CHAPTER 6

Further Inference in the Multiple Regression Model

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain the concepts of restricted and unrestricted sums of squared errors and how they are used to test hypotheses.
2. Use the F-test to test single null hypotheses or joint null hypotheses.
3. Use your computer software to perform an F-test.
4. Test the overall significance of a regression model and identify the components of this test from your computer output.
5. From output of your computer software, locate (a) the sum of squared errors, (b) the F-value for the overall significance of a regression model, (c) the estimated covariance matrix for the least squares estimates, and (d) the correlation matrix for the explanatory variables.
6. Explain the relationship between the finite sample F-test and the large sample χ^2-test, and the assumptions under which each is suitable.
7. Obtain restricted least squares estimates that include nonsample information in the estimation procedure.
8. Explain the properties of the restricted least squares estimator. In particular, how do its bias and variance compare with those of the unrestricted, ordinary, least squares estimator?
9. Explain the differences between models designed for prediction and models designed to estimate a causal effect.
10. Explain what is meant by (a) an omitted variable and (b) an irrelevant variable. Explain the consequences of omitted and irrelevant variables for the properties of the least squares estimator.
11. Explain the concept of a control variable and the assumption necessary for a control variable to be effective.
12. Explain the issues that need to be considered when choosing a regression model.
13. Test for misspecification using RESET.
14. Compute forecasts, standard errors of forecast errors, and interval forecasts from a multiple regression model.
15. Use the Akaike information or Schwartz criteria to select variables for a predictive model.

16. Identify collinearity and explain its consequences for least squares estimation.
17. Identify influential observations in a multiple regression model.
18. Compute parameter estimates for a regression model that is nonlinear in the parameters and explain how nonlinear least squares differs from linear least squares.

KEYWORDS

χ^2-test
AIC
auxiliary regressions
BIC
causal model
collinearity
control variables
F-test
influential observations
irrelevant variables
nonlinear least squares
nonsample information
omitted variable bias
overall significance
prediction
predictive model
RESET
restricted least squares
restricted model
restricted *SSE*
SC
single and joint null hypotheses
unrestricted model
unrestricted *SSE*

Economists develop and evaluate theories about economic behavior. Hypothesis testing procedures are used to test these theories. In Chapter 5, we developed *t*-tests for null hypotheses consisting of a single restriction on one parameter β_k from the multiple regression model, and null hypotheses consisting of a single restriction that involves more than one parameter. In this chapter we extend our earlier analysis to testing a null hypothesis with two or more restrictions on two or more parameters. An important new development for such tests is the *F*-test. A large sample alternative that can be used under weaker assumptions is the χ^2-test.

The theories that economists develop sometimes provide nonsample information that can be used along with the information in a sample of data to estimate the parameters of a regression model. A procedure that combines these two types of information is called restricted least squares. It can be a useful technique when the data are not information-rich—a condition called collinearity—and the theoretical information is good. The restricted least squares procedure also plays a useful practical role when testing hypotheses. In addition to these topics, we discuss model specification for the multiple regression model, prediction, and the construction of prediction intervals. Model specification involves choosing a functional form and choosing a set of explanatory variables.

Critical to the choice of a set of explanatory variables is whether a model is to be used for prediction or causal analysis. For causal analysis, omitted variable bias and selection of control variables is important. For prediction, selection of variables that are highly correlated with the dependent variable is more relevant. We also discuss the problems that arise if our data are not sufficiently rich because the variables are collinear or lack adequate variation, and summarize concepts for detecting influential observations. The use of nonlinear least squares is introduced for models that are nonlinear in the parameters.

6.1 Testing Joint Hypotheses: The *F*-test

In Chapter 5 we showed how to use one- and two-tail *t*-tests to test hypotheses involving

1. A single coefficient
2. A linear combination of coefficients
3. A nonlinear combination of coefficients.

The test for a single coefficient was the most straightforward, requiring only the estimate of the coefficient and its standard error. For testing a linear combination of coefficients, computing the standard error of the estimated linear combination brought added complexity. It uses the variances and covariances of all estimates in the linear combination and can be computationally demanding if done on a hand calculator, especially if there are three or more coefficients in the linear combination. Software will perform the test automatically, however, yielding the standard error, the value of the t-statistic, and the p-value of the test. If assumptions MR1–MR6 hold then t-statistics have exact distributions, making the tests valid for small samples. If MR6 is violated, implying $(e_i|\mathbf{X})$ is no longer normally distributed, or if MR2: $E(e_i|\mathbf{X}) = 0$ is weakened to the conditions $E(e_i) = 0$ and $\text{cov}(e_i, x_{jk}) = 0$, then we need to rely on large sample results that make the tests approximately valid, with the approximation improving as sample size increases.

For testing non-linear combinations of coefficients, one must rely on large sample approximations even if assumptions MR1–MR6 hold, and the delta method must be used to compute standard errors. Derivatives of the nonlinear function and the covariance matrix of the coefficients are required, but as with a linear combination, software will perform the test automatically, computing the standard error for you, as well as the value of the t-statistic and its p-value. In Chapter 5 we gave an example of an interval estimate rather than a hypothesis test for a nonlinear combination, but that example—the optimal level of advertising—showed how to obtain all the ingredients needed for a test. For both hypothesis testing and interval estimation of a nonlinear combination, it is the standard error that requires more effort.

A characteristic of all the t tests in Chapter 5 is that they involve a single conjecture about one or more of the parameters—or, put another way, there is only one "equal sign" in the null hypothesis. In this chapter, we are interested in extending hypothesis testing to null hypotheses that involve multiple conjectures about the parameters. A null hypothesis with multiple conjectures, expressed with more than one equal sign, is called a **joint hypothesis**. An example of a joint hypothesis is testing whether a group of explanatory variables should be included in a particular model. Should variables on socioeconomic background, along with variables describing education and experience, be used to explain a person's wage? Does the quantity demanded of a product depend on the prices of substitute goods, or only on its own price? Economic hypotheses such as these must be formulated into statements about model parameters. To answer the first of the two questions, we set up a null hypothesis where the coefficients of all the socioeconomic variables are equal to zero. For the second question, the null hypothesis would equate the coefficients of prices of all substitute goods to zero. Both are of the form

$$H_0 : \beta_4 = 0, \ \beta_5 = 0, \ \beta_6 = 0 \tag{6.1}$$

where β_4, β_5, and β_6 are the coefficients of the socioeconomic variables, or the coefficients of the prices of substitute goods. The joint null hypothesis in (6.1) contains three conjectures (three equal signs): $\beta_4 = 0$, $\beta_5 = 0$, and $\beta_6 = 0$. A test of H_0 is a joint test for whether all three conjectures hold simultaneously.

It is convenient to develop the test statistic for testing hypotheses such as (6.1) within the context of an example. We return to Big Andy's Burger Barn.

EXAMPLE 6.1 | Testing the Effect of Advertising

The test used for testing a joint null hypothesis is the ***F*-test**. To introduce this test and concepts related to it, consider the Burger Barn sales model given in (5.23):

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.2}$$

Suppose now we wish to test whether *SALES* is influenced by advertising. Since advertising appears in (6.2) as both a linear term *ADVERT* and as a quadratic term $ADVERT^2$, advertising will have no effect on sales if $\beta_3 = 0$ and $\beta_4 = 0$; advertising will have an effect if $\beta_3 \neq 0$ or $\beta_4 \neq 0$ or if both β_3

and β_4 are nonzero. Thus, for this test our null and alternative hypotheses are

$$H_0 : \beta_3 = 0,\ \beta_4 = 0$$

$$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both are nonzero}$$

Relative to the null hypothesis $H_0 : \beta_3 = 0, \beta_4 = 0$, the model in (6.2) is called the **unrestricted model**; the restrictions in the null hypothesis have not been imposed on the model. It contrasts with the **restricted model**, which is obtained by assuming the parameter restrictions in H_0 are true. When H_0 is true, $\beta_3 = 0$ and $\beta_4 = 0$, and *ADVERT* and $ADVERT^2$ drop out of the model. It becomes

$$SALES = \beta_1 + \beta_2 PRICE + e \tag{6.3}$$

The *F*-test for the hypothesis $H_0 : \beta_3 = 0, \beta_4 = 0$ is based on a comparison of the sums of squared errors (sums of squared OLS residuals) from the unrestricted model in (6.2) and the restricted model in (6.3). Our shorthand notation for these two quantities is SSE_U and SSE_R, respectively.

Adding variables to a regression reduces the sum of squared errors—more of the variation in the dependent variable becomes attributable to the variables in the regression and less of its variation becomes attributable to the error. In terms of our notation, $SSE_R - SSE_U \geq 0$. Using the data in the file *andy* to estimate (6.2) and (6.3), we find that $SSE_U = 1532.084$ and $SSE_R = 1896.391$. Adding *ADVERT* and $ADVERT^2$ to the equation reduces the sum of squared errors from 1896.391 to 1532.084.

What the *F*-test does is to assess whether the reduction in the sum of squared errors is sufficiently large to be significant. If adding the extra variables has little effect on the sum of squared errors, then those variables contribute little to explaining variation in the dependent variable, and there is support for a null hypothesis that drops them. On the other hand, if adding the variables leads to a big reduction in the sum of squared errors, those variables contribute significantly to explaining the variation in the dependent variable, and we have evidence against the null hypothesis. The *F*-statistic determines what constitutes a large reduction or a small reduction in the sum of squared errors. It is given by

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \tag{6.4}$$

where J is the number of restrictions or number of hypotheses in H_0, N is the number of observations, and K is the number of coefficients in the unrestricted model.

To use the *F*-statistic to assess whether a reduction in the sum of squared errors is sufficient to reject the null hypothesis, we need to know its probability distribution when the null hypothesis is true. If assumptions MR1–MR6 hold, then, when the **null hypothesis is true**, the statistic F has what is called an *F*-distribution with J numerator degrees of freedom and $(N - K)$ denominator degrees of freedom. Some details about this distribution are given in Appendix B.3.8, with its typical shape illustrated in Figure B.9(a). **If the null hypothesis is not true,** then the difference between SSE_R and SSE_U becomes large, implying that the restrictions placed on the model by the null hypothesis significantly reduce the ability of the model to fit the data. A large value for $SSE_R - SSE_U$ means that the value of F tends to be *large*, so that we *reject* the null hypothesis if the value of the *F*-test statistic becomes too large. What is too large is decided by comparing the value of F to a critical value F_c, which leaves a probability α in the upper tail of the *F*-distribution with J and $N - K$ degrees of freedom. Tables of critical values for $\alpha = 0.01$ and $\alpha = 0.05$ are provided in Statistical Tables 4 and 5. The rejection region $F \geq F_c$ is illustrated in Figure B.9(a).

EXAMPLE 6.2 | The *F*-Test Procedure

Using the hypothesis testing steps introduced in Chapter 3, the *F*-test procedure for testing whether *ADVERT* and $ADVERT^2$ should be excluded from the sales equation is as follows:

1. *Specify the null and alternative hypotheses*: The joint null hypothesis is $H_0 : \beta_3 = 0,\ \beta_4 = 0$. The alternative hypothesis is $H_1 : \beta_3 \neq 0$ or $\beta_4 \neq 0$ or both are nonzero.

2. *Specify the test statistic and its distribution if the null hypothesis is true*: Having two restrictions in H_0 means $J = 2$. Also, recall that $N = 75$, so the distribution of the F-test statistic when H_0 is true is

$$F = \frac{(SSE_R - SSE_U)/2}{SSE_U/(75-4)} \sim F_{(2,71)}$$

3. *Set the significance level and determine the rejection region*: Using $\alpha = 0.05$, the critical value from the $F_{(2,71)}$-distribution is $F_c = F_{(0.95,2,71)}$, giving a rejection region of $F \geq 3.126$. Alternatively, H_0 is rejected if p-value ≤ 0.05.

4. *Calculate the sample value of the test statistic and, if desired, the p-value*: The value of the F-test statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(1896.391 - 1532.084)/2}{1532.084/(75-4)} = 8.44$$

The corresponding p-value is $p = P(F_{(2,71)} > 8.44) = 0.0005$.

5. *State your conclusion*: Since $F = 8.44 > F_c = 3.126$, we reject the null hypothesis that both $\beta_3 = 0$ and $\beta_4 = 0$, and conclude that at least one of them is not zero. Advertising does have a significant effect upon sales revenue. The same conclusion is reached by noting that p-value $= 0.0005 < 0.05$.

You might ask where the value $F_c = F_{(0.95,2,71)} = 3.126$ came from. The F critical values in Statistical Tables 4 and 5 are reported for only a limited number of degrees of freedom. However, exact critical values such as the one for this problem can be obtained for any number of degrees of freedom using your econometric software.

6.1.1 Testing the Significance of the Model

An important application of the F-test is for what is called testing the **overall significance** of a model. In Section 5.5.1, we tested whether the dependent variable y is related to a particular explanatory variable x_k using a t-test. In this section, we extend this idea to a joint test of the relevance of *all* the included explanatory variables. Consider again the general multiple regression model with $(K-1)$ explanatory variables and K unknown coefficients

$$y = \beta_1 + x_2\beta_2 + x_3\beta_3 + \cdots + x_K\beta_K + e \tag{6.5}$$

To examine whether we have a viable explanatory model, we set up the following null and alternative hypotheses:

$$H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0$$
$$H_1 : \textit{At least one} \text{ of the } \beta_k \text{ is nonzero for } k = 2, 3, \ldots, K \tag{6.6}$$

The null hypothesis is a joint one because it has $K-1$ components. It conjectures that each and every one of the parameters β_k, other than the intercept parameter β_1, are simultaneously zero. If this null hypothesis is true, none of the explanatory variables influence y, and thus our model is of little or no value. If the alternative hypothesis H_1 is true, then at least one of the parameters is not zero, and thus one or more of the explanatory variables should be included in the model. The alternative hypothesis does not indicate, however, which variables those might be. Since we are testing whether or not we have a viable explanatory model, the test for (6.6) is sometimes referred to as a **test of the overall significance of the regression model**. Given that the t-distribution can only be used to test a single null hypothesis, we use the F-test for testing the joint null hypothesis in (6.6). The unrestricted model is that given in (6.5). The restricted model, assuming the null hypothesis is true, becomes

$$y_i = \beta_1 + e_i \tag{6.7}$$

The least squares estimator of β_1 in this restricted model is $b_1^* = \sum_{i=1}^{N} y_i/N = \bar{y}$, which is the sample mean of the observations on the dependent variable. The *restricted* sum of squared errors

from the hypothesis (6.6) is

$$SSE_R = \sum_{i=1}^{N}\left(y_i - b_1^*\right)^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 = SST$$

In this one case, in which we are testing the null hypothesis that all the model parameters are zero *except the intercept*, the restricted sum of squared errors is the total sum of squares (SST) from the full unconstrained model. The unrestricted sum of squared errors is the sum of squared errors from the unconstrained model—that is, $SSE_U = SSE$. The number of restrictions is $J = K - 1$. Thus, to test the overall significance of a model, *but not in general*, the F-test statistic can be modified and written as

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} \tag{6.8}$$

The calculated value of this test statistic is compared to a critical value from the $F_{(K-1,\,N-K)}$ distribution. It is used to test the overall significance of a regression model. The outcome of the test is of fundamental importance when carrying out a regression analysis, and it is usually automatically reported by computer software as the F-value.

EXAMPLE 6.3 | Overall Significance of Burger Barns Equation

To illustrate, we test the overall significance of the regression, (6.2), used to explain Big Andy's sales revenue. We want to test whether the coefficients of *PRICE*, *ADVERT*, and $ADVERT^2$ are all zero, against the alternative that at least one of these coefficients is not zero. Recalling that the model is $SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$, the hypothesis testing steps are as follows:

1. We are testing

$$H_0 : \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

against the alternative

$$H_1 : \text{At least one of } \beta_2 \text{ or } \beta_3 \text{ or } \beta_4 \text{ is nonzero}$$

2. If H_0 is true, $F = \dfrac{(SST - SSE)/(4-1)}{SSE/(75-4)} \sim F_{(3,71)}$.
3. Using a 5% significance level, we find the critical value for the F-statistic with (3,71) degrees of freedom is $F_c = 2.734$. Thus, we reject H_0 if $F \geq 2.734$.
4. The required sums of squares are $SST = 3115.482$ and $SSE = 1532.084$ which give an F-value of

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} = \frac{(3115.482 - 1532.084)/3}{1532.084/(75-4)} = 24.459$$

Also, p-value $= P(F \geq 24.459) = 0.0000$, correct to four decimal places.

5. Since $24.459 > 2.734$, we reject H_0 and conclude that the estimated relationship is a significant one. A similar conclusion is reached using the p-value. We conclude that at least one of *PRICE*, *ADVERT*, or $ADVERT^2$ have an influence on sales. Note that this conclusion is consistent with conclusions that would be reached using separate t-tests for the significance of each of the coefficients in (5.25).

Go back and check the output from your computer software. Can you find the F-value 24.459 and the corresponding p-value of 0.0000 that form part of the routine output?

6.1.2 The Relationship Between *t*- and *F*-Tests

A question that may have occurred to you is what happens if we have a null hypothesis which is not a joint hypothesis; it only has one equality in H_0? Can we use an F-test for this case, or do we go back and use a t-test? The answer is when testing a single "equality" null hypothesis (a single restriction) against a "not equal to" alternative hypothesis, either a t-test or an F-test can be used; the test outcomes will be identical. Two-tail t-tests are equivalent to F-tests *when there is*

a single hypothesis in H_0. An F-test cannot be used as an alternative to a one-tail t-test, however. To explore these notions we return to the Big Andy example.

EXAMPLE 6.4 | When are *t*- and *F*-tests equivalent?

In Examples 6.1 and 6.2, we tested whether advertising affects sales by using an F-test to test whether $\beta_3 = 0$ and $\beta_4 = 0$ in the model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.9}$$

Suppose now we want to test whether *PRICE* affects *SALES*. Following the same F-testing procedure, we have $H_0: \beta_2 = 0$, $H_1: \beta_2 \neq 0$, and the restricted model

$$SALES = \beta_1 + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.10}$$

Estimating (6.9) and (6.10) gives $SSE_U = 1532.084$ and $SSE_R = 2683.411$, respectively. The required F-value is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)}$$

$$= \frac{(2683.411 - 1532.084)/1}{1532.084/(75-4)} = 53.355$$

The 5% critical vale is $F_c = F_{(0.95,\,1,\,71)} = 3.976$. Thus, we reject $H_0: \beta_2 = 0$.

Now let us see what happens if we use a t-test for the same problem: $H_0: \beta_2 = 0$ and $H_1: \beta_2 \neq 0$. The results from estimating (6.9) were

$$\underset{\text{(se)}}{\widehat{SALES} = \underset{(6.80)}{109.72} - \underset{(1.046)}{7.640}PRICE + \underset{(3.556)}{12.151}ADVERT - \underset{(0.941)}{2.768}ADVERT^2}$$

The t-value for testing $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$ is $t = 7.640/1.045939 = 7.30444$. The 5% critical value for the t-test is $t_c = t_{(0.975,\,71)} = 1.9939$. We reject $H_0: \beta_2 = 0$ because $7.30444 > 1.9939$. The reason for using so many decimals here will soon become clear. We wish to reduce rounding error to ensure the relationship between the t- and F-tests is correctly revealed.

Notice that the squares of the calculated and critical t-values are identical to the corresponding F-values. That is, $t^2 = (7.30444)^2 = 53.355 = F$ and $t_c^2 = (1.9939)^2 = 3.976 = F_c$. The reason for this correspondence is an exact relationship between the t- and F-distributions. The square of a t random variable with df degrees of freedom is an F random variable with 1 degree of freedom in the numerator and df degrees of freedom in the denominator: $t^2_{(df)} = F_{(1,df)}$. Because of this exact relationship, the p-values for the two tests are identical, meaning that we will always reach the same conclusion whichever approach we take. However, there is no equivalence when using a one-tail t-test when the alternative is an inequality such as $>$ or $<$. Because $F = t^2$, the F-test cannot distinguish between the left and right tails as is needed for a one-tail test. Also, the equivalence between t-tests and F-tests does not carry over when a null hypothesis consists of more than a single restriction. Under these circumstances ($J \geq 2$), an F-test needs to be used.

Summarizing the *F*-Test Procedure

1. The null hypothesis H_0 consists of one or more linear equality restrictions on the model parameters β_k. The number of restrictions is denoted by J. When $J = 1$, the null hypothesis is called a single null hypothesis. When $J \geq 2$, it is called a joint null hypothesis. The null hypothesis may not include any "greater than or equal to" or "less than or equal to" hypotheses.
2. The alternative hypothesis states that one or more of the equalities in the null hypothesis is not true. The alternative hypothesis may not include any "greater than" or "less than" options.
3. The test statistic is the F-statistic in equation (6.24).
4. If assumptions MR1–MR6 hold, and if the null hypothesis is true, F has the F-distribution with J numerator degrees of freedom and $N - K$ denominator degrees of freedom. The null hypothesis is *rejected* if $F \geq F_c$, where $F_c = F_{(1-\alpha,\,J,\,N-K)}$ is the critical value that leaves α percent of the probability in the upper tail of the F-distribution.
5. When testing a single equality null hypothesis, it is perfectly correct to use either the t- or F-test procedure: they are equivalent. In practice, it is customary to test single restrictions using a t-test. The F-test is usually reserved for joint hypotheses.

6.1.3 More General *F*-Tests

So far we have discussed the *F*-test in the context of whether a variable or a group of variables could be excluded from the model. The conjectures made in the null hypothesis were that particular coefficients are equal to zero. The *F*-test can also be used for much more general hypotheses. Any number of conjectures ($J \leq K$) involving linear hypotheses with equal signs can be tested. Deriving the restricted model implied by H_0 can be trickier, but the same general principles hold. The restricted sum of squared errors is still greater than the unrestricted sum of squared errors. In the restricted model, least squares estimates are obtained by minimizing the sum of squared errors subject to the restrictions on the parameters being true, and the unconstrained minimum (SSE_U) is always less than the constrained minimum (SSE_R). If SSE_U and SSE_R are substantially different, assuming that the null hypothesis is true significantly reduces the ability of the model to fit the data; in other words, the data do not support the null hypothesis, and it is rejected by the *F*-test. On the other hand, if the null hypothesis is true, we expect the data to be compatible with the conditions placed on the parameters. We expect little change in the sum of squared errors, in which case the null hypothesis will not be rejected by the *F*-test.

EXAMPLE 6.5 | Testing Optimal Advertising

To illustrate how to obtain a restricted model for a null hypothesis that is more complex than assigning zero to a number of coefficients, we return to Example 5.17 where we found that the optimal amount for Andy to spend on advertising $ADVERT_0$ is such that

$$\beta_3 + 2\beta_4 ADVERT_0 = 1 \tag{6.11}$$

Now suppose that Big Andy has been spending \$1900 per month on advertising and he wants to know whether this amount could be optimal. Does the information from the estimated equation provide sufficient evidence to reject a hypothesis that \$1900 per month is optimal? The null and alternative hypotheses for this test are

$$H_0: \beta_3 + 2 \times \beta_4 \times 1.9 = 1 \quad H_1: \beta_3 + 2 \times \beta_4 \times 1.9 \neq 1$$

After carrying out the multiplication, these hypotheses can be written as

$$H_0: \beta_3 + 3.8\beta_4 = 1 \quad H_1: \beta_3 + 3.8\beta_4 \neq 1$$

How do we obtain the restricted model implied by the null hypothesis? Note that when H_0 is true, $\beta_3 = 1 - 3.8\beta_4$. Substituting this restriction into the unrestricted model in (6.9) gives

$$SALES = \beta_1 + \beta_2 PRICE + (1 - 3.8\beta_4)ADVERT + \beta_4 ADVERT^2 + e$$

Collecting terms and rearranging this equation to put it in a form convenient for estimation yields

$$(SALES - ADVERT) = \beta_1 + \beta_2 PRICE + \beta_4 (ADVERT^2 - 3.8ADVERT) + e \tag{6.12}$$

Estimating this model by least squares with dependent variable $y = (SALES - ADVERT)$ and explanatory variables $x_2 = PRICE$ and $x_3 = (ADVERT^2 - 3.8ADVERT)$ yields the restricted sum of squared errors $SSE_R = 1552.286$. The unrestricted sum of squared errors is the same as before, $SSE_U = 1532.084$. We also have one restriction ($J = 1$) and $N - K = 71$ degrees of freedom. Thus, the calculated value of the *F*-statistic is

$$F = \frac{(1552.286 - 1532.084)/1}{1532.084/71} = 0.9362$$

For $\alpha = 0.05$, the critical value is $F_c = 3.976$. Since $F = 0.9362 < F_c = 3.976$, we do not reject H_0. We conclude that Andy's conjecture, that an advertising expenditure of \$1900 per month is optimal is compatible with the data.

Because there is only one conjecture in H_0, you can also carry out this test using the *t*-distribution. Check it out. For the *t*-value, you should find $t = 0.9676$. The value $F = 0.9362$ is equal to $t^2 = (0.9676)^2$, obeying the relationship between *t*- and *F*-random variables that we mentioned previously. You will also find that the *p*-values are identical. Specifically,

$$p\text{-value} = P(F_{(1,71)} > 0.9362) = P(t_{(71)} > 0.9676) + P(t_{(71)} < -0.9676) = 0.3365$$

The result $0.3365 > 0.05$ leads us to conclude that $ADVERT_0 = 1.9$ is compatible with the data.

