CHAPTER 6

Specification: Choosing the Independent Variables

Before any equation can be estimated, it must be completely *specified.* Specifying an econometric equation consists of three parts: choosing the correct independent variables, the correct functional form, and the correct form of the stochastic error term.

A **specification error** results when any one of these choices is made incorrectly. This chapter is concerned with only the first of these, choosing the variables; the second and third will be taken up in later chapters.

That researchers can decide which independent variables to include in regression equations is a source of both strength and weakness in econometrics. The strength is that the equations can be formulated to fit individual needs, but the weakness is that researchers can estimate many different specifications until they find the one that "proves" their point, even if many other results disprove it. A major goal of this chapter is to help you understand how to choose variables for your regressions without falling prey to the various errors that result from misusing the ability to choose.

The primary consideration in deciding if an independent variable belongs in an equation is whether the variable is essential to the regression on the basis of theory. If the answer is an unambiguous yes, then the variable definitely should be included in the equation, even if it seems to be lacking in statistical significance. If theory is ambivalent or less emphatic, a dilemma arises. Leaving a relevant variable out of an equation is likely to bias the remaining estimates, but including an irrelevant variable leads to higher variances of the estimated coefficients. Although we'll develop statistical tools to help us deal with this decision, it's difficult in practice to be sure that a variable is relevant, and so the problem often remains unresolved.

We devote the fourth section of the chapter to specification searches and the pros and cons of various approaches to such searches. For example, techniques like stepwise regression procedures and sequential specification searches often cause bias or make the usual tests of significance inapplicable, and we do not recommend them. Instead, we suggest trying to minimize the number of regressions estimated and relying as much as possible on theory rather than statistical fit when choosing variables. There are no pat answers, however, and so the final decisions must be left to each individual researcher.

6.1 Omitted Variables

Suppose that you forget to include all the relevant independent variables when you first specify an equation (after all, no one's perfect!). Or suppose that you can't get data for one of the variables that you *do* think of. The result in both these situations is an **omitted variable,** defined as an important explanatory variable that has been left out of a regression equation.

Whenever you have an omitted (or *left-out*) variable, the interpretation and use of your estimated equation become suspect. Leaving out a relevant variable, like price from a demand equation, not only prevents you from getting an estimate of the coefficient of price but also usually causes bias in the estimated coefficients of the variables that are in the equation.

The bias caused by leaving a variable out of an equation is called **omitted variable bias** (or, more generally, **specification bias.**) In an equation with more than one independent variable, the coefficient β_k represents the change in the dependent variable Y caused by a one-unit increase in the independent variable X_k, holding constant the other independent variables in the equation. If a variable is omitted, then it is not included as an independent variable, and it is not held constant for the calculation and interpretation of $\hat{\beta}_k$. This omission can cause bias: It can force the expected value of the estimated coefficient away from the true value of the population coefficient.

Thus, omitting a relevant variable is usually evidence that the entire estimated equation is suspect because of the likely bias in the coefficients of the variables that remain in the equation. Let's look at this issue in more detail.

6.1.1 The Consequences of an Omitted Variable

What happens if you omit an important variable from your equation (perhaps because you can't get the data for the variable or didn't even think of the variable in the first place)? The major consequence of omitting a relevant independent variable from an equation is to cause bias in the regression coefficients that remain in the equation. Suppose that the true regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \qquad (6.1)$$

where ϵ_i is a classical error term. If you omit X_2 from the equation then the equation becomes:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i^* \qquad (6.2)$$

where ϵ_i^* equals

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i} \qquad (6.3)$$

because the stochastic error term includes the effects of any omitted variables, as mentioned in Section 1.2.3. From Equations 6.2 and 6.3, it might seem as though we could get unbiased estimates of β_0 and β_1 even if we left X_2 out of the equation. Unfortunately, this is not the case,[1] because the included coefficients almost surely pick up some of the effect of the omitted variable and therefore will change, causing bias. To see why, take another look at Equations 6.2 and 6.3. The error term ϵ_i^* is not independent of the explanatory variable X_{1i}, as long as X_{1i} and X_{2i} are correlated because if X_{2i} changes, both X_{1i} and ϵ_i^* will change. In other words, if we leave an important variable out of an equation, we violate Classical Assumption III (that the explanatory variables are independent of the error term), unless the omitted variable is uncorrelated with all the included independent variables (which is extremely unlikely). Recall that the correlation between X_1 and X_2 can be measured by the simple correlation coefficient between the two variables (r_{12}) using Equation 5.8.

In general, when there is a violation of one of the Classical Assumptions, the Gauss–Markov Theorem does not hold, and the OLS estimates are not BLUE. Given linear estimators, this means that the estimated coefficients are no longer unbiased or are no longer minimum variance (for all linear unbiased estimators), or both. In such a circumstance, econometricians first deter-

[1] For this to be true, X_1 and X_2 must be perfectly uncorrelated and $E(\beta_2 X_{2i})$ must equal zero, both of which are extremely unlikely.

mine the exact property (unbiasedness or minimum variance) that no longer holds and then suggest an alternative estimation technique that might, in some sense, be better than OLS.

An omitted variable causes Classical Assumption III to be violated in a way that causes bias. The estimation of Equation 6.2 when Equation 6.1 is the truth will cause bias in the estimates of Equation 6.2. This means that:

$$E(\hat{\beta}_1^*) \neq \beta_1 \tag{6.4}$$

Instead of having an expected value equal to the true β_1, the estimate will compensate for the fact that X_2 is missing from the equation. If X_1 and X_2 are correlated and X_2 is omitted from the equation, then the OLS program will attribute to X_1 variations in Y actually caused by X_2, and a biased estimate will result.

To see how a left-out variable can cause bias, picture a production function that states that output (Y) depends on the amount of labor (X_1) and capital (X_2) used. What would happen if data on capital were unavailable for some reason and X_2 was omitted from the equation? In this case, we would be leaving out the impact of capital on output in our model. This omission would almost surely bias the estimate of the coefficient of labor because it is likely that capital and labor are positively correlated (an increase in capital usually requires at least some labor to utilize it and vice versa). As a result, the OLS program would attribute to labor the increase in output actually caused by capital to the extent that labor and capital were correlated. Thus the bias would be a function of the impact of capital on output (β_2) and the correlation between capital and labor.

To generalize for a model with two independent variables, the expected value of the coefficient of an included variable (X_1) when a relevant variable (X_2) is omitted from the equation equals:

$$E(\beta_1) = \beta_1 + \beta_2 \cdot \alpha_1 \tag{6.5}$$

where α_1 is the slope coefficient of the secondary regression that relates X_2 to X_1:

$$X_{2i} = \alpha_0 + \alpha_1 X_{1i} + u_i \tag{6.6}$$

where u_i is a classical error term. α_1 can be expressed as a function of the correlation between X_1 and X_2, the included and excluded variables, or $f(r_{12})$.

Let's take a look at Equation 6.5. It states that the expected value of the included variable's coefficient is equal to its true value plus the omitted variable's true coefficient times a function of the correlation between the included (in) and omitted (om) variables.[2] Since the expected value of an unbiased estimate equals the true value, the right-hand term in Equation 6.5 measures the omitted variable bias in the equation:

$$\boxed{\text{Bias} = \beta_2\alpha_1 \quad \text{or} \quad \text{Bias} = \beta_{om} \cdot f(r_{in,om}) \qquad (6.7)}$$

In general terms, the bias thus equals β_{om}, the coefficient of the omitted variable, times $f(r_{in,om})$, a function of the correlation between the included and omitted variables.

This bias exists unless:

1. the true coefficient equals zero or
2. the included and omitted variables are uncorrelated.

The term $\beta_{om}f(r_{in,om})$ is the amount of specification bias introduced into the estimate of the coefficient of the included variable by leaving out the omitted variable. Although it’s true that there is no bias if the included and excluded variables are uncorrelated, there almost always is some correlation between any two variables in the real world (even if it’s just random), and so bias is almost always caused by the omission of a relevant variable.[3]

6.1.2 An Example of Specification Bias

Consider the following equation for the annual consumption of chicken in the United States. (The data for this example are included in Exercise 5; t-scores differ because of rounding.)

$$\begin{aligned}\hat{Y}_t = 31.5 - &0.73PC_t + 0.11PB_t + 0.23YD_t \qquad (6.8)\\ &(0.08) \qquad (0.05) \qquad (0.02)\\ t = &-9.12 \qquad 2.50 \qquad 14.22\\ \bar{R}^2 = .986 \quad &n = 44 \text{ (annual 1951–1994)}\end{aligned}$$

[2] Equation 6.5 is a conditional expectation that holds when there are exactly two independent variables, but the more general equation is quite similar.

[3] Although the omission of a relevant variable almost always produces bias in the estimators of the coefficients of the included variables, the variances of these estimators are generally lower than they otherwise would be. One method of deciding whether this decreased variance in the distribution of the $\hat{\beta}$s is valuable enough to offset the bias is to compare different estimation techniques with a measure called Mean Square Error (MSE). MSE is equal to the variance plus the square of the bias. The lower the MSE, the better.

where: Y_t = per capita chicken consumption (in pounds) in year t
PC_t = the price of chicken (in cents per pound) in year t
PB_t = the price of beef (in cents per pound) in year t
YD_t = U.S. per capita disposable income (in hundreds of dollars) in year t

This equation is a simple demand for chicken equation that includes the prices of chicken and a close substitute (beef) and an income variable. Note that the signs of the estimated coefficients agree with the signs you would have hypothesized before seeing any regression results.

If we estimate this equation without the price of the substitute, we obtain:

$$\begin{aligned} \hat{Y}_t &= 32.9 - \underset{(0.08)}{0.70}PC_t + \underset{(0.01)}{0.27}YD_t \qquad (6.9)\\ t &= \quad -8.33 \qquad\quad 45.91 \\ \bar{R}^2 &= .984 \qquad n = 44 \text{ (annual 1951–1994)} \end{aligned}$$

Let's compare Equations 6.8 and 6.9 to see if dropping the beef price variable had an impact on the estimated equations. If you compare the overall fit, for example, you can see that $\bar{R}^2$ fell slightly from .986 to .984 when PB was dropped, exactly what we'd expect to occur when a relevant variable is omitted.

More important, from the point of view of showing that an omitted variable causes bias, let's see if the coefficient estimates of the remaining variables changed. Sure enough, dropping PB caused $\hat{\beta}_{PC}$ to go from -0.73 to -0.70 and caused $\hat{\beta}_{YD}$ to go from 0.23 to 0.27. The direction of this bias, by the way, is considered positive because the biased coefficient of PC (-0.70) is more positive (less negative) than the suspected unbiased one (-0.73) and the biased coefficient of YD (0.27) is more positive than the suspected unbiased one of (0.23).

