Chapter 6
Further Inference in the Multiple Regression Model

Walter R. Paczkowski
Rutgers University

Chapter Contents

- Economists develop and evaluate theories about economic behavior
  - Hypothesis testing procedures are used to test these theories
  - 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**

6.1 Joint Hypothesis Testing

- A null hypothesis with multiple conjectures, expressed with more than one equal sign, is called a **joint hypothesis**

1. Example: Should a group of explanatory variables should be included in a particular model?
2. Example: Does the quantity demanded of a product depend on the prices of substitute goods, or only on its own price?

6.1
Testing Joint
Hypotheses

- Both examples are of the form:

Eq. 6.1

$$H_0 : \beta_4 = 0, \beta_5 = 0, \beta_6 = 0$$

  - The joint null hypothesis in Eq. 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

6.1.1
Testing the Effect of Advertising: The *F*-Test

- Consider the model:

Eq. 6.2

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

  – Test whether or not advertising has an effect on sales – but advertising is in the model as two variables

- 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
- The null 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 Eq. 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, β3 = 0 and $\beta_4 = 0$, and *ADVERT* and $ADVERT^2$ drop out of the model

Eq. 6.3

$$SALES = \beta_1 + \beta_2 PRICE + e$$

  – 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 least squares residuals) from the unrestricted model in Eq. 6.2 and the restricted model in Eq. 6.3

  – Shorthand notation for these two quantities is SSE_U and SSE_R, respectively

- The *F*-statistic determines what constitutes a large reduction or a small reduction in the sum of squared errors

Eq. 6.4

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

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

- **If the null hypothesis is true**, then the statistic F has what is called an F-distribution with J numerator degrees of freedom and $N - K$ denominator degrees of freedom
- **If the null hypothesis is not true**, then the difference between SSE_R and SSE_U becomes large
  - The restrictions placed on the model by the null hypothesis significantly reduce the ability of the model to fit the data

- The *F*-test for our sales problem is:
  1. Specify the null and alternative hypotheses:
     - The joint null hypothesis is $H_0 : \beta_3 = 0, \beta_4 = 0$. The alternative hypothesis is $H_0 : \beta_3 \neq 0$ or $\beta_4 \neq 0$ 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 $N = 75$:

$$F = \frac{\left(SSE_R - SSE_U\right)/2}{SSE_U/(75-4)}$$

- The *F*-test for our sales problem is (Continued):
  3. Set the significance level and determine the rejection region
  4. Calculate the sample value of the test statistic and, if desired, the *p*-value

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

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

- The *F*-test for our sales problem is (Continued):
  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

6.1.2
Testing the Significance of the Model

- Consider again the general multiple regression model with (K - 1) explanatory variables and K unknown coefficients

Eq. 6.5

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K + e$$

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

Eq. 6.6

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

- Since we are testing whether or not we have a viable explanatory model, the test for Eq. 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 Eq. 6.6

- The unrestricted model is that given in Eq. 6.5
  - The restricted model, assuming the null hypothesis is true, becomes:

Eq. 6.7

$$y_i = \beta_1 + e_i$$

- The least squares estimator of β_1 in this restricted model is:

$$b_1^* = \sum_{i=1}^{N} y_i \Big/ N = \bar{y}$$

  - The restricted sum of squared errors from the hypothesis Eq. 6.6 is:

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

- Thus, to test the overall significance of a model, *but not in general*, the *F*-test statistic can be modified and written as:

Eq. 6.8

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

- For our problem, note:
  1. We are testing:

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

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

  2. If H_0 is true:

$$F = \frac{(SST - SSE)/(4-1)}{SSE/(75-4)} \sim F_{(3,71)}$$

- For our problem, note (Continued):

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:

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

   - p-value $= P(F \geq 24.459) = 0.0000$

- For our problem, note (Continued):
  5. Since 24.459 > 2.734, we reject H_0 and conclude that the estimated relationship is a significant one
     - Note that this conclusion is consistent with conclusions that would be reached using separate *t*-tests for the significance of each of the coefficients

