Lecture 2: Various Model Selection Criteria

Francis J. DiTraglia

March 4, 2014

1 The Corrected AIC

To derive the TIC and AIC we used asymptotic theory to derive an analytical bias correction. These approximations tend to work well as long n is fairly large relative to p but when this is not the case, they can break down. We'll now consider an alternative that makes stronger assumptions and relies on *exact* small-sample theory rather than asymptotics: the "Corrected" AIC, or AIC_c, of Hurvich and Tsai (1989). Suppose that the true DGP is a linear regression model:

$$\mathbf{y} = X\beta_0 + \boldsymbol{\epsilon}$$

where $\epsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T)$. Then $\mathbf{y}|X \sim N(X\beta_0, \sigma_0^2 \mathbf{I}_T)$ so the likelihood is

$$g(\mathbf{y}|X; \beta_0, \sigma_0^2) = \left(2\pi\sigma_0^2\right)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta_0)'(y - X\beta_0)\right\}$$

and the log-likelihood is

$$\log\left[g(\mathbf{y}|X; \beta_0, \sigma_0^2)\right] = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_0^2) - \frac{1}{2\sigma_0^2}\left(\mathbf{y} - X\beta_0\right)'\left(\mathbf{y} - X\beta_0\right)$$

Now suppose we evaluated the log-likelihood at some *other* parameter values β_1 and σ_1^2. The vector β_1 might, for example, correspond to dropping some regressors from the model by setting their coefficients to zero, or perhaps

adding in some additional regressors. We have

$$\log[f(\mathbf{y}|X;\beta_1,\sigma_1^2)] = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_1^2) - \frac{1}{2\sigma_1^2}(\mathbf{y}-X\beta_1)'(\mathbf{y}-X\beta_1)$$

Since we've specified the density from which the data were generated as well as the density of the approximating model, we can *directly calculate* the KL divergence rather than trying to find a reasonable large sample approximation. It turns out that for this example

$$KL(g;f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} - \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) - 1\right] + \left(\frac{1}{2\sigma_1^2}\right)(\beta_0-\beta_1)'X'X(\beta_0-\beta_1)$$

as you will demonstrate on the problem set. We need to estimate this quantity for it to be of any use in model selection. If let $\widehat{\beta}$ and $\widehat{\sigma}^2$ be the maximum likelihood estimators of β_1 and σ_1^2 and substitute them into the expression for the KL divergence, we have

$$\widehat{KL}(g;f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\widehat{\sigma}^2} - \log\left(\frac{\sigma_0^2}{\widehat{\sigma}^2}\right) - 1\right] + \left(\frac{1}{2\widehat{\sigma}^2}\right)\left(\beta_0-\widehat{\beta}\right)X'X\left(\beta_0-\widehat{\beta}\right)$$

We still have two problems. First, we haven't been entirely clear about what β_1 and σ_1 are. At the moment, they seem to be something like "pseudo-true" values. Second, and more importantly, we don't know β_0 and σ_0^2 so we can't use the preceding expression to compare models.

Hurvich and Tsai (1989) address both of these problems with the assumption that all models under consideration are *at least correctly specified*. That is, while they may include a regressor whose coefficient is in fact zero, they do not exclude any regressors with a non-zero coefficient. This is the same assumption that we used above to reduce TIC to AIC. Under this assumption, β_1 and σ_1^2 *are precisely the same* as β_0 and σ_0^2. More importantly, we can use all of the standard results for the exact finite sample distribution of regression estimators to help us. The idea is to construct an *unbiased* estimator of the

KL divergence. Taking expecations and rearranging slightly, we have

$$\begin{aligned} E\left[\widehat{KL}(g;f)\right] &= \frac{T}{2}\left\{E\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] - \log(\sigma_0^2) + E\left[\log(\widehat{\sigma}^2)\right] - 1\right\} \\ &\quad + \frac{1}{2}E\left[\left(\frac{1}{\widehat{\sigma}^2}\right)\left(\widehat{\beta}-\beta_0\right)X'X\left(\widehat{\beta}-\beta_0\right)\right] \end{aligned}$$

Now, under our assumptions $T\widehat{\sigma}^2/\sigma_0^2 \sim \chi^2_{T-k}$ where k is the number of estimated coefficients in $\widehat{\beta}$. Further, if $Z \sim \chi^2_\nu$ then $E[1/Z] = 1/(\nu-2)$. It follows that

$$E\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] = E\left[\frac{T}{T\widehat{\sigma}^2/\sigma_0^2}\right] = \frac{T}{T-k-2}$$