You may have noticed that our description of this test has deviated slightly from the step-by-step hypothesis testing format introduced in Chapter 3 and used so far in the book.

The same ingredients were there, but the arrangement of them varied. From now on, we will be less formal about following these steps. By being less formal, we can expose you to the type of discussion you will find in research reports, but please remember that the steps were introduced for a purpose: to teach you good habits. Following the steps ensures that you include a description of all the relevant components of the test and that you think about the steps in the correct order. It is **not correct**, for example, to decide on the hypotheses or the rejection region **after** you observe the value of the statistic.

EXAMPLE 6.6 | A One-Tail Test

Suppose that, instead of wanting to test whether the data supports the conjecture "$ADVERT = 1.9$ is optimal," Big Andy wants to test whether the optimal value of $ADVERT$ is greater than 1.9. If he has been spending \$1900 per month on advertising, and he does not want to increase this amount unless there is convincing evidence that the optimal amount is greater than \$1900, he will set up the hypotheses

$$H_0: \beta_3 + 3.8\beta_4 \leq 1 \quad H_1: \beta_3 + 3.8\beta_4 > 1 \tag{6.13}$$

In this case, we can no longer use the F-test. Using a t-test instead, your calculations will reveal $t = 0.9676$. The rejection region for a 5% significane level is reject H_0 if $t \geq 1.667$. Because $0.9676 < 1.667$, we do not reject H_0. There is not enough evidence in the data to suggest the optimal level of advertising expenditure is greater than \$1900.

6.1.4 Using Computer Software

Though it is possible and instructive to compute an F-value by using the restricted and unrestricted sums of squares, it is often more convenient to use the power of econometric software. Most software packages have commands that will automatically compute t- and F-values and their corresponding p-values when provided with a null hypothesis. You should check your software. Can you work out how to get it to test null hypotheses similar to those we constructed? These tests belong to a class of tests called **Wald tests**; your software might refer to them in this way. Can you reproduce the answers we got for all the tests in Chapters 5 and 6?

EXAMPLE 6.7 | Two ($J = 2$) Complex Hypotheses

In this example, we consider a joint test of two of Big Andy's conjectures. In addition to proposing that the optimal level of monthly advertising expenditure is \$1900, Big Andy is planning staffing and purchasing of inputs on the assumption that when $PRICE = \$6$ and $ADVERT = 1.9$, sales revenue will be \$80,000 on average. In the context of our model, and in terms of the regression coefficients β_k, the conjecture is

$$\begin{aligned} &E(SALES|PRICE = 6, ADVERT = 1.9) \\ &= \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 \\ &= \beta_1 + 6\beta_2 + 1.9\beta_3 + 1.9^2\beta_4 \\ &= 80 \end{aligned}$$

Are the conjectures about sales and optimal advertising compatible with the evidence contained in the sample of data? We formulate the joint null hypothesis

$$H_0: \beta_3 + 3.8\beta_4 = 1,\ \beta_1 + 6\beta_2 + 1.9\beta_3 + 3.61\beta_4 = 80$$

The alternative is that at least one of these restrictions is not true. Because there are $J = 2$ restrictions to test jointly, we use an F-test. A t-test is not suitable. Note also that this is an example of a test with two restrictions that are more general than simply omitting variables. Constructing the restricted model requires substituting both of these restrictions into our extended model, which is left as an exercise. Using instead computer output obtained by supplying the two hypotheses directly to the software, we obtain a computed value for the F-statistic of 5.74 and a corresponding p-value of 0.0049. At a 5% significance level, the joint null hypothesis is rejected. As another exercise, use the least squares estimates to predict sales revenue for $PRICE = 6$ and $ADVERT = 1.9$. Has Andy been too optimistic about the level of sales, or too pessimistic?

6.1.5 Large Sample Tests

There are two key requirements for the F-statistic to have the F-distribution in samples of all sizes: (1) assumptions MR1–MR6 must hold and (2) the restrictions in H_0 must be *linear* functions of the parameters $\beta_1, \beta_2, \ldots, \beta_K$. In this section, we are concerned with what test statistics are valid in large samples when the errors are no longer normally distributed or when the strict exogeneity assumption is weakend to $E(e_i) = 0$ and $\text{cov}(e_i, x_{jk}) = 0\ (i \neq j)$. We will also make a few remarks about testing nonlinear hypotheses.

To appreciate the testing alternatives, details about how the F-statistic in (6.4) is constructed are in order. An F random variable is defined as the ratio of two independent chi-square (χ^2) random variables, each divided by their degrees of freedom.[1] That is, if $V_1 \sim \chi^2_{(m_1)}$ and $V_2 \sim \chi^2_{(m_2)}$, and V_1 and V_2 are independent, then

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1, m_2)}$$

In our case, the two independent χ^2 random variables are

$$V_1 = \frac{(SSE_R - SSE_U)}{\sigma^2} \sim \chi^2_{(J)} \quad \text{and} \quad V_2 = \frac{(N-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-K)}$$

If σ^2 were known, V_1 would be a natural candidate for testing whether the difference between SSE_R and SSE_U is sufficiently large to reject a null hypothesis. Because σ^2 is unknown, we use V_2 to eliminate it. Specifically,

$$F = \frac{V_1/J}{V_2/(N-K)} = \frac{\dfrac{(SSE_R - SSE_U)}{\sigma^2}\Big/J}{\dfrac{(N-K)\hat{\sigma}^2}{\sigma^2}\Big/(N-K)} = \frac{(SSE_R - SSE_U)/J}{\hat{\sigma}^2} \sim F_{(J, N-K)} \tag{6.13}$$

Note that $\hat{\sigma}^2 = SSE_U/(N-K)$, and so the result in (6.13) is identical to the F-statistic first introduced in (6.4).

When we drop the normality assumption or weaken the strict exogeneity assumption, the argument becomes slightly different. In this case, V_1 no longer has an exact χ^2-distribution, but we can nevertheless rely on asymptotic theory to say that

$$V_1 = \frac{(SSE_R - SSE_U)}{\sigma^2} \overset{a}{\sim} \chi^2_{(J)}$$

Then, we can go one step further and say that replacing σ^2 by its consistent estimator $\hat{\sigma}^2$ does not change the asymptotic distribution of V_1.[2] That is,

$$\hat{V}_1 = \frac{(SSE_R - SSE_U)}{\hat{\sigma}^2} \overset{a}{\sim} \chi^2_{(J)} \tag{6.14}$$

This statistic is a valid alternative for testing joint linear hypotheses in large samples under less restrictive assumptions, with the approximation improving as sample size increases. At a 5% significance level, we reject H_0 if $\hat{V}_1$ is greater than or equal to the critical value $\chi^2_{(0.95,J)}$, or if the p-value $P\left(\chi^2_{(J)} > \hat{V}_1\right)$ is less than 0.05. In response to an automatic test command, most software will give you values for both F and $\hat{V}_1$. The value for $\hat{V}_1$ will probably be referred to as "chi-square."

Although it is clear that $F = \hat{V}_1/J$, the two test alternatives will not necessarily lead to the same outcome; their p-values will be different. Both are used in practice, and it is possible

[1] See Appendices B.3.6 and B.3.8.

[2] See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Theorem D.16, page 1168 of online Appendix.

that the F-test will provide a better small-sample approximation than $\hat{V}_1$ even under the less restrictive assumptions. As the sample size grows (the degrees of freedom for the denominator of the F-statistic increase), the two tests become identical—their p-values become the same, and their critical values beome equivalent in the sense that $\lim_{N\to\infty} F_{(1-\alpha,J,N-K)} = \chi^2_{(1-\alpha,J)}/J$. Check it out yourself. Suppose $J = 4$ and $\alpha = 0.05$, then from Statistical Table 3, $\chi^2_{(0.95,4)}/4 = 9.488/4 = 2.372$. The F-values are in Statistical Table 4, but it is instructive to use software to provide a few extra values. Doing so, we find $F_{(0.95,\,4,\,60)} = 2.525$, $F_{(0.95,\,4,\,120)} = 2.447$, $F_{(0.95,\,4,\,500)} = 2.390$, $F_{(0.95,\,4,\,1000)} = 2.381$, and $F_{(0.95,\,4,\,10000)} = 2.373$. As $N - K$ increases, the 95th percentile of the F-distribution approaches 2.372.

EXAMPLES 6.2 and 6.5 | Revisited

When testing $H_0: \beta_3 = \beta_4 = 0$ in the equation

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{6.15}$$

we obtain $F = 8.44$ with corresponding p-value $= 0.0005$, and $\chi^2 = 16.88$ with corresponding p-value $= 0.0002$. Because there are two restrictions ($J = 2$), the F-value is half the χ^2-value. The p-values are different because the tests are different.

For testing $H_0: \beta_3 + 3.8\beta_4 = 1$, we obtain $F = 0.936$ with corresponding p-value $= 0.3365$ and $\chi^2 = 0.936$ with corresponding p-value $= 0.3333$. The F- and χ^2-values are equal because $J = 1$, but again the p-values are slightly different.

Testing Nonlinear Hypotheses

Test statistics for joint hypotheses which are nonlinear functions of the parameters are more challenging theoretically,[3] but nevertheless can typically be carried out by your software with relative ease. Only asymptotic results are available, and the relevant test statistic is the chi-square, although you may find that some software also gives an F-value. Another thing to be on lookout for is whether a nonlinear hypothesis can be re-framed as a linear hypothesis to avoid one aspect of the approximation.

EXAMPLE 6.8 | A Nonlinear Hypothesis

In Section 5.7.4, we found that, in terms of the parameters of equation (6.2), the optimal level of advertising is given by

$$ADVERT_0 = \frac{1-\beta_3}{2\beta_4}$$

To test the hypothesis that the optimal level is \$1,900 against the alternative that it is not \$1,900, we can set up the following hypotheses which are nonlinear in the parameters

$$H_0: \frac{1-\beta_3}{2\beta_4} = 1.9 \quad H_1: \frac{1-\beta_3}{2\beta_4} \neq 1.9 \tag{6.16}$$

There are three ways we can approach this problem. The first way is to convert the hypotheses so that they are linear in the parameters. That is, $H_0: \beta_3 + 3.8\beta_4 = 1$ versus $H_1: \beta_3 + 3.8\beta_4 \neq 1$. These are the hypotheses that we tested in Example 6.5. The p-value for the F-test was 0.337.

The second way is to test (6.16) using the t-test value

$$t = \frac{g(b_3, b_4) - 1.9}{\text{se}\left[g(b_3, b_4)\right]}$$

$$= \frac{(1-b_3)/2b_4 - 1.9}{\text{se}\left((1-b_3)/2b_4\right)} = \frac{2.0143 - 1.9}{0.1287} = 0.888$$

The values $g(b_3, b_4) = (1-b_3)/2b_4 = 2.0143$ and $\text{se}\left[g(b_3, b_4)\right] = \text{se}\left((1-b_3)/2b_4\right) = 0.1287$, were found in Example 5.17 for computing an interval estimate for $ADVERT_0$. The third way is to use the χ^2-test for testing (6.16). When we have only a single hypothesis, $\chi^2 = F = t^2 = (0.888)^2 = 0.789$. The F and t^2 critical values correspond, yielding a p-value of 0.377. The χ^2-test is a different test, however. It yields a p-value of 0.374.

[3] See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, pp. 211–212.

Having so many options will undoubtedly leave you wondering what to do. In general, the best strategy is to convert the hypotheses into ones that are linear if that is possible. Otherwise, the t- or χ^2-tests can be used, but the t-test option is not available if $J \geq 2$. The important thing to take away from this section is an appreciation of the different test statistics that appear on your software output—what they mean, where they come from, and the circumstances under which they are exact finite sample tests or asymptotic approximations.

6.2 The Use of Nonsample Information

In many estimation problems we have information over and above the information contained in the sample observations. This nonsample information may come from many places, such as economic principles or experience. When it is available, it seems intuitive that we should find a way to use it. If the nonsample information is correct, and if we combine it with the sample information, the precision with which we can estimate the parameters is improved.

To illustrate how we might go about combining sample and nonsample information, consider a model designed to explain the demand for beer. From the theory of consumer choice in microeconomics, we know that the demand for a good will depend on the price of that good, on the prices of other goods—particularly substitutes and complements—and on income. In the case of beer, it is reasonable to relate the quantity demanded (Q) to the price of beer (PB), the price of liquor (PL), the price of all other remaining goods and services (PR), and income (I). To estimate this demand relationship, we need a further assumption about the functional form. Using "ln" to denote the natural logarithm, we assume, for this case, that the log-log functional form is appropriate:

$$\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) + e \tag{6.17}$$

This model is a convenient one because it precludes infeasible negative prices, quantities, and income, and because the coefficients β_2, β_3, β_4, and β_5 are elasticities. See Section 4.6.

A relevant piece of nonsample information can be derived by noting that if all prices and income go up by the same proportion, we would expect there to be no change in quantity demanded. For example, a doubling of all prices and income should not change the quantity of beer consumed. This assumption is that economic agents do not suffer from "money illusion." Let us impose this assumption on our demand model and see what happens. Having all prices and income change by the same proportion is equivalent to multiplying each price and income by a constant. Denoting this constant by λ and multiplying each of the variables in (6.17) by λ yields

$$\begin{aligned} \ln(Q) &= \beta_1 + \beta_2 \ln(\lambda PB) + \beta_3 \ln(\lambda PL) + \beta_4 \ln(\lambda PR) + \beta_5 \ln(\lambda I) \\ &= \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) \\ &\quad + (\beta_2 + \beta_3 + \beta_4 + \beta_5) \ln(\lambda) + e \end{aligned} \tag{6.18}$$

Comparing (6.17) with (6.18) shows that multiplying each price and income by λ will give a change in $\ln(Q)$ equal to $(\beta_2 + \beta_3 + \beta_4 + \beta_5)\ln(\lambda)$. Thus, for there to be no change in $\ln(Q)$ when all prices and income go up by the same proportion, it must be true that

$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0 \tag{6.19}$$

Thus, we can say something about how quantity demanded should not change when prices and income change by the same proportion, and this information can be written in terms of a specific restriction on the parameters of the demand model. We call such a restriction **nonsample information**. If we believe that this nonsample information makes sense, and hence that the parameter restriction in (6.19) holds, then it seems desirable to be able to obtain estimates that obey this restriction.

To introduce the nonsample information, we solve the parameter restriction $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$ for one of the β_k's. Which one is not important mathematically, but for reasons that will become apparent, we solve for β_4:

$$\beta_4 = -\beta_2 - \beta_3 - \beta_5$$

Substituting this expression into the original model in (6.17) gives

$$\begin{aligned}\ln(Q) &= \beta_1 + \beta_2\ln(PB) + \beta_3\ln(PL) + \left(-\beta_2 - \beta_3 - \beta_5\right)\ln(PR) + \beta_5\ln(I) + e \\ &= \beta_1 + \beta_2\left[\ln(PB) - \ln(PR)\right] + \beta_3\left[\ln(PL) - \ln(PR)\right] + \beta_5\left[\ln(I) - \ln(PR)\right] + e \\ &= \beta_1 + \beta_2\ln\left(\frac{PB}{PR}\right) + \beta_3\ln\left(\frac{PL}{PR}\right) + \beta_5\ln\left(\frac{I}{PR}\right) + e \end{aligned} \tag{6.20}$$

By using the restriction to replace β_4, and using the properties of logarithms, we have constructed the new variables $\ln(PB/PR)$, $\ln(PL/PR)$, and $\ln(I/PR)$. These variables have an appealing interpretation. Because PR represents the price of all other goods and services, (PB/PR) and (PL/PR) can be viewed as the *real* price of beer and the *real* price of liquor, respectively, and (I/PR) can be viewed as *real* income. By applying least squares to the restricted equation (6.20), we obtain the **restricted least squares estimates** $\left(b_1^*, b_2^*, b_3^*, b_5^*\right)$. The restricted least squares estimate for β_4 is given by $b_4^* = -b_2^* - b_3^* - b_5^*$.

EXAMPLE 6.9 | Restricted Least Squares

Observations on Q, PB, PL, PR, and I, taken from a cross section of 30 households are stored in the file *beer*. Using these observations to estimate (6.20), we obtain

$$\begin{aligned}\widehat{\ln(Q)} = &-4.798 - \underset{(0.166)}{1.2994}\ln\left(\frac{PB}{PR}\right) + \underset{(0.284)}{0.1868}\ln\left(\frac{PL}{PR}\right) \\ &+ \underset{(0.427)}{0.9458}\ln\left(\frac{I}{PR}\right)\end{aligned}$$

(se)

and $b_4^* = -(-1.2994) - 0.1868 - 0.9458 = 0.1668$. We estimate the price elasticity of demand for beer as -1.30, the cross-price elasticity of demand for beer with respect to liquor as 0.19, the cross-price elasticity of demand for beer with respect to other goods and services as 0.17, and the income elasticity of demand for beer as 0.95.

Substituting the restriction into the original equation and rearranging it like we did to get (6.20) will always work, but it may not be necessary. Different software has different options for obtaining restricted least squares estimates. Please check what is available in the software of your choice.

What are the properties of the restricted least squares estimation procedure? If assumptions MR1–MR5 hold for the unrestricted model, then the restricted least squares estimator is biased, $E\left(b_k^*\right) \neq \beta_k$, *unless* the constraints we impose are *exactly* true. This result makes an important point about econometrics. A good *economist* will obtain more reliable parameter estimates than a poor one because a good economist will introduce better nonsample information. This is true at the time of model specification as well as later, when constraints might be applied to the model. Nonsample information is not restricted to constraints on the parameters; it is also used for model specification. *Good economic theory* is a very important ingredient in empirical research.

The second property of the restricted least squares estimator is that its variance is smaller than the variance of the least squares estimator, *whether the constraints imposed are true or not*. By combining nonsample information with the sample information, we reduce the variation in the estimation procedure caused by random sampling. This reduction in variance obtained by imposing restrictions on the parameters is not at odds with the Gauss–Markov theorem. The Gauss–Markov result that the least squares estimator is the best linear unbiased estimator applies

to linear and unbiased estimators that use data alone, and no constraints on the parameters. Including additional information with the data gives the added reward of a reduced variance. If the additional nonsample information is correct, we are unambiguously better off; the restricted least squares estimator is unbiased and has lower variance. If the additional nonsample information is incorrect, the reduced variance comes at the cost of bias. This bias can be a big price to pay if it leads to estimates substantially different from their corresponding true parameter values. Evidence on whether or not a restriction is true can be obtained by testing the restriction along the lines of the previous section. In the case of this particular demand example, the test is left as an exercise.

6.3 Model Specification

In what has been covered so far, we have generally taken the role of the model as given. Questions have been of the following type: Given a particular regression model, what is the best way to estimate its parameters? Given a particular model, how do we test hypotheses about the parameters of that model? How do we construct interval estimates for the parameters of a model? What are the properties of estimators in a given model? Given that all these questions require knowledge of the model, it is natural to ask where the model comes from. In any econometric investigation, choice of the model is one of the first steps. In this section, we focus on the following questions: What are the important considerations when choosing a model? What are the consequences of choosing the wrong model? Are there ways of assessing whether a model is adequate?

Three essential features of model choice are (1) choice of functional form, (2) choice of explanatory variables (regressors) to be included in the model, and (3) whether the multiple regression assumptions MR1–MR6, listed in Chapter 5, hold. The implications of some violations of these assumptions have already been discussed. In particular, we have seen how it is necessary to rely on large sample results for inference if the errors are no longer normally distributed (MR6 is violated), or if assumption MR2: $E(e_i|\mathbf{X}) = 0$ is weakened to the alternative assumption that $E(e_i) = 0$ and $\text{cov}(e_i, x_{jk}) = 0$ for $i \neq j$. Later chapters on heteroskedasticity, regression with time-series data, and endogenous regressors deal with violations of MR3, MR4 and $\text{cov}(e_i, x_{jk}) = 0$. In this section, we focus mainly on issues dealing with choice of regressors and also give some consideration to choice of functional form. The properties of alternative functional forms were considered in Sections 2.8, 4.3–4.6, and 5.6. When making a functional-form choice, we need to ask questions such as: How is the dependent variable y likely to respond when the regressors change? At a constant rate? At a decreasing rate? Is it reasonable to assume constant elasticities over the whole range of the data? Are there any patterns in the least squares residuals that suggest an alternative functional form? The use of least squares residuals for assessing the adequacy of a functional form was considered in Section 4.3.4.

For choice of regressors, a fundamental consideration is the purpose of the model—whether it is intended for prediction or for causal analysis. We turn now to that question.

6.3.1 Causality versus Prediction

With causal inference we are primarily interested in the effect of a change in a regressor on the conditional mean of the dependent variable. Is there an effect and, if so, what is its magnitude? We wish to be able to say that a one-unit change in an explanatory variable will cause a particular change in the mean of the dependent variable, other factors held constant. This type of analysis is important for policy work. For example, suppose a government is concerned about educational performance in schools and believes that large class sizes may be the cause of poor performance. Before it spends large sums of money increasing the number of teachers, and building more classrooms, it would want convincing evidence that class size does have an impact on performance. We would need to be able to separate the effect of class size from the effect of other variables

such as socioeconomic background. It may be that large classes tend to be in areas of poor socioeconomic background. Under these circumstances it is important to include all relevant variables so that we can be sure "other factors are held constant" when we measure the effect of class size.

On the other hand, if the purpose of a model is to predict the value of a dependent variable, then, for regressor choice, it is important to choose variables that are highly correlated with the dependent variable and that lead to a high R^2. Whether or not these variables have a direct effect on the dependent variable, and the possible omission of some relevant variables, are less important. Predictive analysis using variables from the increasingly popular field of "big data" is an example of where variables are chosen for their predictive ability rather than to examine causal relationships.

To appreciate the difference in emphasis, and when it matters, suppose the variables $(y_i, x_i, z_i), i = 1, 2, \ldots, N$ are randomly selected from a population satisfying

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i \tag{6.21}$$

We have chosen the notation x for one of the explanatory variables and z for the other explanatory variable to distinguish between what will be an included variable x and an omitted variable z. We assume $E(e_i|x_i, z_i) = 0$ and hence $E(y_i|x_i, z_i) = \beta_1 + \beta_2 x_i + \beta_3 z_i$. Under these assumptions, β_2 and β_3 have the causal interpretations

$$\beta_2 = \frac{\partial E(y_i|x_i, z_i)}{\partial x_i} \quad \beta_3 = \frac{\partial E(y_i|x_i, z_i)}{\partial z_i}$$

That is, β_2 represents the change in the mean of y from a change in x, other factors held constant, and β_3 represents the change in the mean of y from a change in z, other factors held constant The assumption $E(e_i|x_i, z_i) = 0$ is important for these interpretations. It means that changes in x_i or z_i have no impact on the error term. Now suppose that x_i and z_i are correlated, a common occurrence among explanatory variables. Because they are correlated, $E(z_i|x_i)$ will depend on x_i. Let us assume that this dependence can be represented by the linear function

$$E(z_i|x_i) = \gamma_1 + \gamma_2 x_i \tag{6.22}$$

Then, using (6.21) and (6.22), we have

$$\begin{aligned} E(y_i|x_i) &= \beta_1 + \beta_2 x_i + \beta_3 E(z_i|x_i) + E(e_i|x_i) \\ &= \beta_1 + \beta_2 x_i + \beta_3(\gamma_1 + \gamma_2 x_i) \\ &= (\beta_1 + \beta_3\gamma_1) + (\beta_2 + \beta_3\gamma_2)x_i \end{aligned}$$

where $E(e_i|x_i) = E_z[E(e_i|x_i, z_i)] = 0$ by the law of iterated expectations. If knowing x_i or z_i does not help to predict e_i, then knowing x_i does not help to predict e_i either.

Now, we can define $u_i = y_i - E(y_i|x_i)$, $\alpha_1 = \beta_1 + \beta_3\gamma_1$, and $\alpha_2 = \beta_2 + \beta_3\gamma_2$, and write

$$\begin{aligned} y_i &= (\beta_1 + \beta_3\gamma_1) + (\beta_2 + \beta_3\gamma_2)x_i + u_i \\ &= \alpha_1 + \alpha_2 x_i + u_i \end{aligned} \tag{6.23}$$

where $E(u_i|x_i) = 0$ by definition. Application of least squares to (6.23) will yield best linear unbiased estimates of α_1 and α_2. If the objective is to use x_i to predict y_i, we can proceed with this equation without worrying about the omission of z_i. However, because z_i is not held constant, α_2 does not measure the causal effect of x_i on y_i, which is given by β_2. The coefficient α_2 includes the indirect effect of x_i on z_i through γ_2 (which may or may not be causal), followed by the effect of that change in z_i on y_i, through β_3. Note that if $\beta_3 = 0$ (z_i does not effect y_i) or $\gamma_2 = 0$ (z_i and x_i are uncorrelated), then $\alpha_2 = \beta_2$ and estimation of α_2 gives the required causal effect.

