model selection

Jean-Marie Dufour
Edited by Steven N. Durlauf and Lawrence E. Blume

Abstract

The problem of statistical model selection in econometrics and statistics is reviewed. Model selection is interpreted as a decision problem through which a statistical model is selected in order to perform statistical analysis, such as estimation, testing, confidence set construction, forecasting, simulation, policy analysis, and so on. Broad approaches to model selection are described: (1) hypothesis testing procedures, including specification and diagnostic tests; (2) penalized goodness-of-fit methods, such as information criteria; (3) Bayesian approaches; (4) forecast evaluation methods. The effect of model selection on subsequent statistical inference is also discussed.

Keywords

ARMA models; autocorrelation; Bayesian statistics; deterministic models; econometrics; endogeneity; forecasting; forecast evaluation; heteroskedasticity; linear models; model selection; models; parsimony; probability models; serial correlation; specification problems in econometrics; statistical decision theory; statistical inference; stochastic models; structural change; testing; time series analysis

Article

The purpose of econometric analysis is to develop mathematical representations of observable phenomena, which we call models or hypotheses (models subject to restrictions). Such models are then used to perform parameter estimation, test hypotheses, build confidence sets, make forecasts, conduct simulations, analyse policies, and so on. A central feature of modelling activity is the fact that models are usually interpreted as stylized (or simplified) representations that can perform certain tasks – such as prediction – but (eventually) not others, and they are treated as if they were true for certain purposes. Indeed, summarizing and stylizing observed phenomena can be viewed as essential components of modelling activity, which make it useful. This feature is not specific to economics and is shared by other sciences (see Cartwright, 1983).

Models can be classified as either deterministic or stochastic. Deterministic models, which often claim to make arbitrarily precise predictions, can be useful in theoretical activity. However, such models are rarely viewed as appropriate representations of observed data; for example, unless they are highly complex or indeterminate, they are typically logically inconsistent with data. For this reason, models used for econometric analysis are usually stochastic (or statistical).

Formally, a statistical model is a family of probability distributions (or measures) which are proposed to represent observed data. Model selection, in this context, is the task of selecting a family of proposed probability distributions, which will then be used to analyse data and perform other statistical inference operations (such as parameter estimation, hypothesis testing, and so on).

A basic feature of probability models is that they are typically unverifiable: as for any theory that makes an indefinite number of predictions, we can never be sure that the model will not be at odds with new data. Moreover, they are logically unfalsifiable: in contrast with deterministic models, a probabilistic model is usually logically compatible with all the possible observation sets. Consequently, model selection can depend on a wide array of elements, such as the objectives of the model, (economic) theory, the data themselves, and various conventions.

Features which are often viewed as desirable include: (a) simplicity or parsimony (Zellner, Keuzenkamp and McAleer, 2001); (b) the ability to deduce testable (or falsifiable) hypotheses (Popper, 1968); (c) the possibility of interpreting model parameters in terms of economic theory, if not consistency with economic theory; (d) the ability to satisfactorily perform the tasks for which the model is built (prediction, for example); (e) consistency with observed data. It is important to note that these characteristics depend (at least, partially) on conventional elements, such as the objectives of the model, criteria upon which a model will be deemed ‘satisfactory’, and so on. For further discussions of these general issues, the reader may consult Poirier (1994), Morgan and Morrison (1999), Keuzenkamp (2000), Zellner, Keuzenkamp and McAleer (2001) and Dufour (2003).

In this article, we focus on statistical methods for selecting a model on the basis of the available data. Methods for that purpose can be classified in four broad (not mutually exclusive) categories:

1. hypothesis testing procedures, including specification and diagnostic tests;
2. penalized goodness-of-fit methods, such as information criteria;
3. Bayesian approaches;
4. forecast evaluation methods.

