bar{R}^2$ for (6.24) is .090. Thus, even after adjusting for the difference in degrees of freedom, the quadratic model wins out. The quadratic model is also preferred when profit margin is added to each regression.

7/14/99 5:33 PM Page 195



There is an important limitation in using $\bar{R}^2$ to choose between nonnested models: we cannot use it to choose between different functional forms for the dependent variable. This is unfortunate, because we often want to decide on whether $y$ or $\log(y)$ (or maybe some other transformation) should be used as the dependent variable based on goodness-of-fit. But neither $R^2$ nor $\bar{R}^2$ can be used for this. The reason is simple: these $R$-squareds measure the explained proportion of the total variation in whatever dependent variable we are using in the regression, and different functions of the dependent variable will have different amounts of variation to explain. For example, the total variations in $y$ and $\log(y)$ are not the same. Comparing the adjusted $R$-squareds from regressions with these different forms of the dependent variables does not tell us anything about which model fits better; they are fitting two separate dependent variables.

**QUESTION 6.4**

Explain why choosing a model by maximizing $\bar{R}^2$ or minimizing $\hat{\sigma}$ (the standard error of the regression) is the same thing.

**EXAMPLE 6.4**
(CEO Compensation and Firm Performance)

Consider two estimated models relating CEO compensation to firm performance:

$$\widehat{salary} = \underset{(223.90)}{830.63} + \underset{(.0089)}{.0163}\, sales + \underset{(11.08)}{19.63}\, roe \qquad (6.25)$$
$$n = 209,\ R^2 = .029,\ \bar{R}^2 = .020$$

and

$$\widehat{lsalary} = \underset{(0.29)}{4.36} + \underset{(.033)}{.275}\, lsales + \underset{(.0040)}{.0179}\, roe \qquad (6.26)$$
$$n = 209,\ R^2 = .282,\ \bar{R}^2 = .275,$$

where *roe* is the return on equity discussed in Chapter 2. For simplicity, *lsalary* and *lsales* denote the natural logs of *salary* and *sales*. We already know how to interpret these different estimated equations. But can we say that one model fits better than the other?

The $R$-squared for equation (6.25) shows that *sales* and *roe* explain only about 2.9% of the variation in CEO salary in the sample. Both *sales* and *roe* have marginal statistical significance.

Equation (6.26) shows that log(*sales*) and *roe* explain about 28.2% of the variation in log(*salary*). In terms of goodness-of-fit, this much higher $R$-squared would seem to imply that model (6.26) is much better, but this is not necessarily the case. The total sum of squares for *salary* in the sample is 391,732,982, while the total sum of squares for log(*salary*) is only 66.72. Thus, there is much less variation in log(*salary*) that needs to be explained.

At this point, we can use features other than $R^2$ or $\bar{R}^2$ to decide between these models. For example, log(*sales*) and *roe* are much more statistically significant in (6.26) than are *sales* and *roe* in (6.25), and the coefficients in (6.26) are probably of more interest. To be sure, however, we will need to make a valid goodness-of-fit comparison.

7/14/99 5:33 PM Page 196



In Section 6.4, we will offer a goodness-of-fit measure that does allow us to compare models where $y$ appears in both level and log form.

## Controlling for Too Many Factors in Regression Analysis

In many of the examples we have covered, and certainly in our discussion of omitted variables bias in Chapter 3, we have worried about omitting important factors from a model that might be correlated with the independent variables. It is also possible to control for too *many* variables in a regression analysis.

If we overemphasize goodness-of-fit, we open ourselves to controlling for factors in a regression model that should not be controlled for. To avoid this mistake, we need to remember the ceteris paribus interpretation of multiple regression models.

To illustrate this issue, suppose we are doing a study to assess the impact of state beer taxes on traffic fatalities. The idea is that a higher tax on beer will reduce alcohol consumption, and likewise drunk driving, resulting in fewer traffic fatalities. To measure the ceteris paribus effect of taxes on fatalities, we can model *fatalities* as a function of several factors, including the beer *tax*:

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 miles + \beta_3 percmale + \beta_4 perc16\_21 + \ldots,$$

where *miles* is total miles driven, *percmale* is percent of the state population that is male, and *perc16_21* is percent of the population between ages 16 and 21, and so on. Notice how we have not included a variable measuring per capita beer consumption. Are we committing an omitted variables error? The answer is no. If we control for beer consumption in this equation, then how would beer taxes affect traffic fatalities? In the equation

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 beercons + \ldots,$$

$\beta_1$ measures the difference in fatalities due to a one percentage point increase in *tax*, holding *beercons* fixed. It is difficult to understand why this would be interesting. We should not be controlling for differences in *beercons* across states, unless we want to test for some sort of indirect effect of beer taxes. Other factors, such as gender and age distribution, should be controlled for.

The issue of whether or not to control for certain factors is not always clear-cut. For example, Betts (1995) studies the effect of high school quality on subsequent earnings. He points out that, if better school quality results in more education, then controlling for education in the regression along with measures of quality will underestimate the return to quality. Betts does the analysis with and without years of education in the equation to get a range of estimated effects for quality of schooling.

