University of Illinois
Fall 2016

Department of Economics
Roger Koenker

Economics 536
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 describe two 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 (1969) 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[1]

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

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 574 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

[1]Unless otherwise specified, all my logs are natural, i.e., base e.

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.

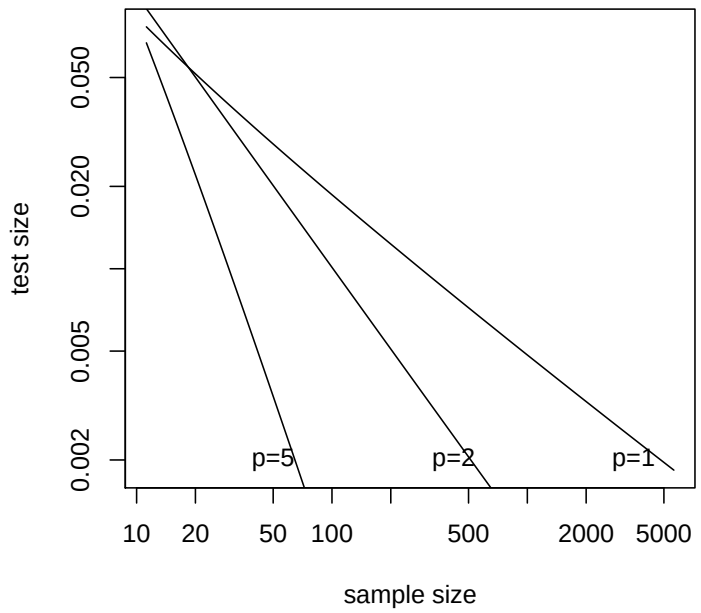

FIGURE 1. Effective Significance Level of SIC Criterion: The figure illustrates the implied significance level of using the Schwarz Criterion for Model Selection in linear regression. In the figure p refers to the number of parameters under consideration, so for example with one parameter considered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

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

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

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. In R, there is a generic function, `AIC` that can be used for most maximum likelihood fitting; it defines the AIC value as $-2\ell_i + kp_i$, so we are *minimizing* not maximizing and the value of the parameter k determines the nature of the penalty. By default $k = 2$, but it can be set to another value at the user's discretion.

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,

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

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

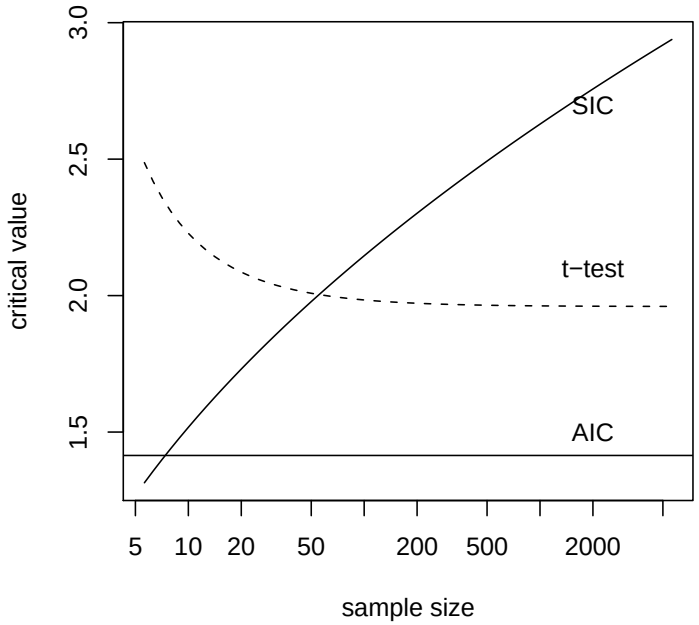

FIGURE 2. Comparison of effective critical value for model selection using SIC, AIC, and conventional t-test: The figure illustrates the implied critical values for SIC and AIC model selection in linear regression for the case of adding a single variable to the regression.

at $\sqrt{2}$ and thus is much more lenient than either of the other procedures in accepting new covariates.

2. Model Selection, Shrinkage and the Lasso

The basic idea of the information criterion approach to model selection is to penalize the likelihood by some function of the parametric dimension of the model in an effort to balance the two objectives of parsimony, or simplicity, of the model and goodness of fit, or fidelity. A variety of fidelity and penalty criterion have been suggested.

