University of Illinois
Fall 1998

Department of Economics
Roger Koenker

Economics 472
Lecture 4

Model Selection and Fishing for Significance

1. Model Selection

Classical hypothesis testing plays a central role in econometrics, but in many applied problems we face a preliminary stage of the analysis in which we need to make decisions about model specification. These decisions are not very well formalized in terms of classical hypothesis testing, and gradually specialized procedures have been developed for this under the rubric "model selection." In this lecture I will try of these procedures and relate them to more classical notions of hypothesis testing.

The framework for model selection can be described as follows. We have a collection of parametric models

$$\{f_i(x, \theta)\}$$

where $\theta \in \Theta_j$ for $j = 1, \ldots, J$. Some linear structure is usually imposed on the parameter space, so typically $\Theta_j = m_j \cap \theta_J$, where m_j is a linear subspace of $\Re^{p_J}$ of dimension p_j and $p_1 < p_2 < \ldots < p_J$. To formally justify some of our subsequent connections to hypothesis testing it would be also necessary to add the requirement that the models are *nested*, i.e., that $\theta_1 \subset \theta_2 \subset \ldots \subset \theta_J$.

Akaike (1970) was the first to offer a unified approach to the problem of model selection. His point of view was to choose a model from the set $\{f_i\}$ which performed well when evaluated on the basis of forecasting performance. His criterion, which has come to be called the Akaike information criterion is,

$$AIC(j) = l_j(\hat{\theta}) - p_j$$

where $l_j(\hat{\theta})$ the log likelihood corresponding to the j^{th} model maximized over $\theta \in \Theta_j$. Akaike's model selection rule was simply to maximize AIC over the j models, that is to choose the model j^* which maximizes $AIC(j)$. This approach seeks to balance improvement in the fit of the model, as measured by the value of the likelihood, with a penalty term, p_j. Thus one often sees this and related procedures referred to as penalized likelihood methods. The trade-off is simply: does the improvement which comes inevitably from expanding the dimensionality of the model compensate for the increased penalty?

Subsequent work by Schwarz (1978) showed that while the AIC approach may be quite satisfactory for selecting a forecasting model it had the unfortunate property that it was inconsistent, in particular, as $n \to \infty$, it tended to choose too large a model with positive probability. Schwarz (1978) formalized the model selection problem from a Bayesian standpoint and showed that as

$n \to \infty$, the modified criterion/footnote Unless otherwise specified, all my logs are natural, i.e., base e.

$$SIC(j) = l_j(\hat{\theta}) - \frac{1}{2} p_j \log n$$

had the property that, presuming that there was a true model, j^*, then $\hat{j} = \text{argmax } S(j)$, satisfied

$$p(\hat{j} = j^*) \to 1.$$

Note that since $\frac{1}{2} \log n > 1$ for $n > 8$, the SIC penalty is larger than the AIC penalty, so SIC tends to pick a smaller model. In effect, by letting the penalty tend to infinity slowly with n, we eliminate the tendency of AIC to choose too large a model.

How does this connect with classical hypothesis testing? It can be shown, in my 476 for example, that under quite general conditions for nested models, that

$$2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i)) \rightsquigarrow \chi^2_{p_j - p_i}$$

for $p_j > p_i = p^*$. That is, when model i is true, and model $p_j > p_i$, twice the log likelihood ratio statistic is approximately χ^2 with degrees of freedom equal to the difference in the parametric dimension of the two models. So classical hypothesis testing would suggest that we should reject an hypothesized smaller model i, in favor of a larger model j iff

$$T_n = 2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))$$

exceeds an appropriately chosen critical value from the $\chi^2_{p_j - p_i}$ table. In contrast Schwarz would choose j over i, iff

$$\frac{2(l_j - l_i)}{p_j - p_i} > \log n$$

The fraction on the left hand side of this inequality may be interpreted as the numerator of an F statistic. Under $H_0: \ j^* = i$, it is simply a χ^2 divided by its degrees of freedom which is an F with $p_j - p_i$ numerator degrees of freedom and ∞ denominator degrees of freedom. Thus, $\log n$ can be interpreted as an implicit critical value for the model selection decision based on SIC.

