Econometric Theory, **21**, 2005, 60–68. Printed in the United States of America.
DOI: 10.1017/S0266466605050048

CHALLENGES FOR ECONOMETRIC MODEL SELECTION

BRUCE E. HANSEN
University of Wisconsin

Standard econometric model selection methods are based on four conceptual errors: parametric vision, the assumption of a true data generating process, evaluation based on fit, and ignoring the impact of model uncertainty on inference. Instead, econometric model selection methods should be based on a semiparametric vision, models should be viewed as approximations, models should be evaluated based on their purpose, and model uncertainty should be incorporated into inference methods. These problems have been examined individually but not jointly, and my view is that future research into econometric model selection should attempt to address all four issues.

1. INTRODUCTION

The theory of econometric model selection is underdeveloped, despite the fact that applied econometrics needs sound model building and selection methods. The state of affairs is illustrated by "A Dialogue Concerning a New Instrument for Econometric Modeling" between Clive Granger and David Hendry (this issue). The dialogue displays and reinforces the following conceptual errors:

(1) parametric vision
(2) assuming a true data generating process
(3) evaluation based on fit
(4) ignoring model uncertainty.

In this note I will discuss these four errors in the context of time-series model selection and suggest alternatives and directions for needed research.

2. SEMIPARAMETRIC MODELING

An econometrician is faced with the time-series observations $\{y_t, x_t\}$. Let I_t denote the set $\{y_t, x_t, y_{t-1}, x_{t-1}, \ldots\}$. The goal is to build a model for (some feature of) the conditional distribution of y_t given I_{t-1}. We will use the term

This research was supported by the National Science Foundation. I thank Peter Phillips and a referee for helpful comments that greatly improved the arguments and exposition. Address correspondence to Bruce E. Hansen, Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706, USA; e-mail: bhansen@ssc.wisc.edu.

  0266-4666/05 $12.00

data generating process (DGP) to denote the true conditional distribution (which may be time-varying) and the label *model* to denote a class of hypothetical DGPs under consideration by the econometrician.

Equation (2) of Granger and Hendry describes PcGets as assuming that the true DGP and the approximating models take the form

$$y_t = \beta' z_t + \varepsilon_t, \tag{1}$$

$$\varepsilon_t | I_{t-1} \sim N(0, \sigma^2), \tag{2}$$

where z_t is a finite subset of the variables in I_{t-1}. This is a parametric structure, for both the DGP and the approximating models.

In contrast, there are several reasonable semiparametric specifications for the distribution of ε_t. The weakest is simply a projection error

$$E(z_t \varepsilon_t) = 0, \tag{3}$$

which, being just-identified, should not be viewed as a model or an assumption, because we can always define β so that (3) holds true. Although this model may be excessively weak, it is worth pointing out that this is one model for which the standard ordinary least squares (OLS) estimator is semiparametrically efficient.

Somewhat stronger than projection is uncorrelatedness, which adds

$$E(x_{t-j} \varepsilon_t) = 0, \tag{4}$$

$$E(y_{t-j} \varepsilon_t) = 0$$

for all $j \geq 1$. This model makes the constructive claim that a finite vector z_t is sufficient to capture the serial correlation in y_t (no further lags of y_t or x_t have nonzero coefficients when included in (1)). It is an overidentified model, and so generically the OLS estimator is asymptotically inefficient relative to the generalized method of moments (GMM) or empirical likeihood (EL). However, the existence of a countably infinite set of moment restrictions makes efficient estimation problematic. Kuersteiner (2002) has made advances toward solving this vexing problem.

A stronger restriction is the assumption of a martingale difference sequence (MDS), which specifies that

$$E(\varepsilon_t | I_{t-1}) = 0. \tag{5}$$

Under this condition (1) can properly be called a regression model, because the index $\beta' z_t$ is the conditional mean of y_t given I_{t-1}.