Thus, to estimate a causal effect of a variable x using least squares, we need to start with a model where all variables that are correlated with x and impact on y are included. An alternative, valuable when data on all such variables are not available, is to use control variables. We discuss their use in Section 6.3.4.

6.3.2 Omitted Variables

As explained in the previous section, if our objective is to estimate a causal relationship, then the possible omission of relevant variables is a concern. In this section, we explore further the impact of omitting important variables. Such omissions are always a possibility. Our economic principles may have overlooked a variable, or lack of data may lead us to drop a variable even when it is prescribed by economic theory.

EXAMPLE 6.10 | Family Income Equation

To introduce the **omitted variable problem**, we consider a sample of married couples where both husbands and wives work. This sample was used by labor economist Tom Mroz in a classic paper on female labor force participation. The variables from this sample that we use in our illustration are stored in the file *edu_inc*. The dependent variable is the logarithm of annual family income *FAMINC* defined as the combined income of husband and wife. We are interested in the impact of level of education, both the husband's education (*HEDU*) and the wife's education (*WEDU*), on family income. The first equation to be estimated is

$$\ln(FAMINC) = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + e \qquad (6.24)$$

Coefficient estimates from this equation, their standard errors, and their *p*-values for testing whether they are significantly different from zero, are given in column (1) of Table 6.1. We estimate that an additional year of education

TABLE 6.1 Estimated Equations for Family Income

	ln(*FAMINC*)				
	(1)	(2)	(3)	(4)	(5)
C	10.264	10.539	10.238	10.239	10.310
HEDU	0.0439	0.0613	0.0448	0.0460	0.0517
(se)	(0.0087)	(0.0071)	(0.0086)	(0.0136)	(0.0133)
[*p*-value]	[0.0000]	[0.0000]	[0.0000]	[0.0007]	[0.0001]
WEDU	0.0390		0.0421	0.0492	
(se)	(0.0116)		(0.0115)	(0.0247)	
[*p*-value]	[0.0003]		[0.0003]	[0.0469]	
KL6			–0.1733	–0.1724	–0.1690
(se)			(0.0542)	(0.0547)	(0.0548)
[*p*-value]			[0.0015]	[0.0017]	[0.0022]
XTRA_X5				0.0054	–0.0321
(se)				(0.0243)	(0.0154)
[*p*-value]				[0.8247]	[0.0379]
XTRA_X6				–0.0069	0.0309
(se)				(0.0215)	(0.0101)
[*p*-value]				[0.7469]	[0.0023]
SSE	82.2648	84.4623	80.3297	80.3062	81.0622
RESET *p*-values					
1 term $(\hat{y}^2)$	0.3374	0.1017	0.1881	0.1871	0.1391
2 terms $(\hat{y}^2, \hat{y}^3)$	0.1491	0.0431	0.2796	0.2711	0.2715

for the husband will increase annual income by 4.4%, and an additional year of education for the wife will increase income by 3.9%. Both estimates are significantly different from zero at a 1% level of significance.[4]

What happens if we now incorrectly omit wife's education from the equation? The resulting estimates are given in column (2) of Table 6.1. Omitting *WEDU* leads to an estimate that suggests the effect of an extra year of education for the husband is 6.1%. The effect of the wife's education has been incorrectly attributed to the husband's education leading to an overstatement of the latter's importance. This change in the magnitude of a coefficient is typical of the effect of incorrectly omitting a relevant variable. Omission of a relevant variable (defined as one whose coefficient is nonzero) leads to an estimator that is biased. Naturally enough, this bias is known as **omitted variable bias**.

Omitted Variable Bias: A Proof To give a general expression for the bias for the case where one explanatory variable is omitted from a model with two explanatory variables, consider the model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$. Suppose that we incorrectly omit z_i from the model and estimate instead $y_i = \beta_1 + \beta_2 x_i + v_i$ where $v_i = \beta_3 z_i + e_i$. Then, the estimator used for β_2 is

$$b_2^* = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \beta_2 + \sum w_i v_i$$

where $w_i = (x_i - \bar{x}) / \sum(x_i - \bar{x})^2$. The second equality in this equation follows from Appendix 2D. Substituting for v_i yields

$$b_2^* = \beta_2 + \beta_3 \sum w_i z_i + \sum w_i e_i$$

Assuming that $E(e_i|\mathbf{x}, \mathbf{z}) = 0$, or alternatively, that (y_i, x_i, z_i) is a random sample and $E(e_i|x_i, z_i) = 0$, the conditional mean of b_2^* is

$$E(b_2^*|\mathbf{x}, \mathbf{z}) = \beta_2 + \beta_3 \sum w_i z_i = \beta_2 + \beta_3 \frac{\widehat{\text{cov}}(x, z)}{\widehat{\text{var}}(x)} \tag{6.25}$$

You are asked to prove this result in Exercise 6.3. Unconditionally, we have

$$E(b_2^*) = \beta_2 + \beta_3 E\left[\frac{\widehat{\text{cov}}(x, z)}{\widehat{\text{var}}(x)}\right] \tag{6.26}$$

and in large samples, under less restrictive conditions,

$$b_2^* \xrightarrow{p} \beta_2 + \beta_3 \frac{\text{cov}(x, z)}{\text{var}(x)} \tag{6.27}$$

Thus, $E(b_2^*) \neq \beta_2$ and b_2^* is not a consistent estimator for β_2. It is biased in small and large samples if $\text{cov}(x, z) \neq 0$. In terms of (6.25)—the result is similar for (6.26) and (6.27)—the bias is given by

$$\text{bias}(b_2^*|\mathbf{x}, \mathbf{z}) = E(b_2^*|\mathbf{x}, \mathbf{z}) - \beta_2 = \beta_3 \frac{\widehat{\text{cov}}(x, z)}{\widehat{\text{var}}(x)} \tag{6.28}$$

We can make four more interesting observations from the results in (6.25)–(6.28).

1. Omitting a relevant variable is a special case of using a restricted least squares estimator where the restriction $\beta_3 = 0$ is not true. It leads to a biased estimator for β_2 but one with a lower variance. In columns (1) and (2) of Table 6.1 the reduction in the standard error for the coefficient of *HEDU* from 0.0087 to 0.0071 is consistent with the lower-variance result.
2. Knowing the sign of β_3 and the sign of the covariance between x and z tells us the direction of the bias. In Example 6.9 we expect a wife's level of education to have a positive effect on family income $(\beta_3 > 0)$, and we expect husband's and wife's levels of education to be

[4]There are a number of other entries in Table 6.1 that we discuss in due course: estimates from other equations and RESET values.

TABLE 6.2 **Correlation Matrix for Variables Used in Family Income Example**

	ln(*FAMINC*)	*HEDU*	*WEDU*	*KL6*	*XTRA_X5*	*XTRA_X6*
ln(*FAMINC*)	1.000					
HEDU	0.386	1.000				
WEDU	0.349	0.594	1.000			
KL6	–0.085	0.105	0.129	1.000		
XTRA_X5	0.315	0.836	0.518	0.149	1.000	
XTRA_X6	0.364	0.821	0.799	0.160	0.900	1.000

positively correlated ($\text{cov}(x, z) > 0$). Thus, we expect an upward bias for the coefficient estimate in (2), as indeed has occurred. The positive correlation between *HEDU* and *WEDU* can be confirmed from the correlation matrix in Table 6.2.

3. The bias in (6.28) can also be written as $\beta_3\hat{\gamma}_2$ where $\hat{\gamma}_2$ is the least squares estimate of γ_2 from the regression equation $E(z|x) = \gamma_1 + \gamma_2 x$. This result is consistent with equation (6.23) where we explained how omitting a relevant variable can lead to an incorrect estimate of a causal effect.
4. The importance of the assumption $E(e_i|\mathbf{x}, \mathbf{z}) = 0$ becomes clear. In the equation $y_i = \beta_1 + \beta_2 x_i + v_i$, we have $E(v_i|x_i) = \beta_3 E(z_i|x_i)$. It is the nonzero value for $E(z_i|x_i)$ that leads to the biased estimator for β_2.

EXAMPLE 6.11 | Adding Children Aged Less Than 6 Years

There are, of course, other variables that could be included as explanators of family income. In column (3) of Table 6.1 we include *KL6*, the number of children less than 6 years old. The larger the number of young children, the fewer the number of hours likely to be worked and hence a lower family income would be expected. The estimated coefficient on *KL6* is negative, confirming this expectation. Also, despite the fact that *KL6* is not highly correlated with *HEDU* and *WEDU*, the coefficient estimates for these variables have increased slightly, indicating that once we hold the number of young children constant, the returns to education for both the wife and the husband are greater, with the greater increase going to the wife whose working hours would be the more likely to be affected by the presence of young children.

6.3.3 Irrelevant Variables

The consequences of omitting relevant variables may lead you to think that a good strategy is to include as many variables as possible in your model. However, doing so will not only complicate your model unnecessarily, it may inflate the variances of your estimates because of the presence of **irrelevant variables**—those whose coefficients are zero because they have no direct effect on the dependent variable.

EXAMPLE 6.12 | Adding Irrelevant Variables

To see the effect of irrelevant variables, we add two artificially generated variables *XTRA_X5* and *XTRA_X6* to the family income equation. These variables were constructed so that they are correlated with *HEDU* and *WEDU* but have no influence on family income. The results from including these two variables are given in column (4) of Table 6.1. What can we observe from these estimates? First, as expected, the coefficients of *XTRA_X5* and *XTRA_X6*

have p-values greater than 0.05. They do indeed appear to be irrelevant variables. Also, the standard errors of the coefficients estimated for all other variables have increased, with p-values increasing correspondingly. The inclusion of irrelevant variables that are correlated with the other variables in the equation has reduced the precision of the estimated coefficients of the other variables. This result follows because, by the Gauss–Markov theorem, the least squares estimator of the correct model is the minimum variance linear unbiased estimator.

Finally, let us check what happens if we retain *XTRA_X5* and *XTRA_X6*, but omit *WEDU*, leading to the results in column (5). The coefficients for *XTRA_X5* and *XTRA_X6* have become significantly different from zero at a 5% level of significance. The irrelevant variables have picked up the effect of the relevant omitted variable. While this may not matter if prediction is the main objective of the exercise, it can lead to very erroneous conclusions if we are trying to identify the causal effects of the included variables.

6.3.4 Control Variables

In the discussion so far, we have not explicitly distinguished between variables whose causal effect is of particular interest and other variables that may simply be in the equation to avoid omitted variable bias in the estimate of the causal coefficient. Variables included in the equation to avoid omitted variable bias in the coefficient of interest are called **control variables**. Control variables may be included in the equation because they have a direct effect on the dependent variable in their own right or because they can act as proxy variables for relevant omitted variables that are difficult to observe. For a control variable to serve its purpose and act as a proxy for an omitted variable, it needs to satisfy a **conditional mean independence** assumption. To introduce this assumption, we return to the equation

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i \tag{6.29}$$

where the observation (y_i, x_i, z_i) is obtained by random sampling and where $E(e_i|x_i, z_i) = 0$. Suppose we are interested in β_2, the causal effect of x_i on y_i, and, although β_3 provides the causal effect of z_i on y_i, we are not concerned about estimating it. Also suppose that z_i is omitted from the equation because it is unobservable or because data on it are too difficult to obtain, leaving the equation

$$y_i = \beta_1 + \beta_2 x_i + v_i$$

where $v_i = \beta_3 z_i + e_i$. If z_i and x_i are uncorrelated, there are no problems. Application of least squares to $y_i = \beta_1 + \beta_2 x_i + v_i$ will yield a consistent estimate for β_2. However, as indicated in (6.28), correlation between z_i and x_i leads to a bias in the least squares estimator for β_2 equal to $\beta_3 \text{cov}(x, z)/\text{var}(x)$.

Now consider another variable q that has the property

$$E(z_i|x_i, q_i) = E(z_i|q_i) \tag{6.30}$$

This property says that once we know q, knowing x does not provide any more information about z. It means that x and z will no longer be correlated once q has been partialled out. We say that z_i and x_i are **conditionally mean independent**. An example will help cement this concept.

When labor economists estimate wage equations they are particularly interested in the returns to education. In particular, what is the causal relationship between more education and higher wages? Other variables such as experience are typically added to the equation, but they are usually not the main focus. One variable that is clearly relevant, but difficult to include because it cannot be observed, is ability. Also, more able people are likely to have more education, and so ability and education will be correlated. Excluding the variable "ability" will bias the estimate of the causal effect of education on wages. Suppose, however, that we have observations on *IQ*. *IQ* will clearly be correlated with both education and ability. Will it satisfy the conditional mean independence assumption? We need to be able to write

$$E(ABILITY|EDUCATION, IQ) = E(ABILITY|IQ)$$

That is, once we know somebody's *IQ*, knowing their level of education does not add any extra information about their ability. Another way to think about it is that education is "as if" it was randomly assigned, once we have taken *IQ* into account. One could argue whether this is a reasonable assumption, but, if it is reasonable, then we can proceed to use *IQ* as a control variable or a proxy variable to replace *ABILITY*.

How a Control Variable Works Returning to equation (6.29), namely, $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$, we can write

$$E(y_i|x_i, q_i) = \beta_1 + \beta_2 x_i + \beta_3 E(z_i|x_i, q_i) + E(e_i|x_i, q_i) \tag{6.31}$$

If the conditional mean independence assumption in (6.30) holds, then $E(z_i|x_i, q_i) = E(z_i|q_i)$. For illustrative purposes, we assume $E(z_i|q_i)$ is a linear function of q_i, say $E(z_i|q_i) = \delta_1 + \delta_2 q_i$. We also need to assume that q_i has no direct effect on y_i, so that $E(e_i|x_i, q_i) = 0$.[5] Inserting these results into (6.31), we have

$$\begin{aligned} E(y_i|x_i, q_i) &= \beta_1 + \beta_2 x_i + \beta_3(\delta_1 + \delta_2 q_i) \\ &= \beta_1 + \beta_3\delta_1 + \beta_2 x_i + \beta_3\delta_2 q_i \\ &= \alpha_1 + \beta_2 x_i + \alpha_2 q_i \end{aligned}$$

where $\alpha_1 = \beta_1 + \beta_3\delta_1$ and $\alpha_2 = \beta_3\delta_2$. Defining $u_i = y_i - E(y_i|x_i, q_i)$, we have the equation

$$y_i = \alpha_1 + \beta_2 x_i + \alpha_2 q_i + u_i$$

Since $E(u_i|x_i, q_i) = 0$ by definition, least squares estimates of α_1, β_2, and α_2 will be consistent. Notice that we have been able to estimate β_2, the causal effect of x on y, but we have not been able to consistently estimate β_3, the causal effect of z on y.

This result holds if q is a perfect proxy for z. We may want to ask what happens if the conditional mean independence assumption does not hold, making q an **imperfect proxy** for z. Suppose

$$E(z_i|x_i, q_i) = \delta_1 + \delta_2 q_i + \delta_3 x_i$$

In this case q is not a perfect proxy because, after controlling for it, $E(z_i|x_i, q_i)$ still depends on x. Using similar algebra, we obtain

$$E(y_i|x_i, q_i) = (\beta_1 + \beta_3\delta_1) + (\beta_2 + \beta_3\delta_3) x_i + \beta_3\delta_2 q_i$$

The bias from using this equation to estimate β_2 is $\beta_3\delta_3$. The bias from omitting z instead of using the control variable is $\beta_3 \text{cov}(x, z)/\text{var}(x)$. Thus, for the control variable to be an improvement over omission of z, we require $\delta_3 < \text{cov}(x, z)/\text{var}(x)$. Now, $\text{cov}(x, z)/\text{var}(x)$ is equal to the coefficient of x in a regression of z on x. Thus, the condition $\delta_3 < \text{cov}(x, z)/\text{var}(x)$ is equivalent to saying that the coefficient of x in a regression of z on x is lower after the inclusion of q. Put another way, after partialling out q, the correlation between x and z is reduced but not eliminated.

EXAMPLE 6.13 | A Control Variable for Ability

To illustrate the use of a control variable, we consider the model

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 ABILITY + e$$

and use data stored in the data file *koop_tobias_87*, a subset of data used by Koop and Tobias.[6] The sample is restricted to white males who are at least 16 years of age and who worked at least 30 weeks and 800 hours during the year. The Koop–Tobias data extend from 1979 to 1993. We use

[5] In Exercise 6.4 you are invited to investigate how this assumption can be relaxed.

[6] G. Koop and J.L. Tobias (2004), "Learning about Heterogeneity in Returns to Schooling", *Journal of Applied Econometrics*, 19, 827–849.

observations from 1987, a total of $N = 1057$. The variables *EDUC* and *EXPER* are numbers of years of education and experience, respectively. The variable *ABILITY* is unobserved, but we have instead the proxy variable *SCORE*, which is constructed from the 10 component tests of the Armed Services Vocational Aptitude Battery, administered in 1980, and standardized for age. Omitting *ABILITY*, the least squares estimated equation is

$$\begin{aligned}\widehat{\ln(WAGE)} &= 0.887 + 0.0728EDUC + 0.01268EXPER \\ \text{(se)} \quad & (0.293)\ (0.0091) \qquad (0.0403) \\ & - 0.00571EXPER^2 \\ & (0.00165)\end{aligned}$$

Including the proxy variable *SCORE*, we obtain

$$\begin{aligned}\widehat{\ln(WAGE)} &= 1.055 + 0.0592EDUC + 0.1231EXPER \\ \text{(se)} \quad & (0.297)\ (0.0101) \qquad (0.0401) \\ & - 0.00538EXPER^2 + 0.0604SCORE \\ & (0.00165) \qquad (0.0195)\end{aligned}$$

The return to an extra year of education drops from 7.3% to 5.9% after including the variable *SCORE*, suggesting that omitting the variable *ABILITY* has incorrectly attributed some of its effect to the level of education. There has been little effect on the coefficients of *EXPER* and $EXPER^2$. The conditional mean independence assumption that has to hold to conclude that extra *EDUC* causes a 5.9% increase in *WAGE* is $E(ABILITY|EDUC, EXPER, SCORE) = E(ABILITY|EXPER, SCORE)$. After allowing for *EXPER* and *SCORE*, knowing *EDUC* does not provide any more information about *ABILITY*. This assumption is required for both the education and experience coefficients to be given a causal interpretation. Finally, we note that the coefficient of the proxy variable *SCORE* cannot be given a causal interpretation.

6.3.5 Choosing a Model

Although choosing a model is fundamental, it is often not an easy task. There is no one set of mechanical rules that can be applied to come up with the best model. The choice will depend on the purpose of the model and how the data were collected, and requires an intelligent application of both theoretical knowledge and the outcomes of various statistical tests. Better choices come with experience. What is important is to recognize ways of assessing whether a model is reasonable or not. The following points are helpful for such an assessment.

1. Is the purpose of the model to identify one or more causal effects or is it prediction? Where causality is the focus, omitted variable bias can invalidate conclusions. Careful selection of control variables, whether they be variables in their own right or proxy variables, is necessary. On the other hand, if prediction is the objective, then the major concern is using variables that have high predictive power because of their correlation with the dependent variable. Omitted variables bias is not a major concern.
2. Theoretical knowledge, expert assessment of likely behavior, and general understanding of the nature of the relationship are important considerations for choosing variables and functional form.
3. If an estimated equation has coefficients with unexpected signs, or unrealistic magnitudes, they could be caused by a misspecification such as the omission of an important variable.
4. Patterns in least squares residuals can be helpful for uncovering problems caused by an incorrect functional form. Some illustrations are given in Section 4.3.4.
5. One method for assessing whether a variable or a group of variables should be included in an equation is to perform significance tests. That is, t-tests for hypotheses such as $H_0 : \beta_3 = 0$ or F-tests for hypotheses such as $H_0 : \beta_3 = \beta_4 = 0$. Such tests can include coefficients of squares and products of variables as tests for a suitable functional form. Failure to reject a null hypotheses that one or more coefficients are zero can be an indication that the variable(s) are irrelevant. However, it is important to remember that failure to reject a null hypothesis can also occur if the data are not sufficiently rich to disprove the hypothesis. More will be said about poor data in Section 6.5. For the moment we note that, when

a variable has an insignificant coefficient, it can either be (a) discarded as an irrelevant variable, or (b) retained because the theoretical reason for its inclusion is a strong one.

6. Have the leverage, studentized residuals, DFBETAS, and DFFITS measures identified any influential observations?[7] If an unusual observation is not a data error, then understanding why it occurred may provide useful information for setting up the model.
7. Are the estimated coefficients robust with respect to alternative specifications? If the model is designed to be a causal one, and estimates of the causal coefficient change dramatically when different specifications of the model are estimated, or different sets of control variables are included, then there is cause for concern.
8. A test known as RESET (Regression Specification Error Test) can be useful for detecting omitted variables or an incorrect functional form. Details of this test are provided in Section 6.3.6.
9. Various model selection criteria, based on maximizing R^2, or minimizing the sum of squared errors (*SSE*), subject to a penalty for too many variables, have been suggested. These criteria are more valuable when a model is designed for prediction rather than causal analysis. For reliable prediction a sum of squared errors that is small relative to the explanatory power of the model is essential. We describe three of these criteria in Section 6.4.1: an adjusted R^2, the Akaike information criterion (AIC), and the Schwarz criterion (SC), also known as the Bayesian information criterion (BIC).
10. A more stringent assessment of a model's predictive ability is to use a "hold-out" sample. A least squares estimated equation is designed to minimize the within-sample sum of squared errors. To check out a model's ability to predict outside the sample, some observations can be withheld from estimation and the model can be assessed on its ability to predict the withheld observations. More details are provided in Section 6.4.1.
11. Following the guidelines in the previous 10 points can almost inevitably lead to revisions of originally proposed models, or to more general experimentation with alternative models. Searching for a model with "significant" estimates and the selective reporting of a finally chosen "significant" model is a questionable practice. Not knowing the search process that led to the selected results makes valid interpretation of the results difficult. Proper reporting of results should include disclosure of all estimated models and the criteria used for model selection.

6.3.6 RESET

Testing for model misspecification is a way of asking whether our model is adequate, or whether we can improve on it. It could be misspecified if we have omitted important variables, included irrelevant ones, chosen a wrong functional form, or have a model that violates the assumptions of the multiple regression model. **RESET** (REgression Specification Error Test) is designed to detect omitted variables and incorrect functional form. It proceeds as follows.

Suppose that we have specified and estimated the regression model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

Let (b_1, b_2, b_3) be the least squares estimates, and let

$$\hat{y} = b_1 + b_2 x_2 + b_3 x_3 \tag{6.32}$$

be the fitted values of y. Consider the following two artificial models:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + e \tag{6.33}$$

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e \tag{6.34}$$

[7] These measures for detecting influential observations are discussed in Sections 4.3.6 and 6.5.3.

In (6.33), a test for misspecification is a test of $H_0: \gamma_1 = 0$ against the alternative $H_1: \gamma_1 \neq 0$. In (6.34), testing $H_0: \gamma_1 = \gamma_2 = 0$ against $H_1: \gamma_1 \neq 0$ and/or $\gamma_2 \neq 0$ is a test for misspecification. In the first case, a t- or an F-test can be used. An F-test is required for the second equation. Rejection of H_0 implies that the original model is inadequate and can be improved. A failure to reject H_0 says that the test has not been able to detect any misspecification.

To understand the idea behind the test, note that $\hat{y}^2$ and $\hat{y}^3$ will be polynomial functions of x_2 and x_3. If you square and cube both sides of (6.32), you will get terms such as $x_2^2, x_3^3, x_2x_3, x_2x_3^2$, and so on. Since polynomials can approximate many different kinds of functional forms, if the original functional form is not correct, the polynomial approximation that includes $\hat{y}^2$ and $\hat{y}^3$ may significantly improve the fit of the model. If it does, this fact will be detected through nonzero values of γ_1 and γ_2. Furthermore, if we have omitted variables and these variables are correlated with x_2 and x_3, then they are also likely to be correlated with terms such as x_2^2 and x_3^2, so some of their effect may be picked up by including the terms $\hat{y}^2$ and/or $\hat{y}^3$. Overall, the general philosophy of the test is if we can significantly improve the model by artificially including powers of the predictions of the model, then the original model must have been inadequate.

EXAMPLE 6.14 | Applying RESET to Family Income Equation

To illustrate RESET we return to the family income equation considered in Examples 6.10–6.12. In those examples specifications with different variables included were estimated, and the results presented in Table 6.1. The full model, without the irrelevant variables, was

$$\ln(FAMINC) = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + \beta_4 KL6 + e$$

Please go back and check Table 6.1, where RESET p-values for both $H_0: \gamma_1 = 0$ and $H_0: \gamma_1 = \gamma_2 = 0$ are presented in the last two rows of the table. The only instance where RESET rejects a model at a 5% significance level is where wife's education has been excluded and the null hypothesis is $H_0: \gamma_1 = \gamma_2 = 0$. Exclusion of *KL6* is not picked up by RESET, most likely because it is not highly correlated with *HEDU* and *WEDU*. Also, when the irrelevant variables *XTRA_X5* and *XTRA_X6* are included, and *WEDU* is excluded, RESET does not pick up the misspecification. The likely cause of this failure is the high correlations between *WEDU* and the two irrelevant variables.

