CHAPTER
7
MODEL SELECTION: CRITERIA AND TESTS

QUESTIONS

7.1. Specification errors occur when the form of the relationship between the dependent variable and the explanatory variables is wrongly specified because of:

1. Exclusion of relevant variables from the model, or
2. Inclusion of redundant variables in the model, or
3. Incorrect functional form (e.g., a linear model is fitted whereas the true model is log-linear), or
4. Wrong specification of the error term.

Notice that one or more of these problems might coexist.

7.2. Specification errors arise because:

1. The researcher is not sure of the theory underlying his research;
2. The researcher is not aware of the previous work in the area;
3. The researcher does not have data on the variables relevant for the model.
4. Of errors of measurement in the data.

7.3. A good econometric model:

1. Should be parsimonious;
2. Should obtain unique estimates of the parameters of the model;
3. Should fit the data at hand reasonably well;
4. Should have the signs of the estimated coefficients in accordance with theoretical expectations, and
5. Should have good (out of sample) predictive power.

7.4. Exclusion of relevant variables; inclusion of irrelevant variables; wrong functional form; wrong specification of the error term. Yes, one or more specification errors can occur simultaneously.

7.5. If a variable(s) is wrongly excluded from a model, the coefficients of the variables included in the model can be biased as well as inconsistent, the error variance is incorrectly estimated, the standard errors of the variables included in the model can be biased, and the conventional hypothesis testing based on the *t* and *F* tests can be unreliable.

7.6. The "relevance" of a variable depends on the theory underlying the model. Thus, in a demand function for money, income of the consumer, the interest rate, etc. are relevant variables but not, say, the amount of bananas produced in Mexico.

7.7. In the presence of the irrelevant variables, the OLS estimators are LUE (linear unbiased estimators) but not BLUE, that is, they are not efficient.

7.8. Since the inclusion of the irrelevant variables increases the standard errors of the coefficients, one may tend to accept the null hypothesis that a particular coefficient is zero, although in fact it may not be. Therefore, one should not include unnecessary variables in the model.

7.9. See answers to questions (7.7) and (7.8) above.

7.10. This is a common problem that one faces in any econometric analysis. Here theory should be the guide to model building. If the empirical results are not in accord with theory, one should be very wary of accepting those results, for in econometric model building our primary objective is to test a theory.

PROBLEMS

7.11. **(*a*)** $\widehat{\ln Y_t} = -7.8439 + 0.7148 \ln X_{2t} + 1.1135 \ln X_{3t}$

$t = (-2.9270) \quad (4.6636) \quad (3.7221) \quad R^2 = 0.9837$

The output-labor and output-capital elasticities are, 0.7148 and 1.1135, respectively, and both are individually statistically significant at the 0.005 level (one-tail test).

(*b*) $\widehat{\ln Y_t} = 2.0696 + 1.2576 \ln X_{2t}$

$t = (4.9541) \quad (18.9061) \quad r^2 = 0.9649$

Since we have excluded the capital input variable from this model, the estimated output-labor elasticity of 1.2576 is a biased estimate of the true elasticity; in (*a*), the true model, this estimate was 0.7148, which is much smaller than 1.2576.

As noted in the chapter, $E(a_2) = B_2 + B_3 b_{32}$, where b_{32} is the slope in the regression of $\ln X_3$ on $\ln X_2$, which in the present example is 0.48747. Using the estimated values in (*a*), we therefore see that:

$$E(a_2) = B_2 + B_3 b_{32} = 0.7148 + 1.1135\ (0.48747) = 1.2576.$$

Therefore, a_2 is biased upward.

(*c*) $\widehat{\ln Y_t} = -19.2380 + 2.4409 \ln X_{3t}$

$t = (-10.8443) \quad (16.4554) \qquad\qquad r^2 = 0.9542$

By excluding the relevant variable, labor, we are again committing a specification error. By the procedure outlined in (*b*), it is easy to show that:

$$E(a_3) = B_3 + B_2 b_{23} = 1.1135 + 0.7148(1.85712) = 2.4409,$$

where $b_{23} = 1.85712$.