We can rewrite the final term similarly:

$$E\left[\left(\frac{1}{\widehat{\sigma}^2}\right)\left(\widehat{\beta}-\beta_0\right)X'X\left(\widehat{\beta}-\beta_0\right)\right] = E\left[\left(\frac{T}{T\widehat{\sigma}^2/\sigma_0^2}\right)\frac{\left(\widehat{\beta}-\beta_0\right)X'X\left(\widehat{\beta}-\beta_0\right)}{\sigma_0^2}\right]$$

Under our assumptions the two terms in the product are independent, so we can break apart the expectation. First, we have

$$E\left[\frac{T}{T\widehat{\sigma}^2/\sigma_0^2}\right] = \frac{T}{T-k-2}$$

as above. For the second part,

$$\frac{\left(\widehat{\beta}-\beta_0\right)X'X\left(\widehat{\beta}-\beta_0\right)}{\sigma_0^2} \sim \chi^2_k$$

and hence

$$E\left[\frac{\left(\widehat{\beta}-\beta_0\right)X'X\left(\widehat{\beta}-\beta_0\right)}{\sigma_0^2}\right] = k$$

Putting all the pieces together,

$$\begin{aligned} E\left[\widehat{KL}(g;f)\right] &= \frac{T}{2}\left\{E\left[\frac{\sigma_0^2}{\widehat{\sigma}^2}\right] + \log(\sigma_0^2) - E\left[\log(\widehat{\sigma}^2)\right] - 1\right\} \\ &\quad + \frac{1}{2}E\left[\left(\frac{1}{\widehat{\sigma}^2}\right)\left(\widehat{\beta}-\beta_0\right)X'X\left(\widehat{\beta}-\beta_0\right)\right] \\ &= \frac{T}{2}\left(\frac{T}{T-k-2} - \log(\sigma_0^2) + E\left[\log(\widehat{\sigma}^2)\right] - 1\right) + \frac{T}{2}\left(\frac{k}{T-k-2}\right) \\ &= \frac{T}{2}\left(\frac{T+k}{T-k-2} - \log(\sigma_0^2) + E\left[\log(\widehat{\sigma}^2)\right] - 1\right) \end{aligned}$$

Since $\log(\widehat{\sigma}^2)$ is an unbiased estimator of $E[\log(\widehat{\sigma}^2)]$, substituting this give us an unbiased estimator of $E\left[\widehat{KL}(g;f)\right]$ as desired. The only terms that vary across candidate models are the first and the third. Moreover, the multiplicative factor of $T/2$ does not affect model selection. Hence, the criterion is

$$AIC_c = \log(\widehat{\sigma}^2) + \frac{T+k}{T-k-2}$$

Note that the way this expression is written, *smaller* values indicate a better model. So how does this compare to the plain-vanilla AIC for normal linear regression? The maximum likelihood estimators for this problem are

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'\mathbf{y} \\ \widehat{\sigma}^2 &= \frac{(\mathbf{y}-X\widehat{\beta})'(\mathbf{y}-X\widehat{\beta})}{T} \end{aligned}$$

It follows that the maximized log-likehood is

$$\begin{aligned} \log\left[f(\mathbf{y}|X;\widehat{\theta})\right] &= -\frac{T}{2}\log(\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2}(y-X\widehat{\beta})'(y-X\widehat{\beta}) \\ &= -\frac{T}{2}\log(\widehat{\sigma}^2) - \frac{T}{2} \end{aligned}$$

by substituting $T\widehat{\sigma}^2$ for the numerator of the second term. Hence, the AIC for

this problem is

$$AIC = 2\left(\ell(\widehat{\theta}) - k\right) = -T\log(\widehat{\sigma}^2) - T - 2k$$

But this way of writing things uses the *opposite* sign convention from AIC_c. It's important to keep track of this, since different authors use different sign conventions for information criteria. To make the AIC comparable with our scaling of the AIC_c, we multiply through by $-1/T$ yielding

$$AIC = \log(\widehat{\sigma}^2) + \frac{T + 2k}{T}$$

where *smaller* values now indicate a better model.

2 Simulation-Based Model selection

We'll start with an iid setting. Below, we'll consider generalizations to a time series setting.

2.1 Bootstrap Model Selection

2.2 Cross-Validation

3 Some other Model Selection Criteria

3.1 Hannan-Quinn

3.2 Final Prediction Error

3.3 Mallow's C_p

4 Time Series Examples

We won't go through all of the specifics here since they're almost identical to the material from above. Some more details can be found in McQuarrie and Tsai (1998). The AR and VAR models are straightforward since, in the conditional formulation, they're just univariate and multivariate regression, respectively.