The fact that the bias is positive could have been guessed before any regressions were run if Equation 6.7 had been used. The specification bias caused by omitting the price of beef is expected[4] to be positive because the expected sign of the coefficient of PB is positive and because the expected correlation between the price of beef and the price of chicken itself is positive:

[4] It is important to note the distinction between expected bias and any actual observed differences between coefficient estimates. Because of the random nature of the error term (and hence the $\hat{\beta}$s), the change in an estimated coefficient brought about by dropping a relevant variable from the equation will not necessarily be in the expected direction. Biasedness refers to the central tendency of the sampling distribution of the $\hat{\beta}$s, not to every single drawing from that distribution. However, we usually (and justifiably) rely on these general tendencies. Note also that Equation 6.8 has three independent variables whereas Equation 6.7 was derived for use with equations with exactly two. However, Equation 6.7 represents a general tendency that is still applicable.

$$\text{Expected bias in } \hat{\beta}_{PC} = \beta_{PB} \cdot f(r_{PC,PB}) = (+) \cdot (+) = (+)$$

Similarly for YD:

$$\text{Expected bias in } \hat{\beta}_{YD} = \beta_{PB} \cdot f(r_{YD,PB}) = (+) \cdot (+) = (+)$$

Note that both correlation coefficients are anticipated to be (and actually are) positive. To see this, think of the impact of an increase in the price of chicken on the price of beef and then follow through the impact of any increase in income on the price of beef.

To sum, if a relevant variable is left out of a regression equation

1. there is no longer an estimate of the coefficient of that variable in the equation, and
2. the coefficients of the remaining variables are likely to be biased.

Although the amount of the bias might not be very large in some cases (when, for instance, there is little correlation between the included and excluded variables), it is extremely likely that at least a small amount of specification bias will be present in all such situations.

6.1.3 Correcting for an Omitted Variable

In theory, the solution to a problem of specification bias seems easy: Simply add the omitted variable to the equation. Unfortunately, that's more easily said than done, for a couple of reasons.

First, omitted variable bias is hard to detect. As mentioned above, the amount of bias introduced can be small and not immediately detectable. This is especially true when there is no reason to believe that you have misspecified the model. Some indications of specification bias are obvious (such as an estimated coefficient that is significant in the direction opposite from that expected), but others are not so clear. Could you tell from Equation 6.9 alone that a variable was missing? The best indicators of an omitted relevant variable are the theoretical underpinnings of the model itself. What variables *must* be included? What signs do you expect? Do you have any notions about the range into which the coefficient values should fall? Have you accidentally left out a variable that most researchers would agree is important? The best way to avoid omitting an important variable is to invest the time to think carefully through the equation before the data are entered into the computer.

A second source of complexity is the problem of choosing which variable to add to an equation once you decide that it is suffering from omitted variable bias. That is, a researcher faced with a clear case of specification bias (like an estimated $\hat{\beta}$ that is significantly different from zero in the unexpected direction) will often have no clue as to what variable could be causing the problem. Some beginning researchers, when faced with this dilemma, will add all the possible relevant variables to the equation at once, but this process leads to less precise estimates, as will be discussed in the next section. Other beginning researchers will test a number of different variables and keep the one in the equation that does the best statistical job of appearing to reduce the bias (by giving plausible signs and satisfactory t-values). This technique, adding a "left-out" variable to "fix" a strange-looking regression result, is invalid because the variable that best corrects a case of specification bias might do so only by chance rather than by being the true solution to the problem. In such an instance, the "fixed" equation may give superb statistical results for the sample at hand but then do terribly when applied to other samples because it does not describe the characteristics of the true population.

Dropping a variable will not help cure omitted variable bias. If the sign of an estimated coefficient is different from expected, it cannot be changed to the expected direction by dropping a variable that has a t-score lower (in absolute value) than the t-score of the coefficient estimate that has the unexpected sign. Furthermore, the sign in general will not likely change even if the variable to be deleted has a large t-score.[5]

If the estimated coefficient is significantly different from our expectations (either in sign or magnitude), then it is extremely likely that some sort of specification bias exists in our model. Although it is true that a poor sample of data or a poorly theorized expectation may also yield statistically significant unexpected signs or magnitudes, these possibilities sometimes can be eliminated.

If an unexpected result leads you to believe that you have an omitted variable, one way to decide which variable to add to the equation is to use expected bias analysis. **Expected bias** is the likely bias that omitting a particular variable would have caused in the estimated coefficient of one of the included variables. It can be estimated with Equation 6.7:

$$\text{Expected bias} = \beta_{om} \cdot f(r_{in,om}) \tag{6.7}$$

If the sign of the expected bias is the same as the sign of your unexpected result, then the variable might be the source of the apparent bias. If the sign of the expected bias is *not* the same as the sign of your unexpected result, how-

[5] Ignazio Visco, "On Obtaining the Right Sign of a Coefficient Estimate by Omitting a Variable from the Regression," *Journal of Econometrics*, February 1978, pp. 115–117.

ever, then the variable is extremely unlikely to have caused your unexpected result. Expected bias analysis should be used only when an equation has obvious bias (like a coefficient that is significant in an unexpected direction) and only when you're choosing between theoretically sound potential variables.

As an example of expected bias analysis, let's return to Equation 6.9, the chicken demand equation without the beef price variable. Let's assume that you had expected the coefficient of β_{PC} to be in the range of -1.0 and that you were surprised by the unexpectedly positive coefficient of PC in Equation 6.9. (As you can see by comparing Equations 6.8 and 6.9, your expectation was reasonable, but you can never be sure of this fact in practice.)

This unexpectedly positive result could not have been caused by an omitted variable with negative expected bias but could have been caused by an omitted variable with positive expected bias. One such variable is the price of beef. The expected bias in $\hat{\beta}_{PC}$ due to leaving out PB is positive since both the expected coefficient of PB and the expected correlation between PC and PB are positive:

$$\text{Expected bias in } \hat{\beta}_{PC} = \beta_{PB} \cdot f(r_{PC,PB}) = (+) \cdot (+) = (+)$$

Hence the price of beef is a reasonable candidate to be omitted variable in Equation 6.9.

Although you can never actually observe bias (since you don't know the true β), the use of this technique to screen potential causes of specification bias should reduce the number of regressions run and therefore increase the statistical validity of the results. This technique will work best when only one (or one kind of) variable is omitted from the equation in question. With a number of different kinds of variables omitted simultaneously, the impact on the equation's coefficients is quite hard to specify.

A brief warning: It may be tempting to conduct what might be called "residual analysis" by examining a plot of the residuals in an attempt to find patterns that suggest variables that have been accidentally omitted. A major problem with this approach is that the coefficients of the estimated equation will possibly have some of the effects of the left-out variable already altering their estimated values. Thus, residuals may show a pattern that only vaguely resembles the pattern of the actual omitted variable. The chances are high that the pattern shown in the residuals may lead to the selection of an incorrect variable. In addition, care should be taken to use residual analysis only to choose between theoretically sound candidate variables rather than to generate those candidates.

6.2 Irrelevant Variables

What happens if you include a variable in an equation that doesn't belong there? This case, **irrelevant variables,** is the converse of omitted variables and can be analyzed using the model we developed in Section 6.1. Whereas the omitted variable model has more independent variables in the true model than in the estimated equation, the irrelevant variable model has more independent variables in the estimated equation than in the true one.

The addition of a variable to an equation where it doesn't belong does not cause bias, but it does increase the variances of the estimated coefficients of the included variables.

6.2.1 Impact of Irrelevant Variables

If the true regression specification is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i \tag{6.10}$$

but the researcher for some reason includes an extra variable,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**} \tag{6.11}$$

the misspecified equation's error term can be seen to be:

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i} \tag{6.12}$$

Such a mistake will not cause bias if the true coefficient of the extra (or irrelevant) variable is zero. In that case, $\epsilon_i = \epsilon_i^{**}$. That is, $\hat{\beta}_1$ in Equation 6.11 is unbiased when $\beta_2 = 0$.

However, the inclusion of an irrelevant variable will increase the variance of the estimated coefficients, and this increased variance will tend to decrease the absolute magnitude of their t-scores. Also, an irrelevant variable usually will decrease the $\bar{R}^2$ (but not the R^2). In a model of Y on X_1 and X_2, the variance of the OLS estimator of β_1 is:

$$\mathrm{VAR}(\hat{\beta}_1) = \frac{\sigma^2}{(1 - r_{12}^2) \cdot \sum (X_1 - \bar{X}_1)^2} \tag{6.13}$$

If the irrelevant variable is not in the equation (or if $r_{12} = 0$), then:

$$\mathrm{VAR}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_1 - \bar{X}_1)^2} \tag{6.14}$$

Thus, although the irrelevant variable causes no bias, it causes problems for the regression because it reduces the precision of the regression.

To see why this is so, try plugging a nonzero value (between +1.0 and −1.0) for r_{12} into Equation 6.13 and note that VAR($\hat{\beta}_1$) has increased when compared to Equation 6.14. The equation with an included variable that does not belong in the equation usually has lower t-scores and a lower $\bar{R}^2$ than it otherwise would. The property holds, by the way, only when $r_{12} \neq 0$, but since this is the case in virtually every sample, the conclusion of increased variance due to irrelevant variables is a valid one. Table 6.1 summarizes the consequences of the omitted variable and the included irrelevant variable cases:

TABLE 6.1 EFFECT OF OMITTED VARIABLES AND IRRELEVANT VARIABLES ON THE COEFFICIENT ESTIMATES

Effect on Coefficient Estimates	Omitted Variable	Irrelevant Variable
Bias	Yes*	No
Variance	Decreases*	Increases*

*Unless $r_{12} = 0$.

