Chapter 8

The Multiple Regression Model: Hypothesis Tests and the Use of Nonsample Information

- An important new development that we encounter in this chapter is using the *F*-distribution to simultaneously test a null hypothesis consisting of two or more hypotheses about the parameters in the multiple regression model.
- The theories that economists develop also 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 and the construction of "prediction" intervals.

- In this chapter we adopt assumptions MR1-MR6, including normality, listed on page 150. If the errors are not normal, then the results presented in this chapter will hold approximately if the sample is large.
- What we discover in this chapter is that a single null hypothesis that may involve one or more parameters can be tested via a *t*-test or an *F*-test. Both are equivalent. A joint null hypothesis, that involves a set of hypotheses, is tested via an *F*-test.

8.1 The *F*-Test

- The *F*-test for a set of hypotheses is based on a comparison of the sum of squared errors from the original, unrestricted multiple regression model to the sum of squared errors from a regression model in which the null hypothesis is assumed to be true.
- To illustrate what is meant by an unrestricted multiple regression model and a model that is restricted by the null hypothesis, consider the Bay Area Rapid Food hamburger chain example where weekly total revenue of the chain (tr) is a function of a price index of all products sold (p) and weekly expenditure on advertising (a).

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + e_t \qquad (8.1.1)$$

- Suppose that we wish to test the hypothesis that changes in price have no effect on total revenue against the alternative that price does have an effect.

The null and alternative hypotheses are: H_0: $\beta_2 = 0$ and H_1: $\beta_2 \neq 0$. The restricted model, that assumes the null hypothesis is true, is

$$tr_t = \beta_1 + \beta_3 a_t + e_t \tag{8.1.2}$$

Setting $\beta_2 = 0$ in the unrestricted model in Equation (8.1.1) means that the price variable P_t does not appear in the restricted model in Equation (8.1.2).

- When a null hypothesis is assumed to be true, we place conditions, or constraints, on the values that the parameters can take, and the sum of squared errors increases. Thus, the sum of squared errors from Equation (8.1.2) will be larger than that from Equation (8.1.1).

- The idea of the *F*-test is that if these sums of squared errors are substantially different, then the assumption that the null hypothesis is true has significantly reduced the ability of the model to fit the data, and thus the data do not support the null hypothesis.
- If the null hypothesis is true, we expect that the data are compatible with the conditions placed on the parameters. Thus, we expect little change in the sum of squared errors when the null hypothesis is true.
- We call the sum of squared errors in the model that assumes a null hypothesis to be true the *restricted sum of squared errors*, or SSE_R, where the subscript *R* indicates that the parameters have been restricted or constrained.
- The sum of squared errors from the original model is the *unrestricted sum of squared errors*, or SSE_U. It is *always* true that $SSE_R - SSE_U \geq 0$. Recall from Equation (6.1.7) that

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Let J be the number of hypotheses. The general F-statistic is given by

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(T-K\right)} \tag{8.1.3}$$

If the null hypothesis is true, then the statistic F has an F-distribution with J numerator degrees of freedom and $T - K$ denominator degrees of freedom.

- If the null hypothesis is not true, then the difference between SSE_R and SSE_U becomes large, implying that the constraints placed on the model by the null hypothesis have a large effect on the ability of the model to fit the data. If the difference $SSE_R - SSE_U$ is

large, the value of F tends to be large. Thus, we *reject* the null hypothesis if the value of the F-test statistic becomes too large.

- We compare 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 $T - K$ degrees of freedom. The critical value for the F-distribution is depicted in Figure 8.1. Tables of critical values for $\alpha = .01$ and $\alpha = .05$ are provided at the end of the book (Tables 3 and 4).
- For the unrestricted and restricted models in Equations (8.1.1) and (8.1.2), respectively, we find

$$SSE_U = 1805.168 \qquad SSE_R = 1964.758$$

  By imposing the null hypothesis H_0: $\beta_2 = 0$ on the model the sum of squared errors has increased from 1805.168 to 1964.758.

- There is a single hypothesis, so $J = 1$ and the F-test statistic is:

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(T-K\right)} = \frac{\left(1964.758 - 1805.168\right)/1}{1805.168/\left(52-3\right)}$$
$$= 4.332$$

- We compare this value to the critical value from an F-distribution with 1 and 49 degrees of freedom. For the $F_{(1,\,49)}$ distribution the $\alpha = .05$ critical value is $F_c = 4.038$. Since $F = 4.332 \geq F_c$ we reject the null hypothesis and conclude that price does have a significant effect on total revenue. The p-value for this test is $p = P[F_{(1,\,49)} \geq 4.332] = .0427$, which is less than $\alpha = .05$, and thus we reject the null hypothesis on this basis as well.
- The p-value can also be obtained using modern software such as EViews. See Table 8.1.
- Recall that we used a t-test to test H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$ in Chapter 7. Indeed, in Table 7.2 the p-value for this t-test is 0.0427, the same as the p-value for the F-test that we just considered.

- When testing one "equality" null hypothesis against a "not equal to" alternative hypothesis, either a *t*-test or an *F*-test can be used and the outcomes will be identical.
- The reason for this is that there 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 distribution $F_{(1,\, df)}$.
- When using a *t*-test for H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$, we found that $t = -2.081$, $t_c = 2.01$, and $p = .0427$. The *F*-value that we have calculated is $F = 4.332 = t^2$ and $F_c = (t_c)^2$. 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. There is no equivalence when using a one-tailed *t*-test since the *F*-test is not appropriate when the alternative is an inequality, ">" or "<."
- We can summarize the elements of an *F*-test as follows:

1. The null hypothesis H_0 consists of one or more (J) equality hypotheses. 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 (8.1.3).
4. If the null hypothesis is true, F has the F-distribution with J numerator degrees of freedom and $T - K$ denominator degrees of freedom. The null hypothesis is *rejected* if $F \geq F_c$, where F_c is the critical value that leaves α% of the probability in the upper tail of the F-distribution.
5. When testing a single equality 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 hypothesis using a t-test. The F-test is usually reserved for joint hypotheses.