4.1 Autoregressive Models

Cross-Validation for AR The way we described it above, CV depended in independence. How can we adapt it for AR models? Roughly speaking, the idea is to use the fact that dependence dies out over time and treat observations that are "far enough apart" as *approximately* independent. Specifically, we choose an integer value h and assume that y_t and y_s can be treated as independent as long as $|s-t| > h$. This idea is called "h-block cross-validation" and was introduced by Burman, Chow & Nolan (1994). As in the iid version of leave-one-out cross-validation, we still evaluate a loss function by predicting *one* witheld observation at a time using a model estimated without it. The difference is that we also omit the h neighboring observations *on each side* when fitting the model. For example, if we choose to evaluate squared-error

loss, the criterion is

$$CV_h(1) = \frac{1}{T-p} \sum_{t=p+1}^{T} \left(y_t - \hat{y}^h_{(t)}\right)^2$$

where

$$\hat{y}^h_{(t)} = \hat{\phi}^h_{1(t)} y_{t-1} + \ldots + \hat{\phi}^h_{1(t)} y_{t-p}$$

and $\hat{\phi}^h_{j(t)}$ denotes the jth parameter estimate from the conditional least-squares estimator with observations $y_{t-h}, \ldots, y_{t+h}$ removed. We still have the question of what h to choose. Here there is a trade-off between making the assumption of independence more plausible and leaving enough observations to get precise model estimates. Intriguingly, the simulation evidence presented in McQuarrie and Tsai (1998) suggests that setting $h = 0$, which yields plain-vanilla leave-one-out CV, works well even in settings with dependence.

The idea of h-block cross-validation can also be adapted to versions of cross-validation other than leave-one-out. For details, see Racine (1997, 2000).

4.2 Vector Autoregression Models

Write without an intercept for simplicity (just demean everything)

$$\begin{array}{rcl} \underset{(q\times 1)}{\mathbf{y}_t} & = & \underset{(q\times q)}{\Phi_1}\, \mathbf{y}_{t-1} + \ldots + \Phi_p \mathbf{y}_{t-p} + \epsilon_t \\ \boldsymbol{\epsilon}_t & \overset{iid}{\sim} & N_q(\mathbf{0}, \Sigma) \end{array}$$

Conditional least squares estimation, sample size, etc.

$$
\begin{aligned}
FPE &= \left|\widehat{\Sigma}_p\right| \left(\frac{T+qp}{T-qp}\right)^q \\
AIC &= \log\left|\widehat{\Sigma}_p\right| + \frac{2pq^2+q(q+1)}{T} \\
AIC_c &= \log\left|\widehat{\Sigma}_p\right| + \frac{(T+qp)q}{T-qp-q-1} \\
BIC &= \log\left|\widehat{\Sigma}_p\right| + \frac{\log(T)pq^2}{T} \\
HQ &= \log\left|\widehat{\Sigma}_p\right| + \frac{2\log\log(T)pq^2}{T}
\end{aligned}
$$

Problems with VAR model selection

1. If we fit p lags, we lose p observations under the conditional least squares estimation procedure.

2. Adding a lag introduces q^2 additional parameters.

Cross-Validation for VARs In principle we could use the same h-block idea here as we did for for the AR example above. However, given the large number of parameters we need to estimate, the sample sizes witholding $2h+1$ observations at a time may be too small for this to work well.

4.3 Corrected AIC for State Space Models

Problem with VARs and state space more generally is that we can easily have sample size small relative to number of parameters. In this case AIC-type criteria don't work well. Suggestions for simulation-based selection.

Cavanaugh & Shumway (1997)

5 Bayesian Information Criterion

Since Frank2 talked about this in his part of the course, I won't discuss this derivation in class but I wanted to provide the details for completeness. As in our derivation of TIC and AIC, we'll consider a setting with an iid sample of scalar random variables $Y_1, \ldots, Y_T$. The results still hold in the more general case, but this simplifies the notation.

5.1 Overview of the BIC

Despite its name, the BIC is *not* a Bayesian procedure. It is a large-sample Frequentist *approximation* to Bayesian model selection:

1. Begin with a uniform prior on the set of candidate models so that it suffices to maximize the Marginal Likelihood.

2. The BIC is a large sample approximation to the Marginal Likelihood:

$$\int \pi(\theta_i) f_i(\mathbf{y}|\theta_i) d\theta_i$$

   where i indexes models M_i in a set $\mathcal{M}$.