6.2.2 An Example of an Irrelevant Variable

Let's return to the equation from Section 6.1 for the annual consumption of chicken and see what happens when we add an irrelevant variable to the equation. The original equation was:

$$\begin{array}{rlllll} \hat{Y}_t = & 31.5 - & 0.73PC_t & + 0.11PB_t & + 0.23YD_t & \quad (6.8) \\ & & (0.08) & (0.05) & (0.02) & \\ t = & & -9.12 & 2.50 & 14.22 & \\ \bar{R}^2 = & .986 & n = 44 \text{ (annual 1951–1994)} & & & \end{array}$$

Suppose you hypothesize that the demand for chicken also depends on R, the interest rate (which, perhaps, confuses the demand for a nondurable good with an equation you saw for a consumer durable). If you now estimate the equation with the interest rate included, you obtain:

$$\begin{array}{rllllll} \hat{Y}_t = & 30.0 - & 0.73PC_t & + 0.12PB_t & + 0.22YD_t & + 0.17R_t & \quad (6.15) \\ & & (0.08) & (0.06) & (0.02) & (0.21) & \\ t = & & -9.10 & 2.08 & 11.05 & 0.82 & \\ \bar{R}^2 = & .985 & n = 44 \text{ (annual 1951–1994)} & & & & \end{array}$$

A comparison of Equations 6.8 and 6.15 will make the theory in Section 6.2.1 come to life. First of all, $\bar{R}^2$ has fallen slightly, indicating the reduction in fit adjusted for degrees of freedom. Second, none of the regression coefficients from the original equation changed significantly; compare these results with the larger differences between Equations 6.8 and 6.9. Further, slight increases in the standard errors of the estimated coefficients took place. Finally, the t-score for the potential variable (the interest rate) is very small, indicating that it is not significantly different from zero. Given the theoretical shakiness of the new variable, these results indicate that it is irrelevant and never should have been included in the regression.

6.2.3 Four Important Specification Criteria

We have now discussed at least four valid criteria to help decide whether a given variable belongs in the equation. We think these criteria are so important that we urge beginning researchers to work through them every time a variable is added or subtracted.

1. *Theory:* Is the variable's place in the equation unambiguous and theoretically sound?
2. t-*Test:* Is the variable's estimated coefficient significant in the expected direction?
3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
4. *Bias:* Do other variables' coefficients change significantly when the variable is added to the equation?

If all these conditions hold, the variable belongs in the equation; if none of them do, the variable is irrelevant and can be safely excluded from the equation. When a typical omitted relevant variable is included in the equation, its inclusion probably will increase $\bar{R}^2$ and change other coefficients. If an irrelevant variable, on the other hand, is included, it will reduce $\bar{R}^2$, have an insignificant t-score, and have little impact on the other variables' coefficients.

In many cases, all four criteria do not agree. It is possible for a variable to have an insignificant t-score that is greater than one, for example. In such a case, it can be shown that $\bar{R}^2$ will go up when the variable is added to the equation and yet the t-score will still be insignificant.

Whenever the four criteria for whether a variable should be included in an equation disagree, the econometrician must use careful judgment and should not rely on a single criterion like $\bar{R}^2$ to determine the specification. Researchers should not misuse this freedom by testing various combinations of variables until they find the results that appear to statistically support the point they want to make. All such decisions are a bit easier when you realize that the single most important determinant of a variable's relevance is its theoretical justification. No amount of statistical evidence should make a theoretical necessity into an "irrelevant" variable. Once in a while, a researcher is forced to leave a theoretically important variable out of an equation for lack of a better alternative; in such cases, the usefulness of the equation is limited.

6.3 An Illustration of the Misuse of Specification Criteria

At times, the four specification criteria outlined in the previous section will lead the researcher to an incorrect conclusion if those criteria are applied blindly to a problem without the proper concern for economic principles or common sense. In particular, a t-score can often be insignificant for reasons other than the presence of an irrelevant variable. Since economic theory is the most important test for including a variable, an example of why a variable should not be dropped from an equation simply because it has an insignificant t-score is in order.

Suppose you believe that the demand for Brazilian coffee in the United States is a negative function of the real price of Brazilian coffee (P_{bc}) and a positive function of both the real price of tea (P_t) and real disposable income in the United States (Y_d).[6] Suppose further that you obtain the data, run the implied regression, and observe the following results:

$$\begin{aligned}\widehat{\text{COFFEE}} = 9.1 + {}&7.8P_{bc} + 2.4P_t + 0.0035Y_d \\ &(15.6) \quad\;\; (1.2) \quad (0.0010) \\ t = {}&0.5 \qquad\;\; 2.0 \qquad 3.5 \\ \bar{R}^2 = {}&.60 \quad n = 25\end{aligned} \tag{6.16}$$

The coefficients of the second and third variables, P_t and Y_d, appear to be fairly significant in the direction you hypothesized, but the first variable, P_{bc}, appears to have an insignificant coefficient with an unexpected sign. If you

[6] This example was inspired by a similar one concerning Ceylonese tea published in Potluri Rao and Roger LeRoy Miller, *Applied Econometrics* (Belmont, California: Wadsworth, 1971), pp. 38–40. This book is now out of print.

think there is a possibility that the demand for Brazilian coffee is perfectly price inelastic (that is, its coefficient is zero), you might decide to run the same equation without the price variable, obtaining:

$$\begin{aligned}\widehat{\text{COFFEE}} &= 9.3 + \underset{(1.0)}{2.6\text{P}_\text{t}} + \underset{(0.0009)}{0.0036\text{Y}_\text{d}} \qquad (6.17)\\ \text{t} &= \phantom{9.3 + {}} 2.6 \qquad\quad 4.0\\ \bar{\text{R}}^2 &= .61 \quad \text{n} = 25\end{aligned}$$

By comparing Equations 6.16 and 6.17, we can apply our four specification criteria for the inclusion of a variable in an equation that were outlined in the previous section:

1. *Theory:* Since the demand for coffee could possibly be perfectly price inelastic, the theory behind dropping the variable seems plausible.
2. *t-Test:* The t-score of the possibly irrelevant variable is 0.5, insignificant at any level.
3. $\bar{R}^2$: $\bar{\text{R}}^2$ increases when the variable is dropped, indicating that the variable is irrelevant. (Since the t-score is less than one, this is to be expected.)
4. *Bias:* The remaining coefficients change only a small amount when P_{bc} is dropped, suggesting that there is little if any bias caused by excluding the variable.

Based upon this analysis, you might conclude that the demand for Brazilian coffee is indeed perfectly price inelastic and that the variable is therefore irrelevant and should be dropped from the model. As it turns out, this conclusion would be unwarranted. Although the elasticity of demand for coffee in general might be fairly low (actually, the evidence suggests that it is inelastic only over a particular range of prices), it is hard to believe that Brazilian coffee is immune to price competition from other kinds of coffee. Indeed, one would expect quite a bit of sensitivity in the demand for Brazilian coffee with respect to the price of, for example, Colombian coffee. To test this hypothesis, the price of Colombian coffee (P_{cc}) should be added to the original Equation 6.16:

$$\begin{aligned}\widehat{\text{COFFEE}} &= 10.0 + \underset{(4.0)}{8.0\text{P}_{\text{cc}}} - \underset{(2.0)}{5.6\text{P}_{\text{bc}}} + \underset{(1.3)}{2.6\text{P}_\text{t}} + \underset{(0.0010)}{0.0030\text{Y}_\text{d}} \qquad (6.18)\\ \text{t} &= \phantom{10.0 + {}} 2.0 \qquad -2.8 \qquad 2.0 \qquad 3.0\\ \bar{\text{R}}^2 &= .65 \quad \text{n} = 25\end{aligned}$$

By comparing Equations 6.16 and 6.18, we can once again apply the four criteria:

1. *Theory:* Both prices should always have been included in the model; their logical justification is quite strong.
2. *t-Test:* The t-score of the new variable, the price of Colombian coffee, is 2.0, significant at most levels.
3. $\bar{R}^2$: $\bar{R}^2$ increases with the addition of the variable, indicating that the variable was an omitted variable.
4. *Bias:* Although two of the coefficients remain virtually unchanged, indicating that the correlations between these variables and the price of Colombian coffee variable are low, the coefficient for the price of Brazilian coffee does change significantly, indicating bias in the original result.

An examination of the bias question will also help us understand Equation 6.7, the equation for bias. Since the expected sign of the coefficient of the omitted variable (P_{cc}) is positive and since the simple correlation coefficient between the two competitive prices ($r_{Pcc,Pbc}$) is also positive, the direction of the expected bias in $\hat{\beta}_{Pbc}$ in the estimation of Equation 6.16 is positive. If you compare Equations 6.16 and 6.18, that positive bias can be seen because the coefficient of P_{bc} is $+7.8$ instead of -5.6. The increase from -5.6 to $+7.8$ may be due to the positive bias that results from leaving out P_{cc}.

The moral to be drawn from this example is that theoretical considerations should never be discarded, even in the face of statistical insignificance. If a variable known to be extremely important from a theoretical point of view turns out to be statistically insignificant in a particular sample, that variable should be left in the equation despite the fact that it makes the results look bad.

Don't conclude that the particular path outlined in this example is the correct way to specify an equation. Trying a long string of possible variables until you get the particular one that makes P_{bc} turn negative and significant is not the way to obtain a result that will stand up well to other samples or alternative hypotheses. The original equation should never have been run without the Colombian coffee variable. Instead, the problem should have been analyzed enough so that such errors of omission were unlikely before any regressions were attempted at all. The more thinking that's done before the first regression is run, and the fewer alternative specifications that are estimated, the better the regression results are likely to be.

6.4 Specification Searches

One of the weaknesses of econometrics is that a researcher can potentially manipulate a data set to produce almost *any* results by specifying different regressions until estimates with the desired properties are obtained. Thus, the integrity of all empirical work is potentially open to question.

Although the problem is a difficult one, it makes sense to attempt to minimize the number of equations estimated and to rely on theory rather than statistical fit as much as possible when choosing variables. Theory, not statistical fit, should be the most important criterion for the inclusion of a variable in a regression equation. To do otherwise runs the risk of producing incorrect and/or disbelieved results. We'll try to illustrate this by discussing three of the most commonly used *incorrect* techniques for specifying a regression equation. These techniques produce the best specification only by chance. At worst, they are possibly unethical in that they misrepresent the methods used to obtain the regression results and the significance of those results.

6.4.1 Data Mining

Almost surely the worst way to choose a specification is to simultaneously try a whole series of possible regression formulations and then choose the equation that conforms the most to what the researcher wants the results to look like. In such a situation, the researcher would estimate virtually every possible combination of the various alternative independent variables, and the choice between them would be made on the basis of the results. This practice of simultaneously estimating a number of combinations of independent variables and selecting the best from them ignores the fact that a number of specifications have been examined before the final one. To oversimplify, if you are 95 percent confident that a regression result didn't occur by chance and you run more than 20 regressions, how much confidence can you have in your result? Since you'll tend to keep regressions with high t-scores and discard ones with low t-scores, the reported t-scores overstate the degree of statistical significance of the estimated coefficients.

