University of Illinois | Department of Economics
Spring 2017 | Econ 574 | Roger Koenker

Lecture 19
"Model Selection"

Often in formulating econometric models we face a preliminary stage of the process in which we must select a preferred parametric model from among a large diverse class of alternatives. This problem bears some similarity to classical hypothesis testing but in many respects is rather different. In this lecture, I will describe an approach due to Schwarz (1978) which captures the essentials of the problem from a Bayesian decision theory point of view and has been found to be extremely useful in practice. In the process we will have a chance to revisit some basic ideas from the first few lectures.

Suppose that we have a collection of models of the exponential family form:

$$f_j(x, \theta) = \exp\{\theta' y(x) - b(\theta)\}$$

where $\theta \in \Theta_j$ for $j = 1, \ldots, J$. The exponential family imposes some linear structure on the problem, but does not seem to be absolutely essential. See, e.g., Machado (1993). Suppose that model j restricts Θ by imposing the condition $\Theta_j = m_j \cap \Theta_J$ where m_j is a linear subspace of dimension p_j in $\Re^{p_J}$, where $p_1 < p_2 < \ldots < p_J$.

We will assume that the investigator has a prior of the form

$$\pi(\theta) = \sum \alpha_j \mu_j(\theta)$$

where α_j is the prior probability that model j is correct and $\mu_j(\theta)$ is the prior distribution (measure) on θ, *given model j is correct*. We will assume μ_j are bounded and locally bounded away from zero on $m_j \cap \Theta$, so $\mu_j(\theta)$ puts positive mass on any open subset of $m_j \cap \Theta = \Theta_j$.

The posterior for θ is then, for iid observations, proportional to

$$\sum \alpha_j \exp\{n(\bar{y}'\theta - b(\theta))\}\mu_j(\theta)$$

where $\bar{y} = n^{-1} \sum_{i=1}^n y(x_i)$ is the sufficient statistic for θ. An interesting preliminary question is: How would we forecast with this family of models? The Bayesian approach to forecasting in this circumstance would lead us to combine the forecasts of the several models according to the posterior probabilities of the models.

But the model selection problem is rather different. In model selection we seek to select one, of several, models and plan to ignore the others in subsequent activities. This is probably theoretically unsound, but practically very convenient – we don't want to be burdened by carrying along a big collection of models - we would like one which adequately represents our statistical best judgement about the problem.

To represent this version of the problem more explicitly, Schwarz assumes that among the collection of models there is a "true" one, say, j_0, and that we face the decision problem of choosing an estimate of j_0 subject the $0-1$ loss

$$\mathcal{L}(\hat{j}, j_0), = \begin{cases} 0 & \text{if } \hat{j} = j_0 \\ 1 & \text{otherwise} \end{cases}$$

This leads, as in other contexts, to the rule

$$\hat{j} = \arg\max\{S_n(\bar{y}, j) = \log \int_{m_j \cap \Theta} \alpha_j \exp\{n(\bar{y}'\theta - b(\theta))\}\mu_i(\theta)d\theta$$

which chooses the j which maximizes the posterior probability among all available models. Note that the integral in the above expression is proportioned to the posterior update of the prior probabilities α_j attached to the various models.

Theorem For fixed $\bar{y}, j$ as $n \to \infty$

$$(*) \qquad S_n(\bar{y}, j) = n \ sup_\theta(\bar{y}\theta - b(\theta)) - \frac{1}{2}p_j \log n + R_n(\bar{y}, j)$$

where $R_n(\bar{y}, j) = O(1)$.

Remark Since the first 2 terms in the above expression are growing with n, it is clear that the remainder is asymptotically negligible and therefore we have the criterion

$$\max_j\{\ell_j(\hat{\theta}) - \frac{1}{2}p_j \log n\}$$

where $\ell_j(\hat{\theta})$ denotes the log likelihood of the j^{th} model at the maximum likelihood estimate for that model and p_j is the dimension of the parameter space for the j^{th} model, i.e., $p_j = \dim\{\Theta_j\}$.

We begin by proving the result in a very restrictive special case.