One way to think about this balance is that combines prior information about the problem with the information contained in the current data. The standard paradigm for doing this is Bayesian updating. Recall that Bayes Theorem asserts that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

It is easy to remember this since the proof is obvious from the usual definition of conditional probability for discrete events. Of course, we need to be cautious when $P(B) = 0$. My usual illustration of this is the following Economics 506 problem:

Q. On a trip from Tomsk to Urbana I changed planes in Moscow, London and Chicago. The probabilities of a missed baggage transfer at these cities was .4, .2 and .1 respectively. Given that my luggage was missing in Urbana, what is the probability that the loss occurred at in Moscow?

A. Let $m_i : i = 1, 2, 3$ denote the simple unconditional loss at each of the three cities, and $M_i : i = 1, 2, 3$ the events that a loss occurs at city i for my trip. Thus:

$$\begin{aligned}
P(M_1) &= 0.4 \\
P(M_2) &= P(m_2 \cap \bar{M}_1) \\
&= P(m_2|\bar{M}_1)P(\bar{M}_1) \\
&= 0.2 \cdot (1 - 0.4) \\
&= 0.12 \\
P(M_3) &= P(m_3 \cap \bar{M}_2 \cap \bar{M}_1) \\
&= P(m_3|\bar{M}_2 \cap \bar{M}_1)P(\bar{M}_2|\bar{M}_1)P(\bar{M}_1) \\
&= 0.1 \cdot 0.8 \cdot 0.6 \\
&= 0.048.
\end{aligned}$$

Now,

$$P(M_1|M_1 \cup M_2 \cup M_3) = \frac{P(M_1)}{P(M_1) + P(M_2) + P(M_3)}$$

since the M_i events are mutually exclusive, so,

$$P(M_1|M_1 \cup M_2 \cup M_3) = \frac{0.4}{0.4 + 0.12 + 0.048} \approx 0.704.$$

In regression settings we have a similar situation. We need to combine the information provided by the data, via the likelihood, with whatever prior information we might have from prior experience or introspection.

A theme of the course will be the trade-off between bias and variance in model selection problems: too simple a model risks serious bias that may distort policy conclusions of the model, too complicated a model risks obscuring the important effects in a cloud of uncertainty. Until now we have tried to balance these risks by selecting a model that represents a compromise between our objectives. This requires a slightly schizophrenic viewpoint. On one hand we appear to believe that there are many possible models for our problem, but in the end only one will be taken seriously. (This is rationalized by Schwarz's 0-1 loss function, but often this all or nothing view of models isn't appropriate.)

We will now consider a new approach to reaching a compromise between simple and complex models. We will begin with a brief exposition of Bayesian methods for linear regression, in non-Bayesian statistical circles these methods are sometimes referred to as "shrinkage methods," or Stein-rule methods. After briefly discussing some connections to random effects models for panel data we will then consider non-parametric regression.

Consider (once more) the linear model,

$$y = X\beta + u$$

If we assume (as usual) that $u \sim \mathcal{N}(0, \sigma^2 I)$, then we have likelihood,

$$\mathcal{L}(y|b) = (2\pi)^{-n/2}\sigma^{-n}\exp\{-\frac{1}{2\sigma^2}(\hat{\beta} - b)'X'X(\hat{\beta} - b)\}$$

Suppose that we have a prior opinion that $\beta \sim \mathcal{N}(\beta_0, \Omega)$, i.e. that β has density

$$\pi(b) = (2\pi)^{-p/2}|\Omega|^{-1/2}\exp\{-\frac{1}{2}(b - \beta_0)'\Omega^{-1}(b - \beta_0)\}$$

Bayes rule says that we should update our prior opinion about β, to obtain,

$$p(b|y) = \frac{\mathcal{L}(y|b) \cdot \pi(b)}{\int \mathcal{L}(y|b)\pi(b)db}.$$

Focusing on the $\exp\{\cdot\}$ term in the numerator we obtain, after some algebra,

$$p(b|y) = \kappa\exp\{\frac{1}{2}(b - \tilde{\beta})'(\sigma^{-2}X'X + \Omega^{-1})(b - \tilde{\beta})\}$$

where κ is a constant independent of b and

$$\tilde{\beta} = (\sigma^{-2}(X'X) + \Omega^{-1})^{-1}(\sigma^{-2}(X'X)\hat{\beta} + \Omega^{-1}\beta_0).$$

This result shows that the posterior distribution in this very simple setting is also Gaussian, and has mean $\tilde{\beta}$. We can elaborate on this quite a lot by, for example treating σ as a parameter on which we also have a prior opinion rather than effectively assuming it is known as we have here. But we will postpone this line of inquiry.