The models (3)–(5) are all semiparametric. As the distribution of ε_t can rarely be determined by economic theory, these are inherently the natural models for economists. The models are semiparametric (rather than nonparametric) because the focus is on the finite-dimensional parameter β. The distinction between the parametric model (2) and the semiparametric models (3)–(5) is critically rele-

https://doi.org/10.1017/S0266466605050048 Published online by Cambridge University Press

vant to model selection. The PcGets selection method exploits the Gaussian assumption in its choice of testable hypotheses, test statistics, and sampling distributions, and all of these choices change in a semiparametric framework.

The parametric modeling approach influences traditional model selection methods also. The Akaike information criterion (AIC) and Bayesian information criterion (BIC), for example, equal twice the negative log-likelihood of the model plus a parameterization penalty. Although the AIC has useful properties in quasi-likelihood settings it is still focused on a parametric likelihood. A natural alternative is to develop a model selection criterion based on the GMM criterion, the EL criterion, or a similar semiparametric criterion. Steps in this direction have been taken by Andrews and Lu (2001) and Hong, Preston, and Shum (2003). This line of research should be pursued.

3. MODELS AS APPROXIMATIONS

In model selection, we posit a finite class of potential models. In most treatments, the true DGP is assumed to be an element of this set of models. Given this assumption, it is natural to ask the model selection procedure to be *consistent*—to select the true DGP with probability approaching one as the sample size increases. In casual conversation, however, econometricians will typically assert that their models are approximations to the DGP. For example, suppose our set of approximating models is the univariate linear autoregressions $\mathrm{AR}(k)$ for $k = 0,1,\ldots,m$, with the error a MDS. Suppose in contrast the true DGP is an $\mathrm{AR}(\infty)$, for example, a $\mathrm{MA}(1)$ process. Or suppose the true conditional mean is a nonlinear autoregression. In either case, the correct view is that the finite-order $\mathrm{AR}(k)$ model class is an approximating model to the true DGP. In particular, the modeling assumption that the error is a MDS is false. (The error in the $\mathrm{AR}(\infty)$ might be a MDS, but the error in the approximating model is not.) Once we accept that the model is an approximation, the concept of consistent model selection ceases to be important. Rather, the goal should be for the selected model to do a good job of approximating the true DGP.

Furthermore, the true DGP is unlikely to be time-invariant and is more likely to be evolving. Quite simply, change happens. It follows that any time-invariant model must be viewed as an approximation to the evolving DGP. Flexibility can be added to models by allowing for coefficient drift, but even these models are fundamentally time-invariant. (For example, if the AR coefficient follows the random walk $\beta_t = \beta_{t-1} + u_t$ with $u_t \sim (0, \Sigma)$, the time-invariant parameters are β_0 and Σ.)

These points have been recognized in the model selection literature. Model selection based on the BIC is consistent when the true DGP is an element of the assumed model class but tends to underselect when the true DGP is not in this class. In contrast, model selection based on the AIC is inconsistent when the true DGP is an element of the assumed model class but has some optimality properties when viewed as an approximation. See Berk (1974), Lewis and Rein-

https://doi.org/10.1017/S0266466605050048 Published online by Cambridge University Press

sel (1985), and Hannan and Deistler (1988). Phillips (1995) has discussed model determination in the context of evolving DGPs.

An analogy for model selection when viewed as approximations can be taken from nonparametric kernel density estimation. The important choice facing the econometrician is the magnitude of the bandwidth. For a fixed bandwidth, we can think of the kernel estimator as estimating a misspecified model, as the expectation of the estimator is a smoothed version of the true density and thus is the approximating model. One way to reduce this bias is to shrink the bandwidth to zero. However, estimation variance increases as the bandwidth decreases, so an optimal choice of bandwidth is intermediate.