There are two important lessons from this example. First, if RESET does not reject a model, that model is not necessarily a good one. Second, RESET will not always discriminate between alternative models. Rejection of the null hypothesis is indicative of a misspecification, but failure to reject the null hypothesis tells us very little.

6.4 Prediction

The prediction or forecasting problem for a regression model with one explanatory variable was introduced in Section 4.1. That material extends naturally to the more general model that has more than one explanatory variable. In this section, we describe that extension, reinforce earlier material, and provide some more general background.

Suppose we have values on $K-1$ explanatory variables represented by $\mathbf{x}_0 = (1, x_{02}, x_{03}, \ldots, x_{0K})$, and that we wish to use this information to predict or forecast a corresponding dependent variable value y_0. In Appendix 4D we learned that the minimum mean square error predictor for y_0 is the conditional expectation $E(y_0|\mathbf{x}_0)$. To make this result operational, we need to make an assumption about the functional form for $E(y_0|\mathbf{x}_0)$, and estimate the parameters on which it depends. In line with the multiple regression model, we assume that the conditional expectation is the linear-in-the-parameters function

$$E(y_0|\mathbf{x}_0) = \beta_1 + \beta_2 x_{02} + \beta_3 x_{03} + \cdots + \beta_K x_{0K} \tag{6.35}$$

Defining $e_0 = y_0 - E(y_0|\mathbf{x}_0)$, we can write

$$y_0 = \beta_1 + \beta_2 x_{02} + \beta_3 x_{03} + \cdots + \beta_K x_{0K} + e_0 \tag{6.36}$$

To estimate the parameters $(\beta_1, \beta_2, \ldots, \beta_K)$ in (6.35), we assume we have $i = 1, 2, \ldots, N$ observations y_i and $\mathbf{x}_i = (1, x_{i2}, x_{i3}, \ldots, x_{iK})$ such that

$$E(y_i|\mathbf{x}_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} \tag{6.37}$$

Define $e_i = y_i - E(y_i|\mathbf{x}_i)$ so that the model used to estimate $(\beta_1, \beta_2, \ldots, \beta_K)$ can be written as

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i \tag{6.38}$$

Equations (6.35)–(6.38) make up the **predictive model**. Equations (6.37) and (6.38) refer to the sample observations used to estimate the parameters. Equation (6.35) is the predictor that would be used if the parameters $(\beta_1, \beta_2, \ldots \beta_K)$ were known. Equation (6.36) incorporates the realized value y_0 and the error e_0. When we think of prediction or forecasting—we use the two terms interchangeably—we naturally think of forecasting outside the sample observations. Under these circumstances y_0 will be unobserved at the time the forecast is made. With time-series data, $\mathbf{x}_0$ will be future values of the explanatory variables for which a forecast is required; for cross-section data it will be values for an individual or some other economic unit that was not sampled. There are, however, instances where we make within-sample predictions or forecasts despite the fact that we have observed realized values for y for those observations. One example is their use in RESET where the regression equation was augmented with the squares and cubes of the within-sample predictions. When we are considering within-sample predictions, $\mathbf{x}_0$ will be identical to one of the $\mathbf{x}_i$, or it can be viewed as generic notation to represent all $\mathbf{x}_i$.

Note that (6.36) and (6.38) do **not** have to be causal models. To have a good predictive model, (y_i, y_0) needs to be highly correlated with the variables in $(\mathbf{x}_i, \mathbf{x}_0)$, but there is no requirement that (y_i, y_0) be caused by $(\mathbf{x}_i, \mathbf{x}_0)$. There is no requirement that all variables that affect y have to be included and there is no such thing as omitted variable bias. In (6.38), we are simply estimating the conditional expectation of the variables that are included. Under these circumstances, the interpretation of (e_i, e_0) is different from its interpretation in a **causal model**. In a causal model e represents the effect of variables omitted from the model; it is important that these effects are isolated from those in the model through the exogeneity assumption. We think of e as part of the data generating process. In a predictive model the coefficients in the conditional expectation can represent the direct effect of included variables and the indirect effect of excluded variables. The error term e is simply the difference between the realized value y and its conditional expectation; it is the **forecasting error** that would occur if $(\beta_1, \beta_2, \ldots, \beta_K)$ were known and did not have to be estimated. It does not take on an "all-other-variables" interpretation.

Application of least squares to (6.35) will yield unbiased estimates of $(\beta_1, \beta_2, \ldots, \beta_K)$ conditional on $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. If we assume further that $\text{var}(e_i|\mathbf{X}) = \sigma^2$ and $E(e_i e_j|\mathbf{X}) = 0$ for $i \neq j$, then the least squares estimator is best linear unbiased conditional on $\mathbf{X}$. Unconditionally, it will be a consistent estimator providing assumptions about the limiting behavior of the explanatory variables hold.[8] Having obtained the least squares estimates (b_1, b_2, …, b_K), we can define an operational predictor for y_0 as (6.35) with the unknown β_k replaced by their estimators. That is,

$$\hat{y}_0 = b_1 + b_2 x_{02} + b_3 x_{03} + \cdots + b_K x_{0K} \tag{6.39}$$

An extra assumption that we need is that $(e_0|\mathbf{x}_0)$ is uncorrelated with $(e_i|\mathbf{X})$ for $i = 1, 2, \ldots, N$ and $i \neq 0$. We also assume $\text{var}(e_0|\mathbf{x}_0) = \text{var}(e_i|\mathbf{X}) = \sigma^2$, an assumption used when deriving the variance of the forecast error.

After replacing the β_k with b_k, the forecast error is given by

$$\begin{aligned} f &= y_0 - \hat{y}_0 \\ &= (\beta_1 - b_1) + (\beta_2 - b_2) x_{02} + (\beta_3 - b_3) x_{03} + \cdots + (\beta_K - b_K) x_{0K} + e_0 \end{aligned} \tag{6.40}$$

There are two components in this forecast error: the errors $(\beta_k - b_k)$ from estimating the unknown parameters, and an error e_0 which is the deviation of the realized y_0 from its conditional mean. The predictor $\hat{y}_0$ is unbiased in the sense that $E(f|\mathbf{x}_0, \mathbf{X}) = 0$ and it is a best linear unbiased

[8] See Section 5.7.1 for an illustration in the case of simple regression.

predictor in the sense that the conditional variance $\text{var}(f|\mathbf{x}_0, \mathbf{X})$ is no greater than that of any other linear unbiased predictor. The conditional variance of the prediction error is

$$\begin{aligned}\text{var}(f|\mathbf{x}_0, \mathbf{X}) &= \text{var}\left[\left(\sum_{k=1}^{K}(\beta_k - b_k)x_{0k}\right)\middle|\mathbf{x}_0, \mathbf{X}\right] + \text{var}(e_0|\mathbf{x}_0, \mathbf{X}) \\ &= \text{var}\left[\left(\sum_{k=1}^{K} b_k x_{0k}\right)\middle|\mathbf{x}_0, \mathbf{X}\right] + \sigma^2 \\ &= \sum_{k=1}^{K} x_{0k}^2 \text{var}(b_k|\mathbf{x}_0, \mathbf{X}) + 2\sum_{k=1}^{K}\sum_{j=k+1}^{K} x_{0k}x_{0j}\text{cov}(b_k, b_j|\mathbf{x}_0, \mathbf{X}) + \sigma^2 \end{aligned} \tag{6.41}$$

In the first line of this equation we have assumed that the covariance between $(\beta_k - b_k)$ and e_0 is zero. This assumption will indeed be true for out-of-sample prediction and where e_0 is uncorrelated with the sample data used to estimate the β_k. For within-sample prediction the situation is more complicated. Strictly speaking, if e_0 is equal to one of the e_i in the sample, then $(\beta_k - b_k)$ and e_0 will be correlated. This correlation will not be large relative to the overall variance of f, however, and tends to get ignored in software calculations. In the second line of (6.41) $\beta_k x_{0k}$ can be treated as a constant and so $\text{var}\left((\beta_k - b_k)x_{0k}|\mathbf{x}_0, \mathbf{X}\right) = \text{var}(b_k x_{0k}|\mathbf{x}_0, \mathbf{X})$. The third line follows from the rule for calculating the variance of a weighted sum in (P.20) of the Probability Primer.

Each of the terms in the expression for $\text{var}(f|\mathbf{x}_0, \mathbf{X})$ involves σ^2. To obtain the estimated variance of the forecast error $\widehat{\text{var}}(f|\mathbf{x}_0, \mathbf{X})$, we replace σ^2 with its estimator $\hat{\sigma}^2$. The standard error of the forecast is given by $\text{se}(f) = \sqrt{\widehat{\text{var}}(f|\mathbf{x}_0, \mathbf{X})}$. If the random errors e_i and e_0 are normally distributed, or if the sample is large, then

$$\frac{f}{\text{se}(f)} = \frac{y_0 - \hat{y}_0}{\sqrt{\widehat{\text{var}}(y_0 - \hat{y}_0|\mathbf{x}_0, \mathbf{X})}} \sim t_{(N-K)} \tag{6.42}$$

Following the steps we have used many times, a $100(1-\alpha)\%$ interval predictor for y_0 is $\hat{y}_0 \pm t_c \text{se}(f)$, where t_c is a critical value from the $t_{(N-K)}$-distribution.

Before providing an example there are two practical considerations worth mentioning. First, in (6.41), the error variance σ^2 is typically much larger than the variance of the other component—that part of the forecast error attributable to estimation of the β_k. Consequently, this latter component is sometimes ignored and $\text{se}(f) = \hat{\sigma}$ is used. Second, the framework presented so far does not capture many of the typical characteristics of time-series forecasting. With time-series forecasting, some of the explanatory variables will usually be lagged values of the dependent variable. This means that the conditional expectation of a y_0 will depend on past values of itself. The sample information contributes to the conditional expectation of y_0. In the above exposition we have treated $\mathbf{x}_0$ as future values of the explanatory variables. The sample information has only contributed to the predictor through the estimation of the unknown β_k. In other words, $E(y_0|\mathbf{x}_0) = E(y_0|\mathbf{x}_0, \mathbf{X}, \mathbf{y})$, where $\mathbf{y}$ is used to denote all observations on the dependent variable. A more general scenario for time-series forecasting where this assumption is relaxed is considered in Chapter 9.

EXAMPLE 6.15 | Forecasting *SALES* for the Burger Barn

We are concerned with finding a 95% prediction interval for *SALES* at Big Andy's Burger Barn when $PRICE_0 = 6$, $ADVERT_0 = 1.9$ and $ADVERT_0^2 = 3.61$. These are the values considered by Big Andy in Example 6.6. In terms of the general notation $\mathbf{x}_0 = (1, 6, 1.9, 3.61)$. The point prediction is

$$\begin{aligned}\widehat{SALES}_0 &= 109.719 - 7.640\,PRICE_0 + 12.1512\,ADVERT_0 \\ &\quad - 2.768\,ADVERT_0^2 \\ &= 109.719 - 7.640 \times 6 + 12.1512 \times 1.9 - 2.768 \\ &\quad \times 3.61 \\ &= 76.974\end{aligned}$$

With the settings proposed by Big Andy, we forecast that sales will be \$76,974.

To obtain a prediction interval, we first need to compute the estimated variance of the forecast error. Using equation (6.41) and the covariance matrix values in Table 6.3, we have

$$\begin{aligned}
\widehat{\text{var}}(f|\mathbf{x}_0,\mathbf{X}) &= \hat{\sigma}^2 + \widehat{\text{var}}(b_1|\mathbf{x}_0,\mathbf{X}) + x_{02}^2\widehat{\text{var}}(b_2|\mathbf{x}_0,\mathbf{X}) \\
&\quad + x_{03}^2\widehat{\text{var}}(b_3|\mathbf{x}_0,\mathbf{X}) + x_{04}^2\widehat{\text{var}}(b_4|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{02}\widehat{\text{cov}}(b_1,b_2|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{03}\widehat{\text{cov}}(b_1,b_3|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{04}\widehat{\text{cov}}(b_1,b_4|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{02}x_{03}\widehat{\text{cov}}(b_2,b_3|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{02}x_{04}\widehat{\text{cov}}(b_2,b_4|\mathbf{x}_0,\mathbf{X}) \\
&\quad + 2x_{03}x_{04}\widehat{\text{cov}}(b_3,b_4|\mathbf{x}_0,\mathbf{X}) \\
&= 21.57865 + 46.22702 + 6^2 \times 1.093988 \\
&\quad + 1.9^2 \times 12.6463 + 3.61^2 \times 0.884774 \\
&\quad + 2 \times 6 \times (-6.426113) \\
&\quad + 2 \times 1.9 \times (-11.60096) \\
&\quad + 2 \times 3.61 \times 2.939026 \\
&\quad + 2 \times 6 \times 1.9 \times 0.300406 \\
&\quad + 2 \times 6 \times 3.61 \times (-0.085619) \\
&\quad + 2 \times 1.9 \times 3.61 \times (-3.288746) \\
&= 22.4208
\end{aligned}$$

TABLE 6.3 Covariance Matrix for Andy's Burger Barn Model

	b_1	b_2	b_3	b_4
b_1	46.227019	−6.426113	−11.600960	2.939026
b_2	−6.426113	1.093988	0.300406	−0.085619
b_3	−11.600960	0.300406	12.646302	−3.288746
b_4	2.939026	−0.085619	−3.288746	0.884774

The standard error of the forecast error is $\text{se}(f) = \sqrt{22.4208} = 4.7351$, and the relevant t-value is $t_{(0.975,\,71)} = 1.9939$, giving a 95% prediction interval of

$$\begin{aligned}
&(76.974 - 1.9939 \times 4.7351,\ \ 76.974 + 1.9939 \times 4.7351) \\
&\quad = (67.533,\ \ 86.415)
\end{aligned}$$

We predict, with 95% confidence, that Big Andy's settings for price and advertising expenditure will yield *SALES* between \$67,533 and \$86,415.

6.4.1 Predictive Model Selection Criteria

In this section we consider three model selection criteria: (i) R^2 and $\overline{R}^2$, (ii) AIC, and (iii) SC (BIC), and describe how a hold-out sample can be used to evaluate a model's predictive or forecast ability. Throughout the section you should keep in mind that we are not recommending blind application of any of these criteria. They should be treated as devices that provide additional information about the relative merits of alternative models, and they should be used in conjunction with the other considerations listed in Section 6.3.5 and in the introduction to Section 6.3.

Choice of a model based exclusively on $\overline{R}^2$, AIC, or SC involves choosing a model that minimizes the sum of squared errors with a penalty for adding extra variables. While these criteria can be used for both predictive and causal models, their goal of minimizing a function of the sum of squared errors rather than focusing on the coefficient, make them more suitable for predictive model selection. Another common feature of the criteria is that they are suitable only for comparing models with the same dependent variable, not for models with different dependent variables such as y and $\ln(y)$. More general versions of the AIC and SC, based on likelihood functions[9],

[9] An introduction to maximum likelihood estimation can be found in Appendix C.8.

are available for models with transformations of the dependent variable, but we do not consider them here.

R^2 and $\bar{R}^2$ In Chapters 4 and 5, we introduced the coefficient of determination $R^2 = 1 - SSE/SST$ as a measure of goodness of fit. It shows the proportion of variation in a dependent variable explained by variation in the explanatory variables. Since it is desirable to have a model that fits the data well, there can be a tendency to think that the best model is the one with the highest R^2. There are at least two problems with this line of thinking. First, if cross-sectional data are being used to estimate a causal effect, then low R^2's are typical and not necessarily a concern. What is more important is to avoid omitted variable bias and to have a sample size sufficiently large to get a reliable estimate of the coefficient of interest.

The second problem is one related to predictive models, namely, that comparing models on the basis of R^2 is only legitimate if the models have the same number of explanatory variables. Adding more variables always increases R^2 even if the variables added have no justification. As variables are added the sum of squared errors SSE goes down and thus R^2 goes up. If the model contains $N - 1$ variables, then $R^2 = 1$.

An alternative measure of goodness of fit called the **adjusted-R^2**, denoted as $\bar{R}^2$, has been suggested to overcome this problem. It is computed as

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

This measure does not always go up when a variable is added because of the degrees of freedom term $N - K$ in the numerator. As the number of variables K increases, SSE goes down, but so does $N - K$. The effect on $\bar{R}^2$ depends on the amount by which SSE falls. While solving one problem, this corrected measure of goodness of fit unfortunately introduces other problems. It loses its interpretation; $\bar{R}^2$ is no longer the proportion of explained variation. Also, it can be shown that if a variable is added to an equation, say with coefficient β_K, then $\bar{R}^2$ will increase if the t-value for testing the hypothesis $H_0 : \beta_K = 0$ is greater than one. Thus, using $\bar{R}^2$ as a device for selecting the appropriate set of explanatory variables is like using a hypothesis test for significance of a coefficient with a critical value of 1, a value much less than that typically used with 5% and 10% levels of significance. Because of these complications, we prefer to report the unadjusted R^2 as a goodness-of-fit measure, and caution is required if $\bar{R}^2$ is used for model selection. Nevertheless, you should be familiar with $\bar{R}^2$. You will see it in research reports and on the output of software packages.

Information Criteria Selecting variables to maximize $\bar{R}^2$ can be viewed as selecting variables to minimize SSE, subject to a penalty for introducing too many variables. Both the AIC and the SC work in a similar way but with different penalties for introducing too many variables. The **Akaike information criterion (AIC)** is given by

$$\text{AIC} = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N} \tag{6.43}$$

and the **Schwarz criterion (SC)**, also known as the **Bayesian information criterion** (**BIC**), is given by

$$\text{SC} = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N} \tag{6.44}$$

In each case the first term becomes smaller as extra variables are added, reflecting the decline in the SSE, but the second term becomes larger because K increases. Because $K\ln(N)/N > 2K/N$

for $N \geq 8$, in reasonable sample sizes the SC penalizes extra variables more heavily than does the AIC. Using these criteria, the model with the smallest AIC, or the smallest SC, is preferred.

To get values of the more general versions of these criteria based on maximized values of the likelihood function you need to add $[1 + \ln(2\pi)]$ to (6.43) and (6.44). It is good to be aware of this fact in case your computer software reports the more general versions. However, although it obviously changes the AIC and SC values, adding a constant does not change the choice of variables that minimize the criteria.

Using a Hold-Out Sample

When a model is designed for prediction or forecasting, we are naturally interested in its ability to forecast dependent variable values that have not yet been observed. To assess a model on this basis, we could make some forecasts and then compare these forecasts with the corresponding realizations after they occur. However, if we are in the model construction phase of an investigation, it is unlikely we would want to wait for extra observations. A way out of this dilemma is to hold back some of the observations from estimation and then evaluate the model on the basis of how well it can predict the omitted observations. Suppose we have a total of N observations of which N_1 are used for estimation and $N_2 = N - N_1$ are held back to evaluate a model's forecasting ability. Thus, we have estimates $(b_1, b_2, \ldots, b_K)$ from observations $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, N_1$ and we calculate the predictions

$$\hat{y}_i = b_1 + b_2 x_{i2} + \cdots + b_K x_{iK} \quad i = N_1 + 1, N_2 + 2, \ldots, N$$

A measure of the model's out-of-sample forecasting ability is the **root mean squared error** (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N_2} \sum_{i=N_1+1}^{N} (y_i - \hat{y}_i)^2}$$

We expect this quantity to be larger than its within-sample counterpart $\hat{\sigma} = \sqrt{\sum_{i=1}^{N_1} (y_i - \hat{y}_i)^2 / (N_1 - K)}$ because the least squares estimation procedure is such that $\sum_{i=1}^{N_1} (y_i - \hat{y}_i)^2$ is minimized. Models can be compared on the basis of their hold-out RMSEs.

EXAMPLE 6.16 | Predicting House Prices

Real estate agents and potential homebuyers are interested in valuing houses or predicting the price of a house with particular characteristics. There are many factors that have a bearing on the price of a house, but for our predictive model we will consider just two, the age of the house in years (*AGE*), and its size in hundreds of square feet (*SQFT*). The most general model we consider is

$$\ln(PRICE) = \beta_1 + \beta_2 AGE + \beta_3 SQFT + \beta_4 AGE^2 + \beta_5 SQFT^2 + \beta_6 AGE \times SQFT + e$$

where *PRICE* is the house price in thousands of dollars. Of interest is whether some or all of the quadratic terms AGE^2, $SQFT^2$, and $AGE \times SQFT$ improve the predictive ability of the model. For convenience, we evaluate predictive ability in terms of ln(*PRICE*) not *PRICE*. We use data on 900 houses sold in Baton Rouge, Louisiana in 2005, stored in the data file *br5*. For a comparison based on the RMSE of predictions (but not the other criteria) we randomly chose 800 observations for estimation and 100 observations for the hold-out sample. After this random selection, the observations were ordered so that the first 800 were used for estimation and the last 100 for predictive assessment.

Values of the criteria for the various models appear in Table 6.4. Looking for the model with the highest $\bar{R}^2$, and the models with the smallest values (or largest negative numbers) for the AIC and SC, we find that all three criteria prefer model 2 where AGE^2 is included, but $SQFT^2$ and $AGE \times SQFT$ are excluded. Using the out-of-sample RMSE criterion, model 6, with $AGE \times SQFT$ included in addition to AGE^2, is slightly favored over model 2.

TABLE 6.4 **Model Selection Criteria for House Price Example**

Model	Variables included in addition to ($SQFT, AGE$)	R^2	$\bar{R}^2$	AIC	SC	RMSE
1	None	0.6985	0.6978	–2.534	–2.518	0.2791
2	AGE^2	0.7207	0.7198*	–2.609*	–2.587*	0.2714
3	$SQFT^2$	0.6992	0.6982	–2.535	–2.513	0.2841
4	$AGE \times SQFT$	0.6996	0.6986	–2.536	–2.515	0.2790
5	AGE^2, $SQFT^2$	0.7208	0.7196	–2.607	–2.580	0.2754
6	AGE^2, $AGE \times SQFT$	0.7210	0.7197	–2.608	–2.581	0.2712*
7	$SQFT^2$, $AGE \times SQFT$	0.7006	0.6993	–2.537	–2.510	0.2840
8	$SQFT^2$, AGE^2, $AGE \times SQFT$	0.7212*	0.7197	–2.606	–2.574	0.2754

*Best model according to each of the criteria.

6.5 Poor Data, Collinearity, and Insignificance

Most economic data that are used for estimating economic relationships are nonexperimental. Indeed, in most cases they are simply "collected" for administrative or other purposes. They are not the result of a planned experiment in which an experimental design is specified for the explanatory variables. In controlled experiments the right-hand-side variables in the model can be assigned values in such a way that their individual effects can be identified and estimated with precision. When data are the result of an uncontrolled experiment, many of the economic variables may move together in systematic ways. Such variables are said to be **collinear**, and the problem is labeled **collinearity**. In this case there is neither a guarantee that the data will be "rich in information" nor that it will be possible to isolate the economic relationship or parameters of interest.

As an example, consider the problem faced by the marketing executives at Big Andy's Burger Barn when they try to estimate the increase in sales revenue attributable to advertising that appears in newspapers *and* the increase in sales revenue attributable to coupon advertising. Suppose that it has been common practice to coordinate these two advertising devices, so that at the same time that advertising appears in the newspapers there are flyers distributed containing coupons for price reductions on hamburgers. If variables measuring the expenditures on these two forms of advertising appear on the right-hand side of a sales revenue equation such as (5.2), then the data on these variables will show a systematic, positive relationship; intuitively, it will be difficult for such data to reveal the separate effects of the two types of ads. Although it is clear that total advertising expenditure increases sales revenue, because the two types of advertising expenditure move together, it may be difficult to sort out their separate effects on sales revenue.

As a second example, consider a production relationship explaining output over time as a function of the amounts of various quantities of inputs employed. There are certain factors of production (inputs), such as labor and capital, that are used in *relatively fixed proportions*. As production increases, the changing amounts of two or more such inputs reflect equiproportionate increases. Proportional relationships between variables are the very sort of systematic relationships that epitomize "collinearity." Any effort to measure the individual or separate effects (marginal products) of various mixes of inputs from such data will be difficult.

It is not just relationships between variables in a sample of data that make it difficult to isolate the separate effects of individual explanatory variables. If the values of an explanatory variable

do not vary or change much within a sample of data, then it is clearly difficult to use that data to estimate a coefficient that describes the effect of change in that variable. It is hard to estimate the effect of change if there has been no change.

6.5.1 The Consequences of Collinearity

The consequences of collinearity and/or lack of variation depend on whether we are examining an extreme case in which estimation breaks down or a bad, but not extreme, case in which estimation can still proceed but our estimates lack precision. In Section 5.3.1, we considered the model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

and wrote the variance of the least squares estimator for β_2 as

$$\text{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{(1 - r_{23}^2)\sum_{i=1}^{N}(x_{i2} - \bar{x}_2)^2} \tag{6.45}$$

where r_{23} is the correlation between x_2 and x_3. Exact or extreme collinearity exists when x_2 and x_3 are perfectly correlated, in which case $r_{23} = 1$ and $\text{var}(b_2|\mathbf{X})$ goes to infinity. Similarly, if x_2 exhibits no variation $\sum(x_{i2} - \bar{x}_2)^2$ equals zero and $\text{var}(b_2|\mathbf{X})$ again goes to infinity. In this case, x_2 is collinear with the constant term. In general, w*henever there are one or more* ***exact*** *linear relationships among the explanatory variables, then the condition of exact collinearity exists. In this case, the least squares estimator is not defined*. We *cannot* obtain estimates of β_k's using the least squares principle. One of our least squares assumptions MR5, which says that the values of x_{ik} are not exact linear functions of the other explanatory variables, is violated.