Does this make sense? Why would it be reasonable to let the critical value tend to infinity? We are used to thinking about fixed significance levels like 5% or 1%, and therefore about fixed critical values, but a little reflection suggests that as $n \to \infty$ we might like to have α, the probability of Type I error, bend to zero. This way we could arrange that *both* Type I and Type II error probabilities tend to zero simultaneously. This is the practical consequence of the Schwarz connection between sample sizes and α-levels based on the SIC choice.

Note that AIC uses a fixed critical value of 2, in contrast to SIC, and this is an immediate explanation of why with positive probability it picks too large a model. Unless the critical value tends to infinity with n, there will always be a positive probability of a Type I error.

1.1. SIC in the linear regression model. Recall that for the Gaussian linear regression model

$$l(\beta, \sigma) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{\mathcal{S}}{2\sigma^2}$$
$$\text{where} \quad \mathcal{S} = (y - X\beta)'(y - X\beta)$$

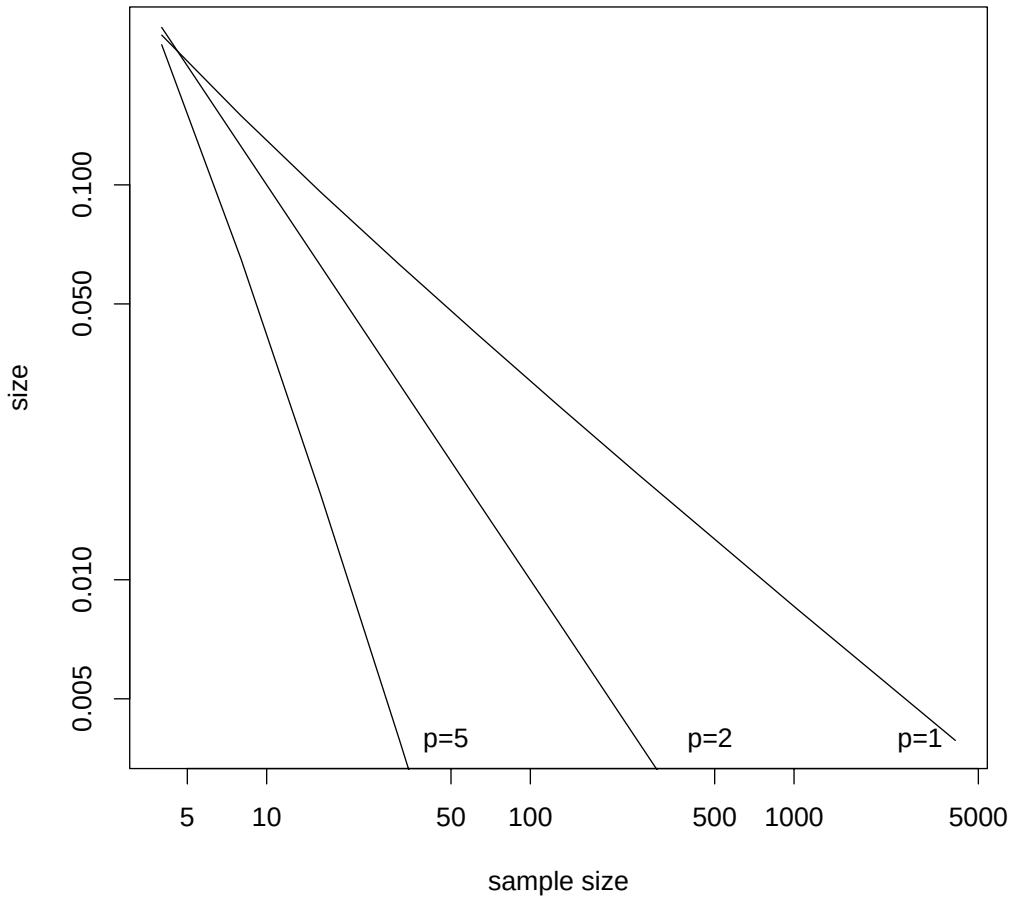

FIGURE 1. Effective Significance Level of SIC Criterion: The figure illustrates the implied significance level of using the Schwarz Criterion for Model Selection in linear regression. In the figure p refers to the number of parameters under consideration, so for example with one parameter considered for deletion, the effective level α of the Schwarz "test" is about .05 at $n = 100$ and about .01 at $n = 1000$.