The three first approaches are meant to be applicable ‘in-sample’, while the last approach stricto sensu requires observations that are not available when the model is selected, but may lead to model revision. (For general reviews of the topic of statistical model selection in econometrics and statistics, see Hocking, 1976; Leamer, 1978; 1983; Draper and Smith, 1981; Judge et al., 1985, chs 7 and 21; Sakamoto, Ishiguro and Kitagawa, 1985; Grasa, 1989; Choi, 1992; Gouriéroux and Monfort, 1995, ch. 22; Charemza and Deadman, 1997; McQuarrie and Tsai, 1998; Burnham and Anderson, 2002; Clements and Hendry, 2002; Miller, 2002; Bhatti, Al-Shanfari and Hossain, 2006). It is also interesting to note that classification techniques in statistics contain results that may be relevant to model selection. This topic, however, goes beyond the scope of the present article (for further discussion, see Krishnaiah and Kanal, 1982).
Model selection and specification errors

Most model selection methods deal in different ways with a trade-off between model realism – which usually suggests considering relatively general, hence complex models – and parsimony. From the viewpoint of estimation, for example, a model which is too simple (or parsimonious) involves specification errors and biases in parameter estimation, while too complex a model leads to parameter estimates with large variances. If the objective is forecasting, it is usually unclear which effect dominates.

For example, let us consider a linear regression model of the form

\[ y_t = x'_t \beta + u_t, \quad t = 1, \ldots, T, \]  

where \( y_t \) is a dependent variable and \( x_{1t}, \ldots, x_{kt} \) are explanatory variables, and \( u_t \) is a random disturbance which is typically assumed to be independent of (or uncorrelated with) the explanatory variables. In the classical linear model, it is assumed that the regressors can be taken as fixed and that the disturbances \( u_1, \ldots, u_T \) are independent and identically distributed (i.i.d.) according to a \( \mathcal{N}(0, \sigma^2) \) distribution. In this context, model selection typically involves selecting the regressors to be included as well as various distributional assumptions to be made upon the disturbances. An especially important version of (1) is the autoregressive model:

\[ y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t, \quad t = 1, \ldots, T. \]  

Then a central model selection issue consists in setting the order \( p \) of the process. In such models, there is typically little theoretical guidance on the order, so data-based model selection rules can be quite useful. A related set-up where model selection is usually based on statistical methods is the class of autoregressive-moving-average (ARMA) models

\[ y_t = \varphi y_{t-1} + \cdots + \varphi_p y_{t-p} + u_t - \theta_1 u_{t-1} - \cdots - \theta_q u_{t-q}. \]  

Hypothesis testing approaches

Since hypothesis tests are based on a wide body of statistical theory (see Lehmann, 1986; Gouriéroux and Monfort, 1995), such procedures are widely used for assessing, comparing and selecting models. Furthermore, econometric models are also based on economic theory which suggests basic elements that can be used for specifying models. This entails a form of asymmetry, in which restrictions suggested by economic theory will be abandoned only if ‘sufficient evidence’ becomes available. Although significance tests are meant to decide whether a given hypothesis (which usually takes the form of a restricted model) is compatible with the data, such procedures can also be used for model selection. It is interesting to note that the methodology originally proposed by Box and Jenkins (1976) for specifying ARMA models was almost exclusively based on significance tests (essentially, autocorrelation tests).

There are two basic ways of using hypothesis tests for that purpose. The first one is forward or specific-to-general approach, in which one starts from a relatively simple model and then checks whether the model can be deemed ‘satisfactory’. This typically involves various specification tests, such as:

1. residual-based tests, including tests for heteroskedasticity, autocorrelation, outliers, distributional assumptions (for example, normality), and so on;
2. tests for unit roots and/or stationarity, to decide whether corrections for integrated variables may be needed;
3. tests for the presence of structural change;
4. tests for the addition of explanatory variables;
5. tests for the functional form used (for example, linearity vs. nonlinearity).

There is a considerable literature on specification tests in econometrics (see Godfrey, 1988; MacKinnon, 1992; Davidson and MacKinnon, 1993). Systematic procedures for adding variables are also known in statistics as forward selection or stepwise regression procedures (Draper and Smith, 1981).

The second way is the backward or general-to-specific approach, in which one starts from a relatively comprehensive model which includes all the relevant variables. This model is then simplified by checking which variables are significant. Backward selection procedures in statistics (Draper and Smith, 1981) and the general-to-specific approach in econometrics (Davidson et al., 1978; Charemza and Deadman, 1997) can be viewed as illustrations of this approach.