The more usual case is one in which correlations between explanatory variables might be high, but not exactly one; variation in explanatory variables may be low but not zero; or linear dependencies between more than two explanatory variables could be high but not exact. These circumstances do *not* constitute a violation of least squares assumptions. By the Gauss–Markov theorem, the least squares estimator is still the best linear unbiased estimator. We might still be unhappy, however, if the best we can do is constrained by the poor characteristics of our data. From (6.45), we can see that when r_{23} is close to one or $\sum(x_i - \bar{x}_2)^2$ is close to zero, the variance of b_2 will be large. A large variance means a large standard error, which means the estimate may not be significantly different from zero and an interval estimate will be wide. The sample data have provided relatively imprecise information about the unknown parameters.

Although (6.45) is only valid for a regression model with two explanatory variables, with a few simple changes we can generalize this equation to gain insights into collinearity in the more general multiple regression model with $K - 1$ explanatory variables. First, recall from Section 4.2.2 that a simple correlation between two variables is the same as the R^2 from the regression of one variable on another, so that $r_{23}^2 = R_{2.}^2$, where $R_{2.}^2$ is the R^2 from the so-called **auxiliary regression** $x_{i2} = \alpha_2 + \alpha_3 x_{i3} + v_i$. Then, another way to write (6.45) is

$$\text{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{\sum(x_{i2} - \bar{x}_2)^2 (1 - R_{2.}^2)} \tag{6.46}$$

The beauty of this equation is that it holds for the general model $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i$, where $R_{2.}^2$ is the R^2 from the auxiliary regression $x_{i2} = \alpha_2 + \alpha_3 x_{i3} + \cdots + \alpha_K x_{iK} + v_i$. The ratio

$$\text{VIF} = 1/(1 - R_{2.}^2)$$

is called the **variance inflation factor**. If $R_{2.}^2 = 0$, indicating no collinearity—no variation in x_2 can be explained by the other explanatory variables—then VIF = 1 and $\text{var}(b_2|\mathbf{X}) = \sigma^2/\sum(x_{i2} - \bar{x}_2)^2$. On the other hand, if $R_{2.}^2 = 0.90$, indicating that 90% of the variation in x_2 can be explained by the other regressors, then VIF = 10 and $\text{var}(b_2|\mathbf{X})$ is ten times

larger than the than it would be if there was no collinearity present. VIF is sometimes used to describe the severity of collinearity in a regression. Auxiliary regression R_k^2's and variance inflation factors can be found for every explanatory variable in a regression; equations analogous to (6.46) hold for each of the coefficient estimates.

By examining $R_{2.}^2$, we can obtain a very informative third representation. The R^2 from the regression $x_{i2} = \alpha_2 + \alpha_3 x_{i3} + \cdots + \alpha_K x_{iK} + v_i$ is the portion of the total variation in x_2 about its mean, $\sum(x_{i2} - \bar{x}_2)^2$, explained by the model. Let the fitted values from the auxiliary regression be $\hat{x}_{i2} = a_2 + a_3 x_{i3} + \cdots + a_K x_{iK}$, where $(a_2, a_3, \ldots, a_K)$ are the least squares estimates of $(\alpha_2, \alpha_3, \ldots, \alpha_K)$. A residual from the auxiliary regression is $x_{i2} - \hat{x}_{i2}$ and its R^2 can be written as

$$R_{2.}^2 = 1 - \frac{\sum(x_{i2} - \hat{x}_{i2})^2}{\sum(x_{i2} - \bar{x}_2)^2}$$

Substituting this into (6.46), we have

$$\text{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{\sum(x_{i2} - \hat{x}_{i2})^2} \tag{6.47}$$

The term $\sum(x_{i2} - \hat{x}_{i2})^2$ is the sum of squared least squares residuals from the auxiliary regression. When collinearity is stronger, with a larger amount of variation in x_2 explained by the other regressors, the smaller $\sum(x_{i2} - \hat{x}_{i2})^2$ becomes and the larger $\text{var}(b_2|\mathbf{X})$ becomes. It is the variation in x_2 that is *not* explained by the other regressors that increases the precision of least squares estimation.

The effects of imprecise estimation resulting from collinearity can be summarized as follows:

1. When estimator standard errors are large, it is likely that the usual t-tests will lead to the conclusion that parameter estimates are not significantly different from zero. This outcome occurs despite possibly high R^2- or F-values indicating significant explanatory power of the model as a whole. The problem is that collinear variables do not provide enough information to estimate their separate effects, even though theory may indicate their importance in the relationship.
2. Estimators may be very sensitive to the addition or deletion of a few observations, or to the deletion of an apparently insignificant variable.
3. Despite the difficulties in isolating the effects of individual variables from such a sample, accurate forecasts may still be possible if the nature of the collinear relationship remains the same within the out-of-sample observations. For example, in an aggregate production function where the inputs labor and capital are nearly collinear, accurate forecasts of output may be possible for a particular ratio of inputs but not for various mixes of inputs.

6.5.2 Identifying and Mitigating Collinearity

Because nonexact collinearity is not a violation of least squares assumptions, it does not make sense to go looking for a problem if there is no evidence that one exists. If you have estimated an equation where the coefficients are precisely estimated and significant, they have the expected signs and magnitudes, and they are not sensitive to adding or deleting a few observations, or an insignificant variable, then there is no reason to try and identify or mitigate collinearity. If there are highly correlated variables, they are not causing you a problem. However, if you have a poorly estimated equation that does not live up to expectations, it is useful to establish why the estimates are poor.

One simple way to detect collinear relationships is to use sample correlation coefficients between pairs of explanatory variables. These sample correlations are descriptive measures of linear association. However, collinear relationships that involve more than two explanatory

variables are better detected using **auxiliary regressions**. If an R_k^2 is high, say greater than 0.8, then a large portion of the variation in x_k is explained by the other regressors, and that may have a detrimental effect on the precision of estimation of β_k. If an auxiliary regression R_k^2 is not high, then the precision of an estimator b_k is not unduly affected by collinearity, although it may still suffer if the variation in x_k is inadequate.

The collinearity problem is that the data do not contain enough "information" about the individual effects of explanatory variables to permit us to estimate all the parameters of the statistical model precisely. Consequently, one solution is to obtain more information and include it in the analysis. One form the new information can take is more, and better, sample data. Unfortunately, in economics, this is not always possible. Cross-sectional data are expensive to obtain, and, with time-series data, one must wait for the data to appear. Alternatively, if new data are obtained via the same nonexperimental process as the original sample of data, then the new observations may suffer the same collinear relationships and provide little in the way of new, independent information. Under these circumstances the new data will help little to improve the precision of the least squares estimates.

A second way of adding new information is to introduce, as we did in Section 6.2, *nonsample* information in the form of restrictions on the parameters. This nonsample information may then be combined with the sample information to provide restricted least squares estimates. The good news is that using nonsample information in the form of linear constraints on the parameter values reduces estimator sampling variability. The bad news is that the resulting restricted estimator is *biased* unless the restrictions are *exactly* true. Thus it is important to use good nonsample information, so that the reduced sampling variability is not bought at a price of large estimator biases.

EXAMPLE 6.17 | Collinearity in a Rice Production Function

To illustrate collinearity we use data on rice production from a cross section of Philippine rice farmers to estimate the production function

$$\ln(PROD) = \beta_1 + \beta_2 \ln(AREA) + \beta_3 \ln(LABOR) + \beta_4 \ln(FERT) + e \tag{6.48}$$

where *PROD* denotes tonnes of freshly threshed rice, *AREA* denotes hectares planted, *LABOR* denotes person-days of hired and family labor and *FERT* denotes kilograms of fertilizer. Data for the years 1993 and 1994 can be found in the file *rice5*. One would expect collinearity may be an issue. Larger farms with more area are likely to use more labor and more fertilizer than smaller farms. The likelihood of a collinearity problem is confirmed by examining the results in Table 6.5, where we have estimated the function using data from 1994 only. These results convey very little information. The 95% interval estimates are very wide, and, because the coefficients of ln(*AREA*) and ln(*LABOR*) are not significantly different from zero, their interval estimates include a negative range. The high auxiliary R^2's and correspondingly high variance inflation factors point to collinearity as the culprit for the imprecise results. Further evidence is a relatively high $R^2 = 0.875$ from estimating (6.48), and a p-value of 0.0021 for the joint test of the two insignificant coefficients, $H_0: \beta_2 = \beta_3 = 0$.

TABLE 6.5 **Rice Production Function Results from 1994 Data**

Variable	Coefficient b_k	se(b_k)	95% Interval Estimate	p-Value*	Auxiliary Regression R^2	VIF
C	−1.9473	0.7385		0.0119		
ln(*AREA*)	0.2106	0.1821	[−0.1573, 0.5786]	0.2543	0.891	9.2
ln(*LABOR*)	0.3776	0.2551	[−0.1379, 0.8931]	0.1466	0.944	17.9
ln(*FERT*)	0.3433	0.1280	[0.0846, 0.6020]	0.0106	0.870	7.7

*p-value for $H_0: \beta_k = 0$ versus $H_1: \beta_k \neq 0$

We consider two ways of improving the precision of our estimates: (1) including non-sample information, and (2) using more observations. For non-sample information, suppose that we are willing to accept the notion of constant returns to scale. That is, increasing all inputs by the same proportion will lead to an increase in production of the same proportion. If this constraint holds, then $\beta_2 + \beta_3 + \beta_4 = 1$. Testing this constraint as a null hypothesis yields a p-value of 0.313; so it is not a constraint that is incompatible with the 1994 data. Substituting $\beta_2 + \beta_3 + \beta_4 = 1$ into (6.48) and rearranging the equation gives

$$\ln\left(\frac{PROD}{AREA}\right) = \beta_1 + \beta_3 \ln\left(\frac{LABOR}{AREA}\right) + \beta_4 \ln\left(\frac{FERT}{AREA}\right) + e \tag{6.49}$$

This equation can be viewed as a "yield" equation. Rice yield per hectare is a function of labor per hectare and fertilizer per hectare. Results from estimating it appear in Table 6.6. Has there been any improvement? The answer is not much! The estimate for β_3 is no longer "insignificant," but that is more attributable to an increase in the magnitude of b_3 than to a reduction in its standard error. The reduction in standard errors is only marginal, and the interval estimates are still wide, conveying little information. The squared correlation between $\ln(LABOR/AREA)$ and $\ln(FERT/AREA)$ is 0.414 which is much less than the earlier auxiliary R^2's, but, nevertheless, the new estimates are relatively imprecise.

As an alternative to injecting non-sample information into the estimation procedure, we examine the effect of including more observations by combining the 1994 data with observations from 1993. The results are given in Table 6.7. Here there has been a substantial reduction in the standard errors, with considerable improvement in the precision of the estimates, despite the fact that the variance inflation factors still remain relatively large. The greatest improvement has been for the coefficient of $\ln(FERT)$, which has the lowest variance inflation factor. The interval estimates for the other two coefficients are still likely to be wider than a researcher would desire, but at least there has been some improvement.

TABLE 6.6 **Rice Production Function Results from 1994 Data with Constant Returns to Scale**

Variable	Coefficient b_k	se(b_k)	95% Interval Estimate	p-Value*
C	−2.1683	0.7065		0.0038
$\ln(AREA)$	0.2262	0.1815	[−0.1474, 0.5928]	0.2197
$\ln(LABOR)$	0.4834	0.2332	[0.0125, 0.9544]	0.0445
$\ln(FERT)$	0.2904	0.1171	[0.0539, 0.5268]	0.0173

*p-value for $H_0: \beta_k = 0$ versus $H_1: \beta_k \neq 0$

TABLE 6.7 **Rice Production Function Results from Data for 1993 and 1994**

Variable	Coefficient	se(b_k)	95% Interval Estimate	p-Value*	Auxiliary Regression R^2	VIF
C	−1.8694	0.4565		0.0001		
$\ln(AREA)$	0.2108	0.1083	[−0.0045, 0.4261]	0.0549	0.870	7.7
$\ln(LABOR)$	0.3997	0.1307	[0.1399, 0.6595]	0.0030	0.901	10.1
$\ln(FERT)$	0.3195	0.0635	[0.1932, 0.4457]	0.0000	0.776	4.5

*p-value for $H_0: \beta_k = 0$ versus $H_1: \beta_k \neq 0$

TABLE 6.8 Statistics for Identifying Influential Observations

Influence Statistic	Formula	Investigative Threshold
Leverage	$h_i = \dfrac{\widehat{\text{var}}(\hat{e}_i) - \hat{\sigma}^2}{\hat{\sigma}^2}$	$h_i > \dfrac{2K}{N}$ or $\dfrac{3K}{N}$
Studentized residual	$\hat{e}_i^{stu} = \dfrac{\hat{e}_i}{\hat{\sigma}(i)(1-h_i)^{1/2}}$	$\left\|\hat{e}_i^{stu}\right\| > 2$
DFBETAS	$\text{DFBETAS}_{ki} = \dfrac{b_k - b_k(i)}{(\hat{\sigma}(i)/\hat{\sigma}) \times \text{se}(b_k)}$	$\left\|\text{DFBETAS}_{ki}\right\| > \dfrac{2}{\sqrt{N}}$
DFFITS	$\text{DFFITS}_i = \left(\dfrac{h_i}{1-h_i}\right)^{1/2} \hat{e}_i^{stu}$	$\left\|\text{DFFITS}_i\right\| > 2\left(\dfrac{K}{N}\right)^{1/2}$

6.5.3 Investigating Influential Observations

In Section 4.3.6, we introduced a number of measures for detecting influential observations. The purpose of having such measures is first to detect whether there may have been a data error, and second, if the accuracy of the data is confirmed, to identify unusual observations that may be worthy of further investigation. Are there observations that can be explained within the context of the proposed model? Are there other factors at work that could have led to the unusual observations?

In Section 4.3.6, the measures were introduced within the context of the simple regression model with one explanatory variable. The same measures are relevant for the multiple regression model, but some of the formulas change slightly to accommodate the extra regressors. Now would be a good time to go back and reread Section 4.3.6. Are you back? Now that you understand the concepts, we can proceed. The important concepts introduced in that section were the leverage of the ith observation, h_i, the studentized residual, $\hat{e}_i^{stu}$, the sensitivity of a coefficient estimate to omission of the ith observation, DFBETAS_{ki}, and the sensitivity of a prediction to omission of the ith observation DFFITS_i. Multiple regression versions of these measures are summarized in Table 6.8 along with conventional thresholds above which further scrutiny of an observation may be warranted. Remember, the purpose is not to throw out unusual observations but to learn from them. They may reveal some important characteristics of the data.

EXAMPLE 6.18 | Influential Observations in the House Price Equation

To illustrate the identification of potentially influential observations, we return to Example 6.16 where, using predictive model selection criteria, the preferred equation for predicting house prices was

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \beta_4 AGE^2 + e$$

In a sample of 900 observations it is not surprising to find a relatively large number of data points where the various influence measures exceed the recommended thresholds. As examples, in Table 6.9 we report the values of the measures for those observations with the three largest DFFITS. It turns out that the other influence measures for these three observations also have large values. In parentheses next to each of the values is the rank of its absolute value. When we check the characteristics of the three unusual observations, we find observation 540 is the newest house in the sample and observation 150 is the oldest house. Observation 411 is both old and large; it is the 10th largest (99th percentile) and the sixth oldest (percentile 99.4) house in the sample. In Exercise 6.20, you are invited to explore further the effect of these observations.

TABLE 6.9 Influence Measures for House Price Equation

Observation	h_i (rank)	$\hat{e}_i^{stu}$ (rank)	DFFITS$_i$ (rank)	DFBETAS$_{ki}$ (rank)		
Threshold	$\frac{2.5K}{N} = 0.011$	2	$2\left(\frac{K}{N}\right)^{1/2} = 0.133$	$\frac{2}{\sqrt{N}} = 0.067$		
				SQFT	*AGE*	*AGE*^{2}
411	0.0319 (10)	−4.98 (1)	0.904 (1)	−0.658 (1)	0.106 (17)	−0.327 (3)
524	0.0166 (22)	−4.31 (3)	0.560 (2)	0.174 (9)	0.230 (2)	−0.381 (2)
150	0.0637 (2)	1.96 (48)	−0.511 (3)	−0.085 (29)	−0.332 (1)	0.448 (1)

6.6 Nonlinear Least Squares

We have discovered how the least squares estimation technique can be used to estimate a variety of nonlinear functions. They include log-log models, log-linear models, and models with quadratic and interaction terms. However, the models we have encountered so far have all been linear in the parameters $\beta_1, \beta_2, \ldots, \beta_K$.[10] In this section we discuss estimation of models that are nonlinear in the parameters. To give an appreciation of what is meant by such a model, it is convenient to begin with the following simple artificial example,

$$y_i = \beta x_{i1} + \beta^2 x_{i2} + e_i \tag{6.50}$$

where y_i is a dependent variable, x_{i1} and x_{i2} are explanatory variables, β is an unknown parameter that we wish to estimate, and the e_i satisfy the multiple regression assumptions MR1–MR5. This example differs from the conventional linear model because the coefficient of x_{i2} is equal to the square of the coefficient of x_{i1}, and the number of parameters is not equal to the number of variables.

How can β be estimated? Think back to Chapter 2. What did we do when we had a simple linear regression equation with two unknown parameters β_1 and β_2? We set up a sum of squared errors function that, in the context of (6.50), is

$$S(\beta) = \sum_{i=1}^{N} \left(y_i - \beta x_{i1} - \beta^2 x_{i2}\right)^2 \tag{6.51}$$

Then we asked what values of the unknown parameters make $S(\beta)$ a minimum. We searched for the bottom of the bowl in Figure 2A.1. We found that we could derive formulas for the minimizing values b_1 and b_2. We called these formulas the least squares estimators.

When we have models that are nonlinear in the parameters, we cannot in general derive formulas for the parameter values that minimize the sum of squared errors function. However, for a given set of data, we can ask the computer to search for the parameter values that take us to the bottom of the bowl. There are many numerical software algorithms that can be used to find minimizing values for functions such as $S(\beta)$. Those minimizing values are known as the **nonlinear least squares estimates**. It is also possible to obtain numerical standard errors that assess the reliability of the nonlinear least squares estimates. Finite sample properties

[10] There have been a few exceptions where we have used notation other than $\beta_1, \beta_2, \ldots, \beta_K$ to denote the parameters.

and distributions of nonlinear least squares estimators are not available, but their large sample asymptotic properties are well established.[11]

EXAMPLE 6.19 | Nonlinear Least Squares Estimates for Simple Model

To illustrate estimation of (6.50), we use data stored in the file *nlls*. The sum of squared error function is graphed in Figure 6.1. Because we only have one parameter, we have a two-dimensional curve, not a "bowl." It is clear from the curve that the minimizing value for β lies between 1.0 and 1.5. From your favorite software, the nonlinear least squares estimate turns out to be $b = 1.1612$. The standard error depends on the degree of curvature of the sum of squares function at its minimum. A sharp minimum with a high degree of curvature leads to a relatively small standard error, while a flat minimum with a low degree of curvature leads to a relatively high standard error. There are different ways of measuring the curvature that can lead to different standard errors. In this example, the "outer-product of gradient" method yields a standard error of $\text{se}(b) = 0.1307$, while the standard error from the "observed-Hessian" method is $\text{se}(b) = 0.1324$.[12] Differences such as this one disappear as the sample size gets larger.

Two words of warning must be considered when estimating a nonlinear-in-the-parameters model. The first is to check that the estimation process has converged to a global minimum. The estimation process is an iterative one where a series of different parameter values are checked until the process converges at the minimum. If your software tells you the process has failed to converge, the output provided, if any, **does not** provide the nonlinear least squares estimates. This might happen if a maximum number of iterations has been reached or there has been a numerical problem that has caused the iterations to stop. A second problem that can occur is that the iterative process may stop at a "local" minimum rather than the "global" minimum. In the example in Figure 6.1, there is a local minimum at $\beta = -2.0295$. Your software will have an option of giving **starting values** to the iterative process. If you give it a starting value of -2, it is highly likely you will end up with the estimate $b = -2.0295$. This value is not the nonlinear least squares estimate, however. The nonlinear least squares estimate is at the global minimum which is the smallest of the minima if more than one exists. How do you guard against ending up at a local minimum? It is wise to try different starting values to ensure you end up at the same place each time. Notice that the curvature at the local minimum in Figure 6.1 is much less than at the global minimum. This should be reflected in a larger "standard error" at the local minimum. Such is indeed the case. We find the outer-product-gradient method yields $\text{se}(b) = 0.3024$, and from the observed-Hessian method we obtain $\text{se}(b) = 0.3577$.

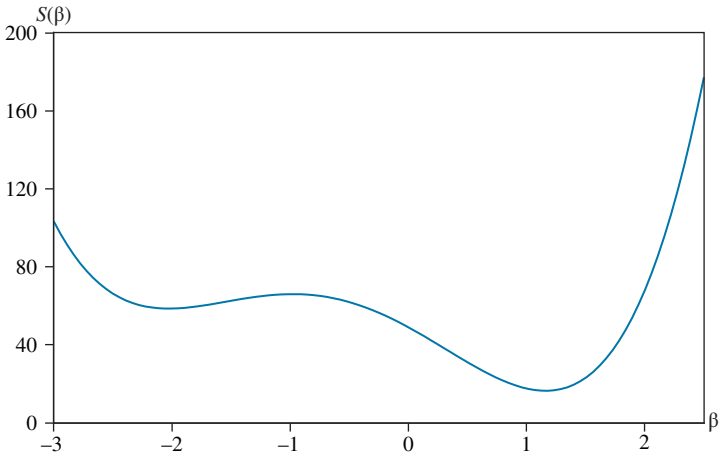

FIGURE 6.1 **Sum of squared errors function for single-parameter example.**

[11] Details of how the numerical algorithms work, how standard errors are obtained, the asymptotic properties of the estimators, and the assumptions necessary for the asymptotic properties to hold, can be found in William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Chapter 7.

[12] These methods require advanced material. See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Section 14.4.6.

EXAMPLE 6.20 | A Logistic Growth Curve

A model that is popular for modeling the diffusion of technological change is the logistic growth curve[13]

$$y_t = \frac{\alpha}{1 + \exp(-\beta - \delta t)} + e_t \tag{6.52}$$

where y_t is the adoption proportion of a new technology. For example, y_t might be the proportion of households who own a computer, or the proportion of computer-owning households who have the latest computer, or the proportion of musical recordings sold as compact disks. In the example that follows, y_t is the share of total U.S. crude steel production that is produced by electric arc furnace technology.

Before considering this example, we note some details about the relationship in equation (6.52). There is only one explanatory variable on the right hand side, namely, time, $t = 1, 2, \ldots, T$. Thus, the logistic growth model is designed to capture the rate of adoption of technological change, or, in some examples, the rate of growth of market share. An example of a logistic curve is depicted in Figure 6.2. In this example, the rate of growth increases at first, to a point of inflection that occurs at $t = -\beta/\delta = 20$. Then the rate of growth declines, leveling off to a saturation proportion given by $\alpha = 0.8$. Since $y_0 = \alpha/(1 + \exp(-\beta))$, the parameter β determines how far the share is below saturation level at time zero. The parameter δ controls the speed at which the point of inflection, and the saturation level, are reached. The curve is such that the share at the point of inflection is $\alpha/2 = 0.4$, half the saturation level.

Traditional technology for steel making, involving blast and oxygen furnaces and the use of iron ore, is being displaced by newer electric arc furnace technology that utilizes scrap steel. This displacement has implications for the suppliers of raw materials such as iron ore. Thus, prediction of the future electric arc furnace share of steel production is of vital importance to mining companies. The file *steel* contains annual data on the electric arc furnace share of U.S. steel production from 1970 to 2015. Using this data to find nonlinear least squares estimates of a logistic growth curve yields the following estimates (standard errors):

$$\hat{\alpha} = 0.8144 \ (0.0511) \quad \hat{\beta} = -1.3777 \ (0.0564)$$
$$\hat{\delta} = 0.0572 \ (0.0043)$$

Quantities of interest are the inflection point at which the rate of growth of the share starts to decline, $-\beta/\delta$; the saturation proportion α; the share at time zero, $y_0 = \alpha/(1 + \exp(-\beta))$; and prediction of the share for various years in the future. In Exercise 6.21, you are invited to find interval estimates for these quantities.