This shows that the estimated elasticity is biased upward by 1.3274.

7.12. **(*a*)** Although most intermediate macroeconomics textbooks discuss the (Keynesian) consumption function as a function of income, there are economists who believe that wealth also is an important determinant of consumption expenditure. Therefore, the choice between Models I and II cannot be decided on purely theoretical grounds.

(*b*) Let $\text{Consumption} = C_1 + C_2 \text{Income} + C_3 \text{Wealth} + w$.

If in a concrete application both C_2 and C_3 turn out to be *individually* statistically significant, then neither Model I or Model II is the correct model. If, however, C_2 is significant and C_3 is not, probably Model I seems appropriate. On the other hand, if C_2 is insignificant but C_3 is significant, Model II may be appropriate. But beware of the problem of multicollinearity that is discussed in Chapter 8.

7.13. Here we commit the error of omitting a relevant variable, the intercept in the present instance. The consequences of omitting a relevant error are discussed in the chapter. Equation (5.40) gives the results of including the intercept in the model. In this particular instance the intercept term turns out to be statistically insignificant. Hence the results given in Equation (5.39) may be appropriate. In general, however, unless there is a strong reason to suppress the intercept, it is best to keep it in the model.

7.14. **(*a*)** $\hat{Y} = 23.9869 - 4.3756 X_3$

$t = (4.5820)\ (-4.2805)$ $r^2 = 0.4134$

(*b*) $\hat{Y} = 3.5318 + 3.9433 X_2 - 2.4994 X_3$

$t = (0.4354)\ (3.0487)\ (-2.3098)$ $R^2 = 0.5724.$

(*c*) That Fama is correct in his statement can be seen from the following regression:

(*i*) $\hat{Y} = -12.2815 + 5.6424 X_2$

$t = (-2.6137)\ (4.9099)$ $r^2 = 0.4811;$

(*ii*) $\hat{X}_2 = 5.1873 - 0.4758 X_3$

$t = (7.5055)\ (-3.5256)$ $r^2 = 0.3234.$

(*d*) First, the regression for 1954-1976 (that is, including 1954 and 1955) is:

$\hat{Y} = -1.3462 + 5.3231 X_2 - 2.6777 X_3$

$t = (-0.1657)\ (4.1037)\ (-2.1202)$ $R^2 = 0.6911.$

Dropping 1954 and 1955 and running the regression for 1956-76, we get:

$\hat{Y} = -11.3627 + 6.0120 X_2 - 1.0744 X_3$

$t = (-1.4726)\ (5.1418)\ (-0.9033)$ $R^2 = 0.7288.$

As a result of omitting just two observations, the regression results have changed dramatically. Inflation now has no statistically discernible effect on real rate of return on common stocks.

(*e*) Introducing $D = 0$ for observations in 1956-1976 and $D = 1$ for observations in 1977-1981, we obtained the following regression:

$$\hat{Y} = -3.3591 + 4.2531 X_2 - 1.6024 X_3 + 1.5156 D$$

t = (-0.3873) (3.3337) (-1.1646) (0.1757) $R^2 = 0.5546$.

Since the dummy coefficient is not statistically significant, there does not seem to be any difference in the behavior of real stock returns between the two periods. Of course, we are assuming that only intercepts differ between the two periods, but not the slopes. But this assumption can be tested by introducing a multiplicative dummy variable.