In any finite sample, any nonparametric estimator can be viewed as an estimator of a particular approximating parametric model. What distinguishes the nonparametric approach from the parametric approach is the attitude, interpretation, and distribution theory. The nonparametric approach acknowledges the bias induced by the approximation, whereas the parametric approach ignores the bias. Ideally, we should view models as approximations to the true DGP and adopt the goal of selecting a model that provides the best approximation to the true DGP. Once we adopt this goal, we are forced to ask: What is meant by the "best approximation"? This leads us to our next topic.

4. FOCUSED EVALUATION

If the goal of model selection is to discover the true DGP, we are led to evaluate models based on their fit. In a parametric setting, the natural measure of fit is the log-likelihood, leading to the popular penalized-likelihood criterion, the AIC and BIC. The program PcGets takes a broader view of fit, examining models from multiple angles and specification checks, but still the goal is to find a model that fits the sample distribution well.

However, once we switch to a view that models are approximations, we might wish to focus attention on the dimensions and features of a model that are important to us for our potential application, not on unessential features. Global fit may be important in some applications but not generically. Often, a model is designed and estimated for a purpose. It follows that when we are explicit about the intended purpose, we should design model selection optimally for this purpose.

In many settings, the goal is to measure some feature of the distribution and thus can typically be written as a low-dimensional function of the model parameters. For example, consider the class of linear autoregressive models

$$y_t = \mu + \beta_1 y_{t-1} + \cdots \beta_k y_{t-k} + \varepsilon_t,$$
$$E(\varepsilon_t | I_{t-1}) = 0.$$

Let $\theta = g(\beta)$ be a real-valued parameter of interest. For example, θ may be an individual β_j, the long-run variance, or an impulse response at a particular hori-

https://doi.org/10.1017/S0266466605050048 Published online by Cambridge University Press

zon. Adopting a loss function such as mean-squared error (MSE) we can express the goal of model selection as the desire to find an estimator $\hat{\theta}$ for θ to minimize the expected loss, for example, $E(\hat{\theta} - \theta)^2$. This is a focused criterion, as it is explicit about the purpose of the model.

Indeed, finite-sample optimal model selection can be quite sensitive to the choice of parameter of interest. We illustrate this with a simple example where the approximating models are AR(k) for $k \in \{0,1,\ldots,k_{\max}\}$, yet the true DGP is a Gaussian ARMA(1,1):

$$y_t = \alpha y_{t-1} + \varepsilon_t - \gamma \varepsilon_{t-1},$$

$$\varepsilon_t \sim N(0,1).$$

Because the DGP is equivalent to an AR(∞), all finite-order AR(k) models are approximations. For the parameter of interest, we consider impulse responses, as these are well defined across the approximating models and the DGP. In the true DGP, the mth impulse response is

$$\theta_m = (\alpha - \gamma)\alpha^{m-1}.$$

In the approximating AR(k) models, the impulse response can be calculated recursively from the coefficients $\beta = (\beta_1,\ldots,\beta_k)$.

We explore the issue of optimal model order k using simulation. A hundred thousand samples were drawn from the ARMA DGP for samples of length $n = 200$ using the ARMA parameters

$$(\alpha,\gamma) \in \{-0.9,-0.7,-0.5,-0.3,-0.1,0.1,0.3,0.5,0.7,0.9\}$$

(100 parameterizations). Note that when $\alpha = \gamma$ the model reduces to $y_t = \varepsilon_t$ so these 10 models are identical. Table 1 displays the MSE-minimizing AR order k for the second and sixth impulse response coefficients. There are some major differences in optimal AR order between the two parameters. To take an extreme case, when $\alpha = 0.5$ and $\beta = 0.9$ the optimal AR orders for the second and sixth impulse responses are 10 and 0, respectively!