In practical work, the backward and forward approaches are typically combined. Both involve a search for a model which is both parsimonious and consistent with the data. However, the results may differ. Specifying a model through significance tests involves many judgements and depends on idiosyncratic decisions. Further, standard hypothesis tests involve the use of typically conventional levels (such as the commonly used five per cent level). The powers of the tests can also have a strong influence on the results.

Penalized goodness-of-fit criteria
As pointed out by Akaike (1974), it is not clear that hypothesis testing is a good basis for model selection. Instead, the problem of model selection may be better interpreted as an estimation problem involving a well-defined loss function. This leads to the topic of goodness-of-fit criteria.

A common way of assessing the performance of a regression model, such as (1), consists in computing the coefficient of determination, that is, the proportion of the dependent variable variance which is ‘explained’ by the model:

\[ R^2 = 1 - \frac{\hat{V}(u)}{\hat{V}(y)} \]  

(4)

where \( \hat{V}(u) = \sum_{t=1}^{T} \hat{\epsilon}_t^2 / T, \) \( \hat{V}(y) = \sum_{t=1}^{T} (y_t - \gamma)^2 / T, \) \( \gamma = \sum_{t=1}^{T} y_t / T \) and \( \hat{\epsilon}_1, \ldots, \hat{\epsilon}_T \) are least squares residuals. This measure, however, has the inconvenient feature that it always increases when a variable is added to the model, even if it is completely irrelevant, and it can be made equal to its maximal value of one by including a sufficient number of regressors (for example, using any set of \( T \) linearly independent regressors).

An early way of avoiding this problem was proposed by Theil (1961, p. 213) who suggested that \( \hat{V}(u) \) and \( \hat{V}(y) \) be replaced by the corresponding unbiased estimators \( \hat{s}^2 = \frac{T}{T-k} \hat{\epsilon}_t^2 / (T-k) \) and \( \hat{s}_y^2 = \frac{T}{T-1} (y_t - \gamma)^2 / (T-1) \). This yields the adjusted coefficient of determination:

\[ R^2 = 1 - \frac{s^2}{s_y^2} = 1 - \frac{T-1}{T-k} (1 - R^2) = \frac{k-1}{T-k} (1 - R^2). \]

(5)

It is easy to see that \( R^2 \) may increase when the number of regressors increases. Note that maximizing \( R^2 \) is equivalent to minimizing the ‘unbiased estimator’ \( s^2 \) of the disturbance variance. Further, if two regression models (which satisfy the assumptions of the classical linear model) are compared, and if one of these is the ‘true’ model, then the value of \( s^2 \) associated with the true model is smaller on average than the one of the other model (see Theil, 1961, p. 543). On the other hand, in large samples, the rule which consists in maximizing \( R^2 \) does not select the true model with a probability converging to one: that is, it is not consistent (see Gouriéroux and Monfort, 1995, section 2.3).

Another approach consists in evaluating the ‘distance’ between the selected model and the true (unknown) model. Let \( f(y) \) the density associated with the postulated model and \( f_0(y) \) the density of the true model, where \( y=(y_1, \ldots, y_T) \). One such distance is the Kullback distance:

\[ I(f, f_0) = \int f(y) \log f(y) - f_0(y) \log f(y) \, dy = E \frac{\log f(y)}{f_0(y)} \]  

Minimizing \( I(f, f_0) \) with respect to \( f \) is equivalent to minimizing \( -E \frac{\log f(y)}{f_0(y)} \). We obtain an information criterion by selecting an ‘estimator’ of \( E \frac{\log f(y)}{f_0(y)} \).

For the case where the model is estimated by maximizing a likelihood function \( L(\Theta) \) over a \( K \times 1 \) parameter vector \( \Theta \), Akaike (1973) suggests that \( L(\hat{\Theta}) \) can be viewed as a natural estimator of \( E \frac{\log f(y)}{f_0(y)} \). However, the fact that \( \Theta \) has been estimated introduces a bias. This bias is (partially) corrected – using an expansion argument – by subtracting the number \( K \) from \( L(\hat{\Theta}) \). This suggests the following information criterion:

\[ AIC_L(\hat{\Theta}_T) = -2L(\hat{\Theta}_T) + 2K \]

(6)

where \( K \) is the dimension of \( \Theta \) (the number of estimated parameters) and multiplication by 2 is introduced to simplify the algebra. Among a given set of models, the one with the lowest \( AIC \) is selected.