8.1.1 The *F*-Distribution: Theory

An F random variable is formed by the ratio of two independent chi-square random variables that have been divided by their degrees of freedom.

If $V_1 \sim \chi^2_{(m_1)}$ and $V_2 \sim \chi^2_{(m_2)}$ and if V_1 and V_2 are independent, then

$$F = \frac{V_1 / m_1}{V_2 / m_2} \sim F_{(m_1,\, m_2)} \tag{8.1.4}$$

- The ***F*-distribution** is said to have m_1 *numerator degrees of freedom* and m_2 *denominator degrees of freedom*. The values of m_1 and m_2 determine the shape of the distribution, which in general looks like Figure 8.1. The range of the random variable is $(0, \infty)$ and it has a long tail to the right.
- When you take advanced courses in econometric theory, you prove that

$$V_1 = \frac{SSE_R - SSE_U}{\sigma^2} \sim \chi^2_{(J)} \tag{8.1.5}$$

$$V_2 = \frac{SSE_U}{\sigma^2} \sim \chi^2_{(T-K)} \tag{8.1.6}$$

and that V_1 and V_2 are independent. The result for V_1 requires the relevant null hypothesis to be true; that for V_2 does not. Note that σ^2 cancels when we take the ratio of V_1 to V_2, yielding

$$F = \frac{V_1/J}{V_2/(T-K)} = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)} \tag{8.1.7}$$

The Chi-square statistic given in the EViews output in Table 8.1 is equal to V_1 with σ^2 replaced by $\hat{\sigma}^2$. It is a large-sample approximation which you will learn more about in advanced courses.

Table 8.1 **EViews Output for Testing Price Coefficient**

Null Hypothesis:	C(2) = 0		
F-statistic	4.331940	Probability	0.042651
Chi-square	4.331940	Probability	0.037404

8.2 Testing the Significance of a Model

- An important application of the *F*-test is for what is called "testing the overall significance of a model." Consider again the general multiple regression model with $(K - 1)$ explanatory variables and K unknown coefficients

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_K x_{tK} + e_t \tag{8.2.1}$$

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

$$H_0: \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0 \tag{8.2.2}$$

H_1: at least one of the β_K is nonzero

- The null hypothesis has $K - 1$ parts, and it is called a joint hypothesis. It states as a conjecture that each and every one of the parameters β_K, other than the intercept parameter β_1, is 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. 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 Equation (8.2.2) is sometimes referred to as a *test of the overall significance of the regression model*.
- The unrestricted model is that given in Equation (8.2.1).
- To test the joint null hypothesis H_0: $\beta_2 = \beta_3 = \ldots = \beta_K = 0$, which actually is $K - 1$ hypotheses, we will use a test based on the *F*-distribution.

- If the joint null hypothesis H_0: $\beta_2 = 0$, $\beta_3 = 0,\ldots, \beta_K = 0$ is true, then the *restricted* model is

$$y_t = \beta_1 + e_t \tag{8.2.3}$$

- The least squares estimator of β_1 in this restricted model is $b_1^* = \dfrac{\sum y_t}{T} = \bar{y}$, which is the sample mean of the observations on the dependent variable.
- The *restricted* sum of squared errors from the hypothesis (8.2.2) is

$$SSE_R = \sum (y_t - b_1^*)^2 = \sum (y_t - \bar{y})^2 = SST$$

- *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, or $SSE_U = SSE$. The number of hypotheses is $J = K - 1$. Thus to *test the overall significance of a model* the *F*-test statistic can be modified as

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(T-K)} \tag{8.2.4}$$

- The calculated value of this test statistic is compared to a critical value from the $F_{(K\text{-}1,\ T\text{-}K)}$ distribution.
- To illustrate, we test the overall significance of the regression used to explain the Bay Area Burger's total revenue. We want to test whether the coefficients of price and of advertising expenditure are both zero, against the alternative that at least one of the coefficients is not zero. Thus, in the model

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + e_t$$

we want to test

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

against the alternative

$$H_1: \beta_2 \neq 0 \text{ or } \beta_3 \neq 0, \text{ or both nonzero.}$$

- The ingredients for this test, and the test statistic value itself, are reported in the Analysis of Variance Table reported by most regression software. The SHAZAM output for the Bay Area Rapid Food data appears in Table 8.2. From this table, we see that $SSE_R = SST = 13581$ and $SSE_U = SSE = 1805.2$.

Table 8.2 **ANOVA Table obtained using SHAZAM**

ANALYSIS OF VARIANCE - FROM MEAN				
	SS	DF	MS	F
REGRESSION	11776.	2.	5888.1	159.828
ERROR	1805.2	49.	36.840	P-VALUE
TOTAL	13581.	51.	266.30	0.000

- The values of *Mean Square* are the ratios of the Sums of Squares values to the degrees of freedom, DF.
- In turn, the ratio of the Mean Squares is the *F*-value for the test of overall significance of the model. For the Bay Area Burger data this calculation is

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(T-K)} = \frac{(13581.35 - 1805.168)/2}{1805.168/(52-3)} = \frac{5888.09}{36.84} = 159.83$$

The 5% critical value for the F statistic with (2, 49) degrees of freedom is $F_c = 3.187$. Since $159.83 > 3.187$, we reject H_0 and conclude that the estimated relationship is a significant one. Instead of looking up the critical value, we could have made our conclusion based on the *p*-value, which is calculated by most software, and is reported in Table 8.2. Our sample of data suggests that price or advertising expenditure or both have an influence on total revenue. Note that this conclusion is consistent with conclusions reached using separate *t*-tests for testing the significance of price and the significance of advertising expenditure in Chapter 7.