To see explicitly how focusing on high $R$-squareds can lead to trouble, consider the housing price example from Section 4.5 that illustrates the testing of multiple hypotheses. In that case, we wanted to test the rationality of housing price assessments. We regressed log(*price*) on log(*assess*), log(*lotsize*), log(*sqrft*), and *bdrms* and tested whether the latter three variables had zero population coefficients while log(*assess*) had a coefficient of unity. But what if we want to estimate a hedonic price model, as in Example 4.8, where the marginal values of various housing attributes are obtained? Should we include log(*assess*) in the equation? The adjusted $R$-squared from the regres-

7/14/99 5:33 PM Page 197



sion with log(*assess*) is .762, while the adjusted *R*-squared without it is .630. Based on goodness-of-fit only, we should include log(*assess*). But this is incorrect if our goal is to determine the effects of lot size, square footage, and number of bedrooms on housing values. Including log(*assess*) in the equation amounts to holding one measure of value fixed and then asking how much an additional bedroom would change another measure of value. This makes no sense for valuing housing attributes.

If we remember that different models serve different purposes, and we focus on the ceteris paribus interpretation of regression, then we will not include the wrong factors in a regression model.

### Adding Regressors to Reduce the Error Variance

We have just seen some examples of where certain independent variables should not be included in a regression model, even though they are correlated with the dependent variable. From Chapter 3, we know that adding a new independent variable to a regression can exacerbate the multicollinearity problem. On the other hand, since we are taking something out of the error term, adding a variable generally reduces the error variance. Generally, we cannot know which effect will dominate.

However, there is one case that is obvious: we should always include independent variables that affect *y* and are *uncorrelated* with all of the independent variables of interest. The reason for this inclusion is simple: adding such a variable does not induce multicollinearity in the population (and therefore multicollinearity in the sample should be negligible), but it will reduce the error variance. In large sample sizes, the standard errors of all OLS estimators will be reduced.

As an example, consider estimating the individual demand for beer as a function of the average county beer price. It may be reasonable to assume that individual characteristics are uncorrelated with county-level prices, and so a simple regression of beer consumption on county price would suffice for estimating the effect of price on individual demand. But it is possible to get a more precise estimate of the price elasticity of beer demand by including individual characteristics, such as age and amount of education. If these factors affect demand and are uncorrelated with price, then the standard error of the price variable will be smaller, at least in large samples.

Unfortunately, cases where we have information on additional explanatory variables that are uncorrelated with the explanatory variables of interest are rare in the social sciences. But it is worth remembering that when these variables are available, they can be included in a model to reduce the error variance without inducing multicollinearity.

## 6.4 PREDICTION AND RESIDUAL ANALYSIS

In Chapter 3, we defined the OLS predicted or fitted values and the OLS residuals. Predictions are certainly useful, but they are subject to sampling variation, since they are obtained using the OLS estimators. Thus, in this section, we show how to obtain confidence intervals for a prediction from the OLS regression line.

From Chapters 3 and 4, we know that the residuals are used to obtain the sum of squared residuals and the *R*-squared, so they are important for goodness-of-fit and testing. Sometimes economists study the residuals for particular observations to learn about individuals (or firms, houses, etc.) in the sample.

7/14/99 5:33 PM Page 198



## Confidence Intervals for Predictions

Suppose we have estimated the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{6.27}$$

When we plug in particular values of the independent variables, we obtain a prediction for $y$, which is an estimate of the *expected value* of $y$ given the particular values for the explanatory variables. For emphasis, let $c_1, c_2, \ldots, c_k$ denote particular values for each of the $k$ independent variables; these may or may not correspond to an actual data point in our sample. The parameter we would like to estimate is

$$\begin{aligned} \theta_0 &= \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k \\ &= \mathrm{E}(y|x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k). \end{aligned} \tag{6.28}$$

The estimator of $\theta_0$ is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k. \tag{6.29}$$

In practice, this is easy to compute. But what if we want some measure of the uncertainty in this predicted value? It is natural to construct a confidence interval for $\theta_0$, which is centered at $\hat{\theta}_0$.

To obtain a confidence interval for $\theta_0$, we need a standard error for $\hat{\theta}_0$. Then, with a large *df*, we can construct a 95% confidence interval using the rule of thumb $\hat{\theta}_0 \pm 2 \cdot \mathrm{se}(\hat{\theta}_0)$. (As always, we can use the exact percentiles in a $t$ distribution.)

How do we obtain the standard error of $\hat{\theta}_0$? This is the same problem we encountered in Section 4.4: we need to obtain a standard error for a linear combination of the OLS estimators. Here, the problem is even more complicated, because all of the OLS estimators generally appear in $\hat{\theta}_0$ (unless some $c_j$ are zero). Nevertheless, the same trick that we used in Section 4.4 will work here. Write $\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k$ and plug this into the equation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

to obtain

$$y = \theta_0 + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \ldots + \beta_k (x_k - c_k) + u. \tag{6.30}$$

In other words, we subtract the value $c_j$ from each observation on $x_j$, and then we run the regression of

$$y_i \text{ on } (x_{i1} - c_1), \ldots, (x_{ik} - c_k), \; i = 1,2, \ldots, n. \tag{6.31}$$

The predicted value in (6.29) and, more importantly, its standard error, are obtained from the *intercept* (or constant) in regression (6.31).

As an example, we obtain a confidence interval for a prediction from a college GPA regression, where we use high school information.