- We used the *F-test* to test whether $\beta_3 = 0$ and $\beta_4 = 0$ in:

Eq. 6.9

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

  – Suppose we want to test if *PRICE* affects *SALES*

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Eq. 6.10

$$SALES = \beta_1 + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$$

- The *F*-value for the restricted model is:

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{\left(2683.411 - 1532.084\right)/1}{1532.084/\left(75-4\right)} = 53.355$$

  - The 5% critical value is $F_c = F_{(0.95, 1, 71)} = 3.976$
  - We reject $H_0 : \beta_2 = 0$

- Using the *t*-test:

$$\widehat{SALES} = 109.72 - 7.640 PRICE + 12.151 ADVERT - 2.768 ADVERT^2$$

$$\text{(se)} \quad (6.80) \quad (1.046) \quad (3.556) \quad (0.941)$$

  – The *t*-value for testing H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$ is $t = 7.640/1.045939 = 7.30444$

  – Its square is $t = (7.30444)^2 = 53.355$, identical to the *F*-value

- The elements of an *F*-test
  1. The null hypothesis H_0 consists of one or more equality restrictions on the model parameters β_k
  2. The alternative hypothesis states that one or more of the equalities in the null hypothesis is not true
  3. The test statistic is the *F*-statistic in (6.4)
  4. If the null hypothesis is true, *F* has the *F*-distribution with *J* numerator degrees of freedom and *N* - *K* denominator degrees of freedom
  5. When testing a single equality null hypothesis, it is perfectly correct to use either the *t*- or *F*-test procedure: they are equivalent

- 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 ($\leq K$) involving linear hypotheses with equal signs can be tested

- Consider the issue of testing:

Eq. 6.11

$$\beta_3 + 2\beta_4 ADVERT_0 = 1$$

  - If $ADVERT_0$ = \$1,900 per month, then:

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

or

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

- Note that when H_0 is true, $\beta_3 = 1 - 3.8\beta_4$ so that:

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

or

Eq. 6.12

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

- 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 an advertising expenditure of $1,900 per month is optimal is compatible with the data

- The *t*-value is $t = 0.9676$
  - $F = 0.9362$ is equal to $t^2 = (0.9676)^2$
  - The *p*-values are identical:

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

- Suppose we have:

Eq. 6.13

$$H_0: \beta_3 + 3.8\beta_4 \leq 1 \qquad H_1: \beta_3 + 3.8\beta_4 > 1$$

- In this case, we can no longer use the *F*-test
  - Because $F = t^2$, the *F*-test cannot distinguish between the left and right tails as is needed for a one-tail test
  - We restrict ourselves to the *t*-distribution when considering alternative hypotheses that have inequality signs such as < or >

- Most software packages have commands that will automatically compute *t*- and *F*-values and their corresponding *p*-values when provided with a null hypothesis
  - These tests belong to a class of tests called **Wald tests**

- Suppose we conjecture that:

$$E(SALES) = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2$$
$$= \beta_1 + 6\beta_2 + 1.9\beta_3 + 1.9^2\beta_4$$
$$= 80$$

– We formulate the joint null hypothesis:

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

– Because there are $J = 2$ restrictions to test jointly, we use an F-test

  - A t-test is not suitable

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

- Consider the log-log functional form for a demand model for beer:

Eq. 6.14

$$\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I)$$

  – 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, say λ:

Eq. 6.15

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

- To have no change in ln(Q) when all prices and income go up by the same proportion, it must be true that:

Eq. 6.16

$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$$

  - We call such a restriction **nonsample information**

- To estimate a model, start with:

Eq. 6.17

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

  - Solve the restriction for one of the parameters, say β_4:

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

- Substituting gives:

Eq. 6.18

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

- To get least squares estimates that satisfy the parameter restriction, called **restricted least squares estimates**, we apply the least squares estimation procedure directly to the restricted model:

Eq. 6.19

$$\widehat{\ln}(Q) = -4.798 - 1.2994 \ln\left(\frac{PB}{PR}\right) + 0.1868 \ln\left(\frac{PL}{PR}\right) + 0.9458 \ln\left(\frac{I}{PR}\right)$$

$$(\text{se}) \qquad (0.166) \qquad (0.284) \qquad (0.427)$$

- Let the restricted least squares estimates in Eq. 6.19 be denoted by b^*_1, b^*_2, b^*_3, and b^*_5
  - To obtain an estimate for β_4, we use the restriction:

$$b_4^* = -b_2^* - b_3^* - b_5^* = -(-1.2994) - 0.1868 - 0.9458 = 0.1668$$

  - By using the restriction *within* the model, we have ensured that the estimates obey the constraint:

$$\mathrm{b}_2^* + \mathrm{b}_3^* + \mathrm{b}_4^* + \mathrm{b}_4^* = 0$$

- Properties of this restricted least squares estimation procedure:
  1. The restricted least squares estimator is biased, unless the constraints we impose are exactly true
  2. 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

6.3
Model Specification

- In any econometric investigation, choice of the model is one of the first steps
  - 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?

- It is possible that a chosen model may have important variables omitted
  - 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

- Consider the model:

Eq. 6.20

$$\begin{aligned}\widehat{FAMINC} &= -5534 + 3132 HEDU + 4523 WEDU \\ (\text{se}) \quad & (11230) \quad (803) \quad (1066) \\ (p\text{-value}) \quad & (0.622) \quad (0.000) \quad (0.000)\end{aligned}$$

- If we incorrectly omit wife's education:

Eq. 6.21

$$\widehat{FAMINC} = -26191 + 5155HEDU$$
$$(\text{se}) \qquad (8541) \quad (658)$$
$$(p\text{-value}) \quad (0.002)(0.000)$$

- Relative to Eq. 6.20, omitting *WEDU* leads us to overstate the effect of an extra year of education for the husband by about $2,000
  - Omission of a relevant variable (defined as one whose coefficient is nonzero) leads to an estimator that is biased
  - This bias is known as **omitted-variable bias**

- Write a general model as:

Eq. 6.22

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

- Omitting x_3 is equivalent to imposing the restriction $\beta_3 = 0$
  - It can be viewed as an example of imposing an incorrect constraint on the parameters

- The bias is:

Eq. 6.23

$$\text{bias}\left(b_2^*\right) = E\left(b_2^*\right) - \beta_2 = \beta_3 \frac{\widehat{\text{cov}}\left(x_2, x_3\right)}{\widehat{\text{var}}\left(x_2\right)}$$

Table 6.1 Correlation Matrix for Variables Used in Family Income Example

	FAMINC	*HEDU*	*WEDU*	*KL6*	X_5	X_6
FAMINC	1.000					
HEDU	0.355	1.000				
WEDU	0.362	0.594	1.000			
KL6	–0.072	0.105	0.129	1.000		
X_5	0.290	0.836	0.518	0.149	1.000	
X_6	0.351	0.821	0.799	0.160	0.900	1.000

- Note that:
  1. $\beta_3 > 0$ because husband's education has a positive effect on family income.
  2. $\widehat{\text{cov}}(x_2, x_3)$ because husband's and wife's levels of education are positively correlated.
  – Thus, the bias is positive

- Now consider the model:

Eq. 6.24

$$\widehat{FAMINC} = -7755 + 3212HEDU + 4777WEDU - 14311KL6$$
$$\begin{array}{lcccc} (se) & (11163) & (797) & (1061) & (5004) \\ (p\text{-value}) & (0.488) & (0.000) & (0.000) & (0.004) \end{array}$$

  - Notice that the coefficient estimates for *HEDU* and *WEDU* have not changed a great deal
    - This outcome occurs because *KL6* is not highly correlated with the education variables

- You to think that a good strategy is to include as many variables as possible in your model.
  - Doing so will not only complicate your model unnecessarily, but may also inflate the variances of your estimates because of the presence of **irrelevant variables**.

- You to think that a good strategy is to include as many variables as possible in your model.
  - Doing so will not only complicate your model unnecessarily, but may also inflate the variances of your estimates because of the presence of **irrelevant variables**.