The above criterion has also been generalized by various authors leading the following general class of criteria:

\[ IC_L(\hat{\Theta}_T) = -2L(\hat{\Theta}_T) + c(T, K)K \]

(7)

where \( c(T, K) \) is a function of \( T \) and \( K \). In the case of Gaussian models, such as (1) or (2) with i.i.d. \( N(0, \sigma^2) \) disturbances, we have

\[ L_T(\hat{\Theta}_T) = \frac{1}{2} \ln(\hat{\sigma}_T^2) + d_T, \]  

where \( d_T \) is a constant which only depends on \( T \), so that minimizing \( IC_L(\hat{\Theta}_T) \) is equivalent to minimizing

\[ IC(\hat{\Theta}_T) = \ln(\hat{\sigma}_T^2) + c(T, K)K \frac{T}{T}. \]

(8)

Alternative values of \( c(T, K) \) which have been proposed include:

1. \( c(T, K) = 2 \) (Akaike, 1969), which yields what is usually called the AIC criterion;
2. \( c(T, K) = \ln(T) \) (Schwarz, 1978);
3. \( c(T, K) = 2 \times T \times \ln(T) \) where \( K + s = T > 1 \) (Hannan and Quinn, 1979);
4. \( c(T, K) = 2 + \frac{2K(K+1)}{T - K - 1} \) (Hurvich and Tsai, 1989), which leads to the \( AIC_c \) criterion.

An especially convenient feature of such information criteria is the fact that they can be applied to both regression models (through(7)) as well as to various nonlinear models (using (6)).

Other related rules include: (a) criteria based on an estimate of the final prediction error, which try to estimate the mean square prediction error
Bayesian model selection

Bayesian model selection involves comparing models through their ‘posterior probabilities’ giving observed data. Suppose we have two models $M_1$ and $M_2$ each of which postulates that the observation vector $y$ follows a probability density which depends on a parameter vector: $p_1(y|\theta_1, M_1)$ under $M_1$, and $p_2(y|\theta_2, M_2)$ under $M_2$, where $\theta_1$ and $\theta_2$ are unknown parameter vectors (which may have different dimensions). Further, each one of the parameter vectors is assigned a ‘prior distribution’ ($p(\theta_1|M_1)$ and $p(\theta_2|M_2)$), and each model a ‘prior probability’ ($p(M_1)$ and $p(M_2)$).

Then one may compute the ‘posterior probability’ of each model $i$ given the data

$$p(M_i|y) = p(M_i) p_1(y|\theta_i, M_1) p(\theta_i|M_i)$$

(8)

This posterior probability of each model provides a direct measure of the ‘plausibility’ of each model. In such contexts, the ratio

$$K_{12} = \frac{p(M_1|y)}{p(M_2|y)}$$

(9)

is called the ‘posterior odds ratio’ of $M_1$ relative to $M_2$.

A rational decision rule for selecting between $M_1$ and $M_2$ then emerges if we can specify a loss function such as

$$L(i, j) = \text{cost of choosing } M_j \text{ when } M_i \text{ is true}.$$  

(10)

If $L(i, i)=0$ for $i=1, 2$, expected loss is minimized by choosing $M_1$ when

$$K_{12} \geq \frac{L(2,1)}{L(1,2)}.$$  

(11)

and $M_2$ when otherwise. In particular, if $L(1,2)=L(2,1)$, expected loss is minimized by choosing the model with the highest posterior probability. Such rules can be extended to problems where more than two models are compared.

The Bayesian approach automatically introduces a penalty for non-parsimony and easily allows the use of decision-theoretic considerations. The main difficulty consists in assigning prior distributions on model parameters and prior probabilities to competing models. For further discussion, see Zellner (1971, ch. 10), Leamer (1978; 1983), Gelman et al. (2003) and Lancaster (2004).

