Adjusted *R*-Squared

Most regression packages will report, along with the R-squared, a statistic called the **adjusted *R*-squared**. Because the adjusted R-squared is reported in much applied work, and because 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

$$R^2 = 1 - (\mathrm{SSR}/n)/(\mathrm{SST}/n), \qquad [6.20]$$

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 σ_y^2 as the population variance of y and let σ_u^2 denote the population variance of the error term, u. (Until now, we have used σ^2 to denote σ_u^2, but it is helpful to be more specific here.) The **population *R*-squared** is defined as $\rho^2 = 1 - \sigma_u^2/\sigma_y^2$; 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 σ_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 σ_y^2. Using these estimators, we arrive at the adjusted R-squared:

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

because $\hat{\sigma}^2 = \mathrm{SSR}/(n-k-1)$. 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, ρ^2, 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, $n - k - 1$. 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 (because 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

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

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 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 4.5, 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 equations 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 4.5). 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 Computer Exercise C5 in Chapter 4. (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:

$$rdintens = \beta_0 + \beta_1 \log(sales) + u. \qquad [6.23]$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u. \qquad [6.24]$$

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 fewer 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.

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 purpose. 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, and are often very different. 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.

EXPLORING FURTHER 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:

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

and

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

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.

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

$$\begin{aligned} miles &= \text{total miles driven.} \\ percmale &= \text{percentage of the state population that is male.} \\ perc16\_21 &= \text{percentage of the population between ages 16 and 21, and so on.} \end{aligned}$$

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

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

β_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.

As a second example, suppose that, for a developing country, we want to estimate the effects of pesticide usage among farmers on family health expenditures. In addition to pesticide usage amounts, should we include the number of doctor visits as an explanatory variable? No. Health expenditures include doctor visits, and we would like to pick up all effects of pesticide use on health expenditures. If we include the number of doctor visits as an explanatory variable, then we are only measuring the effects of pesticide use on health expenditures other than doctor visits. It makes more sense to use number of doctor visits as a dependent variable in a separate regression on pesticide amounts.

The previous examples are what can be called **over controlling** for factors in multiple regression. Often this results from nervousness about potential biases that might arise by leaving out an important explanatory variable. But it is important to remember the ceteris paribus nature of multiple regression. In some cases, it makes no sense to hold some factors fixed precisely because they should be allowed to change when a policy variable changes.

Unfortunately, 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 pursuing 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 change the purpose of the analysis and estimate a *hedonic price model*, which allows us to obtain the marginal values of various housing attributes? Should we include $\log(assess)$ in the equation? The adjusted R-squared from the regression 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 clear: we should always include independent variables that affect y and are *uncorrelated* with all of the independent variables of interest. Why?