8.2.1 The Relationship between Joint and Individual Tests

Why use the *F*-distribution to perform a simultaneous test of H_0: $\beta_2 = 0$, $\beta_3 = 0$? Why not just use separate *t*-tests on each of the null hypotheses H_0: $\beta_2 = 0$ and H_0: $\beta_3 = 0$? The answer relates to the correlation between the least squares estimators. The *F*-test that

tests both hypotheses simultaneously makes allowance for the fact that the least squares estimators b_2 and b_3 are correlated. It is a test for whether the *pair* of values $\beta_2 = 0$ and $\beta_3 = 0$ are consistent with the data. When separate *t*-tests are performed, the possibility that $\beta_2 = 0$ is not considered when testing H_0: $\beta_3 = 0$, and vice versa. It is not a pair of values being tested with *t*-tests, but a consequence about a single parameter at a time. Each *t*-test is treated in isolation from the other, no allowance is made for the correlation between b_2 and b_3. As a consequence, the joint *F*-test at a 5% significance level is not equivalent to separate *t*-tests that each uses a 5% significance level. Conflicting results can occur. For example, it is possible for individual *t*-tests to fail to conclude that coefficients are significantly different from zero, while the *F*-test implies that the coefficients are *jointly* significant. This situation frequently arises when the data are collinear, as described in Section 8.7.

8.3 An Extended Model

We have hypothesized so far in this chapter that total revenue at the Bay Area Rapid Food franchise is explained by product price and advertising expenditures,

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + e_t \tag{8.3.1}$$

One aspect of this model that is worth questioning is whether the *linear* relationship between revenue, price, and advertising expenditure is a good approximation to reality.

- This linear model implies that increasing advertising expenditure will continue to increase total revenue at the same rate, irrespectively of the existing level of revenue and advertising expenditure. That is, the coefficients β_3, that measures the response of $E(tr)$ to a change in a, is constant.

- However, as the level of advertising expenditure increases, we would expect diminishing returns to set in. That is, the increase in revenue that results from advertising grows.
- One way of allowing for diminishing returns to advertising is to include the squared value of advertising, a^2, into the model as another explanatory variable, so

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 + e_t \tag{8.3.2}$$

Adding the term $\beta_4 a_t^2$ to our original specification yields a model in which the response of expected revenue to advertising depends on the level of advertising.

- The response of $E(tr)$ to a is

$$\frac{\Delta E(tr_t)}{\Delta a_t}_{(p \text{ held constant})} = \frac{\partial E(tr_t)}{\partial a_t} = \beta_3 + 2\beta_4 a_t \tag{8.3.3}$$

When a_t increases by one unit ($1,000), and p_t is held constant, $E(tr)$ increases by $(\beta_3 + 2\beta_4 a_t) \times \$1{,}000$.

- To determine the anticipated signs for β_3 and β_4 we note that we would expect the response of revenue to advertising to be positive when $a_t = 0$. That is, we expect that $\beta_3 > 0$. Also, to achieve diminishing returns the response must decline as a_t increases. That is, we expect $\beta_4 < 0$.
- For estimation purposes, the squared value of advertising is "just another variable." That is, we can write Equation (8.3.2) as

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + e_t \qquad (8.3.4)$$

where $y_t = tr_t$, $x_{t2} = p_t$, $x_{t3} = a_t$, and $x_{t3} = a_t^2$.

- The least squares estimates, using the data in Table 7.1, are

$$\hat{tr}_t = 104.81 - 6.582\,p_t + 2.948a_t + 0.0017a_t^2 \qquad \text{(R8.4)}$$
$$\text{(6.58)} \quad \text{(3.459)} \quad \text{(0.786)} \quad \text{(0.0361)} \quad \text{(s.e.)}$$

  What can we say about the addition of a_t^2 to the equation? The first thing to notice is that its coefficient is positive, not negative, as was expected. Second, its *t*-value for the hypothesis H_0: $\beta_4 = 0$ is $t = 0.0017/0.0361 = 0.048$. This very low value indicates that b_4 is not significantly different from zero. If β_4 is zero, there are no diminishing returns to advertising, which is counter to our belief in the phenomenon of diminishing returns. Thus, we conclude that β_4 has been estimated imprecisely and its standard error is too large.

- When economic parameters are estimated imprecisely, one solution is to obtain more and better data. Recall that the variances of the least squares estimators are reduced

by increasing the number of sample observations. Consequently, another 26 weeks of data were collected. These data have been appended to the data in Table 7.1. The ranges of p_t and a_t are wider in this data set, and greater variation in the explanatory variables leads to a reduction in the variances of the least squares estimators, and may help us achieve more precise least squares estimates. This fact, coupled with the fact that we now have a total of 78 observations, rather than 52, gives us a chance of obtaining a more precise estimate of β_4, and the other parameters as well.

- Using the new combining all the data we obtain the following least squares estimated equation

$$\hat{tr}_t = 110.46 - 10.198\,p_t + 3.361\,a_t - 0.0268\,a_t^2 \quad \text{(R8.5)}$$
$$(3.74) \quad (1.582) \qquad (0.422) \quad (0.0159) \qquad \text{(s.e.)}$$

- A comparison of the standard errors in this equation with those in Equation (R8.4) indicates that the inclusion of the additional 26 observations has greatly improved the precision of our estimates. In particular, the estimated coefficient of a_t^2 now has the expected sign. Its t-value of $t = -1.68$ implies that b_4 is significantly different from zero, using a one-tailed test and $\alpha = .05$. The 78 data points we have are compatible with the assumption of diminishing returns to advertising expenditures.