Lemma 1 If $\bar{y}'\theta - b(\theta) = A - \lambda \parallel \theta - \theta_0 \parallel^2$ for some $\lambda > 0, \theta_0 \in m_j$ and $\mu_j(\theta) = \mu$, i.e., Lesbesgue measure on m_j then $(*)$ holds.

Proof Note that

$$Q = \int \alpha_j \exp\{n(A - \lambda \parallel \theta - \theta_0 \parallel^2)\}d\mu = \alpha_j e^{nA} \int \exp\{-n\lambda \parallel \theta - \theta_0 \parallel^2\}d\mu$$

but this integral is proportional to a p_j dimensional normal density with covariance matrix Ω such that

$$\frac{1}{2}\Omega^{-1} = n\lambda I_{p_j}$$

so

$$\Omega = (2n\lambda)^{-1} I_{p_j} \equiv \omega I_{p_j}$$

and $|\Omega| = \omega^{p_j}$ so,

$$\begin{aligned} \int \exp\{-n\lambda \parallel \theta - \theta_0 \parallel^2\}d\mu &= (2\pi)^{p_j/2}|\Omega|^{1/2} \cdot \int (2\pi)^{-p_j/2}|\Omega|^{-1/2} \exp\{\ \}d\mu \\ &= (2\pi)^{p_j/2}(2n\lambda)^{-p_j/2} = (\pi/n\lambda)^{p_j/2} \end{aligned}$$

and therefore

$$Q = \alpha_j e^{nA}(\pi/n\lambda)^{p_j/2}.$$

But of course

$$\sup_{\theta}\{A - \lambda \parallel \theta - \theta_0 \parallel^2\} = A$$

so we have

$$S_n(\bar{y}, j) = nA - \frac{1}{2}p_j \log n + R_n$$

where $R_n = 1/2p_j \log(\pi/\lambda) + \log \alpha_j$ ■

Example In the Gaussian linear model we may write,

$$\parallel y - Xb \parallel^2 = n\hat{\sigma}^2 + (b - \hat{\beta})'(X'X)(b - \hat{\beta})$$

where $\hat{\sigma}^2 = n^{-1} \parallel y - X\hat{\beta} \parallel^2$ and $\hat{\beta} = (X'X)^{-1}X'y$. Thus, here $\hat{\sigma}^2$ plays the role of A, but the norm for $(b - \hat{\beta})$ is not just simply Euclidean. Of course, we can always reparameterize the model so it *would* be Euclidean.

What does the lemma really say? The crucial conclusion to draw from the lemma is the fact that the only asymptotically nonnegligible effect of the prior is the penalty term $1/2p_j \log n$ everything else from the prior, the α_j, the details of $\mu_j(\theta)$ if it were to depend upon θ vanishes without trace because it is all $O_p(1)$ where as the likelihood and the penalty are n-dependent.

Now we extend the domain if relevance of the lemma to more general and realistic situations by establishing two further lemmas.

Lemma 2 If two bounded positive r.v.s U, V agree on the set where either exceeds ρ for some ρ such that $0 < \rho < \sup U$ then

$$\log EU^n - \log EV^n \to 0 \qquad \text{as } n \to \infty.$$

Proof Without loss of generality, suppose that for $U < \rho, V = 0$ so

$$0 \leq U^n - V^n \leq \rho^n$$

Thus

$$EV^n \leq EU^n \leq EV^n + \rho^n = EV^n \left(1 + \frac{\rho^n}{EV^n}\right)$$

Then, since

$$\lim_{n\to\infty} (EV^n)^{1/n} \to \sup V = \sup U > \rho$$

we have

$$\frac{\rho}{(EV^n)^{1/n}} < 1 \Rightarrow \frac{\rho^n}{EV^n} \to 0$$

so

$$\log\left(1 + \frac{\rho^n}{EV^n} \to 0\right)$$ ■

Lemma 3 For some ρ such that $0 < \rho < e^A$, $A = \sup\{y'\theta - b(\theta)\}$, a vector θ_0, and some $\underline{\lambda}, \bar{\lambda}$, if $\exp\{y'\theta - b(\theta)\} > \rho$, then

$$A - \underline{\lambda} \parallel \theta - \theta_0 \parallel^2 < y'\theta - b(\theta) < A - \bar{\lambda} \parallel \theta - \theta_0 \parallel^2 .$$