7/14/99 5:33 PM Page 199



**EXAMPLE 6.5**
(Confidence Interval for Predicted College GPA)

Using the data in GPA2.RAW, we obtain the following equation for predicting college GPA:

$$\widehat{colgpa} = \underset{(0.075)}{1.493} + \underset{(.00007)}{.00149}\ sat - \underset{(.00056)}{.01386}\ hsperc - \underset{(.01650)}{.06088}\ hsize + \underset{(.00227)}{.00546}\ hsize^2 \qquad (6.32)$$
$$n = 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560,$$

where we have reported estimates to several digits to reduce round-off error. What is predicted college GPA, when *sat* = 1,200, *hsperc* = 30, and *hsize* = 5 (which means 500)? This is easy to get by plugging these values into equation (6.32): $\widehat{colgpa} = 2.70$ (rounded to two digits). Unfortunately, we cannot use equation (6.32) directly to get a confidence interval for the expected *colgpa* at the given values of the independent variables. One simple way to obtain a confidence interval is to define a new set of independent variables: *sat0* = *sat* − 1,200, *hsperc0* = *hsperc* − 30, *hsize0* = *hsize* − 5, and $hsizesq0 = hsize^2 - 25$. When we regress *colgpa* on these new independent variables, we get

$$\widehat{colgpa} = \underset{(0.020)}{2.700} + \underset{(.00007)}{.00149}\ sat0 - \underset{(.00056)}{.01386}\ hsperc0 - \underset{(.01650)}{.06088}\ hsize0 + \underset{(.00227)}{.00546}\ hsizesq0$$
$$n = 4{,}137,\ R^2 = .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560.$$

The only difference between this regression and that in (6.32) is the intercept, which is the prediction we want, along with its standard error, .020. It is not an accident that the slope coefficents, their standard errors, *R*-squared, and so on are the same as before; this provides a way to check that the proper transformations were done. We can easily construct a 95% confidence interval for the expected college GPA: 2.70 ± 1.96(.020) or about 2.66 to 2.74. This confidence interval is rather narrow due to the very large sample size.

Because the variance of the intercept estimator is smallest when each explanatory variable has zero sample mean (see Question 2.5 for the simple regression case), it follows from the regression in (6.31) that the variance of the prediction is smallest at the mean values of the $x_j$. (That is, $c_j = \bar{x}_j$ for all $j$.) This result is not too surprising, since we have the most faith in our regression line near the middle of the data. As the values of the $c_j$ get farther away from the $\bar{x}_j$, $\text{Var}(\hat{y})$ gets larger and larger.

The previous method allows us to put a confidence interval around the OLS estimate of $\text{E}(y|x_1,\ldots,x_k)$, for any values of the explanatory variables. But this is not the same as obtaining a confidence interval for a new, as yet unknown, outcome on $y$. In forming a confidence interval for an outcome on $y$, we must account for another very important source of variation: the variance in the unobserved error.

7/14/99 5:33 PM Page 200



Let $y^0$ denote the value for which we would like to construct a confidence interval, which we sometimes call a **prediction interval**. For example, $y^0$ could represent a person or firm not in our original sample. Let $x_1^0, \ldots, x_k^0$ be the new values of the independent variables, which we assume we observe, and let $u^0$ be the unobserved error. Therefore, we have

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \ldots + \beta_k x_k^0 + u^0. \tag{6.33}$$

As before, our best prediction of $y^0$ is the expected value of $y^0$ given the explanatory variables, which we estimate from the OLS regression line: $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \hat{\beta}_2 x_2^0 + \ldots + \hat{\beta}_k x_k^0$. The **prediction error** in using $\hat{y}^0$ to predict $y^0$ is

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0) + u^0 - \hat{y}_0. \tag{6.34}$$

Now, $\mathrm{E}(\hat{y}^0) = \mathrm{E}(\hat{\beta}_0) + \mathrm{E}(\hat{\beta}_1)x_1^0 + \mathrm{E}(\hat{\beta}_2)x_2^0 + \ldots + \mathrm{E}(\hat{\beta}_k)x_k^0 = \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0$, because the $\hat{\beta}_j$ are unbiased. (As before, these expectations are all conditional on the sample values of the independent variables.) Because $u^0$ has zero mean, $\mathrm{E}(\hat{e}^0) = 0$. We have showed that the expected prediction error is zero.

In finding the variance of $\hat{e}^0$, note that $u^0$ is uncorrelated with each $\hat{\beta}_j$, because $u^0$ is uncorrelated with the errors in the sample used to obtain the $\hat{\beta}_j$. By basic properties of covariance (see Appendix B), $u^0$ and $\hat{y}^0$ are uncorrelated. Therefore, the **variance of the prediction error** (conditional on all in-sample values of the independent variables) is the sum of the variances:

$$\mathrm{Var}(\hat{e}^0) = \mathrm{Var}(\hat{y}^0) + \mathrm{Var}(u^0) = \mathrm{Var}(\hat{y}^0) + \sigma^2, \tag{6.35}$$

where $\sigma^2 = \mathrm{Var}(u^0)$ is the error variance. There are two sources of variance in $\hat{e}^0$. The first is the sampling error in $\hat{y}^0$, which arises because we have estimated the $\beta_j$. Because each $\hat{\beta}_j$ has a variance proportional to $1/n$, where $n$ is the sample size, $\mathrm{Var}(\hat{y}^0)$ is proportional to $1/n$. This means that, for large samples, $\mathrm{Var}(\hat{y}^0)$ can be very small. By contrast, $\sigma^2$ is the variance of the error in the population; it does not change with the sample size. In many examples, $\sigma^2$ will be the dominant term in (6.35).