The final formula is yet another application of a general strategy for combining two estimates: we have $\hat{\beta}$ and β_0 and they have covariance matrices $\sigma^2(X'X)^{-1}$ and Ω respectively. They are combined accordingly – weighted by the inverses of the covariance matrices – provided the normality assumptions are reasonable. We will see several other examples of this combination phenomenon later in the course.

Some special cases to consider:

(1) When $\sigma^2 \to 0$ then the likelihood dominates the prior.
(2) When $\sigma^2 \to \infty$ then the prior dominates the likelihood.
(3) As $n \to \infty$, typically we assume $n^{-1}X'X \to D$ a positive definite matrix, so $\sigma^{-2}(X'X) \to n\sigma^{-2}D$ and the factor n causes the likelihood to dominate as n grows.

(4) Suppose

$$\Omega = \begin{bmatrix} w_0 I_p & 0 \\ 0 & w_1 I_q \end{bmatrix}$$

then by varying w's we get some interesting cases. E.g. $w_0 = \infty$ would express the view that we were clueless about the first p elements, but we might still want to shrink the last q elements towards their prior mean.

Some Practicalities

If we diagonalize $\Omega^{-1} = Q'Q$, then

$$\lambda Q\beta \sim \mathcal{N}(\lambda Q\beta_0, \lambda^2 Q\Omega Q') \sim \mathcal{N}(Q\beta_0, \lambda^2 I)$$

and we can now write the model augmented by the prior information as

$$\begin{pmatrix} y \\ \lambda Q\beta_0 \end{pmatrix} = \begin{pmatrix} X \\ \lambda Q \end{pmatrix} \beta + \begin{pmatrix} u \\ v \end{pmatrix}$$

where the vector $(u, v)' \sim \mathcal{N}(0, \tilde{\Omega})$ where

$$\tilde{\Omega} = \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \lambda^2 I \end{bmatrix}.$$

and we can "run" this regression to get $\tilde{\beta}$ or "run" several of these regressions to get the whole "decolletage", $\tilde{\beta}(\lambda)$. Another special case is $\Omega = I$ so $Q = I$; this is often called ridge regression. All coefficients are shrunken toward zero.

Dimensions and Penalties

A natural question at this point would be: What is the relationship between the AIC/SIC approach and the Bayesian/Ridge approach. Is there some way to consider the shrinkage procedure of the latter as producing models of "reduced dimension?" And if so, can we provide an explicit way to measure the dimension of the shrunken model? To briefly consider this let's adopt the simplest form of the Gaussian setup considered above so $\beta_0 = 0$ and we will scale the prior covariance matrix as $\Omega = \lambda\Omega_0$, thereby parameterizing the "contract curve in Figure 3 by λ.

In this case our Bayes estimator produces the "fitted" values,

$$\begin{aligned} \hat{y} &= X\hat{\beta} \\ &= X(\sigma^{-2}X^\top X + \Omega^{-1})^{-1}\sigma^{-2}X^\top X\hat{\beta} \\ &= X(\sigma^{-2}X^\top X + \Omega^{-1})^{-1}\sigma^{-2}X^\top X(X^\top X)^{-1}X^\top y \\ &= X(\sigma^{-2}X^\top X + \Omega^{-1})^{-1}X^\top y/\sigma^2 \\ &\equiv Ly \end{aligned}$$