Proof It is well known that $B(\theta) = \nabla^2 b(\theta)$ is the covariance matrix of y, hence $y'\theta - b(\theta)$ is globally strictly concave and therefore has a unique maximum, say θ_0. Expanding at θ_0, we have

$$F(\theta) = y'\theta - b(\theta) = A - \frac{1}{2}(\theta - \theta_0)' B(\theta^*)(\theta - \theta_0)$$

where $\theta^* = \omega\theta + (1-\omega)\theta_0$ and $\omega \in (0,1)$. Since $x'Ax$ can be bounded by

$$\underline{\lambda}_A \parallel x \parallel^2 \leq x'Ax \leq \bar{\lambda}_A \parallel x \parallel^2$$

we can bound $F(\theta)$ for $N_{\theta_0,M} = \{\theta | \parallel \theta - \theta_0 \parallel < M\}$ where $\underline{\lambda}_A$ and $\bar{\lambda}_A$ are taken as twice the maximum and minimum eigenvalues of $B(\theta^*)$ for θ^* in $N_{\theta_0,M}$.

We may now prove the theorem.

Proof By the local boundedness of $\mu_j(\theta)$, there exists μ_0 such that

$$\exp\{y'\theta - b(\theta)\} d\mu_j(\theta) \geq \exp\{y'\theta - b(\theta)\} d\mu_0.$$

Now let $U = \exp\{\bar{y}'\theta - b(\theta)\}$, choose ρ as in Lemma 3; then for $U > \rho$ by Lemma 3 we have

$$A - \underline{\lambda} \parallel \theta - \theta_0 \parallel^2 < \log U < A - \bar{\lambda} \parallel \theta - \theta_0 \parallel^2$$

for all $\theta \in N_{\theta_0,M}$. Now,

$$EU^n = E \exp\{n(y'\theta - b(\theta))\}$$

and satisfies

$$EV^n < EU^n < EV^n + \rho^n = EV^n \left(1 + \frac{\rho^n}{EV^n}\right) \to EV^n$$

for $\log V = A - \underline{\lambda} \parallel \theta - \theta_0 \parallel^2$. So by Lemma 2, $\log EU^n - \log EV^n \to 0$ ■

Remark In effect we have reduced the general case to the special one of Lemma 1 with λ equal twice the maximum likelihood of $B(\theta)$ in a neighborhood of θ_0. Note in the Gaussian case $B(\theta)$ is independent of θ.

Practicalities of Model Selection

In Akaike (1970) it was suggested that model selection for the purpose of forecasting could be based upon maximizing the criterion,

$$AIC = \ell_j(\hat{\theta}) - p_j.$$

However, the work of Schwarz (1978) shows that as $n \to \infty$,

$$P(j^*_{AIC} > j_0) \nrightarrow 0$$

while if instead we use

$$j^*_{SIC} = \arg\max_j \ell_j(\hat{\theta}) - \frac{1}{2} p_j \log n$$

then

$$P(j^*_{SIC} = j_0) \to 1$$

Since for $n > 8, 1/2 \log n > 1$, Schwarz's SIC tends to select a smaller, more parsimonious, model than does AIC.

It may be helpful to try to relate these model selection criteria to conventional asymptotic theory of hypothesis testing. We have seen in Lecture 11, that under quite general conditions, for $\Theta_j > \Theta_i$ we have

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

SIC says "reject i in favor of the bigger model j" if

$$\ell_j - \ell_i > \frac{1}{2}(p_j - p_i) \log n$$

i.e., if

$$\frac{2(\ell_j - \ell_i)}{p_j - p_i} > \log n.$$

The fraction on the left hand side of this inequality may be seen as the numerator of an F statistic. Under H_0 it is a χ^2 divided by its degrees of freedom. So we can regard $\log n$ as a critical value for an F-test to choose i vs j, in which the denominator degrees of freedom are taken as infinite.

What is happening to the critical value as $n \to \infty$? Clearly, as $n \to \infty$ the α-level of the test (the probability of a Type I error) is tending to zero. Is this reasonable? We are used to thinking about fixed significance levels for tests, like 5%, or 1%, but a little reflection suggests that for $n \to \infty$, we should try to choose $\alpha \to 0$ so that both Type I and Type II error probabilities tend to zero. This is the consequence of the implicit $\log n$ critical value of this critical value for various sample sizes. Not unreasonably, we select a much more stringent α-level for tests in the range $n > 1000$, then we might traditionally consider reasonable.