Under the classical linear model assumptions, the $\hat{\beta}_j$ and $u^0$ are normally distributed, and so $\hat{e}^0$ is also normally distributed (conditional on all sample values of the explanatory variables). Earlier, we described how to obtain an unbiased estimator of $\mathrm{Var}(\hat{y}^0)$, and we obtained our unbiased estimator of $\sigma^2$ in Chapter 3. By using these estimators, we can define the standard error of $\hat{e}^0$ as

$$\mathrm{se}(\hat{e}^0) = \{[\mathrm{se}(\hat{y}_0)]^2 + \hat{\sigma}^2\}^{1/2}. \tag{6.36}$$

Using the same reasoning for the $t$ statistics of the $\hat{\beta}_j$, $\hat{e}^0/\mathrm{se}(\hat{e}^0)$ has a $t$ distribution with $n - (k + 1)$ degrees of freedom. Therefore,

$$\mathrm{P}[-t_{.025} \leq \hat{e}^0/\mathrm{se}(\hat{e}^0) \leq t_{.025}] = .95,$$

where $t_{.025}$ is the 97.5th percentile in the $t_{n-k-1}$ distribution. For large $n - k - 1$, remember that $t_{.025} \approx 1.96$. Plugging in $\hat{e}^0 = y^0 - \hat{y}^0$ and rearranging gives a *95% prediction interval* for $y^0$:

7/14/99 5:33 PM Page 201



$$\hat{y}^0 \pm t_{.025} \cdot \text{se}(\hat{e}^0); \qquad \textbf{(6.37)}$$

as usual, except for small *df*, a good rule of thumb is $\hat{y}^0 \pm 2\text{se}(\hat{e}^0)$. This is wider than the confidence interval for $\hat{y}^0$ itself, because of $\hat{\sigma}^2$ in (6.36); it often is much wider to reflect the factors in $u^0$ that we have not controlled for.

---

**EXAMPLE 6.6**
(Confidence Interval for Future College GPA)

Suppose we want a 95% CI for the future college GPA for a high school student with *sat* = 1,200, *hsperc* = 30, and *hsize* = 5. Remember, in Example 6.5 we obtained a confidence interval for the *expected* GPA; now we must account for the unobserved factors in the error term. We have everything we need to obtain a CI for *colgpa*. $\text{se}(\hat{y}^0) = .020$ and $\hat{\sigma} = .560$ and so, from (6.36), $\text{se}(\hat{e}^0) = [(.020)^2 + (.560)^2]^{1/2} \approx .560$. Notice how small $\text{se}(\hat{y}^0)$ is relative to $\hat{\sigma}$: virtually all of the variation in $\hat{e}^0$ comes from the variation in $u^0$. The 95% CI is $2.70 \pm 1.96(.560)$ or about 1.60 to 3.80. This is a wide confidence interval, and it shows that, based on the factors used in the regression, we cannot significantly narrow the likely range of college GPA.

---

## Residual Analysis

Sometimes it is useful to examine individual observations to see whether the actual value of the dependent variable is above or below the predicted value; that is, to examine the residuals for the individual observations. This process is called **residual analysis**. Economists have been known to examine the residuals from a regression in order to aid in the purchase of a home. The following housing price example illustrates residual analysis. Housing price is related to various observable characteristics of the house. We can list all of the characteristics that we find important, such as size, number of bedrooms, number of bathrooms, and so on. We can use a sample of houses to estimate a relationship between price and attributes, where we end up with a predicted value and an actual value for each house. Then, we can construct the residuals, $\hat{u}_i = y_i - \hat{y}_i$. The house with the most negative residual is, at least based on the factors we have controlled for, the most underpriced one relative to its characteristics. It also makes sense to compute a confidence interval for what the future selling price of the home could be, using the method described in equation (6.37).

Using the data in HPRICE1.RAW, we run a regression of *price* on *lotsize*, *sqrft*, and *bdrms*. In the sample of 88 homes, the most negative residual is −120.206, for the 81st house. Therefore, the asking price for this house is $120,206 below its predicted price.

There are many other uses of residual analysis. One way to rank law schools is to regress median starting salary on a variety of student characteristics (such as median LSAT scores of entering class, median college GPA of entering class, and so on) and to obtain a predicted value and residual for each law school. The law school with the largest residual has the highest predicted value added. (Of course, there is still much uncertainty about how an individual's starting salary would compare with the median for a law school overall.) These residuals can be used along with the costs of attending

7/14/99 5:33 PM Page 202



each law school to determine the best value; this would require an appropriate discounting of future earnings.

Residual analysis also plays a role in legal decisions. A *New York Times* article entitled "Judge Says Pupil's Poverty, Not Segregation, Hurts Scores" (6/28/95) describes an important legal case. The issue was whether the poor performance on standardized tests in the Hartford School District, relative to performance in surrounding suburbs, was due to poor school quality at the highly segregated schools. The judge concluded that "the disparity in test scores does not indicate that Hartford is doing an inadequate or poor job in educating its students or that its schools are failing, because the predicted scores based upon the relevant socioeconomic factors are about at the levels that one would expect." This conclusion is almost certainly based on a regression analysis of average or median scores on socioeconomic characteristics of various school districts in Connecticut. The judge's conclusion suggests that, given the poverty levels of students at Hartford schools, the actual test scores were similar to those predicted from a regression analysis: the residual for Hartford was not sufficiently negative to conclude that the schools themselves were the cause of low test scores.