How can an optimal model be selected? An interesting recent recommendation is the focused information criterion (FIC) of Claeskens and Hjort (2003). Their FIC is an asymptotic estimate of the MSE of the estimator $\hat{\theta}$ of the parameter of interest. Claeskens and Hjort have a general expression for the FIC for a broad class of problems. Consider the special case of the linear model

$$y_t = \mu + \beta' x_t + \varepsilon_t, \tag{6}$$

where x_t is $k \times 1$ and demeaned and the intercept μ is the only parameter that is included in all models. A submodel m of (6) takes the form

$$y_t = \mu + \beta_m' x_{mt} + \varepsilon_t,$$

https://doi.org/10.1017/S0266466605050048 Published online by Cambridge University Press

Table 1. Impulse responses 2 and 6: MSE-minimizing AR order

		β									
		−0.9	−0.7	−0.5	−0.3	−0.1	0.1	0.3	0.5	0.7	0.9
	−0.9	0/0	4/5	4/5	3/5	2/3	2/1	2/1	2/1	2/3	9/4
	−0.7	7/2	0/0	4/3	2/2	2/1	1/2	1/3	4/3	2/3	2/1
	−0.5	8/0	3/0	0/0	1/1	1/1	1/0	1/0	1/0	1/0	2/0
	−0.3	7/0	3/0	0/0	0	1/0	1/0	2/0	3/0	4/0	1/0
α	−0.1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	4/0
	0.1	4/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
	0.3	1/0	4/0	3/0	2/0	1/0	1/0	0/0	0/0	3/0	8/0
	0.5	2/0	1/0	1/0	1/0	1/0	1/0	1/2	0/0	3/0	10/0
	0.7	2/1	2/3	2/3	1/3	1/1	2/1	2/2	4/3	0/0	8/2
	0.9	11/4	2/3	2/1	2/1	2/1	2/2	2/4	3/5	4/5	0/0

where $x_{mt} = S_m' x_t$ and $\beta_m = S_m' \beta$ with S_m a $k \times q$ selector matrix. Let $\hat{\beta} = (X'X)^{-1}(X'Y)$ be the estimator of the full unconstrained model, $\hat{\beta}_m = (X_m' X_m)^{-1}(X_m' Y)$ be the estimator of model m (using standard matrix notation), and $\hat{\sigma}^2$ be the residual variance from the full model.

Recalling that the parameter of interest is $\theta = g(\beta)$, define the $k \times 1$ vector $D = (\partial/\partial\beta) g(\beta)$. Let $\hat{D} = (\partial/\partial\beta) g(\beta^*)$ be an estimator of D using an estimate β^* of β. Claeskens and Hjort (2003) are not specific concerning the choice of β^*. Based on some preliminary simulation results, I suggest using $\beta_m^* = S_m \hat{\beta}_m$, the constrained estimator, which varies across models m. After some manipulations it can be shown that the Claeskens–Hjort FIC for model m is simply

$$FIC_m = (\hat{D}'(\hat{\beta} - S_m \hat{\beta}_m))^2 + 2\hat{\sigma}^2 \hat{D}' S_m (X_m' X_m)^{-1} S_m' \hat{D}.$$

The FIC choice for m is the value that minimizes FIC_m across the models m. The criterion is computationally simple. The only complication over the traditional AIC or BIC is that it involves the partial derivative $\hat{D}$.

Returning to the simulation example introduced previously, 50,000 samples were drawn from each parameterization, the model order selected by the AIC, BIC, and FIC, the selected models estimated. The root MSE for the second and sixth impulse response coefficient was calculated. Typically the AIC had lower MSE than the BIC, so I focus on the comparison between the FIC and AIC. The ratio between their root MSEs is reported in Tables 2 and 3. Values of unity imply equal precision, whereas values over one indicate that the AIC produces a lower MSE, and conversely for values under one.