Akaike, H. (1974). A new look at the statistical identification model, *IEEE Transactions on Automatic Control*, 19, 716-723.

Machado, J.A.F. (1993). Robust model selection and M-estimation, *Econometric Theory*, 9, 478-93.

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

The Lasso as a Model Selection Device

Tibshirani (1996) introduced the "Lasso" as a method of regularization for least squares regression problems, and Chen, Donoho and Saunders (1998) discussed very similar methods in the context of selection of basis functions in models with highly overparameterized basis selection problems. This form of regularization is closely related to the total variation penalties that we have already considered in function estimation.

I'll only briefly mention an application of this ℓ_1 penalty approach to model selection in quantile regression. For further details see Li and Zhu (2008). Consider the penalized quantile regression problem:

$$\min \sum \rho_\tau(y_i - x_i^\top \beta) - \lambda \|\beta\|_1.$$

It is relatively easy to solve this problem since it can be formulated as a QR problem with augmented data. In my quantreg package one can simply use the `method = lasso` option for

`rq`. The advantage of the ℓ_1 penalty over the Gaussian ℓ_2 penalty is that it tends to shrink some coefficients all the way to zero rather than gradually shrink all the coefficients a small amount. There is still the obvious question about how to choose λ, but there is a relatively simple suggestion that seems to work reasonably well in practice. The effective dimension of the model selected for any choice of λ can be defined as,

$$p(\lambda) = \sum \partial \hat{y}_i / \partial y_i$$

which in the QR situation is simply the number of observations interpolated by the fitted function. Li and Zhu suggest that this quantity can then be used in conjunction with the usual SIC or GCV criterion, so for example, one can find λ to minimize:

$$SIC(\lambda) = \log(n^{-1} \sum \rho_\tau(y_i - x_i^\top \beta)) + p(\lambda) \log n/(2n).$$

and show that this performs quite well in simulations relative to minimizing the (infeasible) mean absolute deviation. (This suggestion was made earlier by Koenker, Ng and Portnoy in the context of TV penalized function estimation.

A nice example of how the lasso works in applications, suggested by Donoho and Candes, is the following secret decoder ring problem:
Problem: Transmit $x \in \mathbb{R}^n$ over a noisey channel.

Encoding: Send $y = Ax$ for $A \in \mathbb{R}^{m \times n}$, $m \gg n$, and receive either:

$$\tilde{y} = Ax + u$$

$$\tilde{y} = Ax + u + v$$

where $u \sim$ Gaussian, and v is (sparsely) arbitrarily bad.

Decoding: Set $Q = I - A(A^\top A)^{-1} A^\top$ and do either:

$$\hat{x} = (A^\top A)^{-1} A^\top \tilde{y}$$

$$\hat{x} = (A^\top A)^{-1} A^\top (\tilde{y} - \tilde{v})$$

where: $\tilde{v} = \text{argmin}\{\|v\|_1 \text{ such that } \|Q(\tilde{y} - v)\|_\infty \leq K\}$ The former is termed Gaussian decoding, since it simply uses least squares fitting, while the latter is referred to as Dantzig decoding since it relies on the linear programming solution $\tilde{v}$.

As a numerical example suppose, $n = 256$, $m = 512$, and entries in x, u and A are iid standard Gaussian, and let v be the mixture: $v_i = 0.9\delta_0 + 0.1\delta_{-2y_i}$. Then the next figure illustrates the performance of of three methods: The Gaussian decoding is seriously disrupted by the noise introduced by the v_i, but Dantzig decoding is almost as good as the performance of "an oracle" that knew the values of the v_i and was able to remove their effect.

Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso, *JRSS(B)* 58, 267-288.

Chen, S. Donoho, D. and Saunders, M. (1998) Atomic Decomposition by Basis Pursuit, SIAM J. of Scientific Computing, 20, 33-61.

Koenker, R. Ng, P. and Portnoy, S. (1994) Quantile Smoothing Splines, *Biometrika*, 81,673-680.