**QUESTION 6.5**

How might you use residual analysis to determine which movie actors are overpaid relative to box office production?

## Predicting *y* When log(*y*) Is the Dependent Variable

Since the natural log transformation is used so often for the dependent variable in empirical economics, we devote this subsection to the issue of predicting $y$ when $\log(y)$ is the dependent variable. As a byproduct, we will obtain a goodness-of-fit measure for the log model that can be compared with the $R$-squared from the level model.

To obtain a prediction, it is useful to define $logy = \log(y)$; this emphasizes that it is the log of $y$ that is predicted in the model

$$logy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \qquad (6.38)$$

In this equation, the $x_j$ might be transformations of other variables; for example, we could have $x_1 = \log(sales)$, $x_2 = \log(mktval)$, $x_3 = ceoten$ in the CEO salary example.

Given the OLS estimators, we know how to predict *logy* for any value of the independent variables:

$$\widehat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \qquad (6.39)$$

Now, since the exponential undoes the log, our first guess for predicting $y$ is to simply exponentiate the predicted value for $\log(y)$: $\hat{y} = \exp(\widehat{logy})$. This does not work; in fact, it will systematically *underestimate* the expected value of $y$. In fact, if model (6.38) follows the CLM assumptions MLR.1 through MLR.6, it can be shown that

$$\mathrm{E}(y|\boldsymbol{x}) = \exp(\sigma^2/2)\cdot\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k),$$

where $\boldsymbol{x}$ denotes the independent variables and $\sigma^2$ is the variance of $u$. [If $u \sim \text{Normal}(0,\sigma^2)$, then the expected value of $\exp(u)$ is $\exp(\sigma^2/2)$.] This equation shows that a simple adjustment is needed to predict $y$:

7/14/99 5:33 PM Page 203



$$\hat{y} = \exp(\hat{\sigma}^2/2)\exp(\hat{logy}), \tag{6.40}$$

where $\hat{\sigma}^2$ is simply the unbiased estimator of $\sigma^2$. Since $\hat{\sigma}$, the standard error of the regression, is always reported, obtaining predicted values for $y$ is easy. Because $\hat{\sigma}^2 > 0$, $\exp(\hat{\sigma}^2/2) > 1$. For large $\hat{\sigma}^2$, this adjustment factor can be substantially larger than unity.

The prediction in (6.40) is not unbiased, but it is consistent. There are no unbiased predictions of $y$, and in many cases, (6.40) works well. However, it does rely on the normality of the error term, $u$. In Chapter 5, we showed that OLS has desirable properties, even when $u$ is not normally distributed. Therefore, it is useful to have a prediction that does not rely on normality. If we just assume that $u$ is independent of the explanatory variables, then we have

$$\mathrm{E}(y|\boldsymbol{x}) = \alpha_0\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k), \tag{6.41}$$

where $\alpha_0$ is the expected value of $\exp(u)$, which must be greater than unity.

Given an estimate $\hat{\alpha}_0$, we can predict $y$ as

$$\hat{y} = \hat{\alpha}_0\exp(\hat{logy}), \tag{6.42}$$

which again simply requires exponentiating the predicted value from the log model and multiplying the result by $\hat{\alpha}_0$.

It turns out that a consistent estimator of $\hat{\alpha}_0$ is easily obtained.

**PREDICTING *y* WHEN THE DEPENDENT VARIABLE IS log(*y*):**

(i) Obtain the fitted values $\hat{logy}_i$ from the regression of $logy$ on $x_1, \ldots, x_k$.
(ii) For each observation $i$, create $\hat{m}_i = \exp(\hat{logy}_i)$.
(iii) Now regress $y$ on the single variable $\hat{m}$ *without* an intercept; that is, perform a simple regression through the origin. The coefficient on $\hat{m}$, the only coefficient there is, is the estimate of $\alpha_0$.

Once $\hat{\alpha}_0$ is obtained, it can be used along with predictions of $logy$ to predict $y$. The steps are as follows:

(i) For given values of $x_1, x_2, \ldots, x_k$, obtain $\hat{logy}$ from (6.39).
(ii) Obtain the prediction $\hat{y}$ from (6.42).

---

**EXAMPLE 6.7**
(Predicting CEO Salaries)

The model of interest is

$$\log(salary) = \beta_0 + \beta_1\log(sales) + \beta_2\log(mktval) + \beta_3 ceoten + u,$$

so that $\beta_1$ and $\beta_2$ are elasticities and $100\cdot\beta_3$ is a semi-elasticity. The estimated equation using CEOSAL2.RAW is

7/14/99 5:33 PM Page 204



$$\widehat{lsalary} = \underset{(0.257)}{4.504} + \underset{(.039)}{.163}\, lsales + \underset{(.050)}{.109}\, lmktval + \underset{(.0053)}{.0117}\, ceoten \qquad \textbf{(6.43)}$$
$$n = 177,\ R^2 = .318,$$

where, for clarity, we let *lsalary* denote the log of *salary*, and similarly for *lsales* and *lmktval*. Next, we obtain $\hat{m}_i = \exp(\widehat{lsalary}_i)$ for each observation in the sample. Regressing *salary* on $\hat{m}$ (without a constant) produces $\hat{\alpha}_0 \approx 1.117$.