We see that neither the FIC nor the AIC uniformly dominates the other. For the second impulse response, the FIC-selected model has lower MSE for α close to β, and the AIC-selected model has lower MSE in most other cases. On one

https://doi.org/10.1017/S0266466605050048 Published online by Cambridge University Press

Table 2. Impulse response 2: Root MSE of FIC relative to AIC

		β									
		−0.9	−0.7	−0.5	−0.3	−0.1	0.1	0.3	0.5	0.7	0.9
	−0.9	0.18	1.14	0.98	0.96	1.05	1.04	1.00	1.00	1.01	1.05
	−0.7	1.83	0.18	1.05	1.15	1.15	1.20	1.15	1.17	1.17	1.14
	−0.5	3.17	1.48	0.18	0.81	1.03	1.40	1.55	1.96	2.27	2.21
	−0.3	3.31	2.53	1.13	0.18	0.52	0.77	1.01	1.31	1.54	1.63
α	−0.1	2.28	2.27	1.62	0.81	0.18	0.53	0.92	1.19	1.25	1.20
	0.1	1.14	1.20	1.15	0.88	0.49	0.18	0.84	1.64	2.29	2.28
	0.3	1.61	1.49	1.26	0.96	0.74	0.53	0.18	1.16	2.55	3.3
	0.5	2.30	2.35	2.02	1.61	1.40	1.04	0.84	0.18	1.51	3.15
	0.7	1.23	1.25	1.28	1.25	1.30	1.25	1.23	1.84	0.18	1.83
	0.9	1.07	1.03	1.02	1.02	1.02	1.07	1.01	1.09	1.24	0.18

extreme, for $(\alpha, \beta) = (0.3, -0.9)$, the root MSE of the FIC-selected model is 3.3 times that of the AIC-selected model. On the other extreme, for $\alpha = \beta$, the root MSE of the FIC-selected model is less than one-fifth of that of the AIC-selected model. For the sixth impulse response, the FIC yields lower MSE for most parameterizations. Most notably, for the white noise case $\alpha = \beta$, the root MSE of the FIC estimator is less than $\frac{1}{100}$ of that of the AIC estimator, a dramatic increase in efficiency. The message from Tables 2 and 3 is that the FIC is an intriguing challenger to existing model selection methods and deserves attention and scrutiny.

Table 3. Impulse response 6: Root MSE of FIC relative to AIC

		β									
		−0.9	−0.7	−0.5	−0.3	−0.1	0.1	0.3	0.5	0.7	0.9
	−0.9	0.00	1.61	1.74	0.99	0.95	1.02	1.00	1.03	0.99	0.98
	−0.7	0.48	0.00	0.75	0.97	0.88	0.88	0.97	0.99	1.00	0.98
	−0.5	0.16	0.14	0.00	0.18	0.23	0.49	0.69	0.87	0.93	0.95
	−0.3	0.04	0.03	0.02	0.00	0.02	0.12	0.23	0.41	0.59	0.75
α	−0.1	0.14	0.08	0.04	0.01	0.00	0.01	0.07	0.14	0.29	0.43
	0.1	0.43	0.28	0.14	0.06	0.01	0.00	0.01	0.04	0.08	0.15
	0.3	0.74	0.56	0.39	0.21	0.10	0.10	0.00	0.02	0.03	0.05
	0.5	0.93	0.91	0.84	0.64	0.46	0.24	0.18	0.00	0.13	0.16
	0.7	0.98	0.99	0.98	0.96	0.86	0.87	1.00	0.77	0.00	0.47
	0.9	1.00	1.00	1.02	1.00	1.03	0.95	1.03	1.80	1.54	0.00

https://doi.org/10.1017/S0266466605050048 Published online by Cambridge University Press

Unfortunately, the Claeskens–Hjort FIC has several limitations. One is that their derivation is for parametric (likelihood) problems. The FIC estimates the variance of β_m by $\hat{\sigma}^2(X_m'X_m)^{-1}$, which is generally invalid. As discussed in Section 2, it would be useful to generalize their formula to semiparametric estimators. Second, their analysis is limited in scope because they use a technical assumption that the true model is contained in the set of models under consideration. Although they argue that the latter assumption is not critical to the use of the FIC in practice, for econometric applications it would be constructive to more fully investigate the effects of model approximation, as discussed in Section 3.