Furthermore, such "data mining" and "fishing expeditions" to obtain desired statistics for the final regression equation are potentially unethical methods of empirical research. These procedures include using not only many alternative combinations of independent variables but also many functional forms, lag structures, and what are offered as "sophisticated" or "advanced" estimating techniques. "If you just torture the data long enough, they will confess."[7] In other words, if enough alternatives are tried, the chances of

[7] Thomas Mayer, "Economics as a Hard Science: Realistic Goal or Wishful Thinking?" *Economic Inquiry,* April 1980, p. 175.

obtaining the results desired by the researcher are increased tremendously, but the final result is essentially worthless. The researcher hasn't found any scientific evidence to support the original hypothesis; rather, prior expectations were imposed on the data in a way that is essentially misleading.

6.4.2 Stepwise Regression Procedures

A **stepwise regression** involves the use of a computer program to choose the independent variables to be included in the estimation of a particular equation. The computer program is given a "shopping list" of possible independent variables, and then it builds the equation in steps. It chooses as the first explanatory variable the one that by itself explains the largest amount of the variation of the dependent variable around its mean. It chooses as the second variable the one that adds the most to R^2, given that the first variable is already in the equation. The stepwise procedure continues until the next variable to be added fails to achieve some researcher-specified increase in R^2 (or all the variables are added). The measure of the supposed contribution of each independent variable is the increase in R^2 (which is sometimes called the "R^2 delete") caused by the addition of the variable.

Unfortunately, any correlation among the independent variables (called multicollinearity, which we will take up in more detail in Chapter 8) causes this procedure to be deficient. To the extent that the variables are related, it becomes difficult to tell the impact of one variable from another. As a result, in the presence of multicollinearity, it's impossible to determine unambiguously the individual contribution of each variable enough to say which one is more important and thus should be included first.[8] Even worse, there is no necessity that the particular combination of variables chosen has any theoretical justification or that the coefficients have the expected signs.

Because of these problems, most econometricians avoid stepwise procedures. The major pitfalls are that the coefficients may be biased, the calculated t-values no longer follow the t-distribution, relevant variables may be excluded because of the arbitrary order in which the selection takes place, and

[8] Some programs compute standardized beta coefficients, which are the estimated coefficients for an equation in which all variables have been standardized by subtracting their means from them and by dividing them by their own standard deviations. The higher the beta of an independent variable is in absolute value, the more important it is thought to be in explaining the movements in the dependent variable. Unfortunately, beta coefficients are deficient in the presence of multicollinearity, as are partial correlation coefficients, which measure the correlation between the dependent variable and a given independent variable holding all other independent variables constant.

the signs of the estimated coefficients at intermediate or final stages of the routine may be different from the expected signs. Using a stepwise procedure is an admission of ignorance concerning which variables should be entered.

6.4.3 Sequential Specification Searches

To their credit, most econometricians avoid data mining and stepwise regressions. Instead, they tend to specify equations by estimating an initial equation and then sequentially dropping or adding variables (or changing functional forms) until a plausible equation is found with "good statistics." Faced with knowing that a few variables are relevant (on the basis of theory) but not knowing whether other additional variables are relevant, inspecting $\bar{R}^2$ and *t*-tests for all variables for each specification appears to be the generally accepted practice. Indeed, it would be easy to draw from a casual reading of the previous sections the impression that such a sequential specification search is the best way to go about finding the "truth." Instead, as we shall see, there is a vast difference in approach between a sequential specification search and our recommended approach.

The **sequential specification search** technique allows a researcher to estimate an undisclosed number of regressions and then present a final choice (which is based upon an unspecified set of expectations about the signs and significance of the coefficients) as if it were the only specification estimated. Such a method misstates the statistical validity of the regression results for two reasons:

1. The statistical significance of the results is overestimated because the estimations of the previous regressions are ignored.
2. The set of expectations used by the researcher to choose between various regression results is rarely if ever disclosed.[9] Thus the reader has no way of knowing whether or not all the other regression results had opposite signs or insignificant coefficients for the important variables.

Unfortunately, there is no universally accepted way of conducting sequential searches, primarily because the appropriate test at one stage in the procedure depends on which tests were previously conducted, and also because the tests have been very difficult to invent. One possibility is to reduce the degrees of freedom in the "final" equation by one for each alternative specifica-

[9] As mentioned in Chapter 5, Bayesian regression is a technique for dealing systematically with these prior expectations. For more on this issue, see Edward E. Leamer, *Specification Searches* (New York: Wiley), 1978.

tion attempted. This procedure is far from exact, but it does impose an explicit penalty for specification searches.

More generally, we recommend trying to keep the number of regressions estimated as low as possible; to focus on theoretical considerations when choosing variables, functional forms, and the like; and to document all the various specifications investigated. That is, we recommend combining parsimony (using theory and analysis to limit the number of specifications estimated) with disclosure (reporting all the equations estimated).

There is another side to the story, however. Some researchers feel that the true model will show through if given the chance and that the best statistical results (including signs of coefficients, etc.) are most likely to have come from the true specification. The problem with this philosophy is that the element of chance is ordinarily quite strong in any given application. In addition, reasonable people often disagree as to what the "true" model should look like. As a result, different researchers can look at the same data set and come up with very different "best" equations. Because this can happen, the distinction between good and bad econometrics is not always as clear cut as is implied by the previous paragraphs. As long as researchers have a healthy respect for the dangers inherent in specification searches, they are very likely to proceed in a reasonable way.

The lesson to be learned from this section should be quite clear. Most of the work of specifying an equation should be done before even attempting to estimate the equation on the computer. Since it is unreasonable to expect researchers to be perfect, there will be times when additional specifications must be estimated; however, these new estimates should be thoroughly grounded in theory and explicitly taken into account when testing for significance or summarizing results. In this way, the danger of misleading the reader about the statistical properties of estimates is reduced.

6.4.4 Bias Caused by Relying on the *t*-Test to Choose Variables

In the previous section, we stated that sequential specification searches are likely to mislead researchers about the statistical properties of their results. In particular, the practice of dropping a potential independent variable simply because its t-score indicates that its estimated coefficient is insignificantly different from zero will cause systematic bias in the estimated coefficients (and their t-scores) of the remaining variables.[10]

[10] For a number of better techniques, including sequential or "pretest" estimators and "Stein-rule" estimators, see George G. Judge, W. E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee, *The Theory and Practice of Econometrics* (New York: Wiley, 1985).

Say the hypothesized model for a particular dependent variable is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{6.19}$$

Assume further that, on the basis of theory, we are certain that X_1 belongs in the equation but that we are not as certain that X_2 belongs. Even though we have stressed four criteria to determine whether X_2 should be included, many inexperienced researchers just use the *t*-test on $\hat{\beta}_2$ to determine whether X_2 should be included. If this preliminary *t*-test indicates that $\hat{\beta}_2$ is significantly different from zero, then these researchers leave X_2 in the equation, and they choose Equation 6.19 as their final model. If, however, the *t*-test does *not* indicate that $\hat{\beta}_2$ is significantly different from zero, then such researchers drop X_2 from the equation and consider Y as a function of X_1.

Two kinds of mistakes can be made using such a system. First, X_2 can sometimes be left in the equation when it does not belong there, but such a mistake does not change the expected value of $\hat{\beta}_1$. Second, X_2 can sometimes be dropped from equation when it belongs, and then the estimated coefficient of X_1 will be biased by the value of the true β_2 to the extent that X_1 and X_2 are correlated. In other words, $\hat{\beta}_1$ will be biased every time X_2 belongs in the equation and is left out, and X_2 will be left out every time that its estimated coefficient is not significantly different from zero. That is, the expected value of $\hat{\beta}_1$ will not equal the true β_1, and we will have systematic bias in our equation

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot f(r_{X_1,X_2}) \cdot P \neq \beta_1 \tag{6.20}$$

where P indicates the probability of an insignificant t-score. It is also the case that the t-score of $\hat{\beta}_1$ no longer follows the t-distribution. In other words, the *t*-test is biased by sequential specification searches.

Since most researchers consider a number of different variables before settling on the final model, someone who relies on the *t*-test alone is likely to encounter this problem systematically.

6.4.5 Scanning and Sensitivity Analysis

Throughout this text, we've encouraged you to estimate as few specifications as possible and to avoid depending on fit alone to choose between those specifications. If you read the current economics literature, however, it won't take you long to find well-known researchers who have estimated five or more specifications and then have listed all their results in an academic journal article. What's going on?

In almost every case, these authors have employed one of the two following techniques:

1. Scanning to develop a testable theory
2. Sensitivity analysis

Scanning involves analyzing a data set not for the purpose of testing a hypothesis but for the purpose of developing a testable theory or hypothesis. A researcher who is scanning will run quite a few different specifications, will select the specifications that fit best, and then will analyze these results in the hopes that they will provide clues to a new theory or hypothesis. As a means for stimulating fresh thinking or influencing thinking about substantive issues, scanning may have even more potential than does classical hypothesis testing.[11]

Be careful, however; before you can "accept" a theory or hypothesis, it should be tested on a *different* data set (or in another context) using the hypothesis testing techniques of this text. A new data set must be used because our typical statistical tests have little meaning if the new hypotheses are tested on the old data set; after all, the researcher knows ahead of time what the results will be! The use of such dual data sets is easiest when there is a plethora of data. This sometimes is the case in cross-sectional research projects but rarely is the case for time-series research.

Sensitivity analysis consists of purposely running a number of alternative specifications to determine whether particular results are *robust* (not statistical flukes). In essence, we're trying to determine how sensitive a particular result is to a change in specification. Researchers who use sensitivity analysis run (and report on) a number of different specifications and tend to discount a result that appears significant in some specifications and insignificant in others. Indeed, the whole purpose of sensitivity analysis is to gain confidence that a particular result is significant in a variety of alternative specifications and is not based on a single specification that has been estimated on only one data set. For a simple example of sensitivity analysis, see Exercise 15 at the end of the chapter.

6.5 Lagged Independent Variables

Virtually all the regressions we've studied so far have been "instantaneous" in nature. In other words, they have included independent and dependent variables from the same time period, as in:

[11] For an excellent presentation of this argument, see Lawrence H. Summers, "The Scientific Illusion in Empirical Macroeconomics," *Scandinavian Journal of Economics*, 1991, pp. 129–148. For a number of related points of view, see David F. Hendry, Edward E. Leamer, and Dale J. Poirer, *A Conversation on Econometric Methodology*, Institute of Statistics and Decision Sciences, Duke University, 1989, 144 pages.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \epsilon_t \qquad (6.21)$$

where the subscript t is used to refer to a particular point in time. If all variables have the same subscript, then the equation is instantaneous.