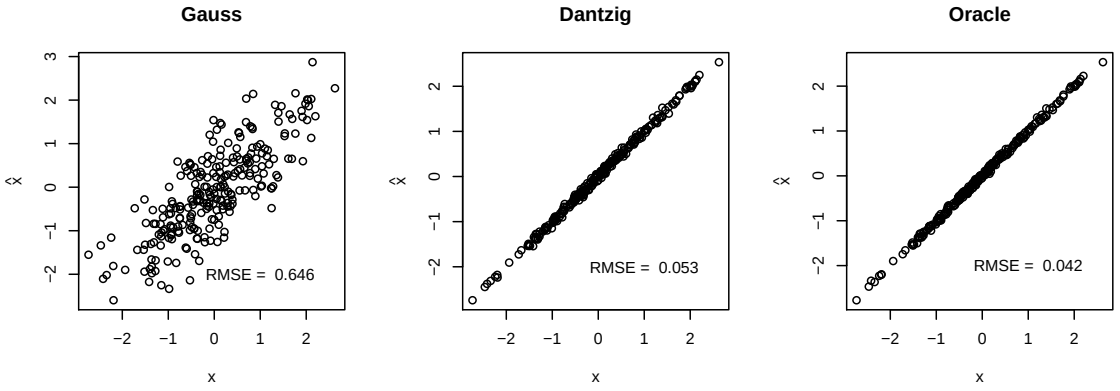

Figure 1: Dantzig decoding (2) achieves almost the same accuracy as if v were known.

Li, Y. and Zhu, J. (2008) L_1 norm quantile Regression, *J. of Comp. and Graphical Statistics*, 17, 163-185.

A Postscript on Screening Regressions, or "Fishing for Fun and Publication"

Freedman (1983) is a rather disturbing paper about the effect of preliminary testing, or model selection, on the validity of subsequent inference. Consider the very simple linear model,

$$(M) \qquad y = X\beta + \mu \qquad\qquad u \sim \mathcal{N}(0, \sigma^2 I)$$

Suppose X is a $n \times p$ matrix satisfying $X'X = I_p$ and suppose that as $n \to \infty, p \to \infty$ such that $p/n \to \rho$ for some $0 < \rho < 1$. Let M_0 denote the version of M in which $\beta = 0$.

Theorem 1 Under M_0, $R_n^2 \to \rho$ and $F_n \to 1$.

Now consider a two step estimator for model M, in which we first estimate the model, to a preliminary screening for "significant variables" and then reestimate using only these variables. Suppose the screening uses level α, denote the critical value for a two-tailed test at level α by λ, and define

$$g(\lambda) = \int_{|z|>\lambda} z^2 d\Phi(z).$$

Theorem 2 Let $q_{n,\alpha}$ be the number of parameters estimated in Stage 2, and $R^2_{n,\alpha}$ and $F_{n,\alpha}$ the R^2, and F statistic for that regression then, under M_0,

$$\begin{aligned} q_{n,\alpha}/n &\to \alpha\rho \\ R^2_{n,\alpha} &\to g(\lambda) \\ F_{n,\alpha} &\to \frac{g(\lambda)}{\alpha} \Big/ \frac{(1-g(\lambda)\rho)}{(1-\alpha\rho)} \end{aligned}$$

Proofs given in Freedman are a bit sketchy and the computations are rather tedious, so I won't provide details.

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

$$R^2_{n,\alpha} \approx .72 \qquad \text{(quite respectable looking!)}$$

$$q_{n,\alpha} \approx \alpha\rho n = \frac{1}{4}\,\frac{1}{2} \cdot 100 = 12.5$$

$$F_{n,\alpha} \approx 4.0 \qquad \text{scary!}$$

$$P[F_{12,88} > 4] \approx .0001$$

These asymptotic predictions are supported by some very limited Monte Carlo. What they say is quite disturbing. Starting from a model in which "nothing matters" we can, in only two sequential testing steps, reach a conclusion which appears to show overwhelming evidence of a significant effect. I have thought for some time that this would be a useful topic for a thesis.

Freedman, D. (1983). A Note on screening regressions, *American Statistician*, 37, 152-155.