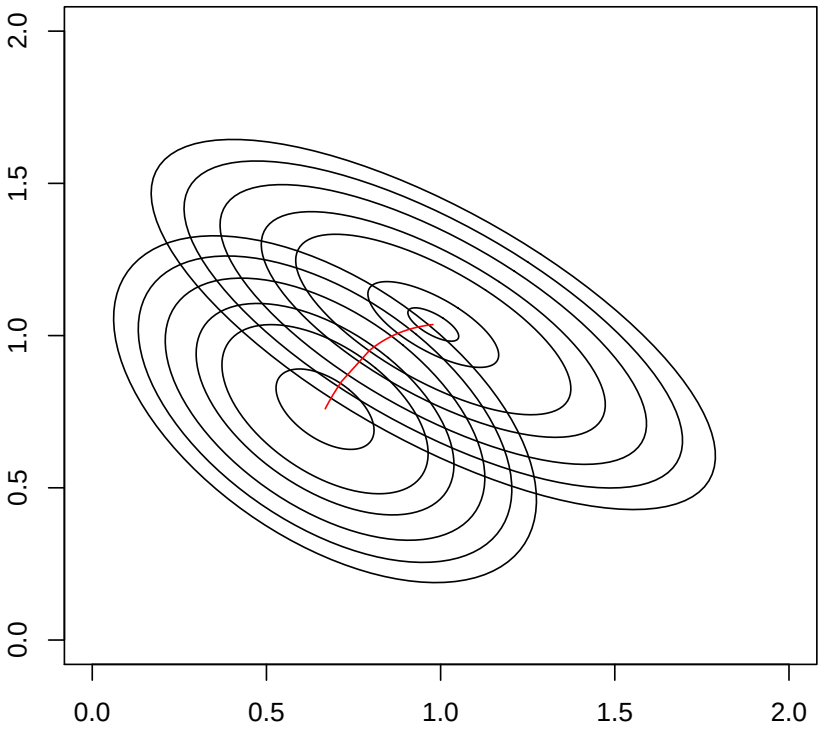

FIGURE 3. Bayesian Regression: Contours of the likelihood and prior are illustrated by the two families of concentric ellipses. The locus of tangencies between the contours of the prior and the contours of the likelihood represent a shrinkage path that connects the mle and the prior mean. Each point on the path can be interpreted as representing an intensity of belief in the prior. And each of these points then represents a posterior mean, $\tilde{\beta}$ reflecting a particular intensity of belief in the prior.

So we have a linear estimator.[2] Now suppose that that the prior was very vague, i.e. uninformative about β so $L = P_X = X(X^\top X)^{-1}X^\top$, then the

[2]There are many uses of the word "linear" that we will encounter. We have the linear model, it can be linear in variables, i.e. covariates, or linear in parameters. Then there are linear estimators, that is the present situation and simply means that the parameter estimates, and therefore the fitted values, $\hat{y}$ are a linear function of the observed response vector, y.

(hopefully familiar!) computation,

$$\mathrm{Tr}(P_X) = \mathrm{Tr}((X^\top X)^{-1}(X^\top X)) = \mathrm{Tr}(I_p) = p$$

tells us that the parametric dimension of the classical linear model is determined by simply summing up the diagonal elements of the usual projection matrix. Now, L is *not* a projection matrix, it is not idempotent, nevertheless, it has become standard practice to use $\mathrm{Tr}(L)$ as a measure of the dimension of such models. We will revisit this relationship at the end of the course when we consider penalty methods for nonparametric smoothing problems.

Other forms of Shrinkage: The Lasso and TV Smoothing

A recent suggestion by Tibshirani (1996) also considered by Donoho and Chen (1998) and Koenker, Ng and Portnoy (1994) replaces the AIC/BIC penalty terms, by the ℓ_1 norm of the estimated parameter vector

$$\mathrm{Pen}(\theta) = \sum_{i=1}^{p} |\theta_i|.$$

Thus, for the leading example of least squares regression, Tibshirani proposes solving

$$\min_\theta \sum (y_i - x_i'\theta)^2 + \lambda \mathrm{Pen}(\theta)$$

for some appropriately chosen λ. He calls this the lasso, for least absolute shrinkage and selection operator. This form of the penalty has the effect of "shrinking" the vector $\hat{\theta}$ toward the origin, but unlike the more conventional ℓ_2 shrinkage penalty also known as "ridge regression" or Stein estimation, the ℓ_1 penalty tends to shrink many of the coordinates of θ all the way to 0, while the ℓ_2 penalty tends to shrink each of the coordinates a little way toward 0, as in the next figure.

Why does the Lasso produce something that behaves much more like model selection while ridge regression with the Gaussian prior not do so? This is illustrated in 2 dimensions in the next figure: the diamond shaped prior contours of the ℓ_1 penalty make it much more likely that we will get corner solutions as "tangents" and this will zero out certain coefficients.