Some beginning researchers, when choosing their independent variables, jump to the mistaken conclusion that all regressions should follow the pattern of Equation 6.21 and contain only variables that come from the same time period. Such a conclusion ignores the fact that not all economic or business situations imply such instantaneous relationships between the dependent and independent variables. In many cases we must allow for the possibility that time might elapse between a change in the independent variable and the resulting change in the dependent variable. The length of this time between cause and effect is called a **lag.** Many econometric equations include one or more *lagged independent variables* like X_{1t-1}, where the subscript $t - 1$ indicates that the observation of X_1 is from the time period previous to time period t, as in the following equation:

$$Y_t = \beta_0 + \beta_1 X_{1t-1} + \beta_2 X_{2t} + \epsilon_t \qquad (6.22)$$

In this equation, X_1 has been lagged by one time period, but the relationship between Y and X_2 is still instantaneous.

For example, think about the process by which the supply of an agricultural product is determined. Since agricultural goods take time to grow, decisions on how many acres to plant or how many eggs to let hatch into egg-producing hens (instead of selling them immediately) must be made months if not years before the product is actually supplied to the consumer. Any change in an agricultural market, such as an increase in the price that the farmer can earn for providing cotton, has a lagged effect on the supply of that product:

$$C_t = f(\overset{+}{P}C_{t-1}, \overset{-}{PF_t}) + \epsilon_t = \beta_0 + \beta_1 PC_{t-1} + \beta_2 PF_t + \epsilon_t \qquad (6.23)$$

where: C_t = the quantity of cotton supplied in year t
PC_{t-1} = the price of cotton in year $t - 1$
PF_t = the price of farm labor in year t

Note that this equation hypothesizes a lag between the price of cotton and the production of cotton, but not between the price of farm labor and the production of cotton. It's reasonable to think that if cotton prices change, farmers won't be able to react immediately because it takes a while for cotton to be planted and to grow.

The meaning of the regression coefficient of a lagged variable is not the same as the meaning of the coefficient of an unlagged variable. The estimated coefficient of a lagged X measures the change in *this year's* Y attributed to a one-unit increase in *last year's* X (holding constant the other Xs in the equation). Thus β_1 in Equation 6.23 measures the extra number of units of cotton that would be produced this year as a result of a one-unit increase in last year's price of cotton, holding this year's price of farm labor constant.

If the lag structure is hypothesized to take place over more than one time period, or if a lagged dependent variable is included on the right-hand side of an equation, the question becomes significantly more complex. Such cases, called *distributed lags*, will be dealt with in Chapter 12.

6.6 An Example of Choosing Independent Variables

It's time to get some experience choosing independent variables. After all, every equation so far in the text has come with the specification already determined, but once you've finished this course you'll have to make all such specification decisions on your own. In future chapters, we'll use a technique called "interactive regression learning exercises" to allow you to make your own actual specification choices and get feedback on your choices. To start, though, let's work through a specification together.

To keep things as simple as possible, we'll begin with a topic near and dear to your heart, your GPA! Suppose a friend who attends a small liberal arts college surveys all 25 members of her econometrics class, obtains data on the variables listed below, and asks for your help in choosing a specification:

GPA_i = the cumulative college grade point average on the *i*th student on a four-point scale

HGPA_i = the cumulative high school grade point average of the *i*th student on a four-point scale

MSAT_i = the highest score earned by the *i*th student on the math section of the SAT test (800 maximum)

VSAT_i = the highest score earned by the *i*th student on the verbal section of the SAT test (800 maximum)

$\text{SAT}_i = \text{MSAT}_i + \text{VSAT}_i$

GREK_i = a dummy variable equal to 1 if the *i*th student is a member of a fraternity or sorority, 0 otherwise

HRS_i = the *i*th student's estimate of the average number of hours spent studying per course per week in college

PRIV_i = a dummy variable equal to 1 if the *i*th student graduated from a private high school, 0 otherwise

JOCK_i = a dummy variable equal to 1 if the *i*th student is or was a member of a varsity intercollegiate athletic team for at least one season, 0 otherwise

lnEX_i = the natural log of the number of full courses that the *i*th student has completed in college.

Assuming that GPA_i is the dependent variable, which independent variables would you choose? Before you answer, think through the possibilities carefully. What are the expected signs of each of the coefficients? How strong is the theory behind each variable? Which variables seem obviously important? Which variables seem potentially irrelevant or redundant? Are there any other variables that you wish your friend had collected?

To get the most out of this example, you should take the time to *write down* the exact specification that you would run:

$$\text{GPA}_i = f(?, ?, ?, ?, ?) + \epsilon$$

It's hard for most beginning econometricians to avoid the temptation of including *all* the above variables in a GPA equation and then dropping any variables that have insignificant t-scores. Even though we mentioned in the previous section that such a specification search procedure will result in biased coefficient estimates, most beginners don't trust their own judgment and tend to include too many variables. With this warning in mind, do you want to make any changes in our proposed specification?

No? OK, let's compare notes. We believe that grades are a function of a student's ability, how hard the student works, and the student's experience taking college courses. Consequently, our specification would be:

$$\text{GPA}_i = f(\overset{+}{\text{HGPA}_i}, \overset{+}{\text{HRS}_i}, \overset{+}{\text{lnEX}_i}) + \epsilon$$

We can already hear you complaining! What about SATs, you say? Everyone knows they're important. How about jocks and Greeks? Don't they have lower GPAs? Don't prep schools grade harder and prepare students better than public high schools?

Before we answer, it's important to note that we think of specification choice as choosing which variables to *include,* not which variables to *exclude.* That is, we don't assume automatically that a given variable should be included in an equation simply because we can't think of a good reason for dropping it.

Given that, however, why did we choose the variables we did? First, we think that the best predictor of a student's college GPA is his or her high school GPA. We have a hunch that once you know HGPA, SATs are redundant, at least at a liberal arts college where there are few multiple choice tests. In addition, we're concerned that possible racial and gender bias in the SAT test makes it a questionable measure of academic potential, but we recognize that we could be wrong on this issue.

As for the other variables, we're more confident. For example, we feel that once we know how many hours a week a student spends studying, we couldn't care less what that student does with the rest of his or her time, so JOCK and GREK are superfluous once HRS is included. Finally, while we recognize that some private schools are superb and that some public schools are not, we'd guess that PRIV is irrelevant; it probably has only a minor effect.

If we estimate this specification on the 25 students, we obtain:

$$\begin{aligned}\widehat{GPA}_i = -0.26 &+ 0.49HGPA_i + 0.06HRS_i + 0.42\ln EX_i \qquad (6.24)\\ &\quad(0.21) \qquad\quad (0.02) \qquad\quad (0.14)\\ t &= 2.33 \qquad\quad 3.00 \qquad\quad 3.00\\ n = 25 &\quad \bar{R}^2 = .585 \quad F = 12.3\end{aligned}$$

Since we prefer this specification on theoretical grounds, since the overall fit seems reasonable, and since each coefficient meets our expectations in terms of sign, size, and significance, we consider this an acceptable equation. The only circumstance under which we'd consider estimating a second specification would be if we had theoretical reasons to believe that we had omitted a relevant variable. The only variable that might meet this description is SAT_i (which we prefer to the individual MSAT and VSAT):

$$\begin{aligned}\widehat{GPA}_i = -0.92 &+ 0.47HGPA_i + 0.05HRS_i \qquad (6.25)\\ &\quad(0.22) \qquad\quad (0.02)\\ t &= 2.12 \qquad\quad 2.50\\ &+ 0.44\ln EX_i + 0.00060SAT_i\\ &\quad(0.14) \qquad\quad (0.00064)\\ t &= 3.12 \qquad\quad 0.93\\ n = 25 &\quad \bar{R}^2 = .583 \quad F = 9.4\end{aligned}$$

Let's use our four specification criteria to compare Equations 6.24 and 6.25:

1. *Theory:* As discussed above, the theoretical validity of SAT tests is a matter of some academic controversy, but they still are one of the most-cited measures of academic potential in this country.
2. t-*Test:* The coefficient of SAT is positive, as we'd expect, but it's not significantly different from zero.

3. $\bar{R}^2$: As you'd expect (since SAT's t-score is under one), $\bar{R}^2$ falls slightly when SAT is added.
4. *Bias:* None of the estimated slope coefficients changes significantly when SAT is added, though some of the t-scores do change because of the increase in the SE($\hat{\beta}$)s caused by the addition of SAT.

Thus, the statistical criteria support our theoretical contention that SAT is irrelevant.

Finally, it's important to recognize that different researchers could come up with different final equations on this topic. A researcher whose prior expectation was that SAT unambiguously belonged in the equation would have estimated Equation 6.25 and accepted that equation without bothering to estimate Equation 6.24.

6.7 Summary

1. The omission of a variable from an equation will cause bias in the estimates of the remaining coefficients to the extent that the omitted variable is correlated with included variables.
2. The bias to be expected from leaving a variable out of an equation equals the coefficient of the excluded variable times a function of the simple correlation coefficient between the excluded variable and the included variable in question.
3. Including a variable in an equation in which it is actually irrelevant does not cause bias, but it will usually increase the variances of the included variables' estimated coefficients, thus lowering their t-values and lowering $\bar{R}^2$.
4. Four useful criteria for the inclusion of a variable in an equation are:
   a. Theory
   b. *t*-Test
   c. $\bar{R}^2$
   d. Bias
5. Theory, not statistical fit, should be the most important criterion for the inclusion of a variable in a regression equation. To do otherwise runs the risk of producing incorrect and/or disbelieved results. For example, stepwise regression routines will generally give biased esti-

mates and will almost always have test statistics that will not follow the distribution necessary to use standard t-tables.

Exercises

(Answers to even-numbered questions are in Appendix A.)

1. Write the meaning of each of the following terms without referring to the book (or your notes), and compare your definition with the version in the text for each:
   a. omitted variable
   b. irrelevant variable
   c. specification bias
   d. stepwise regression
   e. sequential specification search
   f. specification error
   g. the four specification criteria
   h. expected bias
   i. lagged independent variable

2. For each of the following situations, determine the *sign* (and if possible comment on the likely size) of the expected bias introduced by omitting a variable:
   a. In an equation for the demand for peanut butter, the impact on the coefficient of disposable income of omitting the price of peanut butter variable. (*Hint:* Start by hypothesizing signs.)
   b. In an earnings equation for workers, the impact on the coefficient of experience of omitting the variable for age.
   c. In a production function for airplanes, the impact on the coefficient of labor of omitting the capital variable.
   d. In an equation for daily attendance at outdoor concerts, the impact on the coefficient of the weekend dummy variable (1 = weekend) of omitting a variable that measures the probability of precipitation at concert time.