Evaluating at $\hat{\beta}$, and $\hat{\sigma}^2 = \mathcal{S}/n$ we get

$$l(\hat{\beta}, \hat{\sigma}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\hat{\sigma}^2$$

Thus, maximizing SIC

$$l_i - \frac{1}{2}p_i \log(n)$$

is equivalent to minimizing

$$\frac{n}{2}\log\hat{\sigma}_j^2 + \frac{1}{2}p_i\log n$$

or minimizing,

$$\log\hat{\sigma}_j^2 + (p_j/n)\log n.$$

In statistical packages one needs to be careful to check exactly what is being computed before reporting such numbers as SIC.

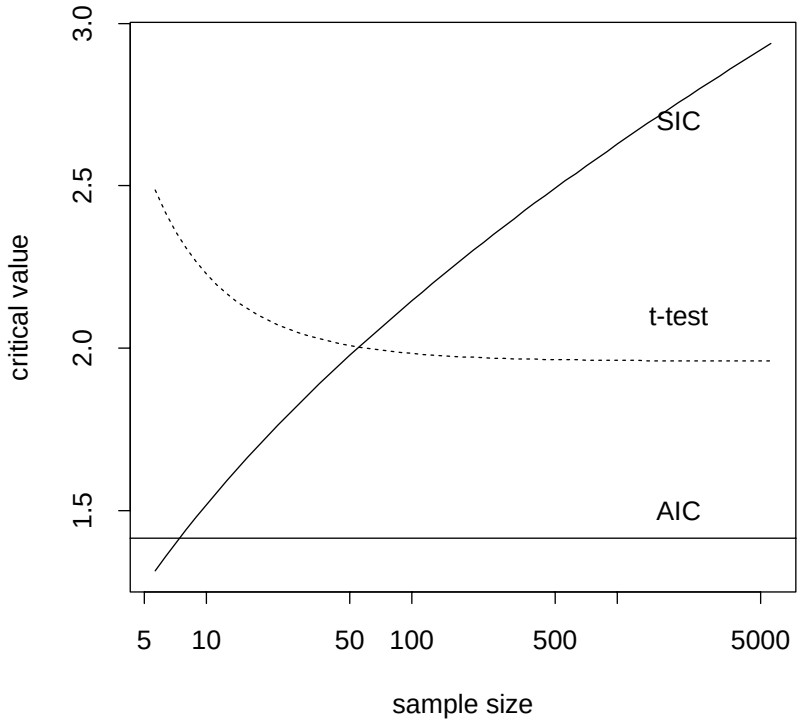

FIGURE 2. Comparison of effective critical value for model selection using SIC, AIC, and conventional t-test: The figure illustrates the implied critical values for SIC and AIC model selection in linear regression for the case of adding a single variable to the regression.

How does this connect to the F test in regression? We "know" that there is generally a close connection between F and LR tests, but how does this work in regression? Note,

$$\begin{aligned} l_i - l_j &= \frac{n}{2}(\log \hat{\sigma}_j^2 - \log \hat{\sigma}_j^2) \\ &= \frac{n}{2}\log(\hat{\sigma}_j^2/\hat{\sigma}_i^2) \\ &= \frac{n}{2}\log\left(1 - \frac{\hat{\sigma}_i^2 - \hat{\sigma}_j^2}{\hat{\sigma}_i^2}\right) \end{aligned}$$

and using the usual Taylor-series approximation for $\log(1 \pm a)$ for a small we have

$$2(l_i - l_j) \approx \frac{n(\hat{\sigma}_j^2 - \hat{\sigma}_i^2)}{\hat{\sigma}_i^2}.$$

Dividing the right hand side by $p_j - p_i$ yields the usual F statistic.

As a final remark, we might observe that in the case that $p_j - p_i = 1$ so we are only considering adding one variable to the regression, we can relate the SIC and AIC rules to conventional hypothesis testing in the following simple way. Recall that in the case of a single linear restriction in the regression the F statistic is simply the square of the the corresponding t statistic. Thus, in the case of the conventional regression t-test, SIC implicitly proposes the critical value, $\sqrt{\log(n)}$ while the AIC uses $\sqrt{2}$. Note that the latter is quite lenient, but this is perhaps reasonable if the final intent is forecasting. Note also that the classical two-sided critical value for the t-test, illustrated by the dotted line, converges to the familiar number 1.96, and crosses the SIC curve at about sample size $n = 50$. In contrast the AIC selection criterion is fixed at $\sqrt{2}$ and thus is much more lenient than either of the other procedures in accepting new covariates.

