Chapter 6
Further Inference in the Multiple Regression Model

Walter R. Paczkowski
Rutgers University

Chapter Contents

- 6.1 Joint Hypothesis Testing
- 6.2 Model Specification
- 6.3 Prediction

6.1
Joint Hypothesis Testing

6.1 Testing Joint Hypotheses

- 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?

- 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

- 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, $\beta_3 = 0$ and $\beta_4 = 0$, and *ADVERT* and $ADVERT^2$ drop out of the model

Eq. 6.2

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

  - 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. 6.2
  - 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. 6.3

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

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 $\beta_4 \neq 0$ or both are nonzero
  2. Specify the test statistic and its distribution if the null hypothesis is true:
     - Having two restrictions in H_0 means $J = 2$
     - Also, recall that $N = 75$:

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

- The *F*-test for our sales problem is (Continued):
  3. Set the significance level and determine the rejection region. If α = 5%, then $F_c = F_{(0.95,2,71)} = 3.13$.
  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 $F = 8.44 > F_c = 3.13$, 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

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

Eq. 6.4

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

- 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

6.1.2
Testing the Significance of the Model

- The unrestricted model is that given in Eq. 6.4
  - 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 (6.3)
  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

- We spend $1,900 per month for advertising. We want to know whether this amount could be optimal.
- Optimal advertising expenditure implies that:

Eq. 6.11

$$\left.\frac{\Delta E(S)}{\Delta A}\right|_{(P\text{ constant})} = \beta_3 + 2\beta_4 ADVERT_0 = 1$$

- Thus we conduct the test:

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

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:

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

- Suppose that we want to test if the optimal level of advertising is larger than $1,900. Then we have:

Eq. 6.13

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

- 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
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

Eq. 6.14

- Consider the model:

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

- If we incorrectly omit wife's education:

Eq. 6.15

$$\begin{aligned} \widehat{FAMINC} &= -26191 + 5155 HEDU \\ (\text{se}) &\quad\; (8541) \quad\; (658) \\ (p\text{-value}) &\quad (0.002)(0.000) \end{aligned}$$

- Relative to Eq. 6.14, 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. 6.16

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

  - 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. 6.17

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

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 wife'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. 6.18

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

  - 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
    - Thus useful results can still be obtained when a relevant variable is omitted if that variable is uncorrelated with the included ones.

- 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:

$$\widehat{FAMINC} = -7759 + 3340HEDU + 5869WEDU - 14200KL6 + 889X_5 - 1067X_6$$
$$(se) \quad (11195)(1250) \quad (2278) \quad (5044) \quad (2242) \quad (1982)$$
$$(p\text{-value}) \quad (0.500)(0.008) \quad (0.010) \quad (0.005) \quad (0.692) \quad (0.591)$$

  – 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. *Adjusted* 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 $R^2 = 1$

6.2.4a
The Adjusted Coefficient of Determination

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

Eq. 6.19

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

- The adjusted-R^2 is no longer the proportion of explained variation
- If a variable is added to the equation with coefficient β_k, then the $\bar{R}^2$ will increase if the t-ratio for testing $H_0 : \beta_k = 0$, is larger than 1.

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

Eq. 6.20

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

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

Eq. 6.21

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

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. 6.22

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

- Now consider the following two artificial models:

Eq. 6.23
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + e$$

Eq. 6.24
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e$$

- In Eq. 6.23 a test for misspecification is a test of $H_0\!:\!\gamma_1 = 0$ against the alternative $H_1\!:\!\gamma_1 \neq 0$
- In Eq. 6.24, 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 (see Eq. 6.18) 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

- Although the RESET test is often useful to pick up poorly specified models it does not discriminate between alternative models. For example, it is possible for the RESET to reject neither of two different functional forms used in a model.

6.3
Prediction

- Consider the model:

Eq. 6.26

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

– 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:

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

- As in Chapter 3 we can calculate the variance of the forecast error, $f = \left(y_0 - \hat{y}_0\right)$, as:

Eq. 6.27

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

- 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

- The estimated variance of the forecast error is:

$$\widehat{\operatorname{var}(f)} = 22.4208$$

  – The standard error of the forecast error is:

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

- The 95% prediction interval is:

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

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

Key Words

Keywords

- AIC
- auxiliary regression
- BIC
- collinearity
- F-test
- irrelevant variables
- 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