We can use this value of $\hat{\alpha}_0$ along with (6.43) to predict *salary* for any values of *sales*, *mktval*, and *ceoten*. Let us find the prediction for *sales* = 5,000 (which means \$5 billion, since *sales* is in millions of dollars), *mktval* = 10,000 (or \$10 billion), and *ceoten* = 10. From (6.43), the prediction for *lsalary* is 4.504 + .163·log(5,000) + .109·log(10,000) + .0117(10) ≈ 7.013. The predicted salary is therefore 1.117·exp(7.013) ≈ 1,240.967, or \$1,240,967. If we forget to multiply by $\hat{\alpha}_0 = 1.117$, we get a prediction of \$1,110,983.

---

We can use the previous method of obtaining predictions to determine how well the model with log($y$) as the dependent variable explains $y$. We already have measures for models when $y$ is the dependent variable: the $R$-squared and the adjusted $R$-squared. The goal is to find a goodness-of-fit measure in the log($y$) model that can be compared with an $R$-squared from a model where $y$ is the dependent variable.

There are several ways to find this measure, but we present an approach that is easy to implement. After running the regression of $y$ on $\hat{m}$ through the origin in step (iii), we obtain the fitted values for this regression, $\hat{y}_i = \hat{\alpha}_0 \hat{m}_i$. Then, we find the sample correlation between $\hat{y}_i$ and the actual $y_i$ in the sample. The *square* of this *can* be compared with the $R$-squared we get by using $y$ as the dependent variable in a linear regression model. Remember that the $R$-squared in the fitted equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

is just the squared correlation between the $y_i$ and the $\hat{y}_i$ (see Section 3.2).

---

**EXAMPLE 6.8**
(Predicting CEO Salaries)

After step (iii) in the preceding procedure, we obtain the fitted values $\widehat{salary}_i = \hat{\alpha}_0 \hat{m}_i$. The simple correlation between $salary_i$ and $\widehat{salary}_i$ in the sample is .493; the square of this value is about .243. This is our measure of how much *salary* variation is explained by the log model; it is *not* the $R$-squared from (6.43), which is .318.

Suppose we estimate a model with all variables in levels:

$$salary = \beta_0 + \beta_1 sales + \beta_2 mktval + \beta_3 ceoten + u.$$

The $R$-squared obtained from estimating this model using the same 177 observations is .201. Thus, the log model explains more of the variation in *salary*, and so we prefer it on goodness-of-fit grounds. The log model is also chosen because it seems more realistic and the parameters are easier to interpret.

---

7/14/99 5:33 PM Page 205



## SUMMARY

In this chapter, we have covered some important multiple regression analysis topics.

Section 6.1 showed that a change in the units of measurement of an independent variable changes the OLS coefficient in the expected manner: if $x_j$ is multiplied by $c$, its coefficient is divided by $c$. If the dependent variable is multiplied by $c$, *all* OLS coefficients are multiplied by $c$. Neither $t$ nor $F$ statistics are affected by changing the units of measurement of any variables.

We discussed beta coefficients, which measure the effects of the independent variables on the dependent variable in standard deviation units. The beta coefficients are obtained from a standard OLS regression after the dependent and independent variables have been transformed into $z$-scores.

As we have seen in several examples, the logarithmic functional form provides coefficients with percentage effect interpretations. We discussed its additional advantages in Section 6.2. We also saw how to compute the exact percentage effect when a coefficient in a log-level model is large. Models with quadratics allow for either diminishing or increasing marginal effects. Models with interactions allow the marginal effect of one explanatory variable to depend upon the level of another explanatory variable.

We introduced the adjusted $R$-squared, $\bar{R}^2$, as an alternative to the usual $R$-squared for measuring goodness-of-fit. While $R^2$ can never fall when another variable is added to a regression, $\bar{R}^2$ penalizes the number of regressors and can drop when an independent variable is added. This makes $\bar{R}^2$ preferable for choosing between nonnested models with different numbers of explanatory variables. Neither $R^2$ nor $\bar{R}^2$ can be used to compare models with different dependent variables. Nevertheless, it is fairly easy to obtain goodness-of-fit measures for choosing between $y$ and $\log(y)$ as the dependent variable, as shown in Section 6.4.

In Section 6.3, we discussed the somewhat subtle problem of relying too much on $R^2$ or $\bar{R}^2$ in arriving at a final model: it is possible to control for too many factors in a regression model. For this reason, it is important to think ahead about model specification, particularly the ceteris paribus nature of the multiple regression equation. Explanatory variables that affect $y$ and are uncorrelated with all the other explanatory variables can be used to reduce the error variance without inducing multicollinearity.

In Section 6.4, we demonstrated how to obtain a confidence interval for a prediction made from an OLS regression line. We also showed how a confidence interval can be constructed for a future, unknown value of $y$.

Occasionally, we want to predict $y$ when $\log(y)$ is used as the dependent variable in a regression model. Section 6.4 explains this simple method. Finally, we are sometimes interested in knowing about the sign and magnitude of the residuals for particular observations. Residual analysis can be used to determine whether particular members of the sample have predicted values that are well above or well below the actual outcomes.