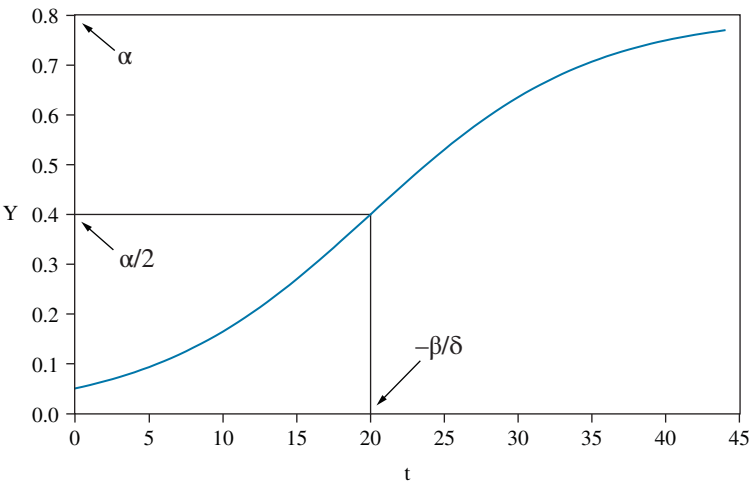

FIGURE 6.2 **A Logistic Growth Curve.**

[13] For other possible models, see Exercises 4.15 and 4.17.

6.7 Exercises

6.7.1 Problems

6.1 When using $N = 50$ observations to estimate the model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$, you obtain $SSE = 2132.65$ and $s_y = 9.8355$.

a. Find R^2.

b. Find the value of the F-statistic for testing $H_0 : \beta_2 = 0, \beta_3 = 0$. Do you reject or fail to reject H_0 at a 1% level of significance?

c. After augmenting this model with the squares and cubes of predictions $\hat{y}_i^2$ and $\hat{y}_i^3$, you obtain $SSE = 1072.88$. Use RESET to test for misspecification at a 1% level of significance.

d. After estimating the model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 z_i^2 + e_i$, you obtain $SSE = 401.179$. What is the R^2 from estimating this model?

e. After augmenting the model in (d) with the squares and cubes of predictions $\hat{y}_i^2$ and $\hat{y}_i^3$, you obtain $SSE = 388.684$. Use RESET to test for misspecification at a 5% level of significance.

6.2 Consider the following model that relates the percentage of a household's budget spent on alcohol *WALC* to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + \beta_5 AGE^2 + e$$

Using 1200 observations from a London survey, this equation was estimated with and without the *AGE* variables included, giving the following results:

$$\underset{\text{(se)}}{\widehat{WALC}} = \underset{}{8.149} + \underset{(0.486)}{2.884 \ln(TOTEXP)} - \underset{(0.382)}{1.217 NK} - \underset{(0.1790)}{0.5699 AGE} + \underset{(0.002332)}{0.005515 AGE^2} \quad \hat{\sigma} = 6.2048$$

$$\underset{\text{(se)}}{\widehat{WALC}} = \underset{}{-1.206} + \underset{(0.482)}{2.152 \ln(TOTEXP)} - \underset{(0.376)}{1.421 NK} \quad \hat{\sigma} = 6.3196$$

a. Use an F-test and a 5% significance level to test whether *AGE* and AGE^2 should be included in the equation.

b. Use an F-test and a 5% significance level to test whether *NK* should be included in the first equation. [*Hint*: $F = t^2$]

c. Use an F-test, a 5% significance level and the first equation to test $H_0 : \beta_2 = 3.5$ against the alternative $H_1 : \beta_2 \neq 3.5$.

d. After estimating the following equation, we find $SSE = 46086$.

$$WALC - 3.5 \ln(TOTEXP) + NK = \beta_1 - (2\beta_5 \times 50) AGE + \beta_5 AGE^2 + e$$

Relative to the original equation with all variables included, for what null hypothesis is this equation the restricted model? Test this null hypothesis at a 5% significance level.

e. What is the χ^2-value for the test in part (d)? In this case, is there a reason why a χ^2-test might be preferred to an F-test?

6.3 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + e_i$, where $E(e_i|\mathbf{X}) = 0$, $\text{var}(e_i|\mathbf{X}) = \sigma^2$, and $E(e_i e_j|\mathbf{X}) = 0$ for $i \neq j$, with $\mathbf{X}$ representing all observations on x and z. Suppose z_i is omitted from the equation, so that we have the least squares estimator for β_2 as

$$b_2^* = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Prove that

a. $b_2^* = \beta_2 + \beta_3 \sum w_i z_i + \sum w_i e_i$, where $w_i = (x_i - \bar{x}) / \sum (x_i - \bar{x})^2$.

b. $E(b_2^* | \mathbf{X}) = \beta_2 + \beta_3 \widehat{\text{cov}}(x, z) / \widehat{\text{var}}(x)$

c. $\text{var}(b_2^* | \mathbf{X}) = \sigma^2 \Big/ \left(N \widehat{\text{var}}(x)\right)$

d. $\text{var}(b_2^* | \mathbf{X}) \leq \text{var}(b_2 | \mathbf{X})$, where b_2 is the least squares estimator with both x and z included. [*Hint*: check out equation (5.13).]

6.4 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 q_i + e_i$, where $E(e_i | \mathbf{X}) = 0$, with $\mathbf{X}$ representing all observations on x, z, and q. Suppose z_i is unobservable and omitted from the equation, but conditional mean independence $E(z_i | x_i, q_i) = E(z_i | q_i)$ holds, with $E(z_i | q_i) = \delta_1 + \delta_2 q_i$.

a. Show that $E(y_i | x_i, q_i) = (\beta_1 + \beta_3 \delta_1) + \beta_2 x_i + (\beta_3 \delta_2 + \beta_4) q_i$.

b. **i.** Is it possible to get a consistent estimate of the causal effect of x_i on y_i?

ii. Is it possible to get a consistent estimate of the causal effect of z_i on y_i?

iii. Is it possible to get a consistent estimate of the causal effect of q_i on y_i?

6.5 Consider the following wage equation where *EDUC* = years of education and *EXPER* = years of experience:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

Suppose that observations on *EXPER* are not available, and so you decide to use the variables *AGE* and AGE^2 instead. What assumptions are sufficient for the least squares estimate for β_2 to be given a causal interpretation?

6.6 Use an F-test to jointly test the relevance of the two variables *XTRA_X5* and *XTRA_X6* for the family income equation in Example 6.12 and Table 6.1.

6.7 In Example 6.15 a prediction interval for *SALES* from Big Andy's Burger Barn was computed for the settings $PRICE_0 = 6$, $ADVERT_0 = 1.9$. Find point and 95% interval estimates for

$$E(SALES | PRICE = 6, ADVERT = 1.9)$$

Contrast your answers with the point and interval predictions that were obtained in Example 6.15. [*Hint*: The easiest way to calculate the standard error for your point estimate is to utilize some of the calculations given in Example 6.15.]

6.8 Consider the wage equation

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EDUC^2 + \beta_4 EXPER + \beta_5 EXPER^2 + \beta_6 (EDUC \times EXPER) + e$$

where the explanatory variables are years of education (*EDUC*) and years of experience (*EXPER*). Estimation results for this equation, and for modified versions of it obtained by dropping some of the variables, are displayed in Table 6.10. These results are from 200 observations in the file *cps5_small*.

a. What restriction on the coefficients of Eqn (A) gives Eqn (B)? Use an F-test to test this restriction. Show how the same result can be obtained using a t-test.

b. What restrictions on the coefficients of Eqn (A) give Eqn (C)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

c. What restrictions on the coefficients of Eqn (B) give Eqn (D)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

d. What restrictions on the coefficients of Eqn (A) give Eqn (E)? Use an F-test to test these restrictions. What question would you be trying to answer by performing this test?

e. Based on your answers to parts (a)–(d), which model would you prefer? Why?

f. Compute the missing AIC value for Eqn (D) and the missing SC value for Eqn (A). Which model is favored by the AIC? Which model is favored by the SC?

TABLE 6.10 Wage Equation Estimates for Exercise 6.8

Variable	Coefficient Estimates and (Standard Errors)				
	Eqn (A)	Eqn (B)	Eqn (C)	Eqn (D)	Eqn (E)
C	0.403	1.483	1.812	2.674	1.256
	(0.771)	(0.495)	(0.494)	(0.109)	(0.191)
$EDUC$	0.175	0.0657	0.0669		0.0997
	(0.091)	(0.0692)	(0.0696)		(0.0117)
$EDUC^2$	−0.0012	0.0012	0.0010		
	(0.0027)	(0.0024)	(0.0024)		
$EXPER$	0.0496	0.0228		0.0314	0.0222
	(0.0172)	(0.0091)		(0.0104)	(0.0090)
$EXPER^2$	−0.00038	−0.00032		−0.00060	−0.00031
	(0.00019)	(0.00019)		(0.00022)	(0.00019)
$EXPER \times EDUC$	−0.001703				
	(0.000935)				
SSE	37.326	37.964	40.700	52.171	38.012
AIC	−1.619	−1.612	−1.562		−1.620
SC		−1.529	−1.513	−1.264	−1.554

6.9 RESET suggests augmenting an existing model with the squares or the squares and higher powers of the predictions $\hat{y}_i$. For example, $\left(\hat{y}_i^2\right)$ or $\left(\hat{y}_i^2, \hat{y}_i^3\right)$ or $\left(\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4\right)$. What would happen if you augmented the model with the predictions $\hat{y}_i$?

6.10 Reconsider Example 6.19 where we used nonlinear least squares to estimate the model $y_i = \beta x_{i1} + \beta^2 x_{i2} + e_i$ by minimizing the sum of squares function $S(\beta) = \sum_{i=1}^{N}\left(y_i - \beta x_{i1} - \beta^2 x_{i2}\right)^2$.

a. Show that $\dfrac{dS}{d\beta} = -2\sum_{i=1}^{N} x_{i1}y_i + 2\beta\left(\sum_{i=1}^{N} x_{i1}^2 - 2\sum_{i=1}^{N} x_{i2}y_i\right) + 6\beta^2\sum_{i=1}^{N} x_{i1}x_{i2} + 4\beta^3\sum_{i=1}^{N} x_{i2}^2$

b. Show that $\dfrac{d^2S}{d\beta^2} = 2\left(\sum_{i=1}^{N} x_{i1}^2 - 2\sum_{i=1}^{N} x_{i2}y_i\right) + 12\beta\sum_{i=1}^{N} x_{i1}x_{i2} + 12\beta^2\sum_{i=1}^{N} x_{i2}^2$

c. Given that $\sum_{i=1}^{N} x_{i1}^2 = 10.422155$, $\sum_{i=1}^{N} x_{i2}^2 = 3.586929$, $\sum_{i=1}^{N} x_{i1}x_{i2} = 4.414097$, $\sum_{i=1}^{N} x_{i1}y_i = 16.528022$, and $\sum_{i=1}^{N} x_{i2}y_i = 10.619469$, evaluate $dS/d\beta$ at both the global minimum $\beta = 1.161207$ and at the local minimum $\beta = -2.029494$. What have you discovered?

d. Evaluate $d^2S/d\beta^2$ at both $\beta = 1.161207$ and $\beta = -2.029494$.

e. At the global minimum, we find $\hat{\sigma}_G = 0.926452$ whereas, if we incorrectly use the local minimum, we find $\hat{\sigma}_L = 1.755044$. Evaluate

$$q = \hat{\sigma}\sqrt{\frac{2}{d^2S/d\beta^2}}$$

at both the global and local minimizing values for β and $\hat{\sigma}$. What is the relevance of these values of q? Go back and check Example 6.19 to see what you have discovered.

6.11 In Example 6.7 we tested the joint null hypothesis

$$H_0: \beta_3 + 3.8\beta_4 = 1,\ \beta_1 + 6\beta_2 + 1.9\beta_3 + 3.61\beta_4 = 80$$

in the model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e_i$$

By substituting the restrictions into the model and rearranging variables, show how the model can be written in a form where least squares estimation will yield restricted least squares estimates.

6.12 This exercise uses data on 850 houses sold in Baton Rouge, Louisiana during mid-2005. We will be concerned with the selling price in thousands of dollars (*PRICE*), the size of the house in hundreds of square feet (*SQFT*), the number of bathrooms (*BATHS*), and the number of bedrooms (*BEDS*). Consider the following conditional expectations

$$E(PRICE|BEDS) = \alpha_1 + \alpha_2 BEDS \quad \text{(XR6.12.1)}$$

$$E(PRICE|BEDS, SQFT) = \beta_1 + \beta_2 BEDS + \beta_3 SQFT \quad \text{(XR6.12.2)}$$

$$E(SQFT|BEDS) = \gamma_1 + \gamma_2 BEDS \quad \text{(XR6.12.3)}$$

$$E(PRICE|BEDS, SQFT, BATHS) = \delta_1 + \delta_2 BEDS + \delta_3 SQFT + \delta_4 BATHS \quad \text{(XR6.12.4)}$$

$$E(BATHS|BEDS, SQFT) = \theta_1 + \theta_2 BEDS + \theta_3 SQFT \quad \text{(XR6.12.5)}$$

a. Express α_1 and α_2 in terms of the parameters $(\beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2)$.
b. Express β_1, β_2, and β_3 in terms of the parameters $(\delta_1, \delta_2, \delta_3, \delta_4, \theta_1, \theta_2, \theta_3)$.
c. Use the information in Table 6.11 and a 1% significance level to test whether

$$E(PRICE|BEDS, SQFT, BATHS) = E(PRICE|BEDS)$$

d. Show that the estimates in Table 6.11 satisfy the expressions you derived in parts (a) and (b).
e. Can you explain why the coefficient of *BEDS* changed sign when *SQFT* was added to equation (XR6.12.1).
f. Suppose that $E(BATHS|BEDS) = \lambda_1 + \lambda_2 BEDS$. Use the results in Table 6.11 to find estimates for λ_1 and λ_2.
g. Use the estimates from part (f) and the estimates for equations (XR6.12.3) and (XR6.12.4) to find estimates of α_1 and α_2. Do they agree with the estimates in Table 6.11?
h. Would you view any of the parameter estimates as causal?

TABLE 6.11 **Estimates for House Price Equations for Exercise 6.12**

	Coefficient Estimates and (Standard Errors)				
	(XR6.12.1) *PRICE*	(XR6.12.2) *PRICE*	(XR6.12.3) *SQFT*	(XR6.12.4) *PRICE*	(XR6.12.5) *BATHS*
C	−71.873	−0.1137	−6.7000	−24.0509	0.67186
	(16.502)	(11.4275)	(1.1323)	(11.7975)	(0.06812)
BEDS	70.788	−28.5655	9.2764	−32.649	0.1146
	(5.041)	(4.6504)	(0.3458)	(4.593)	(0.0277)
SQFT		10.7104		9.2648	0.04057
		(0.3396)		(0.4032)	(0.00202)
BATHS				35.628	
				(5.636)	
SSE	8906627	4096699	41930.6	3911896	145.588

6.13 Do gun buybacks save lives? Following the "Port Arthur massacre" in 1996, the Australian government introduced a gun buyback scheme in 1997. The success of that scheme has been investigated by Leigh and Neill.[14] Using a subset of their data on the eight Australian states and territories for the years

[14]Leigh, A. and C. Neill (2010), "Do Gun Buybacks Save Lives? Evidence from Panel Data?", *American Law and Economics Review*, 12(2), p. 509–557.

1980–2006, with 1996 and 1997 omitted, making a total of $N = 200$ observations, we estimate the following model

$$SUIC\_RATE = \beta_1 + \beta_2 GUNRATE + \beta_3 URATE + \beta_4 CITY + \beta_5 YEAR + e$$

Three equations are considered, one where *SUIC_RATE* denotes the firearm suicide rate, one where it represents the non-firearm suicide rate and one for the overall suicide rate, all measured in terms of deaths per million population. For the years after 1997, the variable *GUNRATE* is equal to the number of guns bought back during 1997, per thousand population; it is zero for the earlier years; *URATE* is the unemployment rate, *CITY* is the proportion of the population living in an urban area and *YEAR* is included to capture a possible trend. The estimated equations are given in Table 6.12.

TABLE 6.12 **Estimates for Gun Buyback Equations for Exercise 6.13**

	Coefficient Estimates and (Standard Errors)		
	Firearm Suicide Rate	**Non-firearm Suicide Rate**	**Overall Suicide Rate**
C	1909	–1871	38.37
	(345)	(719)	(779.76)
GUNRATE	–0.223	0.553	0.330
	(0.069)	(0.144)	(0.156)
URATE	–0.485	1.902	1.147
	(0.534)	(1.112)	(1.206)
CITY	–0.628	0.053	–0.576
	(0.057)	(0.118)	(0.128)
YEAR	–0.920	0.976	0.056
	(0.174)	(0.362)	(0.393)
SSE	29745	129122	151890
SSE_R	50641	131097	175562

a. Is there evidence that the gun buyback has reduced firearm suicides? Has there been substitution away from firearms to other means of suicide? Is there a trend in the suicide rate?

b. Is there evidence that greater unemployment increases the suicide rate?

c. Test jointly whether *URATE* and *CITY* contribute to the each of the equations. The sums of squared errors for the equations without these variables are given in the row SSE_R.

6.14 Do gun buybacks save lives? Following the "Port Arthur massacre" in 1996, the Australian government introduced a gun buyback scheme in 1997. As mentioned in Exercise 6.13, the success of that scheme has been investigated by Leigh and Neill. Using a subset of their data on the eight Australian states and territories for the years 1980–2006, with 1996 and 1997 omitted, making a total of $N = 200$ observations, we estimate the following model

$$HOM\_RATE = \beta_1 + \beta_2 GUNRATE + \beta_3 YEAR + e$$

Three equations are considered, one where *HOM_RATE* is the homicide rate from firearms, one where it represent the non-firearm homicide rate and one for the overall homicide rate, all measured in terms of deaths per million population. For the years after 1997, the variable *GUNRATE* is equal to the number of guns bought back during 1997, per thousand population; it is zero for the earlier years; *YEAR* is included to capture a possible trend. The estimated equations are given in Table 6.13.

a. Is there evidence that the gun buyback has reduced firearm homicides? Has there been an increase or a decrease in the homicide rate?

b. Using a joint test on the coefficients of *GUNRATE* and *YEAR*, test whether each of the homicide rates has remained constant over the sample period.

TABLE 6.13 **Estimates for Gun Buyback Equations for Exercise 6.14**

Coefficient Estimates and (Standard Errors)			
	Firearm Homicide Rate	Non-firearm Homicide Rate	Overall Homicide Rate
C	694	1097	1791
	(182)	(816)	(907)
GUNRATE	0.0181	0.0787	0.0968
	(0.0352)	(0.1578)	(0.1754)
YEAR	–0.346	–0.540	–0.886
	(0.092)	(0.410)	(0.456)
SSE	9017	181087	223842
s_y	7.1832	30.3436	34.0273

6.15 The following equation estimates the dependence of *CANS* (the weekly number of cans of brand A tuna sold in thousands) on the price of brand A in dollars (*PRA*) and the prices of two competing brands B and C (*PRB* and *PRC*). The equation was estimated using 52 weekly observations.

$$\widehat{E}(CANS|PRA, PRB, PRC) = 22.96 - 47.08PRA + 9.30PRB + 16.51PRC \qquad SSE = 1358.7$$

a. When *PRB* and *PRC* are omitted from the equation, the sum of squared errors increases to 1513.6. Using a 10% significance level, test whether the prices of the competing brands should be included in the equation. $\left(F_{(0.9,\,2,48)} = 2.417\right)$

b. Consider the following two estimated equations: $\widehat{E}(PRB|PRA) = 0.5403 + 0.3395PRA$ and $\widehat{E}(PRC|PRA) = 0.7708 + 0.0292PRA$. If *PRB* and *PRC* are omitted from the original equation for *CANS*, by how much will the coefficient estimate for *PRA* change? By how much will the intercept estimate change?

c. Find point and 95% interval estimates of $E(CANS|PRA = 0.91, PRB = 0.91, PRC = 0.90)$ using the original equation. The required standard error is 1.58.

d. Find a point estimate for $E(CANS|PRA = 0.91)$ using the equation you constructed in part (b). Can you suggest why the point estimates in (c) and (d) are different? Are there values for *PRB* and *PRC* for which they would be identical?

e. Find a 95% prediction interval for *CANS* when $PRA = 0.91$, $PRB = 0.91$ and $PRC = 0.90$. If you were a statistical consultant to the supermarket selling the tuna, how would you report this interval?

f. When $\widehat{CANS}^2$ is added to the original equation as a regressor the sum of squared errors decreases to 1198.9. Is there any evidence that the equation is misspecified?

6.16 Using 28 annual observations on output (Y), capital (K), labor (L) and intermediate materials (M) for the U.S manufacturing sector, to estimate the Cobb–Douglas production function

$$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \beta_4 \ln(M) + e$$

gave the following results

$$b_2 = 0.1856 \qquad b_3 = 0.3990 \qquad b_4 = 0.4157 \qquad SSE = 0.05699 \qquad s_{\ln(Y)} = 0.23752$$

The standard deviations of the explanatory variables are $s_{\ln(K)} = 0.28108$, $s_{\ln(L)} = 0.17203$, and $s_{\ln(M)} = 0.27505$. The sums of squared errors from running auxiliary regressions on the explanatory variables are (the subscript refers to the dependent variable in the auxiliary regression)

$$SSE_{\ln(K)} = 0.14216 \qquad SSE_{\ln(L)} = 0.02340 \qquad SSE_{\ln(M)} = 0.04199$$

a. Find (i) the standard errors for b_2, b_3, and b_4, (ii) the R^2's for each of the auxiliary regressions, and (iii) the variance inflation factors for b_2, b_3, and b_4.

b. Test the significance of b_2, b_3, and b_4 using a 5% level of significance.

c. Use a 5% level of significance to test the following hypotheses: (i) $H_0: \beta_2 = 0$, $\beta_3 = 0$, (ii) $H_0: \beta_2 = 0, \beta_4 = 0$, (iii) $H_0: \beta_3 = 0, \beta_4 = 0$, and (iv) $H_0: \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$. The restricted sums of squared errors for the first three hypotheses are (i) $SSE_R = 0.0551$, (ii) $SSE_R = 0.08357$ and (iii) $SSE_R = 0.12064$.

d. Comment on the presence and impact of collinearity.

6.7.2 Computer Exercises

6.17 Reconsider Example 6.16 in the text. In that example a number of models were assessed on their within-sample and out-of-sample predictive ability using data in the file *br5*. Of the models considered, the one with the best within-sample performance, as judged by the $\overline{R}^2$, AIC and SC criteria was

$$\ln(PRICE) = \beta_1 + \beta_2 AGE + \beta_3 SQFT + \beta_4 AGE^2 + e \tag{XR6.17}$$

In this exercise we investigate whether we can improve on this function by adding the number of bathrooms (*BATHS*) and the number of bedrooms (*BEDROOMS*). Estimate the equations required to fill in the following table. The models have been numbered from 9 to 12 as extensions of those in Table 6.3. Model 2 is the same as equation (XR6.17). For the subsequent models extra variables are added, with model 12 being the last one considered. For the RMSE values, use the last 100 observations as the hold-out sample. Discuss the results. Include in your discussion a comparison with the results in Table 6.3.

Model	Variables included in addition to those in (XR6.17)	$\overline{R}^2$	AIC	SC	RMSE
2	None				
9	*BATHS*				
10	*BATHS, BEDROOMS*				
11	*BATHS, BEDROOMS* × *SQFT*				
12	*BATHS, BEDROOMS* × *SQFT, BATHS* × *SQFT*				

6.18 Consider Example 6.17 where the rice production function

$$\ln(PROD) = \beta_1 + \beta_2 \ln(AREA) + \beta_3 \ln(LABOR) + \beta_4 \ln(FERT) + e$$

was estimated using data from the file *rice5*.

a. Using data from 1994 only, contrast the outcomes of the following hypothesis tests.

i. $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$,

ii. $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$,

iii. $H_0: \beta_2 = \beta_3 = 0$ versus $H_1: \beta_2 \neq 0$ or $\beta_3 \neq 0$ or both β_2 and β_3 are nonzero.

b. Show that the restricted model corresponding to the restriction $\beta_2 + \beta_3 + \beta_4 = 1$ is given by

$$\ln\left(\frac{PROD}{AREA}\right) = \beta_1 + \beta_3 \ln\left(\frac{LABOR}{AREA}\right) + \beta_4 \ln\left(\frac{FERT}{AREA}\right) + e$$

c. Some output from estimating the equation in part (b) using 1994 data is given in Table 6.6. It includes point and interval estimates for β_2, $\text{se}(b_2)$, and a p-value for testing $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$. Describe how these results can be obtained and verify that they are correct.

d. Estimate a constant-returns-to-scale production function using data from both 1993 and 1994. Compare the standard errors and 95% interval estimates with those in Table 6.7 where both years of data were used, but constant returns to scale was not imposed. Include all coefficients in your comparison. What are the auxiliary R^2's for the two variables in the restricted model?