3. Consider the following annual model of the death rate (per million population) due to coronary heart disease in the United States (Y_t):

$$
\begin{aligned}
\hat{Y}_t = 140 + {} & 10.0C_t + 4.0E_t - 1.0M_t \\
& \quad (2.5) \qquad (1.0) \qquad (0.5) \\
t = {} & \quad 4.0 \qquad\;\; 4.0 \qquad -2.0 \\
n = 31\ (1950-1980) & \qquad \bar{R}^2 = .678
\end{aligned}
$$

where: C_t = per capita cigarette consumption (pounds of tobacco) in year t
E_t = per capita consumption of edible saturated fats (pounds of butter, margarine, and lard) in year t
M_t = per capita consumption of meat (pounds) in year t

a. Create and test appropriate null hypotheses at the 10 percent level. What, if anything, seems to be wrong with the estimated coefficient of M?

b. The most likely cause of a coefficient that is significant in the unexpected direction is omitted variable bias. Which of the following variables could possibly be an omitted variable that is causing $\hat{\beta}_M$'s unexpected sign? Explain.

B_t = per capita consumption of hard liquor (gallons) in year t
F_t = the average fat content (percentage) of the meat that was consumed in year t
W_t = per capita consumption of wine and beer (gallons) in year t
R_t = per capita number of miles run in year t
H_t = per capita open-heart surgeries in year t
O_t = per capita amount of oat bran eaten in year t

c. If you had to choose one variable to add to the equation, what would it be? Explain your answer. (*Hint:* You're not limited to the variables listed in part b above.)

4. Assume that you've been hired by the surgeon general of the United States to study the determinants of smoking behavior and that you estimate the following cross-sectional model based on data from 1988 for all 50 states (standard errors in parentheses)[12]:

$$\begin{array}{llllll} \hat{C}_i = 100 & -\ 9.0E_i & +\ 1.0I_i & -\ 0.04T_i & -\ 3.0V_i & +\ 1.5R_i \quad (6.26) \\ & (3.0) & (1.0) & (0.04) & (1.0) & (0.5) \\ t = & -\ 3.0 & 1.0 & -\ 1.0 & -\ 3.0 & 3.0 \end{array}$$

$$\bar{R}^2 = .50 \quad n = 50 \text{ (states)}$$

where: C_i = the number of cigarettes consumed per day per person in the *i*th state

[12] This question is generalized from a number of similar studies, including John A. Bishop and Jang H. Yoo, "The Impact of the Health Scare, Excise Taxes, and Advertising on Cigarette Demand and Supply," *Southern Economic Journal*, January 1988, pp. 402–411.

E_i = the average years of education for persons over 21 in the *i*th state

I_i = the average income in the *i*th state (thousands of dollars)

T_i = the tax per package of cigarettes in the *i*th state (cents)

V_i = the number of video ads against smoking aired on the three major networks in the *i*th state.

R_i = the number of radio ads against smoking aired on the five largest radio networks in the *i*th state

a. Develop and test (at the 5 percent level) appropriate hypotheses for the coefficients of the variables in this equation.

b. Do you appear to have any irrelevant variables? Do you appear to have any omitted variables? Explain your answer.

c. Let's assume that your answer to part b was yes to both. Which problem is more important to solve first, irrelevant variables or omitted variables? Why?

d. One of the purposes of running the equation was to determine the effectiveness of antismoking advertising on television and radio. What is your conclusion?

e. The surgeon general decides that tax rates are irrelevant to cigarette smoking and orders you to drop them from your equation. Given the following results, use our four specification criteria to decide whether you agree with her conclusion. Carefully explain your reasoning (standard errors in parentheses).

$$\begin{aligned} \hat{C}_i = 101 - &\underset{(3.0)}{9.1E_i} + \underset{(0.9)}{1.0I_i} - \underset{(1.0)}{3.5V_i} + \underset{(0.5)}{1.6R_i} \qquad (6.27) \\ &\bar{R}^2 = .50 \quad n = 50 \text{ (states)} \end{aligned}$$

5. The data set in Table 6.2 is the one that was used to estimate the chicken demand examples of Sections 6.1.2 and 6.2.2.

a. Use these data to reproduce the specifications in the chapter. (filename CHICK6)

b. Find data for the price of another substitute for chicken and add that variable to your version of Equation 6.8. Analyze your results. In particular, apply the four criteria for the inclusion of a variable to determine whether the price of the substitute is an irrelevant or previously was an omitted variable.

6. You have been retained by the "Expressive Expresso" company to help them decide where to build their next "Expressive Expresso" store.

TABLE 6.2 DATA FOR THE CHICKEN DEMAND EQUATION

Year	Y	PC	PB	YD
1951	21.8	25.0	28.7	14.86
1952	22.1	22.1	24.3	15.39
1953	21.9	22.1	16.3	16.11
1954	22.8	16.8	16.0	16.19
1955	21.3	18.6	15.6	17.04
1956	24.4	16.0	14.9	17.87
1957	25.4	13.7	17.2	18.51
1958	28.0	14.0	21.9	18.84
1959	28.7	11.0	22.6	19.68
1960	28.0	12.2	20.4	20.14
1961	30.0	10.1	20.2	20.67
1962	30.0	10.2	21.3	21.56
1963	30.7	10.0	19.9	22.30
1964	31.1	9.2	18.0	23.89
1965	33.4	8.9	19.9	25.47
1966	35.5	9.7	22.2	27.20
1967	36.3	7.9	22.3	28.83
1968	36.4	8.2	23.4	31.02
1969	38.1	9.7	26.2	33.03
1970	40.1	9.1	27.1	35.51
1971	40.1	7.7	29.0	38.12
1972	41.5	9.0	33.5	40.82
1973	39.7	15.1	42.8	45.63
1974	39.6	9.7	35.6	49.42
1975	38.8	9.9	32.3	53.83
1976	41.9	12.9	33.7	58.57
1977	42.7	12.0	34.5	63.84
1978	44.8	12.4	48.5	71.24
1979	48.3	13.9	66.1	78.90
1980	48.4	11.0	62.4	86.97
1981	50.4	11.1	58.6	96.03
1982	51.5	10.3	56.7	101.33
1983	52.6	12.7	55.5	107.77
1984	54.5	15.9	57.3	119.14
1985	56.3	14.8	53.7	125.94
1986	58.1	12.5	52.6	132.13
1987	61.9	11.0	61.1	138.53
1988	63.8	9.2	66.6	148.84
1989	67.5	14.9	69.5	157.74
1990	70.4	9.3	74.6	166.89
1991	73.5	7.1	72.7	171.82
1992	76.8	8.6	71.3	180.32
1993	78.9	10.0	72.6	185.64
1994	80.5	7.6	66.7	192.59

Sources: U.S. Department of Agriculture. *Agricultural Statistics;* U.S. Bureau of the Census. *Historical Statistics of the United States,* U.S. Bureau of the Census. *Statistical Abstract of the United States.*
Note: filename CHICK6

You decide to run a regression on the sales of the 30 existing "Expressive Expresso" stores as a function of the characteristics of the locations they are in and then use the equation to predict the sales at the various locations you are considering for the newest store. You end up estimating (standard errors in parentheses):

$$\hat{Y}_i = 30 + \underset{(0.02)}{0.1X_{1i}} + \underset{(0.01)}{0.01X_{2i}} + \underset{(1.0)}{10.0X_{3i}} + \underset{(1.0)}{3.0X_{4i}}$$

where: Y_i = average daily sales (in hundreds of dollars) of the *i*th store
X_{1i} = the number of cars that pass the *i*th location per hour
X_{2i} = average income in the area of the *i*th store
X_{3i} = the number of tables in the *i*th store
X_{4i} = the number of competing shops in the area of the *i*th store

a. Hypothesize expected signs, calculate the correct t-scores, and test the significance at the 1 percent level for each of the coefficients.
b. What problems appear to exist in the equation? What evidence of these problems do you have?
c. What suggestions would you make for a possible second run of this admittedly hypothetical equation? (*Hint:* Before recommending the inclusion of a potentially left-out variable, consider whether the exclusion of the variable could possibly have caused any observed bias.)

7. Discuss the topic of specification searches with various members of your econometrics class. What is so wrong with not mentioning previous (probably incorrect) estimates? Why should readers be suspicious when researchers attempt to find results that support their hypotheses? Who would try to do the opposite? Do these concerns have any meaning in the world of business? In particular, if you're not trying to publish a paper, couldn't you use any specification search techniques you want to find the best equation?

8. Suppose you run a regression explaining the number of hamburgers that the campus fast-food store (let's call it "The Cooler") sells per day as a function of the price of their hamburgers (in dollars), the weather (in degrees F), the price of hamburgers at a national chain nearby (also in dollars), and the number of students (in thousands) on campus that day. Assume that The Cooler stays open whether or not

school is in session (for staff, etc.). Unfortunately, a lightning bolt strikes the computer and wipes out all the memory and you cannot tell which independent variable is which! Given the following regression results (standard errors in parentheses):

$$\begin{aligned}\hat{Y}_i = 10.6 + &\underset{(2.6)}{28.4X_{1i}} + \underset{(6.3)}{12.7X_{2i}} + \underset{(0.61)}{0.61X_{3i}} - \underset{(5.9)}{5.9X_{4i}} \\ &\bar{R}^2 = .63 \quad n = 35\end{aligned}$$

a. Attempt to identify which result corresponds to which variable.
b. Explain your reasoning for part a above.
c. Develop and test hypotheses about the coefficients assuming that your answer to part a is correct. What suggestions would you have for changes in the equation for a rerun when the computer is back up again?

9. Most of the examples in the text so far have been demand-side equations or production functions, but economists often also have to quantify supply-side equations that are not true production functions. These equations attempt to explain the production of a product (for example, Brazilian coffee) as a function of the price of the product and various other attributes of the market that might have an impact on the total output of growers.
   a. What sign would you expect the coefficient of price to have in a supply-side equation? Why?
   b. What other variables can you think of that might be important in a supply-side equation?
   c. Many agricultural decisions are made months (if not a full year or more) before the results of those decisions appear in the market. How would you adjust your hypothesized equation to take account of these lags?
   d. Given all the above, carefully specify the exact equation you would use to attempt to explain Brazilian coffee production. Be sure to hypothesize the expected signs, be specific with respect to lags, and try to make sure you have not omitted an important independent variable.