8.4 Testing Some Economic Hypotheses

Using the expanded model for Bay Area Rapid Food total revenue in Equation (8.3.2) and the $T = 78$ observations, we can test some interesting economic hypotheses and illustrate the use of *t*- and *F*-tests in economic analysis.

8.4.1 The Significance of Advertising

- Our expanded model is

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 + e_t \tag{8.4.1}$$

- How would we test whether advertising has an effect upon total revenue? If either β_3 or β_4 are not zero then advertising has an effect upon revenue.

- Based on one-tailed *t*-tests we can conclude that individually, β_3 and β_4, are not zero, and of the correct sign.
- But the question we are now asking involves both β_3 and β_4, and thus a joint test is appropriate. The joint test will use the *F*-statistic in Equation (8.1.3) to test $H_0 : \beta_3 = 0,\ \beta_4 = 0$.
- Compare the unrestricted model in Equation (8.4.1) to the restricted model, which assumes the null hypothesis is true. The restricted model is

$$tr_t = \beta_1 + \beta_2 p_t + e_t \tag{8.4.2}$$

The elements of the test are:

1. The joint null hypothesis $H_0 : \beta_3 = 0,\ \beta_4 = 0$.
2. The alternative hypothesis $H_1 : \beta_3 \neq 0$, or $\beta_4 \neq 0$, or both are nonzero.

3. The test statistic is $F = \dfrac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$ where $J = 2$, $T = 78$ and $K = 4$. SSE_U = 2592.301 is the sum of squared errors from Equation (8.4.1). SSE_R = 20907.331 is the sum of squared errors from Equation (8.4.2).
4. If the joint null hypothesis is true, then $F \sim F_{(J,T-K)}$. The critical value F_c comes from the $F_{(2,74)}$ distribution, and for the $\alpha = .05$ level of significance it is 3.120.
5. The value of the F-statistic is $F = 261.41 > F_c$ and 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, implying that advertising has a significant effect upon total revenue.

8.4.2 The Optimal Level of Advertising

Economic theory tells us that we should undertake all those actions for which the marginal benefit is greater than the marginal cost. This optimizing principle applies to

the Bay Area Rapid Food franchise as it attempts to choose the optimal level of advertising expenditure.

- From Equation (8.3.3) the marginal benefit from another unit of advertising is the increase in total revenue:

$$\left.\frac{\Delta E(tr_t)}{\Delta a_t}\right|_{(p \text{ held constant})} = \beta_3 + 2\beta_4 a_t$$

- The marginal cost of another unit of advertising is the cost of the advertising plus the cost of preparing additional products sold due to effective advertising. If we ignore the latter costs, advertising expenditures should be increased to the point where the marginal benefit of \$1 of advertising falls to \$1, or where

$$\beta_3 + 2\beta_4 a_t = 1$$

- Using the least squares estimates for β_3 and β_4 in (R8.5) we can *estimate* the optimal level of advertising from

$$3.361 + 2(-.0268)\hat{a}_t = 1$$

  Solving, we obtain $\hat{a}_t = 44.0485$, which implies that the optimal weekly advertising expenditure is \$44,048.50.
- Suppose that the franchise management, based on experience in other cities, thinks that \$44,048.50 is too high, and that the optimal level of advertising is actually about \$40,000. We can test this conjecture using either a *t*- or *F*-test.
- The null hypothesis we wish to test is H_0: $\beta_3 + 2\beta_4(40) = 1$ against the alternative that H_1: $\beta_3 + 2\beta_4(40) \neq 1$. The test statistic is

$$t = \frac{(b_3 + 80b_4) - 1}{\text{se}(b_3 + 80b_4)}$$

which has a $t_{(74)}$ distribution if the null hypothesis is true. The only tricky part of this test is calculating the denominator of the *t*-statistic. Using the properties of variance developed in Chapter 2.5.2,

$$\text{var}(b_3 + 80b_4) = \text{var}(b_3) + 80^2\text{var}(b_4) + 2(80)\text{cov}(b_3, b_4) = .76366$$

where the estimated variances and covariance are provided by your statistical software.

- Then, the calculated value of the *t*-statistic is

$$t = \frac{1.221 - 1}{\sqrt{.76366}} = .252$$

The critical value for this two-tailed test comes from the $t_{(74)}$ distribution. At the α = .05 level of significance t_c = 1.993, and thus we cannot reject the null hypothesis that the optimal level of advertising is \$40,000 per week.

- Alternatively, using an *F*-test, the test statistic is $F = \dfrac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$ where $J = 1$, $T = 78$ and $K = 4$. $SSE_U = 2592.301$ is the sum of squared errors from the full unrestricted model in Equation (8.4.1).
- SSE_R is the sum of squared errors from the restricted model in which it is assumed that the null hypothesis is true. The restricted model is

$$tr_t = \beta_1 + \beta_2 p_t + (1 - 80\beta_4)a_t + \beta_4 a_t^2 + e_t$$

- Rearranging this equation by collecting terms, to put it in a form that is convenient for estimation, we have

$$(tr_t - a_t) = \beta_1 + \beta_2 p_t + \beta_4 (a_t^2 - 80a_t) + e_t$$

- Estimating this model by least squares yields the restricted sum of squared errors SSE_R = 2594.533. The calculated value of the F-statistic is

$$F = \frac{(2594.533 - 2592.301)/1}{2592.302/74} = .0637$$

- The value F = .0637 is $t^2 = (.252)^2$, obeying the relationship between t- and F-random variables that we mentioned previously. The critical value F_c comes from the $F_{(1, 74)}$ distribution. For α = .05 the critical value is F_c = 3.970.