6.19 Consider the following expenditure share equation where *WFOOD* is the proportion of household total expenditure allocated to food, *TOTEXP* is total weekly household expenditure in British pounds (£), and *NK* is the number of children in the household. Conditions MR1–MR5 are assumed to hold. We will be using data from the file *london5*.

$$WFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 \left[NK \times \ln(TOTEXP)\right] + e$$

a. For a household with the median total expenditure of £90, show that the change in $E(WFOOD|TOTEXP, NK)$ from adding an extra child is $\beta_3 + \beta_4 \ln(90)$.

b. For a household with two children, show that the change in $E(WFOOD|TOTEXP, NK)$ from an increase in total expenditure from £80/week to £120/week is $\beta_2 \ln(1.5) + 2\beta_4 \ln(1.5)$.

c. For a household with two children and total expenditure of £90/week, show that

$$E(WFOOD|TOTEXP, NK) = \beta_1 + \beta_2 \ln(90) + 2\beta_3 + 2\beta_4 \ln(90)$$

d. Consider the following three statements:

A. $\beta_3 + \beta_4 \ln(90) = 0.025$

B. $\beta_2 \ln(1.5) + 2\beta_4 \ln(1.5) = -0.04$

C. $\beta_1 + \beta_2 \ln(90) + 2\beta_3 + 2\beta_4 \ln(90) = 0.37$

We will be concerned with using F and χ^2 tests to test the following three null hypotheses: $H_0^{(1)}$: A is true; $H_0^{(2)}$: A and B are true; $H_0^{(3)}$: A and B and C are true. The alternative hypothesis in each case is that $H_0^{(i)}$ is not true.

What are the relationships between the F and χ^2 tests for each of the three hypotheses? For $H_0^{(1)}$, what is the relationship between the t and F tests?

e. Find the p-values for the F and χ^2 tests for $H_0^{(1)}$, $H_0^{(2)}$, and $H_0^{(3)}$, first using the first 100 observations in *london5*, then using the first 400 observations, and then using all 850 observations.

f. Comment on how changing the sample size, and adding more hypotheses, affects the results of the tests. Are there any dramatic differences between the F-test outcomes and the χ^2-test outcomes?

6.20 In Example 6.18, using 900 observations from the data file *br5*, we identified three potentially influential observations in the estimation of the model

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \beta_4 AGE^2 + e$$

Those observations were numbers 150, 411 and 540.

a. Estimate the model with (i) all observations, (ii) observation 150 excluded, (iii) observation 411 excluded, (iv) observation 540 excluded, and (v) observations 150, 411, and 540 excluded. Report the results and comment on their sensitivity to the omission of the observations.

b. Using the estimates from all observations, find the forecast errors corresponding to the within sample predictions at observations 150, 411, and 540.

c. Using the estimates obtained when observation 150 is excluded, find the out-of-sample forecast error for observation 150.

d. Using the estimates obtained when observation 411 is excluded, find the out-of-sample forecast error for observation 411.

e. Using the estimates obtained when observation 540 is excluded, find the out-of-sample forecast error for observation 540.

f. Using the estimates obtained when observations 150, 411, and 540 are excluded, find the out-of-sample forecast errors for observations 150, 411, and 540.

g. Compare the forecast errors obtained in parts (b)–(f) and comment on their sensitivity to the omission of the observations.

6.21 Reconsider Example 6.20 where a logistic growth curve for the share of U.S. steel produced by electric arc furnace (EAF) technology was estimated. The curve is given by the equation

$$y_t = \frac{\alpha}{1 + \exp(-\beta - \delta t)} + e_t$$

a. Find 95% interval estimates for the following:

i. The saturation level α.

ii. The inflection point $t_I = -\beta/\delta$ at which the rate of growth starts to decline. What years does the interval correspond to?

iii. The EAF share in 1969.

iv. The predicted EAF shares from 2016 to 2050. Plot the predictions and their 95% bounds. Comment on how far the predictions are from the saturation level and on the behavior of the 95% bounds.

b. Use a 5% significance level to test the joint null hypothesis that the saturation level is 0.85 and the point of inflection is at $t_I = 25$. Set up the null hypothesis for the point of inflection so that it is linear in the parameters β and δ. Given the interval estimates you found in (a)(i) and (a)(ii), does

the result surprise you? What extra information does the test use that was not used in (a)(i) and (a)(ii)?

c. Estimate the model with the restrictions implied by the null hypothesis in (b) imposed. Find the sum of squared errors and test the null hypothesis with an F-test that uses the restricted and unrestricted sums of squared errors. How does this result compare with that from the automatic test command that you used for part (b)?

6.22 To examine the quantity theory of money, Brumm[15] specifies the equation

$$INFLAT = \beta_1 + \beta_2 MONEY + \beta_3 OUTPUT + e$$

where *INFLAT* is the growth rate of the general price level, *MONEY* is the growth rate of the money supply, and *OUTPUT* is the growth rate of national output. According to theory we should observe that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$. The data in the data file *brumm* are on 76 countries for the year 1995.

a. Using a 5% significance level, test

i. the *strong* joint hypothesis that $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_3 = -1$.

ii. the *weak* joint hypothesis $\beta_2 = 1$ and $\beta_3 = -1$.

b. Using the DFFITS criterion, find the four most influential observations.

c. Repeat the two tests with the four most influential observations omitted. Does omission of these four observations change the test outcome?

d. Moroney[16] has argued that β_2 is likely to be different for different countries. Suppose that $\beta_2 = \alpha_1 + \alpha_2 MONEY + \alpha_3 OUTPUT$. Substitute this equation into the original model and, omitting the same four influential observations, estimate the new model.

e. Repeat the two tests for the model estimated in (d) for a hypothetical country with the sample median values $MONEY = 16.35$ and $OUTPUT = 2.7$.

6.23 For two inputs X_1 and X_2 and output Y, a constant elasticity of substitution (CES) production function is given by

$$Y = \alpha\left[\delta X_1^{-\rho} + (1-\delta) X_2^{-\rho}\right]^{-\eta/\rho}$$

where $\alpha > 0$ is an efficiency parameter, $\eta > 0$ is a returns to scale parameter, $\rho > -1$ is a substitution parameter, and $0 < \delta < 1$ is a distribution parameter that relates the share of output to each of the two inputs. The elasticity of substitution between the two inputs is given by $\varepsilon = 1/(1+\rho)$. If $\eta = 1$ and $\rho = 0$ ($\varepsilon = 1$), then the CES production function simplifies to the constant-returns-to-scale Cobb–Douglas production function $Y = \alpha X_1^{\delta} X_2^{1-\delta}$.[17] Using the data in the file *rice5*, define $Y = PROD/AREA$, $X_1 = LABOR/AREA$ and $X_2 = FERT/AREA$.

a. Using nonlinear least squares, estimate the following log form of the CES function

$$\ln(Y) = \beta - \frac{\eta}{\rho}\ln\left[\delta X_1^{-\rho} + (1-\delta) X_2^{-\rho}\right] + e$$

where $\beta = \ln(\alpha)$. Report your results and standard errors. [*Hint*: If you run into difficulties, try using 0.5 as the starting value for all of your parameters.]

b. Find 95% interval estimates for α, η, ε, and δ.

c. Using a 5% significance level, test the null hypothesis $H_0: \eta = 1$, $\rho = 0$ against the alternative $H_1: \eta \neq 1$ or $\rho \neq 0$. Does a constant-returns-to-scale Cobb–Douglas function appear to be adequate?

6.24 Using the data in the file *br5*, find least squares estimates of the following house-price relationships for houses sold in Baton Rouge during 2005.

$$\ln(PRICE) = \alpha_1 + \alpha_2 BEDROOMS + e_1$$
$$\ln(PRICE) = \beta_1 + \beta_2 BEDROOMS + \beta_3 SQFT + e_2$$
$$SQFT = \gamma_1 + \gamma_2 BEDROOMS + u_1$$

[15] Brumm, H.J. (2005) "Money Growth, Output Growth, and Inflation: A Reexamination of the Modern Quantity Theory's Linchpin Prediction," *Southern Economic Journal*, 71(3), 661–667.

[16] Moroney, J.R. (2002), "Money Growth, Output Growth and Inflation: Estimation of a Modern Quantity Theory," *Southern Economic Journal*, 69(2), 398–413.

[17] Proving this result requires some advanced calculus. You need to take natural logarithms of both sides, set $\eta = 1$ and use l'Hôpital's rule to take limits as $\rho \to 0$.

a. Report the coefficient estimates and their standard errors.

b. Show how the estimates $(\hat{\alpha}_1, \hat{\alpha}_2)$ can be found from the parameter estimates in the other two equations. How does the interpretation of $\hat{\beta}_2$ differ from the interpretation of $\hat{\alpha}_2$? What would you characterize as the omitted variable bias when estimating α_2? Is there evidence that *BEDROOMS* has a direct effect on ln(*PRICE*)?

c. Estimate the equation $\ln(PRICE) = \theta_1 + \theta_2 SQFT + e_3$. Compare the estimates $\hat{\theta}_2$ and $\hat{\beta}_3$. What was the effect of omitting *BEDROOMS* on the estimated coefficient for *SQFT*? What assumption about e_3 is necessary for θ_2 to be given the causal interpretation: an increase in house size of 100 square feet leads to a θ_2 increase in ln(*PRICE*), when all other variables are held constant?

d. We will investigate whether this assumption might be violated. Estimate the following equation and report the results

$$\ln(PRICE) = \delta_1 + \delta_2 SQFT + \delta_3 AGE + \delta_4 AGE^2 + e_4$$

e. A comparison of this equation with that in part (c) suggests $e_3 = \delta_3 AGE + \delta_4 AGE^2 + e_4$. Assume $E(e_4|SQFT, AGE) = 0$. We wish to investigate whether $E(e_3|SQFT) = 0$. Show that $E(e_3|SQFT) = 0$ if $\delta_3 = \delta_4 = 0$ **or** if $E(AGE|SQFT) = 0$ and $E(AGE^2|SQFT) = 0$.

f. Test the hypothesis $H_0 : \delta_3 = \delta_4 = 0$ at a 5% significance level.

g. Estimate the equations $AGE = \lambda_1 + \lambda_2 SQFT + u_2$ and $AGE^2 = \phi_1 + \phi_2 SQFT + u_3$. Use a 5% significance level to test the hypotheses $H_0 : \lambda_2 = 0$ and $H_0 : \phi_2 = 0$.

h. What do you conclude about the assumption $E(e_3|SQFT) = 0$?

6.25 Using the data in the file *br5*, estimate the equation

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \beta_4 AGE^2 + e$$

where *PRICE* is the selling price in thousands of dollars for houses sold in Baton Rouge, Louisiana, in 2005, *SQFT* is the size of each house in hundreds of square feet and *AGE* is the age of each house in years.

a. Report the coefficient estimates and their standard errors.

b. Graph the estimate of $E[\ln(PRICE)|SQFT = 22, AGE]$ against *AGE*. (In the sample the median and average values for *SQFT* are 21.645 and 22.737, respectively.)

c. In part (b), you will have noticed that the higher-priced houses are the very new ones and the very old ones. Using a 5% significance level test the joint null hypothesis that (i) two houses of the same size, a 5-year old house and an 80-year old house, have the same expected log-price, *and* (ii) a 5-year old house with 2000 square feet has the same expected log-price as a 30-year old house with 2800 square feet.

d. Using a 5% significance level, test the joint null hypothesis that (i) houses start becoming more expensive with age when they are 50 years old, *and* (ii) a 2200 square feet house that is 50 years old has an expected log-price that corresponds to $100,000.

e. Add the variables *BATHS* and $SQFT \times BEDROOMS$ to the model with coefficients β_5 and β_6, respectively. Estimate this model and report the results.

f. Using a 5% significance level, test whether adding these two variables has improved the predictive ability of the model.

g. You are building a new 2300 square-feet house ($AGE = 0$) with three bedrooms and two bathrooms. Adding one extra bedroom and bathroom will increase its size by 260 square feet. Estimate the increase in value of the house from the extra bedroom and bathroom. (Use the natural predictor.)

h. What do you estimate will be the extra value of the house in 20 years' time?

6.26 Each morning between 6:30AM and 8:00AM Bill leaves the Melbourne suburb of Carnegie to drive to work at the University of Melbourne. The time it takes Bill to drive to work (*TIME*), depends on the departure time (*DEPART*), the number of red lights that he encounters (*REDS*), and the number of trains that he has to wait for at the Murrumbeena level crossing (*TRAINS*). Observations on these variables for the 249 working days in 2015 appear in the data file *commute5*. *TIME* is measured in minutes. *DEPART* is the number of minutes after 6:30AM that Bill departs. Consider the equation

$$TIME = \beta_1 + \beta_2 DEPART + \beta_3 REDS + \beta_4 TRAINS + e$$

and suppose assumptions MR1–MR5 hold.

a. Test the following joint hypotheses using a 5% significance level:
  i. The expected delay from a red light is 1.8 minutes *and* the expected delay from a train is 3.2 minutes.
  ii. The expected delay from a red light is 2 minutes *and* the expected delay from a train is 3 minutes.
  iii. The expected delay from a train is 3.5 minutes *and* the delay from a train is double that from a red light.
  iv. The expected delay from a train is 3.5 minutes *and* the delay from a train is double that from a red light *and* leaving at 7:30AM instead of 7:00AM makes the trip 10 minutes longer.

b. Bill suspects that the later he leaves, the more likely he is to encounter a train. Test this hypothesis at a 5% significance level using estimates from the model

$$E(TRAINS|DEPART, REDS) = \alpha_1 + \alpha_2 DEPART + \alpha_3 REDS$$

Is there any evidence of a relationship between the number of trains and the number of red lights?

c. Show that

$$E(TIME|DEPART, REDS) = \left(\beta_1 + \beta_4\alpha_1\right) + \left(\beta_2 + \beta_4\alpha_2\right) DEPART + \left(\beta_3 + \beta_4\alpha_3\right) REDS$$

d. Regress *TIME* on *DEPART* and *REDS* to get estimates for $\delta_1 = \beta_1 + \beta_4\alpha_1$, $\delta_2 = \beta_2 + \beta_4\alpha_2$, and $\delta_3 = \beta_3 + \beta_4\alpha_3$. Using these estimates and those from (a) and (c), show that $\hat{\delta}_1 = b_1 + b_4\hat{\alpha}_1$, $\hat{\delta}_2 = \left(b_2 + b_4\hat{\alpha}_2\right)$, and $\hat{\delta}_3 = b_3 + b_4\hat{\alpha}_3$, where b_k denotes an OLS estimate from the original equation.

e. Interpret b_2 and $\hat{\delta}_2$. Why are they different? How would you characterize any omitted variable bias?

6.27 It has been claimed that an extra year of experience increases wage by 0.8% and that an extra year of education is worth 14 extra years of experience. Doing the calculation, this would mean an extra year of education increases wage by 11.2%. We will investigate this hypothesis using data in the file *cps5_small*. Only those observations for which years of education exceeds 7 will be used. Perform all tests at a 5% level of significance.

a. Estimate the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$ and jointly test the claims about the marginal effects of *EDUC* and *EXPER*.

b. Use RESET to test the adequacy of the model; perform the test with the squares of the predictions *and* the squares and cubes of the predictions.

c. After estimating the model

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EDUC^2 + \beta_5 EXPER^2 + \beta_6(EDUC \times EXPER) + e$$

jointly test the claims about the marginal effects of *EDUC* and *EXPER* at the following levels of *EDUC* and *EXPER*:
  i. $EDUC = 10$, $EXPER = 5$
  ii. $EDUC = 14$, $EXPER = 24$
  iii. $EDUC = 18$, $EXPER = 40$

d. Use RESET to test the adequacy of the model; perform the test with the squares of the predictions *and* the squares and cubes of the predictions.

e. How would you respond to the claim about the marginal effects of *EDUC* and *EXPER*?

6.28 Using time-series data on five different countries, Atkinson and Leigh[18] investigate the impact of the marginal tax rate paid by high-income earners on the level of inequality. A subset of their data can be found in the file *inequality*.

a. Using data on Australia, estimate the equation $SHARE = \beta_1 + \beta_2 TAX + e$ where *SHARE* is the percentage income share of the top 1% of incomes, and *TAX* is the median marginal tax rate (as a percentage) paid on wages by the top 1% of income earners. Interpret your estimate for β_2. Would you interpret this as a causal relationship?

[18] Atkinson, A.B. and A. Leigh (2013), "The Distribution of Top Incomes in Five Anglo-Saxon Countries over the Long Run," *Economic Record*, 89, 1–17.

b. It is generally recognized that inequality was high prior to the great depression, then declined during the depression and World War II, increasing again toward the end of the sample period. To capture this effect, estimate the following model with a quadratic trend

$$SHARE = \alpha_1 + \alpha_2 TAX + \alpha_3 YEAR + \alpha_4 YEAR^2 + e$$

where *YEAR* is defined as $1 = 1921, 2 = 1922, \ldots, 80 = 2000$. Interpret the estimate for α_2. Has adding the trend changed the effect of the marginal tax rate? Can the change in this estimate, or lack of it, be explained by the correlations between *TAX* and *YEAR* and *TAX* and $YEAR^2$?

c. In what year do you estimate that expected *SHARE* will be smallest? Find a 95% interval estimate for this year. Does the actual year with the smallest value for *SHARE* fall within the interval?

d. The top marginal tax rate in 1974 was 64%. Test the hypothesis that, in the year 2000, the expected income share of the top 1% would have been 6% if the marginal tax rate had been 64% at that time.

e. Test jointly the hypothesis in (d) *and* that a marginal tax rate of 64% in 1925 would have led to an expected income share of 6% for the top 1% of income earners.

f. Add the growth rate (*GWTH*) to the equation in part (b) and reestimate. Interpret the estimated coefficient for *TAX*.

g. Using the equation estimated in part (f), estimate the year when *SHARE* will be smallest? Find a 95% interval estimate for this year. Does the actual year with the smallest value for *SHARE* fall within the interval?

h. Using the equation estimated in part (f), test the hypothesis that, in the year 2000, the expected income share of the top 1% would have been 6% if the marginal tax rate had been 64% at that time.

i. Using the equation estimated in part (f), test jointly the hypothesis in (h) *and* that a marginal tax rate of 64% in 1925 would have led to an expected income share of 6% for the top 1% of income earners.

j. Has adding the variable *GWTH* led to substantial changes to your estimates and test results? Can the changes, or lack of them, be explained by the correlations between *GWTH* and the other variables in the equation?

6.29 Using time-series data on five different countries, Atkinson and Leigh investigate the impact of the marginal tax rate paid by high-income earners on the level of inequality. A subset of their data can be found in the file *inequality*.

a. Using data on the United States, estimate the equation $\ln(SHARE) = \beta_1 + \beta_2 TAX + e$ where *SHARE* is the percentage income share of the top 1% of incomes, and *TAX* is the median marginal tax rate (as a percentage) paid on wages by the top 1% of income earners. Interpret your estimate for β_2. Would you interpret this as a causal relationship?

b. It is generally recognized that inequality was high prior to the great depression, then declined during the depression and World War II, increasing again toward the end of the sample period. To capture this effect, estimate the following model with a quadratic trend

$$\ln(SHARE) = \alpha_1 + \alpha_2 TAX + \alpha_3 YEAR + \alpha_4 YEAR^2 + e$$

where *YEAR* is defined as $1 = 1921, 2 = 1922, \ldots, 80 = 2000$. Interpret the estimate for α_2. Has adding the trend changed the effect of the marginal tax rate? Can the change in this estimate, or lack of it, be explained by the correlations between *TAX* and *YEAR* and *TAX* and $YEAR^2$?

c. In what year do you estimate that *SHARE* will be smallest? Find a 95% interval estimate for this year. Does the actual year with the smallest value for *SHARE* fall within the interval?

d. The top marginal tax rate in 1974 was 50%. Test the hypothesis that, in the year 2000, the expected log income share of the top 1% would have been ln(12) if the marginal tax rate had been 50% at that time.

e. Test jointly the hypothesis in (d) *and* that a marginal tax rate of 50% in 1925 would have led to an expected log income share of log(12) for the top 1% of income earners.

f. Add the growth rate (*GWTH*) to the equation in part (b) and re-estimate. Has adding this variable *GWTH* led to substantial changes to your estimates and test results? Can the changes, or lack of them, be explained by the correlations between *GWTH* and the other variables in the equation?

g. Using the results from part (f), find point and 95% interval estimates for the marginal tax rate that would be required to reduce the income share of the top 1% to 12% in 2001, assuming $GWTH_{2001} = 3$.

6.30 Consider a translog production function where output is measured as firm sales and there are three inputs: capital, labor, and materials. This function can be written as

$$LSALES = \beta_C + \beta_K K + \beta_L L + \beta_M M + \beta_{KK} K^2 + \beta_{LL} L^2 + \beta_{MM} M^2 + \beta_{KL}(K \times L) + \beta_{KM}(K \times M) + \beta_{LM}(L \times M) + e$$

where *LSALES* is the log of sales, and K, L, and M are the logs of capital, labor and materials, respectively. The translog function is often known as a flexible functional form, intended to approximate a variety of possible functional forms. There are two hypotheses that are likely to be of interest:

$$H_0^{(1)}: \beta_{KK} = 0, \ \beta_{LL} = 0, \ \beta_{MM} = 0, \ \beta_{KL} = 0, \ \beta_{KM} = 0, \ \beta_{LM} = 0$$

(A Cobb–Douglas function is adequate)

$$H_0^{(2)}: \begin{cases} \beta_K + \beta_L + \beta_M = 1 \\ 2\beta_{KK} + \beta_{KL} + \beta_{KM} = 0 \\ \beta_{KL} + 2\beta_{LL} + \beta_{LM} = 0 \\ \beta_{KM} + \beta_{LM} + 2\beta_{MM} = 0 \end{cases} \quad \text{(constant returns to scale)}$$

The data file *chemical_small* contains observations on 1200 firms in China's chemical industry, taken in the year 2006. It is a subset of the data used by Baltagi, Egger, and Kesina[19].

a. Use these data to estimate the translog production function. Are all the coefficient estimates significant at a 5% level of significance?

b. Test $H_0^{(1)}$ at a 5% level of significance.

c. Test $H_0^{(2)}$ at a 5% level of significance. What would be the test outcome if you used a 1% level of significance?

d. Does RESET suggest the translog function is adequate?

e. Estimate the model with the restrictions implied by constant returns to scale $\left(H_0^{(2)}\right)$ imposed. Obtain estimates and standard errors for all 10 coefficients.

f. Compare the estimates and standard errors from parts (a) and (e).

g. Does RESET suggest the restricted model is adequate?

6.31 Everaert and Pozzi[20] develop a model to examine the predictability of consumption growth in 15 OECD countries. Their data is stored in the file *oecd*. The variables used are growth in real per capita private consumption (*CSUMPTN*), growth in real per capita government consumption (*GOV*), growth in per capita hours worked (*HOURS*), growth in per capita real disposable labor income (*INC*), and the real interest rate (*R*). Using only the data for Japan, answer the following questions:

a. Estimate the following model and report the results

$$CSUMPTN = \beta_1 + \beta_2 HOURS + \beta_3 GOV + \beta_4 R + \beta_5 INC + e$$

Are there any coefficient estimates that are not significantly different from zero at a 5% level?

b. The coefficient β_2 could be positive or negative depending on whether hours worked and private consumption are complements or substitutes. Similarly, β_3 could be positive or negative depending on whether government consumption and private consumption are complements or substitutes. What have you discovered? What does a test of the hypothesis $H_0: \beta_2 = 0, \beta_3 = 0$ reveal?

c. Re-estimate the equation with *GOV* omitted and, for the coefficients of the remaining variables, comment on any changes in the estimates and their significance.

d. Estimate the equation

$$GOV = \alpha_1 + \alpha_2 HOURS + \alpha_3 R + \alpha_4 INC + v$$

and use these estimates to reconcile the estimates in part (a) with those in part (c).

---

[19] Baltagi, B.H., P. H. Egger and M. Kesina (2016), "Firm-level Productivity Spillovers in China's Chemical Industry: A Spatial Hausman-Taylor Approach," *Journal of Applied Econometrics*, 31(1), 214–248.

[20] Everaert, G. and L. Ponzi (2014), "The Predictability of Aggregate Consumption Growth in OECD Countries: A Panel Data Analysis," *Journal of Applied Econometrics*, 29(3), 431–453.

e. Re-estimate the models in parts (a) and (c) with the year 2007 omitted and use each of the estimated models to find point and 95% interval forecasts for consumption growth in 2007.

f. Which of the two models, (a) or (c), produced the more accurate forecast for 2007?

6.32 In their study of the prices of Californian and Washington red wines, Costanigro, Mittelhammer and McCluskey[21] categorize the wines into commercial, semipremium, premium, and ultrapremium. Their data for premium wines are stored in the file *wine1*; those for ultrapremium wines are in the file *wine2*. We will be concerned with the variables *PRICE* (bottle price, CPI adjusted), *SCORE* (score out of 100 given by the Wine Spectator Magazine), *AGE* (years of aging), and *CASES* (number of cases produced in thousands).

a. What signs would you expect on the coefficients $(\beta_2, \beta_3, \beta_4)$ in the following model? Why?

$$\ln(PRICE) = \beta_1 + \beta_2 SCORE + \beta_3 AGE + \beta_4 CASES + e$$

b. Estimate separate equations for premium and ultrapremium wine, and discuss the results. Do the coefficients have the expected signs? If not is there an alternative explanation? Is *SCORE* more important for premium wines or ultrapremium wines? Is *AGE* more important for premium wines or ultrapremium wines?

c. Find point and 95% interval estimates for

i. $E[\ln(PRICE)|SCORE = 90, AGE = 2, CASES = 2]$ for premium wines, and

ii. $E[\ln(PRICE)|SCORE = 93, AGE = 3, CASES = 1]$ for ultrapremium wines.

