Lecture 9: Econometric Methodology and Model Selection

R.G. Pierse

1 Introduction

In standard econometric textbooks and in this lecture course, the statistical model underlying the data is generally assumed to be known. In reality, of course, it is not the case that the applied econometrician knows the 'true' model at the outset of the analysis. The investigator will start with a specification that seems reasonable *a priori*, on the basis of economic theory. However, this specification will be modified in the light of the coefficient estimates and diagnostic tests that come out of the data analysis. The final model will be the outcome of a specification search that involves both theory and the data.

1.1 Data Mining

'Data mining' is a pejorative term that often used to be applied to empirical analysis in econometrics that allows the data to select the model. Clearly, selecting the model solely on the basis of fitting the data, will lead to a model that will often be economic nonsense, and whose good statistical properties, high R^2, significant t-ratios, etc., are largely spurious, because they have been self-selected.

However, selecting a model, completely ignoring the information coming from the data, is just as silly. For example, imposing economic restrictions that are rejected by the data, or ignoring evidence of autocorrelation in regression residuals will lead to biased coefficient estimates. In many cases, the data will clearly suggest that a particular model specification is wrong. The investigator can then use a combination of economic insight and data-based information.

There are many areas of model specification in which economic theory does not tell us what to do. In particular, it does not often suggest the correct functional form to use, and often specifies only the long-run equilibrium and not the short-run dynamics of an economic relationship. In these cases, we have no *a priori* view and must look to the data to give us the answer.

2 Specification Searches

Leamer (1978) distinguishes six types of specification search undertaken in the process of model selection:

Types of Search	**Purpose**
(1) Hypothesis testing search	Choosing a 'true' model
(2) Interpretive search	Interpreting a model with correlated variables
(3) Simplification search	Constructing a 'fruitful' model
(4) Proxy Variable search	Choosing between different measures of variables
(5) Data selection search	Selecting the appropriate data set
(6) Post-data model construction	Improving an existing model

The differences between these six types of search can be very minor. However, they can be useful as a means of organisation of ideas.

The six search types will be illustrated using a model for the demand for oranges, estimated on data for 150 households, taken from Leamer (1978):

$$\begin{aligned} \log O_i &= 6.2 + 0.85\log Y_i - 0.67\log P_i \quad , \quad R^2 = 0.15 \\ & \ (1.1)\ (0.21) \qquad\quad (0.13) \end{aligned} \tag{2.1}$$

where O is the quantity of oranges purchased, Y is income, and P is the price of oranges, and coefficient standard errors are given in parentheses.

2.1 Hypothesis Testing Search

An example of a hypothesis testing search would be a test of the economic restriction of a unit price elasticity in (2.1). Imposing this restriction results in the equation

$$\begin{aligned} \log O_i + \log P_i &= 7.2 + 0.96\log Y_i \quad , \quad R^2 = 0.14 \\ & \ (1.0)\ (0.20) \end{aligned}$$

and an F-test *rejects* this restriction at the 5% level.

2.2 Data Selection Search

Next the investigator estimates separate regressions for the North and the South, to test the hypothesis that the demand may be different in the two regions, obtaining:

$$\begin{aligned} \log O_i &= 7.3 + 0.89\log Y_i - 0.60\log P_i \quad , \quad R^2 = 0.18 \\ & \ (1.9)\ (0.41) \qquad\quad (0.25) \end{aligned}$$

and

$$\begin{aligned}\log O_i &= \underset{(2.2)}{7.0} + \underset{(0.31)}{0.82} \log Y_i - \underset{(0.26)}{1.10} \log P_i \quad , \quad R^2 = 0.19\end{aligned}$$

respectively. The hypothesis that the coefficients of the income and price variables are different is not rejected at the 5% level of significance.

2.3 Proxy variable Search

Believing that expenditure E might be a better proxy of income than Y, the investigator estimates the regression:

$$\begin{aligned}\log O_i &= \underset{(1.0)}{5.2} + \underset{(0.18)}{1.1} \log E_i - \underset{(0.16)}{0.45} \log P_i \quad , \quad R^2 = 0.18\end{aligned} \; .$$

As a result of this proxy variable search, the income variable has become more significant and the R^2 has increased.

2.4 Post-data model construction

Noting the low values for R^2 in the previous regressions, the investigator tries adding the price of a substitute product, grapefruit, PG, to the regression. The resulting regression is

$$\begin{aligned}\log O_i &= \underset{(1.0)}{3.1} + \underset{(0.83)}{0.83} \log E_i + \underset{(0.15)}{0.01} \log P_i - \underset{(0.60)}{0.56} \log PG_i \quad , \quad R^2 = 0.20\end{aligned} \; .$$

This is an example of post-data model construction, that is revising the model in the light of previous results. Note that although the R^2 has increased, the two price elasticities are now both statistically insignificant and have the wrong sign.

2.5 Interpretive Search

The researcher now recalls the homogeneity postulate of demand theory and re-estimates the regression imposing this restriction (that the coefficients on E, P, and PG should sum to zero):

$$\begin{aligned}\log O_i &= \underset{(0.9)}{4.2} + \underset{(0.19)}{0.52} \log E_i - \underset{(0.14)}{0.61} \log P_i + \underset{(0.31)}{0.09} \log PG_i \quad , \quad R^2 = 0.19\end{aligned} \; .$$

This regression is the result of an interpretative search. The restriction is accepted by the data, and results in both price coefficients having the correct signs and the coefficient on P becoming significant.

2.6 Simplification Search

Finally, noting that the price coefficient on grapefruit is insignificant, and that the income and price elasticities are close in absolute value, the researcher estimates the following model:

$$\begin{aligned} \log O_i &= 3.7 + 0.58 \log(E_i/P_i) \quad , \quad R^2 = 0.18 \\ & (0.8) \ (0.18)\,. \end{aligned}$$

The objective of this final search is to obtain a simple *parsimonious* model with a small number of parameters.