7.15. **(*a*)** The regression results for the four modes are as follows:

A:	$\widehat{\ln Y}_t = 1.5536 + 0.9976 \ln X_{2t} - 0.3328 \ln X_{3t}$ t = (17.370) (52.606) (-13.795) $R^2 = 0.9942$
B:	$\widehat{\ln Y}_t = 1.5932 + 0.8353 \ln X_{2t} + 0.1758 \ln X_{2(t-1)} - 0.3526 \ln X_{3t}$ t = (12.219) (3.045) (0.652) (-12.511) $R^2 = 0.9942$
C:	$\widehat{\ln Y}_t = 1.6295 + 1.0058 \ln X_{2t} - 0.2363 \ln X_{3t} - 0.1208 \ln X_{3(t-1)}$ t = (17.008) (52.027) (-3.951) (-1.920) $R^2 = 0.9950$
D:	$\widehat{\ln Y}_t = 1.2490 + 0.6713 \ln X_{2t} - 0.2704 \ln X_{3t} + 0.3332 \ln Y_{(t-1)}$ t = (11.599) (6.593) (-9.469) (3.404) $R^2 = 0.9964$

(*b*) Omission of relevant variable bias.

(*c*) The income and price elasticities are as follows:

Model	Income Elasticity	Price Elasticity
A	0.9976	-0.3328
B	(0.8353 + 0.1758) = 1.0111	-0.3526

C	1.0058	(-0.2363) + (-0.1208) = -0.3571
D	0.6713 / 0.6668 = 1.0067	(-0.2704) / 0.6668 = -0.4055

(***d***) In the CLRM it is assumed that the explanatory variables are nonstochastic, that is, their values are fixed in repeated sampling. But if the lagged value of the dependent variable is one of the explanatory variables, this assumption cannot be met. As a result, as shown in Chapter 12, the usual OLS procedure may not be valid.

7.16. The results of the Cobb-Douglas production function including the trend variable are as follows:

$$\hat{\ln} Y_t = 4.9443 - 0.1218 \ln X_{2t} + 0.4034 \ln X_{3t} + 0.1181 X_{4t}$$

$$t = (1.2285) \quad (-0.4753) \quad (1.3947) \quad (3.6023)$$

$$R^2 = 0.9925$$

The trend variable is statistically significant at the 5% level. By not including the trend variable in the original model, we have committed the specification error of excluding a relevant variable. The consequence in the present example are clearly visible. Neither the labor input not the capital input seem to have any impact on output when the trend variable (perhaps denoting technology) is included in the model.

In this example, what is happening is that there is a significant trend in *Y*, X_2, and X_3. Therefore, what this regression shows is the relationship between output and the two inputs after the (common) trend in them has been removed. In other words, this regression gives the short-run relationship between output and labor and capital inputs, which in the present instance is not statistically significant.

Caution: The practice of introducing the trend variable in a regression has now come under scrutiny. The regression results presented here assume that the trend is *deterministic* and not *stochastic*. On this, see Chapter 12 where this topic is discussed at some length.

7.17. **(*a*)** The *EViews* output of the log-linear model is as follows:

Dependent Variable: LOG(Y)
Sample: 1960 1982

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.189792	0.155715	14.06283	0.0000
LOG(X2)	0.342555	0.083266	4.113970	0.0007
LOG(X3)	-0.504592	0.110894	-4.550212	0.0002
LOG(X4)	0.148545	0.099673	1.490334	0.1535
LOG(X5)	0.091105	0.100716	0.904568	0.3776
R-squared	0.982313			

Note 1: A crucial point to remember: In *EViews*, the logarithmic functions are functions to the base *e* (natural logarithms). *EViews* uses the "log" term for such logarithms, even though natural logarithms are denoted with "ln" in general practice. So, in *EViews*, "log" stands for natural logarithm. If you want to convert a natural logarithm into one with the base 10, you should use the relationship: $\log_{10} x = \log_e x / \log_e 10$.

Note 2: In this example, there is a sixth variable, X_6, which is not used here as it is a composite of X_4 and X_5.