8.4.3 The Optimal Level of Advertising and Price

- Weekly total revenue is expected to be \$175,000 if advertising is \$40,000, and p = \$2. In the context of our model,

$$\begin{aligned} E(tr_t) &= \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 \\ &= \beta_1 + \beta_2(2) + \beta_3(40) + \beta_4(40)^2 \\ &= 175 \end{aligned}$$

- Are this conjecture *and* the conjecture that optimal advertising is \$40,000 compatible with the evidence contained in the sample of data? We now formulate the two joint hypotheses

$$H_0: \beta_3 + 2\beta_4(40) = 1, \qquad \beta_1 + 2\beta_2 + 40\beta_3 + 1600\beta_4 = 175$$

- The alternative is that at least one of these hypotheses is not true. Because there are $J = 2$ hypotheses to test jointly we will use an F-test.
- Constructing the restricted model will now require substituting both of these hypotheses into our extended model. The test statistic is $F = \dfrac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$, where $J = 2$. The computed value of the F-statistic is $F = 1.75$.
- The critical value for the test comes from the $F_{(2, 74)}$ distribution and is $F_c = 3.120$. Since $F < F_c$ we do not reject the null hypothesis, and conclude that the sample data are compatible with the hypothesis that the optimal level of advertising is \$40,000 per week and that if the price is \$2 the total revenue will be, on average, \$175,000 per week.

8.5 The Use of Nonsample Information

In many estimation and inference 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 would 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 will be 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 (p_B), the price of other liquor (p_L), the price of all other remaining goods and services (p_R), and income (m). We write this relationship as

$$q = f(p_B, p_L, p_R, m) \tag{8.5.1}$$

- We assume the log-log functional form is appropriate for this demand relationship

$$\ln(q) = \beta_1 + \beta_2 \ln(p_B) + \beta_3 \ln(p_L) + \beta_4 \ln(p_R) + \beta_5 \ln(m) \tag{8.5.2}$$

  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.

- 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."

- 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 Equation (8.5.2) by λ, yields

$$\begin{aligned} \ln(q) &= \beta_1 + \beta_2 \ln(\lambda p_B) + \beta_3 \ln(\lambda p_L) + \beta_4 \ln(\lambda p_R) + \beta_5 \ln(\lambda m) \\ &= \beta_1 + \beta_2 \ln p_B + \beta_3 \ln p_L + \beta_4 \ln p_R + \beta_5 \ln m + (\beta_2 + \beta_3 + \beta_4 + \beta_5)\ln(\lambda) \end{aligned} \qquad (8.5.3)$$

- Comparing Equation (8.5.2) with Equation (8.5.3) 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

$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0 \qquad (8.5.4)$$

- 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 Equation (8.5.4) holds, then it seems desirable to be able to obtain estimates that obey this restriction.
- To obtain estimates that obey Equation (8.5.4), begin with the multiple regression model

$$\ln(q) = \beta_1 + \beta_2\ln(\lambda p_B) + \beta_3\ln(\lambda p_L) + \beta_4\ln(\lambda p_R) + \beta_5\ln(\lambda m) + e_t \qquad (8.5.5)$$

and a sample of data consisting of thirty years of annual data on beer consumption collected from a randomly selected household.

- 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. For reasons explained shortly we solve for β_4:

$$\beta_4 = -\beta_2 - \beta_3 - \beta_5 \tag{8.5.6}$$

Substituting this expression into the original model in Equation (8.5.5) gives

$$\begin{aligned}
\ln(q_t) &= \beta_1 + \beta_2 \ln(p_{Bt}) + \beta_3 \ln(p_{Lt}) + (-\beta_2 - \beta_3 - \beta_5)\ln(p_{Rt}) + \beta_5 \ln(m_t) + e_t \\
&= \beta_1 + \beta_2 \left(\ln(p_{Bt}) - \ln(p_{Rt})\right) + \beta_3 \left(\ln(p_{Lt}) - \ln(p_{Rt})\right) \\
&\quad + \beta_5 \left(\ln(m_t) - \ln(p_{Rt})\right) + e_t \qquad (8.5.7) \\
&= \beta_1 + \beta_2 \ln\left(\frac{p_{Bt}}{p_{Rt}}\right) + \beta_3 \ln\left(\frac{p_{Lt}}{p_{Rt}}\right) + \beta_5 \ln\left(\frac{m_t}{p_{Rt}}\right) + e_t
\end{aligned}$$

- We have used the parameter restriction to eliminate the parameters β_4 and in so doing, and in using the properties of logarithms, we have constructed the new variables $\ln(P_{Bt}/P_{Rt})$, $\ln(P_{Lt}/P_{Rt})$, and $\ln(m_t/P_{Rt})$. The last line in Equation (8.5.7) is our "restricted" model.
- To get "restricted least squares estimates," we apply the least squares estimation to the restricted model

$$\ln \hat{q}_t = \underset{(3.714)}{-4.798} \underset{(0.166)}{-1.2994 \ln\left(\frac{p_{Bt}}{p_{Rt}}\right)} + \underset{(0.284)}{0.1868 \ln\left(\frac{p_{Lt}}{p_{Rt}}\right)} + \underset{(0.427)}{0.9458 \ln\left(\frac{m_t}{p_{Rt}}\right)} \quad \text{(R8.8)}$$

- Let the restricted least squares estimates in Equation (R8.8) be denoted as b_k^*. In Equation (R8.8) we have estimates of β_1, β_2, β_3 and β_5. To obtain an estimate of β_4 we use the restriction (8.5.6)

$$\begin{aligned} b_4^* &= -b_2^* - b_3^* - b_5^* \\ &= -(-1.2994) - 0.1868 - 0.9458 \\ &= 0.1668 \end{aligned}$$

- By using the restriction *within* the model, we have ensured that the estimates obey the constraint, so that $b_2^* + b_3^* + b_4^* + b_5^* = 0$.
- What are the properties of this "restricted" least squares estimation procedure? First, the restricted least squares estimator is biased, and $E(b_k^*) \neq \beta_k$, 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 and later, when constraints might be applied to the model. 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. The 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, with no constraints on the parameters.
- By incorporating the additional information with the data, we usually give up unbiasedness in return for reduced variances.