10. If you think about the previous question, you'll realize that the *same* dependent variable (quantity of Brazilian coffee) can have different expected signs for the coefficient of the *same* independent variable (the price of Brazilian coffee), depending on what other variables are in the regression.

a. How is this possible? That is, how is it possible to expect different signs in demand-side equations from what you would expect in supply-side ones?

b. Given that we will not discuss how to estimate simultaneous equations until Chapter 14, what can be done to avoid the "simultaneity bias" of getting the price coefficient from the demand equation in the supply equation and vice versa?

c. What can you do to systematically ensure that you do not have supply-side variables in your demand equation or demand-side variables in your supply equation?

11. You've been hired by "Indo," the new Indonesian automobile manufacturer, to build a model of U.S. car prices in order to help the company undercut our prices. Allowing Friedmaniac zeal to overwhelm any patriotic urges, you build the following model of the price of 35 different American-made 1996 U.S. sedans (standard errors in parentheses):

$$\text{Model A: } \hat{P}_i = \underset{}{3.0} + \underset{(0.07)}{0.28W_i} + \underset{(0.4)}{1.2T_i} + \underset{(2.9)}{5.8C_i} + \underset{(0.20)}{0.20L_i}$$

$$\bar{R}^2 = .92$$

where: P_i = the list price of the *i*th car (thousands of dollars)
W_i = the weight of the *i*th car (hundreds of pounds)
T_i = a dummy equal to 1 if the *i*th car has an automatic transmission, 0 otherwise
C_i = a dummy equal to 1 if the *i*th car has cruise control, 0 otherwise
L_i = the size of the engine of the *i*th car (in liters)

a. Your firm's pricing expert hypothesizes positive signs for all the slope coefficients in Model A. Test her expectations at the 95 percent level of confidence.

b. What econometric problems appear to exist in Model A? In particular, does the size of the coefficient of C cause any concern? Why? What could be the problem?

c. You decide to test the possibility that L is an irrelevant variable by dropping it and rerunning the equation, obtaining Model T below. Which model to you prefer? Why? (*Hint:* Be sure to use our four specification criteria.)

d. In answering part c, you surely noticed that the $\bar{R}^2$ figures were identical. Did this surprise you? Why or why not?

$$
\begin{aligned}
\text{Model T: } \hat{P} = 18 &+ \underset{(0.07)}{0.29W_i} + \underset{(0.03)}{1.2T_i} + \underset{(2.9)}{5.9C_i} \\
\bar{R}^2 &= .92
\end{aligned}
$$

12. Determine the sign (and, if possible, comment on the likely size) of the bias introduced by leaving a variable out of an equation in each of the following cases:
    a. In an annual equation for corn yields per acre (in year t), the impact on the coefficient of rainfall in year t of omitting average temperature that year. (*Hint:* Drought and cold weather both hurt corn yields.)
    b. In an equation for daily attendance at Los Angeles Lakers' home basketball games, the impact on the coefficient of the winning percentage of the opponent (as of the game in question) of omitting a dummy variable that equals 1 if the opponent's team includes a superstar.
    c. In an equation for annual consumption of apples in the United States, the impact on the coefficient of the price of bananas of omitting the price of oranges.
    d. In an equation for student grades on the first midterm in this class, the impact on the coefficient of total hours studied (for the test) of omitting hours slept the night before the test.

13. Suppose that you run a regression to determine whether gender or race has any significant impact on scores on a test of the economic understanding of children.[13] You model the score of the *i*th student on the test of elementary economics (S_i) as a function of the composite score on the Iowa Tests of Basic Skills of the *i*th student, a dummy variable equal to 1 if the *i*th student is female (0 otherwise), the average number of years of education of the parents of the *i*th student, and a dummy variable equal to 1 if the *i*th student is nonwhite (0 otherwise). Unfortunately, a rainstorm floods the computer center and makes it impossible to read the part of the computer output that identifies which variable is which. All you know is that the regression results are (standard errors in parentheses):

$$
\begin{aligned}
\hat{S}_i = 5.7 &- \underset{(0.63)}{0.63X_{1i}} - \underset{(0.88)}{0.22X_{2i}} + \underset{(0.08)}{0.16X_{3i}} + \underset{(0.01)}{0.12X_{4i}} \\
n &= 24 \quad \bar{R}^2 = .54
\end{aligned}
$$

[13] These results have been jiggled to meet the needs of this question, but this research actually was done. See Stephen Buckles and Vera Freeman, "Male-Female Differences in the Stock and Flow of Economic Knowledge," *Review of Economics and Statistics,* May 1983, pp. 355–357.

a. Attempt to identify which result corresponds to which variable. Be specific.

b. Explain the reasoning behind your answer to part a above.

c. Assuming that your answer is correct, create and test appropriate hypotheses (at the 5 percent level) and come to conclusions about the effects of gender and race on the test scores of this particular sample.

d. Did you use a one-tailed or two-tailed test in part c above? Why?

14. William Sander[14] estimated a 50-state cross-sectional model of the farm divorce rate as part of an effort to determine whether the national trend toward more divorces could be attributed in part to increases in the earning ability of women. His equation was (t-scores in parentheses):

$$\hat{Y}_i = -4.1 + \underset{(3.3)}{0.003P_i} + \underset{(1.5)}{0.06L_i} - \underset{(-0.6)}{0.002A_i} + \underset{(13.5)}{0.76N_i}$$

where: Y_i = the farm divorce rate in the *i*th state
P_i = the population density of the *i*th state
L_i = the labor force participation of farm women in the *i*th state
A_i = farm assets held by women in the *i*th state
N_i = the rural nonfarm divorce rate in that state

a. Develop and test hypotheses about the slope coefficients of Sander's equation at the 5 percent level.

b. What (if any) econometric problems (out of omitted variables and irrelevant variables) appear to exist in this equation? Justify your answer.

c. What one specification change in this equation would you suggest? Be specific.

d. Use our four specification criteria to decide whether you believe L is an irrelevant variable. The equation without L (t-scores again in parentheses) was:

$$\hat{Y}_i = -2.5 + \underset{(4.3)}{0.004P_i} - \underset{(-1.3)}{0.004A_i} + \underset{(14.8)}{0.79N_i}$$

(*Hint:* We don't provide $\bar{R}^2$ for these equations, but you can determine whether it went up or down anyway. How?)

[14] William Sander, "Women, Work, and Divorce," *The American Economic Review*, June 1985, pp. 519–523.

15. Look back again at Exercise 16 in Chapter 5, the equation on international price discrimination in pharmaceuticals. In that cross-sectional study, Schut and VanBergeijk estimated two equations in addition to the one cited in the exercise.[15] These two equations tested the possibility that CV_i, total volume of consumption of pharmaceuticals in the *i*th country, and N_i, the population of the *i*th country, belonged in the original equation, Equation 5.17, repeated here:

$$\begin{aligned} \hat{P}_i = \ & 38.22 + \underset{\substack{(0.21)\\ 6.69}}{1.43GDPN_i} - \underset{\substack{(0.22)\\ -2.66}}{0.6CVN_i} + \underset{\substack{(6.12)\\ 1.19}}{7.31PP_i} \qquad (5.17)\\ & -\underset{\substack{(6.93)\\ -2.25}}{15.63DPC_i} - \underset{\substack{(7.16)\\ -1.59}}{11.38IPC_i} \end{aligned}$$

(t = values given beneath the standard errors)

$$n = 32 \text{ (national 1975 data)} \quad \bar{R}^2 = .775 \quad F = 22.35$$

where:
P_i = the pharmaceutical price level in the *i*th country divided by that of the United States
$GDPN_i$ = per capita domestic product in the *i*th country divided by that of the United States
CVN_i = per capita volume of consumption of pharmaceuticals in the *i*th country divided by that of the United States
PP_i = a dummy variable equal to 1 if patents for pharmaceutical products are recognized in the *i*th country and equal to 0 otherwise
DPC_i = a dummy variable equal to 1 if the *i*th country applied strict price controls and 0 otherwise
IPC_i = a dummy variable equal to 1 if the *i*th country encouraged price competition and 0 otherwise

a. Using EViews (or your own computer program) and datafile DRUG5 (or Table 5.1), estimate these two equations. That is, estimate:
   i. Equation 5.17 with CV_i added, and
   ii. Equation 5.17 with N_i added

[15] Frederick T. Schut and Peter A. G. VanBergeijk, "International Price Discrimination: The Pharmaceutical Industry," *World Development*, 1986, pp. 1141–1150.

b. Use our four specification criteria to determine whether CV and N are irrelevant or omitted variables. (*Hint:* The authors expected that prices would be lower if market size was larger because of possible economies of scale and/or enhanced competition.)

c. Why didn't the authors run Equation 5.17 with *both* CV and N included? (*Hint:* While you can estimate this equation yourself, you don't have to do so to answer the question.)

d. Why do you think that the authors reported all three estimated specifications in their results when they thought that Equation 5.17 was the best?

6.8 Appendix: Additional Specification Criteria

So far in this chapter, we've suggested four criteria for choosing the independent variables (economic theory, $\overline{R}^2$, the *t*-test, and possible bias in the coefficients). Sometimes, however, these criteria don't provide enough information for a researcher to feel confident that a given specification is best. For instance, there can be two different specifications that both have excellent theoretical underpinnings. In such a situation, many econometricians use additional, often more formal, specification criteria to provide comparisons of the properties of the alternative estimated equations.

The use of formal specification criteria is not without problems, however. First, no test, no matter how sophisticated, can "prove" that a particular specification is the true one. The use of specification criteria, therefore, must be tempered with a healthy dose of economic theory and common sense. A second problem is that more than 20 such criteria have been proposed; how do we decide which one(s) to use? Because many of these criteria overlap with one another or have varying levels of complexity, a choice between the alternatives is a matter of personal preference.

In this section, we'll describe the use of three of the most popular specification criteria, J. B. Ramsey's RESET test, Akaike's Information Criterion, and the Schwarz Criterion. Our inclusion of just these techniques does not imply that other tests and criteria are not appropriate or useful. Indeed, the reader will find that most other formal specification criteria have quite a bit in common with at least one of the techniques that we include. We think that you'll be more able to use and understand other formal specification criteria[16] once you've mastered these three.