Do the intervals overlap, or is there a clear price distinction between the two classes?

d. Using the "corrected predictor"—see Section 4.5.3—predict the prices for premium and ultrapremium wines for the settings in parts c(i) and c(ii), respectively.

e. Suppose that you are a wine producer choosing between producing 1000 cases of ultrapremium wine that has to be aged three years and is likely to get a score of 93, and 2000 cases of premium wine that is aged two years and is likely to get a score of 90. Which choice gives the higher expected bottle price? Which choice gives the higher expected revenue? (There are 12 bottles in a case of wine.)

6.33 In this exercise we reconsider the premium wine data in the file *wine1*. Please see Exercise 6.32 and *wine1.def* for details.

a. Estimate the following equation using (i) only cabernet wines, (ii) only pinot wines, and (iii) all other varieties:

$$\ln(PRICE) = \beta_1 + \beta_2 SCORE + \beta_3 AGE + \beta_4 CASES + e$$

Using casual inspection, do you think separate equations are needed for the different varieties?

b. We can develop an F-test to test whether there is statistical evidence to suggest the coefficients in the three equations are different. The unrestricted sum of squared errors for such a test is

$$SSE_U = SSE_{CABERNET} + SSE_{PINOT} + SSE_{OTHER}$$

Compute SSE_U.

c. What is the total number of parameters from the three equations? How many parameters are there when we estimate one equation for all varieties? How many parameter restrictions are there if we restrict corresponding coefficients for all varieties to be equal?

d. Estimate one equation for all varieties. This is the restricted model where corresponding coefficients for the different varieties are assumed to be equal.

e. Using a 5% significance level, test whether there is evidence to suggest there should be different equations for different varieties. What is the null hypothesis for this test? Develop some notation that enables you to state the null hypothesis clearly and precisely.

---

[21] Costanigro, M., R.C. Mittelhammer and J.J.McCluskey (2009), "Estimating Class Specific Parametric Models Under Class Uncertainty: Local Polynomial Regression Clustering in an Hedonic Analysis Of Wine Markets" *Journal of Applied Econometrics*, 24(7), 1117–1135.

Appendix 6A The Statistical Power of *F*-Tests

In Appendix 3B, we explored the factors that lead us to reject a null hypothesis about the slope parameter in a simple regression using a *t*-test. The probability of rejecting a false null hypothesis is positively related to the magnitude of the hypothesis error, and the total variation in the explanatory variable, and inversely related to the size of σ^2, the error variance. These are components of the noncentrality parameter, (3B.2), for the *t*-statistic, (3B.1), when the null hypothesis is false.

Here we show that the factors that lead us to reject a false joint null hypothesis are much the same. Consider the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ under assumptions SR1–SR6. We will test the joint null hypothesis $H_0 : \beta_1 = c_1, \beta_2 = c_2$ using an *F*-test. In practice the test is carried out using (6.4) in the usual way. To study the power of the *F*-test we will test an equivalent joint null hypothesis $H_0 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}, \quad \beta_2 = c_2$. If the first pair of hypotheses is true then the second pair of hypotheses is true and vice versa. They are completely equivalent. This is not what you would do in practice but this approach will lead us to a form of the *F*-test that is theoretically useful. In the following steps, we will derive the *F*-statistic by combining test statistics for the separate hypotheses $H_0^1 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$ and $H_0^2 : \beta_2 = c_2$. There are quite a few steps, but do not get discouraged. Each step is small and the reward at the end is substantial. Now is a good time to review Appendix 3B on *t*-tests when the null hypothesis is false, Appendix B.3.6, on the chi-square distribution, Appendix B.3.7, on the *t*-distribution, and Appendix B.3.8, on the *F*-distribution.

If we were going to test the first hypothesis, $H_0^1 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$, what test statistic would we use? Most commonly we use a *t*-test for a single hypothesis. For the present, however, assume that we know the error variance σ^2 so that we also know the true variances and covariance of the least squares estimators that are given in equations (2.14)–(2.16). The test statistic is

$$Z_0^1 = \frac{b_1 + b_2\bar{x} - (c_1 + c_2\bar{x})}{\sqrt{\text{var}(b_1 + b_2\bar{x})}} = \frac{\bar{y} - (c_1 + c_2\bar{x})}{\sqrt{\sigma^2/N}} \tag{6A.1}$$

with Z_0^1 denoting the statistic for the null hypothesis H_0^1. We obtained the second equality by taking advantage of the properties of the least squares estimators, recognizing that $b_1 + b_2\bar{x} = \bar{y}$, and $\text{var}(\bar{y}) = \sigma^2/N$, as shown in Appendix C, equation (C.6). If the null hypothesis is true, Z_0^1 has a standard normal distribution, $N(0,1)$. Our objective is to study testing $H_0^1 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$ when it is not true. To accomplish this rewrite Z_0^1 by adding and subtracting $(\beta_1 + \beta_2\bar{x})$ to the numerator in (6A.1) yielding

$$\begin{aligned} Z_0^1 &= \frac{b_1 + b_2\bar{x} - (\beta_1 + \beta_2\bar{x}) + (\beta_1 + \beta_2\bar{x}) - (c_1 + c_2\bar{x})}{\sqrt{\sigma^2/N}} \\ &= \frac{(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}}{\sqrt{\sigma^2/N}} + \frac{(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}}{\sqrt{\sigma^2/N}} \\ &= Z_1 + \delta_1 \end{aligned} \tag{6A.2}$$

The first term, Z_1, has a standard normal distribution; it is the test statistic calculated using the true parameter values,

$$Z_1 = \frac{(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}}{\sqrt{\sigma^2/N}} \sim N(0,1) \tag{6A.3}$$

The second term

$$\delta_1 = \frac{(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}}{\sqrt{\sigma^2/N}} \tag{6A.4}$$

is the specification error in the hypothesis $H_0^1: \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$. If the null hypothesis is true then $\delta_1 = 0$. If the null hypothesis H_0^1 is not true, then $\delta_1 \neq 0$, and we must account for the fact that δ_1 depends on the sample values, $\mathbf{x}$. In Appendix B.3.6 we define noncentral chi-square random variables. The random variable $Z_0^1|\mathbf{x} = Z_1 + \delta_1 \sim N(\delta_1, 1)$ and $V_0^1|\mathbf{x} = \left(Z_0^1|\mathbf{x}\right)^2 = \left(Z_1 + \delta_1\right)^2 \sim \chi^2_{(1,\delta_1^2)}$ has a noncentral chi-square distribution with one degree of freedom, and noncentrality parameter $\delta = \delta_1^2$. If the null hypothesis is true then $\delta_1 = 0$ and V_0^1 has the chi-square distribution, $V_0^1 \sim \chi^2_{(1,\delta_1^2=0)} = \chi^2_{(1)}$.

The second piece of the puzzle is similar to the first and follows the steps in Appendix 3B. To test $H_0^2: \beta_2 = c_2$, assuming σ^2 is known, use the test statistic

$$Z_0^2 = \frac{b_2 - c_2}{\sqrt{\text{var}(b_2)}} = \frac{b_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \tag{6A.5}$$

If the null hypothesis is true, Z_0^2 has a standard normal distribution, $N(0,1)$. Our objective is to study testing $H_0^2: \beta_2 = c_2$ when it is not true. To accomplish this rewrite Z_0^2 by adding and subtracting β_2 to the numerator, obtaining

$$Z_0^2 = \frac{b_2 - \beta_2 + \beta_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} + \frac{\beta_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} = Z_2 + \delta_2 \tag{6A.6}$$

The first term, Z_2, has a standard normal distribution; it is the test statistic calculated using the true parameter value

$$Z_2 = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \sim N(0, 1) \tag{6A.7}$$

The second term

$$\delta_2 = \frac{\beta_2 - c_2}{\sqrt{\sigma^2 / \sum(x_i - \bar{x})^2}} \tag{6A.8}$$

is the specification error in the hypothesis $H_0^2: \beta_2 = c_2$. If the null hypothesis is true then $\delta_2 = 0$; if the null hypothesis H_0^2 is not true, then $\delta_2 \neq 0$. The random variable $Z_0^2|\mathbf{x} = Z_2 + \delta_2 \sim N(\delta_2, 1)$ and $V_0^2|\mathbf{x} = \left(Z_0^2|\mathbf{x}\right)^2 = \left(Z_2 + \delta_2\right)^2 \sim \chi^2_{(1,\delta_2^2)}$ has a noncentral chi-square distribution with one degree of freedom, and noncentrality parameter $\delta = \delta_2^2$. If the null hypothesis is true, then $\delta_2 = 0$ and V_0^2 has the chi-square distribution, $V_0^2 \sim \chi^2_{(1,\delta_2^2=0)} = \chi^2_{(1)}$.

What is the distribution of $V_1 = V_0^1 + V_0^2 = \left(Z_1 + \delta_1\right)^2 + \left(Z_2 + \delta_2\right)^2$? If Z_1 and Z_2 are statistically independent then $V_1|\mathbf{x} \sim \chi^2_{(2,\delta)}$ with noncentrality parameter $\delta = \delta_1^2 + \delta_2^2$. Because Z_1 and Z_2 are normally distributed random variables, we can prove they are independent by showing that their correlation, or covariance, is zero. Their covariance is

$$\text{cov}(Z_1, Z_2) = E\left\{ \left[Z_1 - E(Z_1)\right]\left[Z_2 - E(Z_2)\right] \right\} = E(Z_1 Z_2)$$

because Z_1 and Z_2 have zero mean, $E(Z_1) = E(Z_2) = 0$. We will show that $E(Z_1 Z_2|\mathbf{x}) = 0$ from which it follows that $E(Z_1 Z_2) = 0$.

$$E(Z_1 Z_2|\mathbf{x}) = E\left\{\left[\frac{(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}}{\sqrt{\sigma^2/N}}\right]\left[\frac{b_2 - \beta_2}{\sqrt{\sigma^2/\sum(x_i - \bar{x})^2}}\right]\middle|\mathbf{x}\right\}$$

$$= E\left\{\frac{\sqrt{N}}{\sigma}\left[(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}\right]\frac{\sqrt{\sum(x_i - \bar{x})^2}}{\sigma}(b_2 - \beta_2)\middle|\mathbf{x}\right\} \tag{6A.9}$$

$$= \frac{\sqrt{N}\sqrt{\sum(x_i - \bar{x})^2}}{\sigma^2}E\left\{\left[(b_1 - \beta_1) + (b_2 - \beta_2)\bar{x}\right](b_2 - \beta_2)|\mathbf{x}\right\}$$

The key component in the last equality is, using (2.15) and (2.16),

$$E\left[(b_1 - \beta_1)(b_2 - \beta_2) + (b_2 - \beta_2)^2\bar{x}\middle|\mathbf{x}\right] = \left[\text{cov}(b_1, b_2|\mathbf{x}) + \bar{x}\text{var}(b_2|\mathbf{x})\right]$$

$$= \frac{-\bar{x}\sigma^2}{\sum(x_i - \bar{x})^2} + \frac{\bar{x}\sigma^2}{\sum(x_i - \bar{x})^2} = 0$$

Since the covariance between Z_1 and Z_2 is zero, they are statistically independent. Thus, $V_1|\mathbf{x} \sim \chi^2_{(2,\delta)}$ where $\delta = \delta_1^2 + \delta_2^2$ and

$$\delta = \delta_1^2 + \delta_2^2 = \left[\frac{(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}}{\sqrt{\sigma^2/N}}\right]^2 + \left[\frac{\beta_2 - c_2}{\sqrt{\sigma^2/\sum(x_i - \bar{x})^2}}\right]^2$$

$$= N\left\{\frac{\left[(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}\right]^2}{\sigma^2}\right\} + \frac{(\beta_2 - c_2)^2\sum_{i=1}^N(x_i - \bar{x})^2}{\sigma^2} \tag{6A.10}$$

The final step is to use V_2 from Section 6.1.5, and that V_1 and V_2 are statistically independent. Following a similar procedure to that in (6.13), we form the F-ratio

$$F|\mathbf{x} = \frac{V_1/2}{V_2/(N-2)} \sim F_{(2,N-2,\delta)}$$

In Figure B.9b we show that increases in the noncentrality parameter δ shifts the F-density to the right, increasing the probability that it exceeds the appropriate critical value F_c, and increasing the probability of rejecting a false null hypothesis.

Examining the noncentrality parameter δ in (6A.10) we first note that $\delta \geq 0$, and $\delta = 0$ only if the joint null hypothesis $H_0 : \beta_1 + \beta_2\bar{x} = c_1 + c_2\bar{x}$, $\beta_2 = c_2$, or $H_0 : \beta_1 = c_1$, $\beta_2 = c_2$, is true. The factors that cause δ to increase are as follows:

1. The magnitude of the hypothesis error. In this example the hypothesis error includes two components, $\left[(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}\right]^2$ and $(\beta_2 - c_2)^2$. The larger these specification errors the higher the probability that the null hypothesis will be rejected. The first term is related to the intercept parameter where the errors in hypotheses about both β_1 and β_2 are contributors, as well as the sample mean, $\bar{x}$. If the sample mean $\bar{x} = 0$, then only the magnitude of $(\beta_1 - c_1)^2$ matters.

2. The sample size, N. As the sample size N increases the value of δ increases not only because it multiplies the first component of δ but also because the data variation $\sum_{i=1}^{N}(x_i - \bar{x})^2$ increases, or at worst stays the same, as N increases. This is very reassuring and a reason to prefer larger samples to smaller ones. The probability of rejecting a false hypothesis approaches one as $N \to \infty$.
3. The variation in the explanatory variable. In the simple regression model the data variation $\sum_{i=1}^{N}(x_i - \bar{x})^2$ is directly related to the probability of rejecting the joint null hypothesis. The larger the data variation, the smaller the variance of b_2, and the more likely we are to detect the discrepancy between β_2 and the hypothesized value c_2.
4. The error variance σ^2. The smaller the error variance, the smaller the uncertainty in the model, and the larger δ becomes, and the higher the probability of rejecting a false joint hypothesis.

For a numerical example we use values arising from the simulation experiment used in Appendix 2H and Appendix 3B. In the first Monte Carlo sample, data file *mc1_fixed_x*, the x-values consist of $x_i = 10$, $i = 1, \ldots, 20$ and $x_i = 20$, $i = 21, \ldots, 40$. The sample mean is $\bar{x} = 15$ so that $\sum(x_i - \bar{x})^2 = 40 \times 5^2 = 1000$. Also, $\sigma^2 = 2500$. The true parameter values in the simulation experiment are $\beta_1 = 100$ and $\beta_2 = 10$. We now test the joint hypothesis $H_0: \beta_1 = 100, \beta_2 = 9$ against the alternative $H_1: \beta_1 \neq 100$ and/or $\beta_2 \neq 9$. At the 5% level of significance we reject the joint null hypothesis if the F-test statistic is greater than the critical value $F_{(0.95,\,2,\,38)} = 3.24482$. You can confirm that the calculated value of the F-statistic is 4.96, so that, at the 5% level of significance, we correctly reject $H_0: \beta_1 = 100, \beta_2 = 9$.

The noncentrality parameter is

$$\delta = N\left\{\frac{\left[(\beta_1 - c_1) + (\beta_2 - c_2)\bar{x}\right]^2}{\sigma^2}\right\} + \frac{(\beta_2 - c_2)^2\sum_{i=1}^{N}(x_i - \bar{x})^2}{\sigma^2}$$

$$= 40\left\{\frac{[(100 - 100) + (10 - 9)\,15]^2}{2500}\right\} + \frac{(10 - 9)^2 1000}{2500} = \frac{(40 \times 15^2) + 1000}{2500}$$

$$= 4$$

The probability of rejecting the joint null hypothesis is the probability that a value from a noncentral F-distribution with noncentrality parameter $\delta = 4$ will exceed $F_{(0.95,\,2,\,38)} = 3.24482$. The test power is $P\left[F_{(m_1=2,\ m_2=38,\ \delta=4)} > 3.24482\right] = 0.38738$.

As another illustration let us test the null hypothesis $H_0: \beta_2 = 9$ against $H_1: \beta_2 \neq 9$ using an F-test. The test critical value is the 95th percentile of the F-distribution, $F_{(0.95,\,1,\,38)} = 4.09817$. The calculated F-test value is 4.91 which exceeds the 5% critical value, so once again we correctly reject the null hypothesis. The noncentrality parameter of the F-distribution for this single hypothesis is the square of δ_2 in (6A.8),

$$\delta = \delta_2^2 = \frac{(\beta_2 - c_2)^2}{\sigma^2/\sum(x_i - \bar{x})^2} = \frac{1}{2500/1000} = 0.4$$

Thus the probability of rejecting the null hypothesis $H_0: \beta_2 = 9$ versus $H_1: \beta_2 \neq 9$ when the true value of $\beta_2 = 10$ is $P\left[F_{(m_1=1,\ m_2=38,\ \delta=0.4)} > 4.09817\right] = 0.09457$.

We note three lessons from this exercise. First, using an F-test, the probability of rejecting the joint hypothesis $H_0: \beta_1 = 100, \beta_2 = 9$ is greater than the probability of rejecting the single hypothesis $H_0: \beta_2 = 9$. Second, in Appendix 3B we found that the probability of rejecting

$H_0: \beta_2 = 9$ versus $H_1: \beta_2 > 9$ using a one-tail t-test was 0.15301, with noncentrality parameter 0.63246. The power of a one-tail test, when it can be appropriately used, is greater than the power of a two-tail test. Third, when using a two-tail t-test the rejection probability must be computed with care because the noncentral t-distribution is not symmetric about zero. The probability of rejecting the hypothesis is

$$P\left(t_{(38,\,0.63246)} \le -1.686\right) + \left[1 - P\left(t_{(38,\,0.63246)} \ge 1.686\right)\right] = 0.0049866 + 0.0895807 = 0.09457$$

Appendix 6B Further Results from the FWL Theorem

In Section 5.2.4, we saw that, from the FWL theorem, the least squares estimate of a coefficient of a particular explanatory variable, say x_2, can be obtained by "partialing out" the effects of the other variables on x_2 and on y, and running a regression with the partialed-out versions of y and x_2. We now consider some further results from the FWL theorem. In particular, we show how the variance of the least squares estimator can be written in terms of a simple expression that depends on x_2 and the partialed-out version of x_2.

Consider the multiple regression model with two explanatory variables, $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$. Partial-out x_3 using the Frisch–Waugh–Lovell (FWL) approach. First, the auxiliary regression of y on x_3 is $y_i = a_1 + a_3 x_{i3} + r_i$ and the least squares residual is $\ddot{y}_i = y_i - \tilde{a}_1 - \tilde{a}_3 x_{i3} = y_i - \tilde{y}_i$, where $\tilde{y}_i = \tilde{a}_1 + \tilde{a}_3 x_{i3}$ is the fitted value from the auxiliary regression. The auxiliary regression of x_2 on x_3 is $x_{i2} = c_1 + c_3 x_{i3} + r_{i2}$ and the least squares residual is $\ddot{x}_{i2} = x_{i2} - \tilde{c}_1 - \tilde{c}_3 x_{i3} = x_{i2} - \tilde{x}_{i2}$, where $\tilde{x}_{i2} = \tilde{c}_1 + \tilde{c}_3 x_{i3}$ is the fitted value from the auxiliary regression for x_2. The FWL theorem says that by estimating the model $\ddot{y}_i = \beta_2 \ddot{x}_{i2} + \ddot{e}_i$, we can obtain the same least squares estimator as from the full model. Because the partialed-out model has no explicit intercept, the least squares estimator is

$$b_2 = \sum \ddot{x}_{i2}\ddot{y}_i \Big/ \sum \ddot{x}_{i2}^2 = \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) \Big/ \sum \left(x_{i2} - \tilde{x}_{i2}\right)^2$$

Note that

- $\tilde{x}_{i2}$ is an estimate of $E\left(x_2 | x_3\right)$ and $\tilde{y}_i$ is an estimate of $E\left(y | x_3\right)$. Thus, when x_3 has been partialed out, we use the conditional means in $b_2 = \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) \Big/ \sum \left(x_{i2} - \tilde{x}_{i2}\right)^2$. When x_3 has not been partialed out we use the unconditional means. A similar statement holds for the variance.
- If we replace $\tilde{y}_i$ by $\bar{y}_i$ and replace $x_{i2} - \tilde{x}_{i2}$ by $x_i - \bar{x}_i$, we have the usual expression for the least squares estimator in the simple regression model.
- Further note that the OLS estimator b_2 in the multiple regression model depends on x_2 and y after removing the linear influence of x_3. In addition, the formula above is valid when the multiple regression model contains any number of variables, with the understanding that $\tilde{y}_i$ and $\tilde{x}_{i2}$ are fitted values from auxiliary regressions containing all explanatory variables except x_2. Very neat!

Let us take the numerator $\sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right)$ and work with it.

$$\begin{aligned} \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) &= \sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{a}_1 - \tilde{a}_3 x_{i3}\right) \\ &= \sum \left(x_{i2} - \tilde{x}_{i2}\right) y_i - \tilde{a}_1 \sum \left(x_{i2} - \tilde{x}_{i2}\right) - \tilde{a}_3 \sum \left(x_{i2} - \tilde{x}_{i2}\right) x_{i3} \end{aligned}$$

The term $\sum \left(x_{i2} - \tilde{x}_{i2}\right) = 0$ because it is the sum of least squares residuals from the auxiliary regression that includes an intercept. Also $\sum \left(x_{i2} - \tilde{x}_{i2}\right) x_{i3} = 0$ because least squares residuals are uncorrelated with model explanatory variables. See Exercises 2.1 and 2.3. Thus

$$\sum \left(x_{i2} - \tilde{x}_{i2}\right)\left(y_i - \tilde{y}_i\right) = \sum \left(x_{i2} - \tilde{x}_{i2}\right) y_i$$

The resulting simplified estimator b_2 is

$$b_2 = \sum \ddot{x}_{i2}\ddot{y}_i / \sum \ddot{x}_{i2}^2 = \sum (x_{i2} - \tilde{x}_{i2}) y_i / \sum (x_{i2} - \tilde{x}_{i2})^2$$

Computationally this is very nice because it is the estimated least squares coefficient from the model $y_i = \beta_2 \ddot{x}_{i2} + \ddot{e}_i$, where $\ddot{x}_{i2} = x_{i2} - \tilde{x}_{i2}$ is a least squares residual.

Now, as in Chapter 2, we can make theoretical progress by further work on the computational form of the least squares estimator. Substitute $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$ into the computational form and simplify.

$$\begin{aligned} b_2 &= \frac{\sum (x_{i2} - \tilde{x}_{i2}) y_i}{\sum (x_{i2} - \tilde{x}_{i2})^2} = \frac{\sum (x_{i2} - \tilde{x}_{i2})(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i)}{\sum (x_{i2} - \tilde{x}_{i2})^2} \\ &= \frac{1}{\sum (x_{i2} - \tilde{x}_{i2})^2} \left[\sum (x_{i2} - \tilde{x}_{i2})(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i) \right] \\ &= \frac{1}{\sum (x_{i2} - \tilde{x}_{i2})^2} \left[\beta_1 \sum (x_{i2} - \tilde{x}_{i2}) + \beta_2 \sum (x_{i2} - \tilde{x}_{i2}) x_{i2} + \beta_3 \sum (x_{i2} - \tilde{x}_{i2}) x_{i3} + \sum (x_{i2} - \tilde{x}_{i2}) e_i \right] \end{aligned}$$

Again $\sum (x_{i2} - \tilde{x}_{i2}) = 0$ and $\sum (x_{i2} - \tilde{x}_{i2}) x_{i3} = 0$. Now, being clever and using $\sum (x_{i2} - \tilde{x}_{i2}) = 0$, we can say

$$\sum (x_{i2} - \tilde{x}_{i2}) x_{i2} = \sum (x_{i2} - \tilde{x}_{i2}) x_{i2} - \tilde{x}_{i2} \sum (x_{i2} - \tilde{x}_{i2}) = \sum (x_{i2} - \tilde{x}_{i2})^2$$

Plugging all this in, we have

$$b_2 = \beta_2 + \frac{\sum (x_{i2} - \tilde{x}_{i2}) e_i}{\sum (x_{i2} - \tilde{x}_{i2})^2}$$

Then, if errors are homoskedastic and serially uncorrelated

$$\begin{aligned} \operatorname{var}(b_2 | \mathbf{X}) &= \operatorname{var}\left[\frac{\sum (x_{i2} - \tilde{x}_{i2}) e_i}{\sum (x_{i2} - \tilde{x}_{i2})^2} \middle| \mathbf{X} \right] = \frac{\sum (x_{i2} - \tilde{x}_{i2})^2 \operatorname{var}(e_i | \mathbf{X})}{\left[\sum (x_{i2} - \tilde{x}_{i2})^2 \right]^2} = \frac{\sum (x_{i2} - \tilde{x}_{i2})^2 \sigma^2}{\left[\sum (x_{i2} - \tilde{x}_{i2})^2 \right]^2} \\ &= \frac{\sigma^2}{\sum (x_{i2} - \tilde{x}_{i2})^2} \end{aligned}$$