8.6 Model Specification

- 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 model assumptions MR1-MR6, on page 150, hold.

8.6.1 Omitted and Irrelevant Variables

- Even with sound economic principles and logic, it is possible that a chosen model may have important variables omitted or irrelevant variables included.

- To introduce the *omitted-variable* problem, suppose that, in a particular industry, the wage rate of employees W_t, depends on their experience E_t and their motivation M_t, such that we can write

$$W_t = \beta_1 + \beta_2 E_t + \beta_3 M_t + e_t \tag{8.6.1}$$

- However, data on motivation are not available. So, instead, we estimate the model

$$W_t = \beta_1 + \beta_2 E_t + v_t \tag{8.6.2}$$

- By estimating Equation (8.6.2) we are imposing the restriction $\beta_3 = 0$ when it is not true. The least squares estimator for β_1 and β_2 will generally be biased, although it will have lower variance. One occasion when it will not be biased is when the omitted

variable (M_t) is uncorrelated with the included variables (E_t). Uncorrelated explanatory variables are rare, however.

- The possibility of omitted-variable bias means one should take care to include all important relevant variables. It also means that, if an estimated equation has coefficients with unexpected signs, or unrealistic magnitudes, a possible cause of these strange results is the omission of an important variable.
- 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: $\beta_3 = \beta_4 = 0$. However, it is important to remember that there are two possible reasons for a test outcome that does not reject a zero null hypothesis as follows:
  1. The corresponding variables have no influence on *y* and can be excluded from the model.

2. The corresponding variables are important ones for inclusion in the model, but the data are not sufficiently good to reject H_0. That is, the data are not sufficiently rich to prove that the variables are important.

- Because the "insignificance" of a coefficient can be caused by (1) or (2), you must be cautious about the following rules that throw out variables with insignificant coefficients. You could be excluding an irrelevant variable, but you also could be inducing omitted-variable bias in the remaining coefficient estimates.
- 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*.
- To see clearly what is meant by an irrelevant variable, suppose that the correct specification is

$$W_t = \beta_1 + \beta_2 E_t + \beta_3 M_t + e_t \qquad (8.6.3)$$

but we estimate the model

$$W_t = \beta_1 + \beta_2 E_t + \beta_3 M_t + \beta_4 C_t + e_t$$

where C_t is the number of children of the t-th employee, and where, in reality, $\beta_4 = 0$. Then, C_t is an irrelevant variable. Including it does *not* make the least squares estimator biased, but it does mean the variances of b_1, b_2 and b_3 will be greater than those obtained by estimating the correct model in Equation (8.6.3). This result follows because, by the Gauss-Markov theorem, the least squares estimator of Equation (8.6.3) is the minimum-variance linear unbiased estimator of β_1, β_2, and β_3. The inflation of the variables will not occur if C_t is uncorrelated with E_t and M_t. Note, however, that

even though the number of children is unlikely to influence the wage rate, it could be correlated with experience.

8.6.1a Omitted Variable Bias: A Proof

- Suppose the true model is $y = \beta_1 + \beta_2 x + \beta_3 h + e$, but we estimate the model $y = \beta_1 + \beta_2 x + e$, omitting h from the model. Then we use the estimator

$$b_2^* = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2} = \frac{\sum (x_t - \bar{x}) y_t}{\sum (x_t - \bar{x})^2} \quad \text{(Note: } \sum (x_t - \bar{x}) = \sum (x_t - \bar{x})\bar{y} = 0\text{)}$$

$$= \beta_2 + \beta_3 \sum w_t h_t + \sum w_t e_t$$

where

$$w_t = \frac{x_t - \bar{x}}{\sum (x_t - \bar{x})^2}$$

So,

$$E(b_2^*) = \beta_2 + \beta_3 \sum w_t h_t \neq \beta_2$$

Taking a closer look, we find that

$$\begin{aligned} \sum w_t h_t &= \frac{\sum (x_t - \bar{x}) h_t}{\sum (x_t - \bar{x})^2} = \frac{\sum (x_t - \bar{x})(h_t - \bar{h})}{\sum (x_t - \bar{x})^2} \\ &= \frac{\sum (x_t - \bar{x})(h_t - \bar{h}) / (T-1)}{\sum (x_t - \bar{x})^2 / (T-1)} = \frac{\hat{\text{cov}}(x_t, h_t)}{\hat{\text{var}}(x_t)} \end{aligned}$$

Consequently,

$$E(b_2^*) = \beta_2 + \beta_3 \frac{\hat{\text{cov}}(x_t, h_t)}{\hat{\text{var}}(x_t)} \neq \beta_2$$

- Knowing the sign of β_3 and the sign of the covariance between x_t and h_t tells us the direction of the bias. Also, while omitting a variable from the regression usually biases the least squares estimator, if the sample covariance, or sample correlation, between x_t and the omitted variable h_t is zero, then the least squares estimator in the misspecified model is still unbiased.