[16] In particular, the likelihood ratio test, versions of which will be covered in Section 12.2, can be used as a specification test. For an introductory level summary of six other specification criteria, see Ramu Ramanathan, *Introductory Econometrics* (Fort Worth: Harcourt Brace Jovanovich, 1998, pp. 164–166).

6.8.1 Ramsey's Regression Specification Error Test (RESET)

One of the most-used formal specification tests other than $\overline{R}^2$ is the Ramsey Regression Specification Test (RESET).[17] The **Ramsey RESET test** is a general test that determines the likelihood of an omitted variable or some other specification error by measuring whether the fit of a given equation can be significantly improved by the addition of $\hat{Y}^2$, $\hat{Y}^3$, and $\hat{Y}^4$ terms.

What's the intuition behind RESET? The additional terms act as proxies for any possible (unknown) omitted variables or incorrect functional forms. If the proxies can be shown by the *F*-test to have improved the overall fit of the original equation, then we have evidence that there is some sort of specification error in our equation. As we'll learn in Chapter 7, the $\hat{Y}^2$, $\hat{Y}^3$, and $\hat{Y}^4$ terms form a *polynomial* functional form. Such a polynomial is a powerful curve-fitting device that has a good chance of acting as a proxy for a specification error if one exists. If there is no specification error, then we'd expect the coefficients of the added terms to be insignificantly different from zero because there is nothing for them to act as a proxy for.

The Ramsey RESET test involves three steps:

1. Estimate the equation to be tested using OLS:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \tag{6.28}$$

2. Take the $\hat{Y}_i$ values from Equation 6.28 and create $\hat{Y}_i^2$, $\hat{Y}_i^3$, and $\hat{Y}_i^4$ terms. Then add these terms to Equation 6.28 as additional explanatory variables and estimate the new equation with OLS:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + \beta_5 \hat{Y}_i^4 + \epsilon_i \tag{6.29}$$

3. Compare the fits of Equations 6.28 and 6.29 using the *F*-test. If the two equations are significantly different in overall fit, we can conclude that it's likely that equation 6.28 is misspecified.

While the Ramsey RESET test is fairly easy to use, it does little more than signal *when* a major specification error might exist. If you encounter a significant Ramsey RESET test, then you face the daunting task of figuring out exactly *what* the error is! Thus, the test often ends up being more useful in "supporting" (technically, not refuting) a researcher's contention that a given

[17] J. B. Ramsey, "Tests for Specification Errors in Classical Linear Squares Regression Analysis," *Journal of the Royal Statistical Society,* 1969, pp. 350–371.

specification has no major specification errors than it is in helping find an otherwise undiscovered flaw.[18]

As an example of the Ramsey RESET test, let's return to the chicken demand model of this chapter to see if RESET can detect the known specification error (omitting the price of beef) in Equation 6.9. Step one involves running the original equation without PB.

$$\begin{aligned} \hat{Y}_t &= 32.9 - 0.70PC_t + 0.27YD_t \\ &\qquad\qquad (0.08) \qquad (0.01) \\ t &= \qquad -8.33 \qquad 45.91 \\ \bar{R}^2 = .984 \quad n &= 44 \text{ (annual 1951−1994)} \quad RSS = 185.66 \end{aligned} \tag{6.9}$$

For step two, we take $\hat{Y}_t$ from Equation 6.9, calculate $\hat{Y}_t^2$, $\hat{Y}_t^3$, and $\hat{Y}_t^4$, and then reestimate Equation 6.9 with the three new terms added in:

$$\begin{aligned} Y_t &= 23.80 - 0.59PC_t + 0.36YD_t + 0.02\hat{Y}_t^2 \\ &\qquad\qquad (1.71) \qquad (0.71) \qquad (0.08) \\ t &= \qquad -0.34 \qquad 0.50 \qquad +0.29 \\ &\quad - 0.007\hat{Y}_t^3 + 0.000055\hat{Y}_t^4 + e_t \\ &\qquad (0.0011) \qquad (0.000054) \\ t &= \qquad -0.68 \qquad +1.02 \\ \bar{R}^2 = .987 \quad n &= 44 \text{ (annual 1951–1994)} \quad RSS = 138.41 \end{aligned} \tag{6.30}$$

In step three, we compare the fits of the two equations by using the *F*-test. Specifically, we test the hypothesis that the coefficients of all three of the added terms are equal to zero:

$$H_0: \hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 0$$
$$H_A: \text{otherwise}$$

The appropriate F-statistic to use is one that is presented in more detail in Section 7.7:

$$F = \frac{(RSS_M - RSS)/M}{RSS/(n - K - 1)} \tag{6.31}$$

[18] The particular version of the Ramsey RESET test we describe in this section is only one of a number of possible formulations of the test. For example, some researchers delete the $\hat{Y}^4$ term from Equation 6.29. In addition, versions of the Ramsey RESET test are useful in testing for functional form errors (to be described in Chapter 7) and serial correlation (to be described in Chapter 9).

where RSS_M is the residual sum of squares from the restricted equation (Equation 6.9), RSS is the residual sum of squares from the unrestricted equation (Equation 6.30), M is the number of restrictions (3), and $(n - K - 1)$ is the number of degrees of freedom in the unrestricted equation (38):

$$F = \frac{(185.66 - 138.41)/3}{138.41/38} = 4.32$$

The critical F-value to use, 2.86, is found in Statistical Table B-2 at the 5 percent level of significance with 3 numerator and 38 denominator[19] degrees of freedom. Since 17.23 is greater than 2.86, we can reject the null hypothesis that the coefficients of the added variables are jointly zero, allowing us to conclude that there is indeed a specification error in Equation 6.9. Such a conclusion is no surprise, since we know that the price of beef was left out of the equation. Note, however, that the Ramsey RESET test tells us only that a specification error is likely to exist in Equation 6.9; it does not specify the details of that error.

6.8.2 Akaike's Information Criterion and the Schwarz Criterion

A second category of formal specification criteria involves adjusting the summed squared residuals (RSS) by one factor or another to create an index of the fit of an equation. The most popular criterion of this type is $\bar{R}^2$, but a number of interesting alternatives have been proposed.

Akaike's Information Criterion (AIC) and the *Schwarz Criterion* (SC) are methods of comparing alternative specifications by adjusting RSS for the sample size (n) and the number of dependent variables (K).[20] These criteria can be used to augment our four basic specification criteria when we try to decide if the improved fit caused by an additional variable is worth the decreased degrees of freedom and increased complexity caused by the addition. Their equations are:

$$AIC = \text{Log}(RSS/n) + 2(K + 1)/n \tag{6.32}$$

$$SC = \text{Log}(RSS/n) + \text{Log}(n)(K + 1)/n \tag{6.33}$$

[19] Statistical Table B-2 does not list 38 numerator degrees of freedom, so, as mentioned in footnote 15 of Chapter 5, you must interpolate between 30 and 40 numerator degrees of freedom to get the answer. In this case, some researchers would note that the calculated F-value exceeds both critical F-values and wouldn't bother with the interpolation. If you'd like more information about this kind of *F*-test, see Section 7.7.

[20] H. Akaike, "Likelihood of a Model and Information Criteria," *Journal of Econometrics*, 1981, pp. 3–14 and G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, 1978, pp. 461–464.

To use AIC and SC, estimate two alternative specifications and calculate AIC and SC for each equation. The lower AIC or SC are, the better the specification. Note that even though the two criteria were developed independently to maximize different object functions, their equations are quite similar. Both criteria tend to penalize the addition of another explanatory variable more than $\bar{R}^2$ does. As a result, AIC and SC will quite often[21] be minimized by an equation with fewer independent variables than the ones that maximize $\bar{R}^2$.

Let's apply Akaike's Information Criterion and the Schwarz Criterion to the same chicken demand example we used for Ramsey's RESET. To see if AIC and/or SC can detect the specification error we already know exists in Equation 6.9 (the omission of the price of beef), we need to calculate AIC and SC for equations with and without the price of beef. The equation with the lower AIC and SC values will, other things being equal, be our preferred specification.

The original chicken demand model, Equation 6.8, was:

$$\begin{array}{rl} \hat{Y}_t = 31.5 - & 0.73PC_t + 0.11PB_t + 0.23YD_t \qquad (6.8) \\ & (0.08) \qquad (0.05) \qquad (0.02) \\ t = & -9.12 \qquad\ 2.50 \qquad 14.22 \end{array}$$

$$\bar{R}^2 = .986 \quad n = 44 \text{ (annual 1951−1994)} \quad RSS = 160.59$$

Plugging the numbers from Equation 6.8 into Equations 6.32 and 6.33, AIC and PC can be seen to be:

$$AIC = Log(160.59/44) + 2(4)/44 = 1.48$$

$$SC = Log(160.59/44) + Log(44)4/44 = 1.64$$

Equation 6.9 (repeated in Section 6.8.1), which omits the price of beef, has an RSS of 185.66 with $K = 2$. Thus:

$$AIC = Log(185.66/44) + 2(3)/44 = 1.58$$

$$SC = Log(185.66/44) + Log(44)3/44 = 1.70$$

For AIC, $1.48 < 1.58$, and for SC, $1.64 < 1.70$, so both Akaike's Information

[21] Using a Monte Carlo study, Judge *et al.* showed that (given specific simplifying assumptions) a specification chosen by maximizing $\bar{R}^2$ is over 50 percent more likely to include an irrelevant variable than is one chosen by minimizing AIC or SC. See George C. Judge, R. Carter Hill, W. E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics* (New York: Wiley, 1988), pp. 849–850. At the same time, minimizing AIC or SC will omit a relevant variable more frequently than will maximizing $\bar{R}^2$.

Criterion and the Schwarz Criterion provide evidence that Equation 6.8 is preferable to Equation 6.9. That is, the price of beef appears to belong in the equation. In practice, these calculations may not be necessary because AIC and SC are automatically calculated by some regression software packages, including EViews.

As it turns out, then, all three new specification criteria indicate the presence of a specification error when we leave the price of beef out of the equation. This result is not surprising, since we purposely left out a theoretically justified variable, but it provides an example of how useful these criteria could be when we're less than sure about the underlying theory.

Note that AIC and SC require the researcher to come up with a particular alternative specification, whereas Ramsey's RESET does not. Such a distinction makes RESET easier to use, but it makes AIC and SC more informative if a specification error is found. Thus our additional specification criteria serve different purposes. RESET is most useful as a general test of the existence of a specification error, whereas AIC and SC are more useful as means of comparing two or more alternative specifications.