Forecast evaluation

In view of the fact that forecasting is one of the most common objectives for building econometric models, alternative models are often assessed by studying post-sample forecasts. Three types of assessments are typically considered in such contexts: (a) tests of predictive failure; (b) descriptive measures of forecast performance, which can be compared across models; (c) tests of predictive ability.

A test of predictive failure involves testing whether the prediction errors associated with a model are consistent with the model. This suggests testing whether forecasts are ‘unbiased’ or ‘too large’ to be consistent with the model. The well-known predictive test for structural change proposed by Chow (1960) is an early example of such an approach. (For further discussion and extensions, see Box and Tiao, 1976; Dufour, 1980; Pesaran, Smith and Yeo, 1985; Dufour, Ghysels and Hall, 1994; Dufour and Ghysels, 1996; Clements and Hendry, 1998.)

Common measures of forecast performance involve mean errors, mean square errors, mean absolute errors, and so on (see Theil, 1961; Diebold, 2004). Although commonly used, such measures are mainly descriptive. They can usefully be complemented by tests of predictive ability. Such procedures test whether the difference between expected measures of forecast performance is zero (or less than zero) against an alternative where it is different from zero (or larger than zero). Tests of this type were proposed, among others, by Meese and Rogoff (1988), Diebold and Mariano (1995), Harvey, Leybourne and Newbold (1997), West (1996), West and McCracken (1998) and White (2000) (for reviews, see also Mariano, 2002; McCracken and West, 2002).

It is important to note that predictive performance and predictive accuracy depend on two features: first, whether the theoretical model used is close to the unknown data distribution and, second, the ability to estimate accurately model parameters (hence on sample size available for estimating these). For a given sample size, a false but parsimonious model may well have better predictive ability than the ‘true’ model.

model selection : The New Palgrave Dictionary of Economics

http://www.dictionaryofeconomics.com/article?id=pde2008_M00014354

2008-09-18 09:47 AM
**Post-model selection inference**

An important issue often raised in relation with model selection is its effect on the validity of inference – such as estimation, tests and confidence sets – obtained after a process of model selection (or pretesting). This issue is subtle and complex. Not surprisingly, both positive and negative assessments can be found.

On the positive side, it has been observed that pretesting (or model selection) does allow one to produce so-called ‘super-efficient’ (or Hodges) estimators, whose asymptotic variance can be at least as low as the Cramér–Rao efficiency bound and lower at certain points (see Le Cam, 1953). This may be viewed as a motivation for using consistent pretesting.

Furthermore, consistent model selection does not affect the asymptotic distributions of various estimators and test statistics, so the asymptotic validity of inferences based on a model selected according to such a rule is maintained (see Pötscher, 1991; Dufour, Ghysels and Hall, 1994).

On the negative side, it is important to note that these are only asymptotic results. In particular, these are pointwise convergence results, not uniform convergence results, so they may be quite misleading concerning what happens in finite samples (for some examples, see Dufour, 1997; Pötscher, 2002). For estimation, there is a considerable literature on the finite-sample distribution of pretest estimators, which can be quite different of their limit distributions (Judge and Bock, 1978; Danilov and Magnus, 2004). For a critical discussion of the effect of model selection on tests and confidence sets, see Lee and Pötscher (2005).

**Conclusion**

The problem of model selection is one of the most basic and challenging problems of statistical analysis in econometrics. Much progress has been done in recent years in developing better model selection procedures and for understanding the consequences of model selection. But model building remains largely an art in which subjective judgements play a central role. Developing procedures applicable to complex models, which may involve a large number of candidate variables, and allowing for valid statistical inference in the presence of model selection remain difficult issues to which much further research should be devoted.

**See Also**

- Bayesian statistics
- econometrics
- endogeneity and exogeneity
- forecasting
- heteroskedasticity and autocorrelation corrections
- linear models
- models
- serial correlation and serial dependence
- specification problems in econometrics
- statistical decision theory
- statistical inference
- structural change
- testing
- time series analysis

**Bibliography**


**How to cite this article**