- Consider the model:

$$\widehat{FAMINC} = -7759 + 3340HEDU + 5869WEDU - 14200KL6 + 889X_5 - 1067X_6$$

$$\begin{array}{lcccccc} (se) & (11195) & (1250) & (2278) & (5044) & (2242) & (1982) \\ (p\text{-value}) & (0.500) & (0.008) & (0.010) & (0.005) & (0.692) & (0.591) \end{array}$$

  – The inclusion of irrelevant variables has reduced the precision of the estimated coefficients for other variables in the equation

- Some points for choosing a model:
  1. Choose variables and a functional form on the basis of your theoretical and general understanding of the relationship
  2. 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
  3. One method for assessing whether a variable or a group of variables should be included in an equation is to perform significance tests

- Some points for choosing a model (Continued):
  4. Consider various model selection criteria
  5. The adequacy of a model can be tested using a general specification test known as RESET

- There are three main model selection criteria:
  1. R^2
  2. *AIC*
  3. *SC* (*BIC*)

- A common feature of the criteria we describe is that they are suitable only for comparing models with the same dependent variable, not models with different dependent variables like y and ln(y)

- The problem is that R^2 can be made large by adding more and more variables, even if the variables added have no justification
  - Algebraically, it is a fact that 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 $R^2 = 1$

6.3.4a
The Adjusted Coefficient of Determination

Eq. 6.25

- An alternative measure of goodness of fit called the adjusted-R^2, denoted as $\bar{R}^2$:

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

- The **Akaike information criterion (*AIC*)** is given by:

Eq. 6.26

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

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

Eq. 6.27

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

Table 6.2 Goodness-of-Fit and Information Criteria for Family Income Example

Included Variables	R^2	$\bar{R}^2$	AIC	SC
$HEDU$	0.1258	0.1237	21.262	21.281
$HEDU, WEDU$	0.1613	0.1574	21.225	21.253
$HEDU, WEDU, KL6$	0.1771	0.1714	21.211	21.248
$HEDU, WEDU, KL6, X5, X6$	0.1778	0.1681	21.219	21.276

- A model could be misspecified if:
  - we have omitted important variables
  - included irrelevant ones
  - chosen a wrong functional form
  - have a model that violates the assumptions of the multiple regression model

- RESET (**RE**gression **S**pecification **E**rror **T**est) is designed to detect omitted variables and incorrect functional form

- Suppose we have the model:

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

  – Let the predicted values of *y* be:

Eq. 6.28

$$\hat{y} = b_1 + b_2 x_2 + \mathrm{b}_3 x_3$$

- Now consider the following two artificial models:

Eq. 6.29
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + e$$

Eq. 6.30
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_1 \hat{y}^{3s} + e$$

- In Eq. 6.29 a test for misspecification is a test of $H_0: \gamma_1 = 0$ against the alternative $H_1: \gamma_1 \neq 0$
- In Eq. 6.30, 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

- Applying RESET to our problem (Eq. 6.24), we get:

$$H_0 : \gamma_1 = 0 \qquad F = 5.984 \qquad p\text{-}value = 0.015$$

$$H_0 : \gamma_1 = \gamma_2 = 0 \quad F = 3.123 \qquad p\text{-}value = 0.045$$

  – In both cases the null hypothesis of no misspecification is rejected at a 5% significance level

6.4

Poor Data Quality, Collinearity, and Insignificance

- 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**

- Consider the model:

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

  – The variance of the least squares estimator for β_2 is:

Eq. 6.31

$$\text{var}(b_2) = \frac{\sigma^2}{\left(1 - r_{23}^2\right)\sum_{i=1}^{n}(x_2 - \bar{x}_2)^2}$$

- Exact or extreme collinearity exists when x_2 and x_3 are perfectly correlated, in which case $r_{23} = 1$ and var(b_2) goes to infinity
  - Similarly, if x_2 exhibits no variation $\sum(x_2 - \bar{x}_2)^2$ equals zero and var(b_2) again goes to infinity
    - In this case x_2 is collinear with the constant term