## KEY TERMS

Adjusted $R$-Squared
Beta Coefficients
Interaction Effect
Nonnested Models
Population $R$-Squared
Predictions

7/14/99 5:33 PM Page 206



Prediction Error
Prediction Interval
Quadratic Functions
Residual Analysis
Standardized Coefficients
Variance of the Prediction Error

## PROBLEMS

**6.1** The following equation was estimated using the data in CEOSAL1.RAW:

$$\widehat{\log(salary)} = \underset{(.324)}{4.322} + \underset{(.033)}{.276}\log(sales) + \underset{(.0129)}{.0215}\,roe - \underset{(.00026)}{.00008}\,roe^2$$
$$n = 209,\ R^2 = .282.$$

This equation allows *roe* to have a diminishing effect on log(*salary*). Is this generality necessary? Explain why or why not.

**6.2** Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimates from the regression of $y_i$ on $x_{i1}, \ldots, x_{ik}$, $i = 1,2, \ldots, n$. For nonzero constants $c_1, \ldots, c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, \ldots, c_k x_{ik}$, $i = 1,2, \ldots, n$, are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0$, $\tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0/c_k)\hat{\beta}_k$. (*Hint*: Use the fact that the $\hat{\beta}_j$ solve the first order conditions in (3.13), and the $\tilde{\beta}_j$ must solve the first order conditions involving the rescaled dependent and independent variables.)

**6.3** Using the data in RDCHEM.RAW, the following equation was obtained by OLS:

$$\widehat{rdintens} = \underset{(0.429)}{2.613} + \underset{(.00014)}{.00030}\,sales - \underset{(.0000000037)}{.0000000070}\,sales^2$$
$$n = 32,\ R^2 = .1484.$$

(i) At what point does the marginal effect of *sales* on *rdintens* become negative?
(ii) Would you keep the quadratic term in the model? Explain.
(iii) Define *salesbil* as sales measured in billions of dollars: *salesbil* = *sales*/1,000. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the *R*-squared. [*Hint*: Note that $salesbil^2 = sales^2/(1{,}000)^2$.]
(iv) For the purpose of reporting the results, which equation do you prefer?

**6.4** The following model allows the return to education to depend upon the total amount of both parents' education, called *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

(i) Show that, in decimal form, the return to another year of education in this model is

$$\Delta\log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

What sign do you expect for $\beta_2$? Why?
(ii) Using the data in WAGE2.RAW, the estimated equation is

7/14/99 5:33 PM Page 207



$$\widehat{\log(wage)} = \underset{(0.13)}{5.65} + \underset{(.010)}{.047}\ educ + \underset{(.00021)}{.00078}\ educ\cdot pareduc +$$
$$\underset{(.004)}{.019}\ exper + \underset{(.003)}{.010}\ tenure$$
$$n = 722,\ R^2 = .169.$$

(Only 722 observations contain full information on parents' education.) Interpret the coefficient on the interaction term. It might help to choose two specific values for *pareduc*—for example, *pareduc* = 32 if both parents have a college education, or *pareduc* = 24 if both parents have a high school education—and to compare the estimated return to *educ*.

(iii) When *pareduc* is added as a separate variable to the equation, we get:

$$\widehat{\log(wage)} = \underset{(0.38)}{4.94} + \underset{(.027)}{.097}\ educ + \underset{(.017)}{.033}\ pareduc - \underset{(.0012)}{.0016}\ educ\cdot pareduc$$
$$+ \underset{(.004)}{.020}\ exper + \underset{(.003)}{.010}\ tenure$$
$$n = 722,\ R^2 = .174.$$

Does the estimated return to education now depend positively on parent education? Test the null hypothesis that the return to education does not depend on parent education.

**6.5** In Example 4.2, where the percentage of students receiving a passing score on a 10th grade math exam (*math10*) is the dependent variable, does it make sense to include *sci11*—the percentage of 11th graders passing a science exam—as an additional explanatory variable?

**6.6** When $atndrte^2$ and *ACT*·*atndrte* are added to the equation estimated in (6.19), the *R*-squared becomes .232. Are these additional terms jointly significant at the 10% level? Would you include them in the model?

**6.7** The following three equations were estimated using the 1,534 observations in 401K.RAW:

$$\widehat{prate} = \underset{(0.78)}{80.29} + \underset{(0.52)}{5.44}\ mrate + \underset{(.045)}{.269}\ age - \underset{(.00004)}{.00013}\ totemp$$
$$R^2 = .100,\ \bar{R}^2 = .098.$$

$$\widehat{prate} = \underset{(1.95)}{97.32} + \underset{(0.51)}{5.02}\ mrate + \underset{(.044)}{.314}\ age - \underset{(0.28)}{2.66}\ \log(totemp)$$
$$R^2 = .144,\ \bar{R}^2 = .142.$$

$$\widehat{prate} = \underset{(0.78)}{80.62} + \underset{(0.52)}{5.34}\ mrate + \underset{(.045)}{.290}\ age - \underset{(.00009)}{.00043}\ totemp$$
$$+ \underset{(.0000000010)}{.0000000039}\ totemp^2$$
$$R^2 = .108,\ \bar{R}^2 = .106.$$

Which of these three models do you prefer. Why?

7/14/99 5:33 PM Page 208



## COMPUTER EXERCISES

**6.8** Use the data in HPRICE3.RAW, only for the year 1981, to answer the following questions. The data are for houses that sold during 1981 in North Andover, MA; 1981 was the year construction began on a local garbage incinerator.