5. MODEL UNCERTAINTY

The act of model selection can have unwanted distributional implications. For example, Pötscher (1991) shows that the distributions of estimators and test statistics are dramatically affected by the act of model selection. Ideally, these distortions from standard theory should not be ignored in inference. Unfortunately, incorporating the effect of model selection on inference methods is very challenging, and effective methods have yet to be developed. To illustrate the difficulty of the problem, Leeb and Pötscher (2003a, 2003b) show that no estimator of either the unconditional or conditional distribution of the post-model-selection estimator can be uniformly consistent. Methods to successfully surmount this obstacle are an important topic for future research.

From another angle, it is possible to argue that model selection itself is a misguided goal. It is quite common to find that confidence intervals from different plausible models are nonintersecting, raising considerable inferential uncertainty. Fundamentally, the uncertainty concerning the choice of model is not reflected in conventional asymptotic and bootstrap confidence intervals.

Although the problem is obvious, the solution is not. One proposal is the method of Bayesian model averaging, which has grown in interest over the past decade. Basically, the relevant models are estimated, and then the posteriors are averaged. See Hoeting, Madigan, Raftery, and Volinsky (1999) for a review. Unfortunately, as with all Bayesian methods, there are many arbitrary decisions regarding priors and unfortunate paradoxes involving parameter transformation, rendering practical use of their methods difficult. In a recent contribution, Hjort and Claeskens (2003) propose a frequentist form of model averaging. This is a welcome addition and should be pursued as a viable supplement to model selection methods.

6. CONCLUSION

My recommendations: Econometric DGPs should be semiparametric, not parametric; models should be viewed as approximations, and econometric theory should take this seriously; model selection criteria should ideally be based on

https://doi.org/10.1017/S0266466605050048 Published online by Cambridge University Press

the purpose of the model, not on a global measure of fit; finally, the impact of model selection on inference should not be ignored. Unfortunately, although each of these goals has been studied and solutions proposed, we currently do not have concrete model selection and inference methods that simultaneously achieve all four goals. Considerable theoretical work will be required to merge and extend the existing theory and methods. My personal view is that this difficult task is well worth the effort, as achieving the goal could be of great value to applied econometric practice.

REFERENCES

Andrews, D.W.K. & B. Lu (2001) Consistent model and moment selection criteria for GMM estimation with applications to dynamic panel models. *Journal of Econometrics* 101, 123–164.

Berk, K.N. (1974) Consistent autoregressive spectral estimates. *Annals of Statistics* 2, 489–502.

Claeskens, G. & N.L. Hjort (2003) The focused information criterion. *Journal of the American Statistical Association* 98, 900–916.

Granger, C. & D. Hendry (2005) A dialogue concerning a new instrument for econometric modeling. *Econometric Theory* (this issue).

Hannan, E.J. & M. Deistler (1988) *The Statistical Theory of Linear Systems*. Wiley.

Hjort, N.L. & G. Claeskens (2003) Frequentist model average estimators. *Journal of the American Statistical Association* 98, 879–899.

Hoeting, J.A., D. Madigan, A.E. Raftery, & C.T. Volinsky (1999) Bayesian model averaging: A tutorial. *Statistical Science* 14, 382–417.

Hong, H., B. Preston, & M. Shum (2003) Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory* 19, 923–943.

Kuersteiner, G.M. (2002) Efficient instrumental variables estimation for autoregressive models with conditional heteroskedasticity. *Econometric Theory* 18, 547–583.

Leeb, H. & B.M. Pötscher (2003a) Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators? Working paper, Department of Statistics, University of Vienna.

Leeb, H. & B.M. Pötscher (2003b) Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators? Working paper, Department of Statistics, University of Vienna.

Lewis, R. & G. Reinsel (1985) Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* 16, 393–411.

Phillips, P.C.B. (1995) Econometric model determination. *Econometrica* 64, 763–812.

Pötscher, B.M. (1991) The effect of model selection on inference. *Econometric Theory* 7, 163–185.