Another variant of this approach is to use the ℓ_1 fidelity criterion,

$$\min \sum |y_i - x_i'\theta| + \lambda \mathrm{Pen}(\theta)$$

This has been studied by several authors, most recently by Wang, Li, and Jiang (*JBES*, 2007, 347-355). The variant studied in Koenker, Ng and Portnoy (1994) is a non-parametric version of this. A crucial question with all such "shrinkage" methods is: how are we to select λ? There are many suggestions, none authoritative. Cynics might suggest that the lasso just replaces an old problem, model selection, by a new problem, λ-selection, however this isn't quite fair. For one thing, the choice of a λ drastically reduces the scope of the combinatorial problem of choosing a set of covariates.

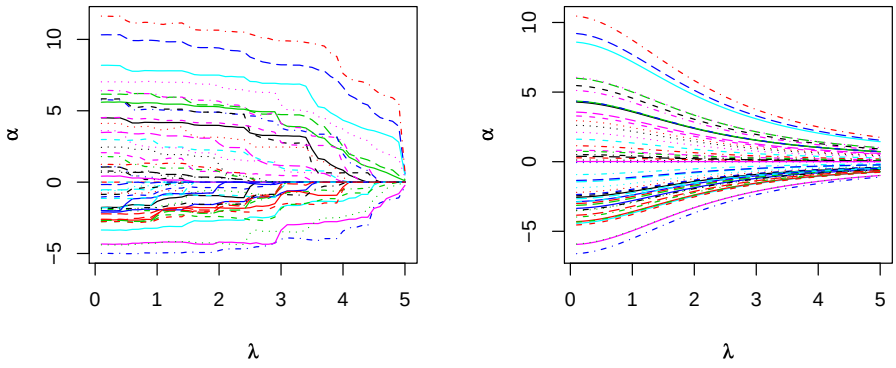

FIGURE 4. Lasso vs Ridge Shrinkage

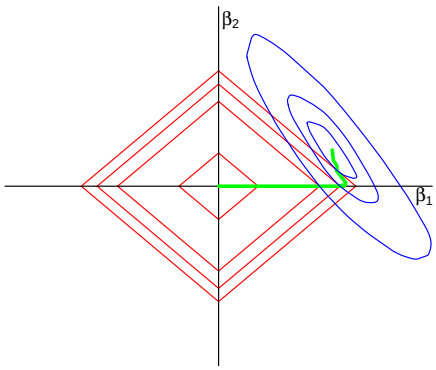

FIGURE 5. Lasso Shrinkage path

Yet another variant is the so–called Dantzig Selector of Candès and Tao (2006?).

$$\min_{\theta}\{\text{Pen}(\theta) \;\; s.t. \parallel X'(y - X\theta) \parallel_{\infty} \leq \lambda\}.$$

At first this looks quite different.[3] But note that the constraint ensures that we are close to a solution to the least squares problem

$$\min \parallel y - X\theta \parallel_2^2$$

so it isn't really *that* different than (*) above.

A nice example of how the Dantzig selector works is the following "toy" communication problem studied in Candès and Randall (2006).

Problem 0: A message $x \in \Re^n$ needs to be transmitted through a noisy channel. If the noise is Gaussian we may consider the following strategy: Let A be a $m \times n$ matrix, $m \gg n$, with (approximately) orthogonal columns. We transmit $Ax \in \Re^m$ and $y = Ax + u$ is received with u iid Gaussian. The receiver then solves the least squares problem

$$\hat{x} = \arg\min_x \parallel y - Ax \parallel_2^2$$

This solution has strong optimality properties under Gaussian assumptions. In effect, using the longer message, y instead of x allows us to remove much of the noise introduced by transmission. But what if the noise isn't Gaussian?

Problem 1: Now suppose we modify the problem so that there is an additional error component, instead of receiving y we receive $\tilde{y} = y + \nu$ where ν is sparse, i.e., most coordinates are 0, but otherwise arbitrary. Consider solving the Dantzig Selection problem

$$\hat{\nu} = \arg\min\{\parallel \nu \parallel s.t. \parallel M_A(\tilde{y} - \nu) \parallel_\infty < K\}$$

where $M_A = I - A(A'A)^{-1}A'$. Given $\hat{\nu}$ we can estimate y, by $\hat{y} = \tilde{y} - \hat{\nu}$ and compute

$$\begin{aligned} \hat{x} &= \operatorname{argmin} \parallel \hat{y} - Ax \parallel_2^2 \\ &= (A'A)^{-1}A'\hat{y} \end{aligned}$$

It turns out that this is *almost* as good as knowing y in the first place as the following figure illustrates.