(*b*) Using EViews, we obtain the following linear model:

Dependent Variable: Y
Sample: 1960 1982

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	37.23236	3.717695	10.01490	0.0000
X2	0.005011	0.004893	1.024083	0.3194
X3	-0.611174	0.162849	-3.753010	0.0015
X4	0.198409	0.063721	3.113734	0.0060
X5	0.069503	0.050987	1.363144	0.1896
R-squared	0.942580			

(*c*) We cannot directly compare the two models for reasons stated in the text. To choose between the two models, we can use the MWD test discussed in the text. After constructing variables Z_1 and Z_2 in the manner described in the text, we obtain the following regressions:

Dependent Variable: Y Sample: 1960 1982				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	34.44709	2.137981	16.11197	0.0000
X2	0.002799	0.002775	1.008372	0.3274
X3	-0.489682	0.093635	-5.229701	0.0001
X4	0.162059	0.036315	4.462567	0.0003
X5	0.090554	0.028884	3.135118	0.0060
Z1	-50.13320	7.941861	-6.312526	0.0000
R-squared	0.982829			

Since the coefficient of Z_1 is statistically significant, we reject the linear model.

Dependent Variable: LOG(Y) Sample: 1960 1982				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.164940	0.168893	12.81841	0.0000
LOG(X2)	0.340388	0.085330	3.989087	0.0009
LOG(X3)	-0.475355	0.131241	-3.621994	0.0021
LOG(X4)	0.129179	0.110943	1.164372	0.2604
LOG(X5)	0.094041	0.103256	0.910762	0.3752
Z2	-2.36E+09	5.33E+09	-0.443187	0.6632
R-squared	0.982515			

Note: -2.36E+09 and 5.33E+09 represent scientific notation.

Since the coefficient of Z_2 is not statistically significant, we do not reject the hypothesis that the true model is log-linear.

7.18. **(*a*)** The *EViews* results are as follows:

Dependent Variable: Y Sample: 1968 1987				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-655.0402	138.0686	-4.744308	0.0002
X	0.522760	0.089616	5.833365	0.0000
TIME	-20.70755	4.910361	-4.217114	0.0007
TIME*TIME	0.223554	0.129155	1.730898	0.1027
R-squared	0.970396			

(*b*) Since the time-squared term is borderline statistically significant (using a one-tail test, it is significant at the 10% level), model (7.13) is mis-specified.

(*c*) In the present case we have omitted a significant variable from the model. As noted in the text, the presence of such a specification error leads not only to biased but also inconsistent estimates of the regression model that omits relevant variable(s). This can be seen from the regression results presented above.

7.19. If we include all the variables in the model as a startup model, we get the following *EViews* results:

Dependent Variable: MAP
Sample: 1 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	132.8086	48.79243	2.721910	0.0297
SPP	-0.001082	0.002687	-0.402874	0.6991
STR	2.794246	2.415948	1.156584	0.2854
EDU	0.795938	0.472978	1.682824	0.1363
MINCOME	0.000175	0.000316	0.554112	0.5968
DUM	21.60799	13.81711	1.563857	0.1618
R-squared	0.947735			

Since none of the explanatory variables is statistically significant, we have to rethink the initial model. It seems that we have multicollinearity in the variables.

(*b*), **(*c*)** and **(*d*)** Using only STR, EDU, and DUM as explanatory variables, we obtain the following results:

Dependent Variable: MAP
Sample: 1 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	114.0212	25.80188	4.419103	0.0017
STR	3.764235	1.555686	2.419663	0.0386
EDU	0.844971	0.272621	3.099430	0.0127
DUM	24.14025	10.42762	2.315029	0.0459
R-squared	0.944921			

You can try other variations. In the preceding regression all the variables are individually statistically significant. But the positive value of the STR

coefficient would suggest that, *ceteris paribus*, the higher the student-teacher ratio, the higher the MAP. This is counter-intuitive.

It would seem both social and economic factors are important in the MAP test outcome.