8.6.2 Testing for Model Misspecification: The RESET Test

- The RESET test (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

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + e_t \tag{8.6.4}$$

Let (b_1, b_2, b_3) be the least squares estimates and the predicted values of the y_t be

$$\hat{y}_t = b_1 + b_2 x_{t2} + b_3 x_{t3} \tag{8.6.5}$$

- Then consider the following two artificial models

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \gamma_1 \hat{y}_t^2 + e_t \tag{8.6.6}$$

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \gamma_1 \hat{y}_t^2 + \gamma_2 y_t^3 + e_t \tag{8.6.7}$$

- In Equation (8.6.6) a test for misspecification is a test of H_0: $\gamma_1 = 0$ against the alternative H_1: $\gamma_1 \neq 0$. In Equation (8.6.7), testing H_0: $\gamma_1 = \gamma_2 = 0$ against H_1: $\gamma_1 \neq 0$ or $\gamma_2 \neq 0$ is a test for misspecification. Rejection of H_0 implies the original model is inadequate and can be improved. A failure to reject H_0 says the test has not been able to detect any misspecification.
- The idea behind the test is a general one. Note that $\hat{y}_t^2$ and $\hat{y}_t^3$ will be polynomial functions of x_{t2} and x_{t3}. Thus, if the original model is not the correct functional form, the polynomial approximation that includes $\hat{y}_t^2$ and $\hat{y}_t^3$ may significantly improve the fit of the model and 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_{t2} and x_{t3}, then some of their effect may be picked up by including the term $\hat{y}_t^2$ and/or $\hat{y}_t^3$. 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.

- As an example of the test, consider the beer demand example used in Section 8.5 to illustrate the inclusion of non-sample information. The log-log model that we specified earlier is

$$\ln(q_t) = \beta_1 + \beta_2 \ln(p_{Bt}) + \beta_3 \ln(p_{Lt}) + \beta_4 \ln(p_{Rt}) + \beta_5 \ln(m_t) + e_t \qquad (8.6.8)$$

Estimating this model, and then augmenting it with the squares of the predictions, and the squares and cubes of the predictions, yields the RESET test results in the top half of Table 8.4. The *F*-values are quite small and their corresponding *p*-values of 0.93 and 0.70 are well above the conventional significance level of 0.05. There is no evidence from the RESET test to suggest the log-log model is inadequate.

Table 8.4 RESET Test Results for Beer Demand Example

Ramsey RESET Test: LOGLOG Model			
F-statistic (1 term)	0.0075	Probability	0.9319
F-statistic (2 terms)	0.3581	Probability	0.7028
Ramsey RESET Test: LINEAR Model			
F-statistic (1 term)	8.8377	Probability	0.0066
F-statistic (2 terms)	4.7618	Probability	0.0186

- Now, suppose that we had specified a linear model instead of a log-log model as follows:

$$q_t = \beta_1 + \beta_2 p_{Bt} + \beta_3 p_{Lt} + \beta_4 p_{Rt} + \beta_5 m_t + e_t \qquad (8.6.9)$$

- Augmenting this model with the squares and then the squares and cubes of the predictions $\hat{q}_t$ yields the RESET test results in the bottom half of Table 8.4. The *p*-values of 0.0066 and 0.0186 are below 0.05 suggesting the linear model is inadequate.

8.7 Collinear Economic Variables

- 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**, or **multicollinearity** when several variables are involved. In this case there is no 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 Bay Area Rapid Food marketing executives when trying to estimate the increase in the total revenue attributable to advertising that appears in newspaper and the increase in total revenue attributable to coupon advertising. Suppose it has been common practice to coordinate these two advertising devices, so that at the same time advertising appears in the newspapers, and 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 total revenue like Equation (7.1.2), 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. Because the two types of advertising expenditure move together, it may be difficult to sort out their separate effects on total 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 amounts of two, or more, such inputs reflect proportionate increases. Proportionate 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.
- We should also note at this point that 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 in an economic or statistical model. A related problem exists when the values of an explanatory variable do not vary or change much within the sample of data. When an explanatory variable exhibits little variation, then it is difficult to isolate its impact. In Chapter 7.3.1, we noted that the more variation in an explanatory variable, the more precisely its coefficient can be estimated. Lack of variation leads to estimator impression. This problem also falls within the context of "collinearity."

8.7.1 The Statistical Consequences of Collinearity

- The consequences of collinear relationships among explanatory variables in an econometric model may be summarized as follows:

  1. Whenever there are one or more *exact* linear relationships among the explanatory variables, then *the condition of exact collinearity, or exact multicollinearity, exists*.

In this case the least squares estimator is not defined. We *cannot* obtain estimates of the β_k's using the least squares principle. This is indicated in Equation (7.3.1). If there is an exact linear relationship between x_{t2} and x_{t3}, for example, then the correlation between them is $r_{23} = \pm 1$, and the variance of b_2 is undefined, since 0 appears in the denominator. The same is true of the covariance and the formulas for b_2 and b_3. Recall that Equation (7.3.1) is

$$\text{var}(b_2) = \frac{\sigma^2}{\sum (x_{t2} - \bar{x}_2)(1 - r_{23}^2)} \tag{7.3.1}$$

and the formula of the covariance between b_2 and b_3 is

$$\text{cov}(b_2, b_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{(x_{t2} - \bar{x}_2)^2}\sqrt{(x_{t3} - \bar{x}_3)^2}}$$

2. When *nearly* exact linear dependencies among the explanatory variables exist, some of the variances, standard errors and covariances of the least squares estimators may be large. We have noted the effect on estimator variance of a high correlation between two explanatory variables in Chapter 7.3.1. Large standard errors for the least squares estimators imply high sampling variability, estimated coefficients that are unstable to small changes in the sample or model specification, interval estimates that are wide, and relatively imprecise information provided by the sample data about the unknown parameters.
3. 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 economic theory may indicate their importance in the relationship.