3. Omitted Variable Bias and Irrelevant Variable Variance Inflation

Much of statistical theory (and therefore practice) boils down to intelligent treatment of bias-variance tradeoffs. (This presumes that science brings to the table problems that are sufficiently well specified so as to make this tradeoff clear, much of what econometricians, and statisticians *do* involves struggling with this specification stage.)

Consider the following stylized situation in regression in which we want to compare the models,

$$y = X\beta + Z\gamma + u \qquad \text{(long-model)}$$

[3]Here $\parallel x \parallel_\infty = \max_i |x|$

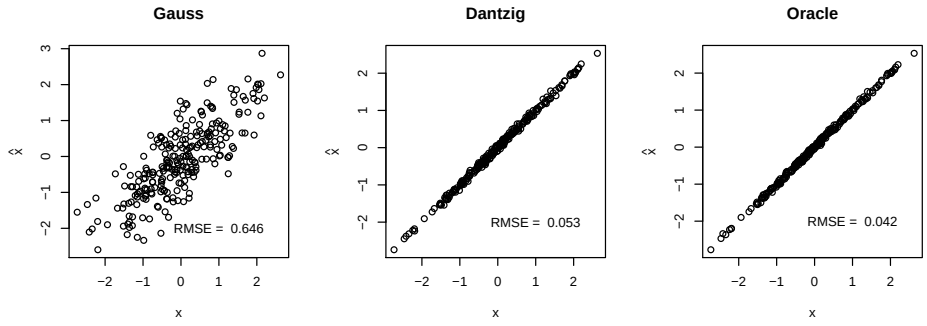

FIGURE 6. The Dantzig Selector Decoding: Estimated values of the signal x plotted versus the true values for three procedures: Gauss ignores the ν component, Dantzig uses the Dantzig Selector, and Oracle removes ν as if there were no contamination component.

and

$$y = X\beta + v \qquad \text{(short-model)}$$

If we assume that the long model is true and we estimate the short model instead we have committed the sin of "omitted variables" and we must pay the price in terms of bias in the estimate the parameter vector, β.

Taking expectations, conditional on $[X\vdots Z]$ we have

$$\begin{aligned} E\hat{\beta}_S &= E(X'X)^{-1}X'y \\ &= E(X'X)^{-1}X'(X\beta + Z\gamma + u) \\ &= \beta + (X'X)^{-1}X'Z\gamma \end{aligned}$$

and we find that the bias associated with estimation of β using the short regression is

$$G\gamma = (X'X)^{-1}X'Z\gamma$$

where the matrix G is what would be obtained by regressing the columns of Z on the columns of X. Note that this bias vanishes if $\gamma = 0$, or if X is orthogonal to Z.

3.1. **Example.** An important example of this so-called omitted variable bias, one that is relevant to the problem set on gasoline demand concerns the bias of estimating a static model when a dynamic one is really called for. Suppose the correct specification is

$$y_t = \alpha + \sum_{i=0}^{p} \beta_i x_{t-i} + u_t$$

where x_t is (scalar) exogenous variable. In our naïvety we estimate instead the static model,

$$y_t = \alpha + \beta_0 x_t + v_t$$

What can we say about the relationship between our estimate of β_0 in the static model and the coefficients of the dynamic model?

At first sight you might expect that the static model estimate, $\hat{\beta}_0$, should estimate the "impact effect" of changes in x_t, that is the coefficient β_0. However, our omitted variable bias expression looks like this

$$E\hat{\beta}_0 = \beta_0 + \sum_{i=1}^{p} g_i \beta_i$$

where g_i is slope coefficient of the obtained in a regression of x_{t-i} on x_t, and an intercept. If x_t is strongly trended, then these g_i will tend to be close to one and consequently $E\hat{\beta}_0$ will be close to $\sum_{i=0}^{p} \beta_i$, that is the long-run effect estimated from the dynamic model. Thus in the gasoline data you should not be surprised to find that the long run elasticities estimated in your dynamic specifications are not too far removed from the elasticities you get from the simple static form of the model. Of course, this conclusion depends crucially on the g_i's, so in other applications the situation could be quite different.

