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 this chapter we adopt assumptions MR1-MR6, including normality.
- 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.
- 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

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

- 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

- The sum of squared errors from the original model is the *unrestricted sum of squared errors*, or SSE_U. It is *always* true that $SSE_R - SSE_U \geq 0$.
- 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.
- 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.
- For the unrestricted and restricted models in equations (8.1.1) and (8.1.2), respectively, we find

$$SSE_U = 1805.168 \quad SSE_R = 1964.758$$

- The F-test statistic is:

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

- 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.

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

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

- 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 $F_{(1,df)}$.

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

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

- 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 courses in econometric theory, you prove that

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

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

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

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

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

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

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

$$\begin{aligned} &H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0 \\ &H_1 : \textit{at least one} \text{ of the } \beta_k \text{ is nonzero} \end{aligned} \tag{8.2.2}$$

- The null hypothesis has K–1 parts, and it is called a joint hypothesis.
- 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 (8.4.2) is sometimes referred to as a *test of the overall significance of the regression model*.
- 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

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

- 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

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

- 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

8.3 An Extended Model

We have hypothesized

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + e_t \qquad (8.3.1)$$

- As the level of advertising expenditure increases, we would expect diminishing returns to set in.

- 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

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 + e_t \qquad (8.3.2)$$

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

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

- We expect that $\beta_3 > 0$. Also, to achieve diminishing returns the response must decline as a_t increases. That is, we expect $\beta_4 < 0$.
- The least squares estimates, using the data in Table 7.1, are
-

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

- Another 26 weeks of data were collected.

- Combining all the data we obtain the following least squares estimated equation

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

- 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 $\alpha = .05$.

8.4 Testing Some Economic Hypotheses

8.4.1 The Significance of Advertising

- Our expanded model is

$$tr_t = \beta_1 + \beta_2 p_t + \beta_3 a_t + \beta_4 a_t^2 + e_t \qquad (8.4.1)$$

- 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.
- The joint test will use the *F*-statistic in (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 (8.4.1). SSE_R = 20907.331 is the sum of squared errors from (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 α = .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

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

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

- 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.

- 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

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

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

- 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 (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

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

- Rearranging 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 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,

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

- We now formulate the two joint hypotheses

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

- Because there are J=2 hypotheses to test jointly we will use an F-test.

- 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

8.5 The Use of Nonsample Information

- 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

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

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

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

- A relevant piece of nonsample information 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 λ

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

- 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

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

- 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.

$$\beta_4 = -\beta_2 - \beta_3 - \beta_5 \qquad (8.5.6)$$

- Substituting this expression into the original model gives

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

- To get "restricted least squares estimates," we apply the least squares estimation to the restricted model

$$\begin{aligned}\ln \hat{q}_t = \; & -4.798 - 1.2994 \ln\left(\frac{p_{Bt}}{p_{Rt}}\right) + 0.1868 \ln\left(\frac{p_{Lt}}{p_{Rt}}\right) + 0.9458 \ln\left(\frac{m_t}{p_{Rt}}\right) \qquad (R8.8) \\ & (3.714)\;\; (0.166) \qquad\qquad (0.284) \qquad\qquad (0.427)\end{aligned}$$

- Let the restricted least squares estimates in equation 8.7.8 be denoted as b_k^*.

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

- The restricted least squares ***estimator*** is biased, and $E\left(b_k^*\right) \neq \beta_k$, <u>unless</u> the constraints we impose are *exactly* true.
- 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 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?

8.6.1 Omitted and Irrelevant Variables

- Suppose that, in a particular industry, the wage rate of employees W_t, depends on their experience E_t and their motivation M_t,

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

- 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 (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.
- 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 it may inflate the variances of your estimates because of the presence of *irrelevant variables*.
- 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 (8.6.3).

8.6.1*a* 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

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

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

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

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

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

- 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_2 and the sign of the covariance between x_t and h_t tells us the direction of the bias.
- 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 the predicted values of the y_t be

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

- 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 (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 (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.
- 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.

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

- Estimating this model, and then augmenting it with squares of the predictions, and squares and cubes of the predictions, yields the RESET test results in the top half of Table 8.5.

- 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 8.5 **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.

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

- 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 8.5. 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.
- Consider the problem faced by Bay Area Rapid Food. Suppose it has been common practice to coordinate these two advertising devices, so that at the same time

advertising appears in the newspapers there are flyers distributed containing coupons for price reductions on hamburgers.

- 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.
- Consider a production relationship. There are certain factors of production (inputs), such as labor and capital, that are *used in relatively fixed proportions*.
- Proportionate relationships between variables are the very sort of systematic relationships that epitomize "collinearity."
- 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.

8.7.1 The Statistical Consequences of Collinearity

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.
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.
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.
4. Estimators 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.

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.
- 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

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

- 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.
- 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.
- We may add structure to the problem by introducing *nonsample* information in the form of restrictions on the parameters.

8.8 Prediction

- Consider

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

where the e_t are uncorrelated random variables with mean 0 and variance σ^2.

- Given a set of values for the explanatory variables, $\begin{pmatrix} 1 & x_{02} & x_{03} \end{pmatrix}$, the prediction problem is to predict the value of the dependent variable y_0, which is given by

$$y_0 = \beta_1 + x_{02}\beta_2 + x_{03}\beta_3 + e_0 \qquad (8.8.2)$$

- 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

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

- 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 $\text{var}(f) = \text{var}\left(y_0 - \hat{y}_0\right)$.

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

- 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

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

- 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 $t_{(T-K)}$ distribution.