(i) To study the effects of the incinerator location on housing price, consider the simple regression model

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

where *price* is housing price in dollars and *dist* is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for $\beta_1$ if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.

(ii) To the simple regression model in part (i), add the variables log(*inst*), log(*area*), log(*land*), *rooms*, *baths*, and *age*, where *inst* is distance from the home to the interstate, *area* is square footage of the house, *land* is the lot size in square feet, *rooms* is total number of rooms, *baths* is number of bathrooms, and *age* is age of the house in years. Now what do you conclude about the effects of the incinerator? Explain why (i) and (ii) give conflicting results.

(iii) Add $[\log(inst)]^2$ to the model from part (ii). Now what happens? What do you conclude about the importance of functional form?

(iv) Is the square of log(*dist*) significant when you add it to the model from part (iii)?

**6.9** Use the data in WAGE1.RAW for this exercise.

(i) Use OLS to estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

and report the results using the usual format.

(ii) Is $exper^2$ statistically significant at the 1% level?

(iii) Using the approximation

$$\%\Delta \hat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

(iv) At what value of *exper* does additional experience actually lower predicted log(*wage*)? How many people have more experience in this sample?

**6.10** Consider a model where the return to education depends upon the amount of work experience (and vice versa):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u.$$

(i) Show that the return to another year of education (in decimal form), holding *exper* fixed, is $\beta_1 + \beta_3 exper$.

7/14/99 5:33 PM Page 209



(ii) State the null hypothesis that the return to education does not depend on the level of *exper*. What do you think is the appropriate alternative?

(iii) Use the data in WAGE2.RAW to test the null hypothesis in (ii) against your stated alternative.

(iv) Let $\theta_1$ denote the return to education (in decimal form), when *exper* = 10: $\theta_1 = \beta_1 + 10\beta_3$. Obtain $\hat{\theta}_1$ and a 95% confidence interval for $\theta_1$. (*Hint*: Write $\beta_1 = \theta_1 - 10\beta_3$ and plug this into the equation; then rearrange. This gives the regression for obtaining the confidence interval for $\theta_1$.)

**6.11** Use the data in GPA2.RAW for this exercise.

(i) Estimate the model

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

where *hsize* is size of graduating class (in hundreds), and write the results in the usual form. Is the quadratic term statistically significant?

(ii) Using the estimated equation from part (i), what is the "optimal" high school size? Justify your answer.

(iii) Is this analysis representative of the academic performance of *all* high school seniors? Explain.

(iv) Find the estimated optimal high school size, using log(*sat*) as the dependent variable. Is it much different from what you obtained in part (ii)?

**6.12** Use the housing price data in HPRICE1.RAW for this exercise.

(i) Estimate the model

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

and report the results in the usual OLS format.

(ii) Find the predicted value of log(*price*), when *lotsize* = 20,000, *sqrft* = 2,500, and *bdrms* = 4. Using the methods in Section 6.4, find the predicted value of *price* at the same values of the explanatory variables.

(iii) For explaining variation in *price*, decide whether you prefer the model from part (i) or the model

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u.$$

**6.13** Use the data in VOTE1.RAW for this exercise.

(i) Consider a model with an interaction between expenditures:

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u.$$

What is the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed? What is the partial effect of *expendA* on *voteA*? Is the expected sign for $\beta_4$ obvious?

(ii) Estimate the equation in part (i) and report the results in the usual form. Is the interaction term statistically significant?

(iii) Find the average of *expendA* in the sample. Fix *expendA* at 300 (for \$300,000). What is the estimated effect of another \$100,000 dollars spent by Candidate B on *voteA*? Is this a large effect?

7/14/99 5:33 PM Page 210



(iv) Now fix *expendB* at 100. What is the estimated effect of $\Delta expendA = 100$ on *voteA*. Does this make sense?

(v) Now estimate a model that replaces the interaction with *shareA*, Candidate A's percentage share of total campaign expenditures. Does it make sense to hold both *expendA* and *expendB* fixed, while changing *shareA*?

(vi) (Requires calculus) In the model from part (v), find the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed. Evaluate this at $expendA = 300$ and $expendB = 0$ and comment on the results.

**6.14** Use the data in ATTEND.RAW for this exercise.

(i) In the model of Example 6.3, argue that

$$\Delta stndfnl/\Delta priGPA \approx \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte.$$

Use equation (6.19) to estimate the partial effect, when $priGPA = 2.59$ and $atndrte = .82$. Interpret your estimate.

(ii) Show that the equation can be written as

$$stndfnl = \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - .82) + u,$$

where $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(.82)$. (Note that the intercept has changed, but this is unimportant.) Use this to obtain the standard error of $\hat{\theta}_2$ from part (i).

**6.15** Use the data in HPRICE1.RAW for this exercise.

(i) Estimate the model

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

and report the results in the usual form, including the standard error of the regression. Obtain predicted price, when we plug in $lotsize = 10{,}000$, $sqrft = 2{,}300$, and $bdrms = 4$; round this price to the nearest dollar.

(ii) Run a regression that allows you to put a 95% confidence interval around the predicted value in part (i). Note that your prediction will differ somewhat due to rounding error.

(iii) Let $price^0$ be the unknown future selling price of the house with the characteristics used in parts (i) and (ii). Find a 95% CI for $price^0$ and comment on the width of this confidence interval.