What about the opposite case? Suppose the short model is correct, but we mistakenly estimate the long model instead. Now, there is no bias problem

$$\begin{aligned} E\hat{\beta}_L &= E(X'M_Z X)^{-1} X'M_Z y \\ &= E(X'M_Z X)^{-1} X'M_Z (X\beta + u) \\ &= \beta \end{aligned}$$

However, there *is* a price to be paid for our extravagance of estimating the parameters, γ, when we didn't need to; this price comes in the form of variance inflation in the estimate of β. To see this we need to do some computations.

Another variant of the omitted variable bias result stated above is that,

Prop. $\hat{\beta}_S = \hat{\beta}_L + G\hat{\gamma}_L$

Pf.

$$\begin{aligned} \hat{\beta}_S &= (X'X)^{-1} X'y \\ &= (X'X)^{-1} X'(\hat{y}_L + \hat{u}_L) \\ &= (X'X)^{-1} X'(X\hat{\beta}_L + Z\hat{\gamma}_L) \\ &= \hat{\beta}_L + G\hat{\gamma}_L \square \end{aligned}$$

And consequently we have

Prop. Assuming $V(y) = E(y - Ey)(y - Ey)^\top = \sigma^2 I$, $V(\hat{\beta}_L) = V(\hat{\beta}_S) + GV(\hat{\gamma}_L))G'$

Pf. Writing, $\hat{\beta}_L = \hat{\beta}_S - G\hat{\gamma}_L$ And noting that

$$\text{Cov } (\hat{\beta}_S, \hat{\gamma}_L) = E(\hat{\gamma}_L - \gamma)(\hat{\beta}_S - \beta) = E(Z^\top M_X Z)^{-1} Z^\top M_X u u^\top X (X^\top X)^{-1} = 0$$

the result follows immediately. □

The result implies that the variability of the long estimate always exceeds the variability of the short estimate. Note that the variability effect is somewhat more benign than the prior bias effect in the sense that it tends to zero as the sample size grows, whereas the bias effect persists irrespective of the sample size. Now, if we admit that neither of our extreme cases holds, that $\gamma \neq 0$, and $G \neq 0$, then the choice between $\hat{\beta}_S$ and $\hat{\beta}_L$ becomes one of trying to carefully weigh the relative magnitudes of the bias and variance. For this we would need to be more explicit about the loss function describing the cost of making an error in estimation of β.

4. Fishing for Significance

The last 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). For a more technical version of some of the same material see Leeb and Pötscher(2005).

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

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

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

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

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

$$\begin{aligned}
&E(Z^2|\ |z| > \lambda) \approx 2.9 \\
&R^2_{n,\alpha} \cong g(\lambda) \approx 0.72 \cdot 0.5 \approx 0.36 \\
&F_{n,\alpha} \cong \left(\frac{g(\lambda}{\alpha}\right) \\
&\frac{(1-g(\lambda)\rho)}{(1-\alpha\rho)} \approx 4.0 \\
&Eq_{n,\alpha} = \alpha\rho n = .25 \cdot .50 \cdot 100 \approx 12.5 \\
&F_{12,88,.05} = 1.88 \\
&P(F_{12,88} > 4.0) \approx .0001 \qquad \square
\end{aligned}$$

Proof of Theorem 2 is a really good exercise for 574. 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

Akaike, H. (1969) Fitting autoregressive models for prediction, *Annals of Institute of Stat. Math*, 21, 243-7.

Candès, E. Tao, T. (2007) The Dantzig selector: Statistical estimation when p is much larger than n ANNALS OF STATISTICS, 35, 2313-2351.

Candès EJ and P. Randall. (2009) Highly robust error correction by convex programming. IEEE Trans. Inform. Theory, 54 2829-2840. ...

Chen SS, DL Donoho, MA Saunders, (2001) Atomic Decomposition by Basis Pursuit. SIAM Review 43, 129- 139.

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

Koenker, R. and Ng, P. and Portnoy, S., (1994) Quantile smoothing splines, Biometrika, 81, 673-680,

Leeb, H. and Pötscher, B. W. (2005) Model Selection and Inference: Facts and Fiction, *Econometric Theory*, 21, 21-59.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., 58, 267-288).

Wang, H and Li, G and Jiang, G, 2007. "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso," Journal of Business & Economic Statistics, 25, 347-355.