3 Hendry Methodology

An approach to model selection that is known as *general-to-specific* or *top-down* was formulated at the LSE in the 1970s. This has now come to be associated particularly with the writings of David Hendry (Hendry (1979), Hendry and Richard (1983), Hendry (1995)) and Grayham Mizon (1977), even though it is more correctly attributed to J.D. Sargan.

The essence of the 'Hendry' approach is *intended overparameterisation* with *data-based simplification.* This is contrasted with the prevailing methodology which comprised *excessive presimplification with inadequate diagnostic testing.* This latter methodology is characterised by the following stylised schema:

1) Start with theories which are *drastic* abstractions of reality
2) Formulate *highly parsimonious* relationships to represent the theories
3) Estimate using techniques that are optimal *only* if the model is correct
4) Test a *few* assumptions using conventional statistics (e.g. DW)
5) Revise the specification in the light of evidence
6) Reestimate the model accordingly

This approach is a *specific-to-general* or *bottoms-up* approach and has two main drawbacks: (i) every test is conditional on assumptions which are only tested later. If these are rejected, then all earlier inference is invalid. (ii) The sequence of tests is unstructured and may lead to premature selection of a model that is too restricted.

By contrast, the Hendry approach starts with a very general dynamic model, which is deliberately *overparameterised.* In the case of a single regressor, X, this would be the *autoregressive distributed lag* (ADL) model

$$Y_t = a_0 + a_1 Y_{t-1} + \cdots + a_{p+1} Y_{t-p-1} + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_{l+1} X_{t-l-1} + u_t$$

where the order of p and l is chosen to be large enough to ensure no serial correlation in the residuals. Note that it is important that this general model includes *all the relevant regressors.*

The general model is then progressively simplified through a sequence of tests. These tests may be motivated either by economic theory, or by the data itself. The advantages of this strategy are: (i) inference at every stage is valid (ii) the testing sequence is structured. At every stage in the sequence of tests, the model is a restricted case of the initial general model and so a joint test can always be formulated of *all* the restrictions that have been imposed so far.

Hendry and Richard (1983) suggest six criteria that should be met by an econometric model:

Criteria	**Meaning**
(1) Data admissibility	Model predictions must be consistent with the data
(2) Theory consistency	Model must make economic sense
(3) Exogenous regressors	Regressors should be uncorrelated with the error term
(4) Parameter constancy	Parameter estimates should be stable
(5) Data coherency	The residuals must be purely random (no pattern)
(6) Encompassing	The model should explain the results of all rival models

Each of these criteria can be tested to see how far the model conforms with the ideal.

4 Tests of Non-Nested Hypotheses

Consider the following two models:

$$\text{Model A} : Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

and

$$\text{Model B} : Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \ .$$

We say that Model B is *nested* within Model A because it is a special case of Model A corresponding to the restriction that $\beta_4 = 0$. All the tests considered so far within this course have been forms of *nested test.*

Now consider the two alternative models:

$$\text{Model C} : Y_t = \alpha_1 + \alpha_2 X_{2t} + u_t$$

and

$$\text{Model D} : Y_t = \beta_1 + \beta_2 Z_{2t} + u_t$$

where X and Z are different variables. We say that Models C and D are *non-nested* because neither is a special case of the other.

Many approaches have been formulated to testing non-nested hypotheses. One will be considered here: this is the J-test of Davidson and MacKinnon (1981). This proceeds as follows:

1) Estimate Models C and D, deriving the fitted values $\widehat{Y}_t^c$ and $\widehat{Y}_t^d$ respectively.

2) Add the variable $\widehat{Y}_t^d$ to Model C and re-estimate to get:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 \widehat{Y}_t^d + u_t$$

and test the hypothesis that $\alpha_3 = 0$ using a t-test. If the hypothesis is not rejected, then Model C is preferred to Model D, since the variables in Model D, as represented by the regressor $\widehat{Y}_t^d$, have no additional explanatory power over and above the variables in Model C.

3) Add the variable $\widehat{Y}_t^c$ to Model D and re-estimate to get:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 \widehat{Y}_t^c + u_t$$

and test the hypothesis that $\beta_3 = 0$ using a t-test. If this hypothesis is not rejected, then Model D is preferred to Model C.

There are four possible outcomes from this testing procedure, as shown in the table:

	$\alpha_3 = 0$	$\alpha_3 = 0$
$\beta_3 = 0$	**Do not reject**	**Reject**
Do not reject	Accept both C and D	Accept D, reject C
Reject	Accept C, reject D	Reject both C and D

If both, or neither model is rejected, then no clear answer is possible.

References

[1] Davidson, R. and J.G. MacKinnon (1981), ‘Several tests for model specification in the presence of alternative hypotheses’, *Econometrica*, 49, 781–793.

[2] Hendry, D.F. (1979), ‘Predictive failure and economic modelling in macroeconomics: the transactions demand for money’, in P. Ormerod (ed.) *Economic Modelling*, Heinemann, London.

[3] Hendry, D.F. (1995), *Dynamic Econometrics*, Oxford University Press, Oxford.

[4] Hendry, D.F. and J-F. Richard (1983), ‘The econometric analysis of economic time series’, *International Statistical Review*, 51, 3–33.

[5] Leamer, E.E. (1978), *Specification Searches: ad hoc inference with nonexperimental data*, John Wiley, New York.

[6] Mizon, G.E. (1977), ‘Model selection procedures’, in M.J.Artis and A.R. Nobay (eds.), *Studies in Current Economic Analysis*, Basil Blackwell, Oxford.