7.20. Econometrically speaking, the Supreme Court's decision is incorrect, for the consequences of excluding relevant variables can be serious. Of course, the defendants in this case simply argued that the plaintiff's model had not included all the relevant variables. If the defendants were serious, they should have presented their own regression results to buttress their argument that as a result of omitting the relevant variables the results submitted by the plaintiffs were seriously biased. In the absence of such evidence, the Supreme Court did the best it could.

7.21. ***(a)*** *Minitab* results for a linear model are:

```
Regression Analysis: Output versus Worker Hrs, Expend

The regression equation is
Output = 233622 + 48.0 Worker Hrs + 9.95 Expend

Predictor     Coef  SE Coef      T      P
Constant    233622  1250364   0.19  0.853
Worker Hrs  47.987    7.058   6.80  0.000
Expend      9.9519   0.9781  10.17  0.000

S = 6300694   R-Sq = 98.1%   R-Sq(adj) = 98.0%

Analysis of Variance

Source          DF           SS           MS        F      P
Regression       2  9.87319E+16  4.93659E+16  1243.51  0.000
Residual Error  48  1.90554E+15  3.96987E+13
Total           50  1.00637E+17
```

(b) Results from the log-linear model are:

```
Regression Analysis: ln Output versus ln Worker Hrs, ln Expend
The regression equation is
ln Output = 3.89 + 0.468 ln Worker Hrs + 0.521 ln Expend
```

```
Predictor          Coef  SE Coef     T      P
Constant         3.8876   0.3962  9.81  0.000
ln Worker Hrs   0.46833  0.09893  4.73  0.000
ln Expend       0.52128  0.09689  5.38  0.000

S = 0.266752   R-Sq = 96.4%   R-Sq(adj) = 96.3%

Analysis of Variance

Source           DF      SS      MS       F      P
Regression        2  91.925  45.962  645.93  0.000
Residual Error   48   3.416   0.071
Total            50  95.340
```

(c) For the MWD test,

H_0: Linear Model: Y is a linear function of the *X*'s
H_1: Log-linear Model: ln *Y* is a linear function of the *X*'s or log of *X*'s

Results are:

```
Regression Analysis: Output versus ln Worker Hrs, ln Expend,
Z1

The regression equation is
Output = - 4.07E+08 + 25000768 ln Worker Hrs + 9825097 ln
Expend + 1.71E+08 Z1

Predictor            Coef   SE Coef      T      P
Constant       -407118647  43875137  -9.28  0.000
ln Worker Hrs    25000768   9335749   2.68  0.010
ln Expend         9825097   8822973   1.11  0.271
Z1              170696243  35598822   4.79  0.000

S = 24253901   R-Sq = 72.5%   R-Sq(adj) = 70.8%

Analysis of Variance

Source           DF           SS           MS      F      P
Regression        3  7.29896E+16  2.43299E+16  41.36  0.000
Residual Error   47  2.76478E+16  5.88252E+14
Total            50  1.00637E+17
```

Since the *Z* coefficient is statistically significant, we can reject H_0.

The second output reveals:

```
Regression Analysis: ln Output versus ln Worker Hrs, ln
Expend, Z2

The regression equation is
ln Output = 4.13 + 0.543 ln Worker Hrs + 0.438 ln Expend -
0.000000 Z2


Predictor             Coef     SE Coef      T      P
Constant            4.1322      0.4372   9.45  0.000
ln Worker Hrs       0.5428      0.1141   4.76  0.000
ln Expend           0.4381      0.1160   3.78  0.000
Z2             -0.00000003  0.00000002  -1.28  0.205


S = 0.264962   R-Sq = 96.5%   R-Sq(adj) = 96.3%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       3  92.040  30.680  437.01  0.000
Residual Error  47   3.300   0.070
Total           50  95.340
```

Since the Z coefficient is *not* statistically significant here, we cannot reject H_1. Therefore, the log-linear model appears to be a better choice.