- In general, whenever 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

- The effects of this imprecise information are:
  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
  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. Accurate forecasts may still be possible if the nature of the collinear relationship remains the same within the out-of-sample observations

- A regression of *MPG* on *CYL* yields:

$$\begin{array}{lll} \widehat{MPG} & = & 42.9 - 3.558CYL \\ (\text{se}) & & (0.83) \quad (0.146) \\ (p\text{-value}) & & (0.000) \quad (0.000) \end{array}$$

  – Now add *ENG* and *WGT*:

$$\begin{array}{llllll} \widehat{MPG} & = & 44.4 - 0.268CYL & - 0.0127ENG & - 0.00571WGT \\ (\text{se}) & & (1.5) \quad (0.413) & (0.0083) & (0.00071) \\ (p\text{-value}) & & (0.000) \quad (0.517) & (0.125) & (0.000) \end{array}$$

- 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, in some cases in which collinear relationships involve more than two of the explanatory variables, the collinearity may not be detected by examining pairwise correlations

- Try an auxiliary model:

$$x_2 = a_1 x_1 + a_3 x_3 + \cdots + a_K x_K + error$$

  - If R^2 from this artificial model is high, above 0.80, say, the implication is that a large portion of the variation in x_2 is explained by variation in the other explanatory variables

- 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.
- A second way of adding new information is to introduce nonsample information in the form of restrictions on the parameters

6.5 Prediction

- Consider the model:

Eq. 6.32

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

– The prediction problem is to predict the value of the dependent variable y_0, which is given by:

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

– The best linear unbiased predictor is:

$$\hat{y}_0 = b_1 + x_{02} b_2 + x_{03} \mathrm{b}_3$$

Eq. 6.33

- The variance of the forecast error, $f = \left(y_0 - \hat{y}_0\right)$, is:

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

- For our example, suppose $PRICE_0 = 6$, $ADVERT_0 = 1.9$, and $ADVERT^2_0 = 3.61$:

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

  - We forecast sales will be $76,974

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 estimated variance of the forecast error is:

$$\widehat{\operatorname{var}}(f) = \hat{\sigma}^2 + \widehat{\operatorname{var}}(b_1) + x_{02}^2\widehat{\operatorname{var}}(b_2) + x_{03}^2\widehat{\operatorname{var}}(b_3) + x_{04}^2\widehat{\operatorname{var}}(b_4)$$
$$+2x_{02}\widehat{\operatorname{cov}}(b_1,b_2) + 2x_{03}\widehat{\operatorname{cov}}(b_1,b_3) + 2x_{04}\widehat{\operatorname{cov}}(b_1,b_4)$$
$$+2x_{02}x_{03}\widehat{\operatorname{cov}}(b_2,b_3) + 2x_{02}x_{04}\widehat{\operatorname{cov}}(b_2,b_4) + 2x_{03}x_{04}\widehat{\operatorname{cov}}(b_3,b_4)$$
$$= 21.57865 + 46.22702 + 6^2 \times 1.093988 + 1.9^2 \times 12.6463 + 3.61^2 \times 0.884774$$
$$+2\times 6\times(-6.426113) + 2\times 1.9\times(-11.60096) + 2\times 3.61\times 2.939026$$
$$+2\times 6\times 1.9\times 0.300407 + 2\times 6\times 3.61\times(-0.085619)$$
$$+2\times 1.9\times 3.61\times(-3.288746)$$
$$= 22.4208$$

  - The standard error of the forecast error is:

$$\operatorname{se}(f) = \sqrt{22.4208} = 4.7351$$

- The 95% prediction interval is:

$$\left(76.974 - 1.9939 \times 4.7351, 76.974 + 1.9939 \times 4.7351\right) = \left(67.533, 86.415\right)$$

  - We predict, with 95% confidence, that the settings for price and advertising expenditure will yield *SALES* between \$67,533 and \$86,415