4. Estimates may be very sensitive to the addition or deletion of a few observations, or the deletion of an apparently insignificant variable.
5. 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 new (future) 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.

8.7.2 Identifying and Mitigating Collinearity

- One simple way to detect collinear relationships is to use sample correlation coefficients between pairs of explanatory variables. A rule of thumb is that a correlation coefficient between two explanatory variables greater than 0.8 or 0.9

indicates a strong linear association and a potentially harmful collinear relationship. The problem with examining only pairwise correlations is that the collinearity relationships may involve more than two of the explanatory variables, which may or may not be detected by examining pairwise correlations.

- A second simple and effective procedure for identifying the presence of collinearity is to estimate so-called "auxiliary regressions." In these least squares regressions the left-hand-side variable is one of the *explanatory* variables, and the right-hand-side variables are all the remaining explanatory variables. For example, the auxiliary regression for x_{t2} is

$$x_{t2} = a_1 x_{t1} + a_3 x_{t3} + \ldots + a_K x_{tK} + error$$

If the R^2 from this artificial model is high, above .80, the implication is that a large portion of the variation in x_{t2} is explained by variation in the other explanatory variables. In Chapter 7.3.1 we made the point that it is variation in a variable that is

not associated with any other explanatory variable that is valuable in improving the precision of the least squares estimator b_2. If the R^2 from the auxiliary regression is not high, then the variation in x_{t2} is not explained by the other explanatory variables, and the estimator b_2's precision is not affected by this problem.

- 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.

- We may add structure to the problem by introducing, as we did in Section 8.5, *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.

8.8 Prediction

The prediction problem for a linear statistical model with one explanatory variable was covered in depth in Chapter 5. The results in that chapter extend naturally to the more general model that has more than one explanatory variable. Let us summarize these results.

- Consider a linear statistical model with an intercept term and two explanatory variables, x_2 and x_3. That is,

$$y_t = \beta_1 + x_{t2}\beta_2 + x_{t3}\beta_3 + e_t \qquad (8.8.1)$$

where the e_t are uncorrelated random variables with mean 0 and variance σ^2. That is, $e_t \sim N(0, \sigma^2)$.

- Given a set of values for the explanatory variables, (1 x_{02} x_{03}), the prediction problem is to predict the value of the dependent variable y_0, which is given by

$$y_0 = \beta_1 + x_{02}\beta_2 + x_{03}\beta_3 + e_0 \tag{8.8.2}$$

- In this prediction problem we are assuming that the parameter values determining y_0 are the same as those in Equation (8.8.1) describing the original sample of data. Furthermore, the random error e_0 we assume to be uncorrelated with each of the sample errors e_t and to have the same mean, 0, and variance, σ^2.
- Under these assumptions, the best linear unbiased predictor of y_0 is given by

$$\hat{y}_0 = b_1 + x_{02}b_2 + x_{03}b_3 \tag{8.8.3}$$

where the b_k's are the least squares estimators.

- This predictor is unbiased in the sense that the average value of the forecast error is zero. That is, if $f = (y_0 - \hat{y}_0)$ is the forecast error then $E(f) = 0$. The predictor is best in that for any other linear and unbiased predictor of y_0, the variance of the forecast error is larger than $\text{var}(f) = \text{var}(y_0 - \hat{y}_0)$.
- The variance of forecast error $(y_0 - \hat{y}_0)$ contains two components. One component arises because b_1, b_2, and b_3 are the estimates of the true parameters. The other component occurs because e_0 is random. An expression for $\text{var}(y_0 - \hat{y}_0)$ is obtained by computing

$$\begin{aligned}
\text{var}(f) &= \text{var}[(\beta_1 + \beta_2 x_{02} + \beta_3 x_{03} + e_0) - (b_1 + b_2 x_{02} + b_3 x_{03})] \\
&= \text{var}(e_0 - b_1 - b_2 x_{02} - b_3 x_{03}) \\
&= \text{var}(e_0) + \text{var}(b_1) + x_{02}^2 \text{var}(b_2) + x_{03}^2 \text{var}(b_3) \\
&\quad + 2x_{02}\text{cov}(b_1, b_2) + 2x_{03}\text{cov}(b_1, b_3) + 2x_{02}x_{03}\text{cov}(b_2, b_3)
\end{aligned} \tag{8.8.4}$$

- To obtain var(f) we have used the facts that the unknown parameters and the values of the explanatory variables are constants, and that e_0 is uncorrelated with the sample data, and thus is uncorrelated with the least squares estimators b_k. Then var(e_0) = σ^2 and the remaining variances and covariances of the least squares estimators are obtained using the rule for calculating the variance of a weighted sum in Equation (2.5.8). Each of these terms involves σ^2 which we replace with its estimator $\hat{\sigma}^2$ to obtain the estimated variance of the forecast error $\hat{\text{var}}(f)$. The square root of this quantity is the standard error of the forecast, $\text{se}(f) = \sqrt{\hat{\text{var}}(f)}$.
- If the random errors e_t and e_0 are normally distributed, or if the sample is large, then

$$\frac{f}{\text{se}(f)} = \frac{y_0 - \hat{y}_0}{\sqrt{\hat{\text{var}}\left(y_0 - \hat{y}_0\right)}} \sim t_{(T-K)} \tag{8.8.5}$$

- Consequently, a 100(1−α)% interval predictor for y_0 is $\hat{y}_0 \pm t_c \text{se}(f)$, where t_c is a critical value from the $t_{(T-K)}$ distribution.
- Thus, we have shown that the methods for prediction in the model with $K = 3$ are straightforward extensions of the results from the simple linear regression model. If $K > 3$, the methods extend similarly.

Exercise

8.3	8.5	8.7	8.10	8.11
8.13	8.14			