2. Fishing for Significance

The second part of this lecture concerns the difficulties associated with preliminary testing and model selection from the point of view of eventual inference about the selected model. This is an old topic which has received considerable informal attention but it is rather rare to find serious formal consideration of it. My discussion will be based largely on Freedman (1983).

Freedman, early in his career, was a leading light in probability theory and wrote several fundamental books on Markov Chains. Later, he began to take an interest in matters more applied and statistical in nature. One of his earlier ventures in this direction was a project to evaluate the swarm of "energy models" which emerged from the 1973 oil shock. These were models which purported to "explain" energy demand and how we might control it.

Freedman's model of energy models is highly stylized, and mildly ironic. He presumes a model of the form

$$(*) \qquad y_i = x_i\beta_0 + u_i$$

which u_i iid $\mathcal{N}(0, \sigma^2)$. The matrix $X = (x_i)$ is n by p and satisfies $X'X = I_p$. And $p \to \infty$ as $n \to \infty$ so that $p/n \to \rho$ for some $0 < \rho < 1$. That is, as the sample size grows the modeler introduces new explanatory variables in such a way that the ratio p/n tends to a constant. Further, he assumes that $\beta_0 = 0$.

Theorem 1: For model (*), $R_n^2 \to \rho$ and $F_n \to 1$.

Proof: The usual F_n statistic for the model, since $\beta_0 = 0$, is really distributed as F so $EF_n = (n-p)/(n-p-2)$ which tends to 1. However, recall that

$$F_n = \frac{n-p-1}{p} \cdot \frac{R_n^2}{1-R_n^2}$$

so

$$R_n^2 = F/\left(\frac{n-p-1}{p} + F\right)$$

and thus since $F \to 1$ we have that $R_n^2 \to \rho$.

This result is rather trivial and is just a warm up for a more interesting question which really reveals David Freedman's model for energy economists. Consider the following sequential estimation strategy: all p variables are tried initially, those attaining α-level of significance in a standard t-test are retained, say, $q_{n,\alpha}$, of them, then the model is reestimated with only these variables. Let $R^2_{n,\alpha}$ and $F_{n,\alpha}$ denote the R^2 and F statistics for this second stage regression.

Theorem 2: For model (*) $R^2_{n,\alpha} \to g(\lambda_\alpha)$ and $F_{n,\alpha} \to \left(\frac{g(\lambda_\alpha}{\alpha}\right) / \left(\frac{1-g(\lambda)\rho}{1-\alpha\rho}\right)$ where

$$g(\lambda) = \int_{|g|>\lambda} z^2 \phi(z) dz$$

and λ is chosen so $\Phi(\lambda = 1 - \alpha/2$.

Example: Suppose $n = 100, p = 50$, so $\rho = 1/2$. Set $\alpha = .25$ so $\lambda = 1.15$, and $g(\lambda) = .72$ then

$$E(Z^2 |\ |z| > \lambda) \approx 2.9$$
$$R^2_{n,\alpha} \cong g(\lambda) \approx .72$$
$$F_{n,\alpha} \cong \left(\frac{g(\lambda}{\alpha}\right)$$
$$\frac{(1-g(\lambda)\rho)}{(1-\alpha\rho)} \approx 4.0$$
$$Eq_{n,\alpha} = \alpha\rho n = .25 \cdot .50 \cdot 100 \approx 12.5$$
$$F_{12,88,.05} = 1.88$$
$$P(F_{12,88} > 4.0) \equiv .0001 \qquad \square$$

Proof of Theorem 2 is really good exercise for 476. For purposes of 472 the example is sufficient to warn you that the consequences of preliminary testing are serious and you need to adjust your expectations and significance levels in light of such activity. I'll say a little more about this when we talk about the bootstrap.

References

Freedman, D. (1983) A note on screening regression equation, *American Statistician*, 37, 152-56.

Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, 6, 461-64.