- The point forecast and the point estimate are both the same:

$$\widehat{SALES}_0 = \widehat{E}(SALES_0) = 76.974$$

  – But:

$$\operatorname{se}\left(\widehat{E}(SALES_0)\right) = \sqrt{\widehat{\operatorname{var}}(f) - \hat{\sigma}^2} = \sqrt{22.4208 - 21.5786} = 0.9177$$

  – A 95% confidence interval for $\widehat{E}(SALES_0)$ is:

$$(76.974 - 1.9939 \times 0.9177, 76.974 + 1.9939 \times 0.9177) = (75.144, 78.804)$$

Keywords

- AIC
- auxiliary regression
- BIC
- collinearity
- F-test
- irrelevant variables
- nonsample information
- omitted variables
- omitted variable bias
- overall significance
- prediction
- RESET
- restricted least squares
- restricted model
- restricted SSE
- SC
- single and joint null hypothesis
- testing many parameters
- unrestricted model
- unrestricted SSE

Key Words

Appendices

- The *F*-statistic is defined as:

Eq. 6A.1

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

- We can also show that:

Eq. 6A.2

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

- For sufficiently large sample:

Eq. 6A.3

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

- But we can also show that:

Eq. 6A.4

$$V_2 = \frac{(N-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-K)}$$

- From the book's appendix, we know that:

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

- Therefore:

Eq. 6A.5

$$\frac{\left. \dfrac{\left(SSE_R - SSE_U\right)}{\sigma^2} \middle/ J \right.}{\left. \dfrac{\left(N-K\right)\hat{\sigma}^2}{\sigma^2} \middle/ \left(N-K\right) \right.} = \frac{\left[SSE_R - SSE_U\right] / J}{\hat{\sigma}^2} \sim F_{(J, N-K)}$$

- A little reflection shows that:

$$F = \frac{\hat{V}_1}{J}$$

- When testing

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

in the equation

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

we get

$$F = 8.44 \qquad p\text{-value} = .0005$$

$$\chi^2 = 16.88 \qquad p\text{-value} = .0002$$

- Testing

$$H_0: \beta_3 + 3.8\beta_4 = 1$$

we get

$$F = .936 \qquad p\text{-value} = .3365$$

$$\chi^2 = .936 \qquad p\text{-value} = .3333$$

- Consider the model:

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

  – Now suppose we incorrectly omit x_{i3} and estimate:

$$y_i = \beta_1 + \beta_2 x_{i2} + v_i$$

- Notice the new disturbance term
  - It's

$$v_i = \beta_3 x_{i3} + e_i$$

- The estimator for β_2 is:

Eq. 6B.1

$$b_2^* = \frac{\sum (x_{i2} - \bar{x}_2)(y_i - \bar{y})}{\sum (x_{i2} - \bar{x}_2)^2} = \beta_2 + \sum w_i v_i$$

where

$$w_i = \frac{(x_{i2} - \bar{x}_2)}{\sum (x_{i2} - \bar{x}_2)^2}$$

- Substituting for v_i yields:

$$b_2^* = \beta_2 + \beta_3 \sum w_i x_{i3} + \sum w_i e_i$$

where

$$w_i = \frac{\left(x_{i2} - \bar{x}_2\right)}{\sum \left(x_{i2} - \bar{x}_2\right)^2}$$

- Hence: $E\left(b_2^*\right) = \beta_2 + \beta_3 \sum w_i x_{i3}$

$$= \beta_2 + \beta_3 \frac{\sum \left(x_{i2} - \bar{x}_2\right) x_{i3}}{\sum \left(x_{i2} - \bar{x}_2\right)^2}$$

$$= \beta_2 + \beta_3 \frac{\sum \left(x_{i2} - \bar{x}_2\right)\left(x_{i3} - \bar{x}_3\right)}{\sum \left(x_{i2} - \bar{x}_2\right)^2}$$

$$= \beta_2 + \beta_3 \frac{\widehat{\operatorname{cov}}\left(x_2, x_3\right)}{\widehat{\operatorname{var}}\left(x_2\right)} \neq \beta_2$$