3. As usual when Bayesian procedures are subjected to Frequentist asymptotics, the priors on parameters vanish in the limit.

4. We proceed by a *Laplace Approximation* to the Marginal Likelihood

5.2 Laplace Approximation

For the moment simplify the notation by suppressing dependence on M_i. We want to approximate:

$$\int \pi(\theta) f(\mathbf{y}|\theta) d\theta$$

This is actually a common problem in applications of Bayesian inference:

- Notice that $\pi(\theta)f(\mathbf{y}|\theta)$ is the *kernel* of some probability density, i.e. the density without its normalizing constant.
- *How do we know this?* By Bayes' Rule

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{y}|\theta)}{\int \pi(\theta)f(\mathbf{y}|\theta)d\theta}$$

  is a proper probability density and the denominator is *constant* with respect to θ. (The parameter has been "integrated out.")

- In Bayesian inference, we specify $\pi(\theta)$ and $f(\mathbf{y}|\theta)$, so $\pi(\theta)f(\mathbf{y}|\theta)$ is known. But to calculate the posterior we need to *integrate* to find the normalizing constant.
- Only in special cases (e.g. conjugate families) can we find the exact normalizing constant. Typically some kind of approximation is needed:
  - Importance Sampling
  - Markov-Chain Monte Carlo (MCMC)
  - *Laplace Approximation*

The Laplace Approximation is an *analytical approximation* based on Taylor Expansion arguments. In Bayesian applications, the expansion is carried out around the posterior mode, i.e. the mode of $\pi(\theta)f(\mathbf{y}|\theta)$, but we will expand around the Maximum likelihood estimator.

Proposition 5.1 (Laplace Approximation)**.**

$$\int \pi(\theta)f(\mathbf{y}|\theta)d\theta \approx \frac{\exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})(2\pi)^{p/2}}{n^{p/2}\left|J(\hat{\theta})\right|^{1/2}}$$

Where $\hat{\theta}$ *is the* maximum likelihood estimator*,* p *the dimension of* θ *and*

$$J(\hat{\theta}) = -\frac{1}{n}\frac{\partial^2 \log f(\mathbf{y}|\hat{\theta})}{\partial\theta\partial\theta'}$$

Proof. A rigorous proof of this result is complicated. The following is a sketch. First write $\ell(\theta)$ for $\log f(\mathbf{y}|\theta)$ so that

$$\pi(\theta)f(\mathbf{y}|\theta) = \pi(\theta)\exp\{\log f(\mathbf{y}|\theta)\} = \pi(\theta)\exp\{\log \ell(\theta)\}$$

By a second-order Taylor Expansion around the MLE $\hat{\theta}$

$$\ell(\theta) = \ell(\hat{\theta}) + \frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell \tag{1}$$

since the derivative of $\ell(\theta)$ is zero at $\hat{\theta}$ by the definition of MLE. A first-order expansion is sufficient for $\pi(\theta)$ because the derivative does not vanish at $\hat{\theta}$

$$\pi(\theta) = \pi(\hat{\theta}) + \frac{\partial\pi(\hat{\theta})}{\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\pi \tag{2}$$

Substituting Equations 1 and 2,

$$\begin{aligned}
\int \pi(\theta)f(\mathbf{y}|\theta)d\theta &= \int \exp\left\{\ell(\hat{\theta}) + \frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\} \\
&\qquad \times \left[\pi(\hat{\theta}) + \left(\theta - \hat{\theta}\right)' \frac{\partial\pi(\hat{\theta})}{\partial\theta} + R_\pi\right] d\theta \\
&= \exp\left\{\ell(\hat{\theta})\right\}(I_1 + I_2 + I_3)
\end{aligned}$$

where

$$\begin{aligned} I_1 &= \pi(\hat{\theta}) \int \exp\left\{ \frac{1}{2} \left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \left(\theta - \hat{\theta}\right) + R_\ell \right\} d\theta \\ I_2 &= \frac{\partial \pi(\hat{\theta})}{\partial\theta'} \int \left(\theta - \hat{\theta}\right) \exp\left\{ \frac{1}{2} \left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \left(\theta - \hat{\theta}\right) + R_\ell \right\} d\theta \\ I_3 &= \int R_\pi \, \exp\left\{ \frac{1}{2} \left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \left(\theta - \hat{\theta}\right) + R_\ell \right\} d\theta \end{aligned}$$

Under certain regularity conditions (not the standard ones!) we can treat R_ℓ and R_π as approximately equal to zero for large n uniformly in θ, so that

$$\begin{aligned} I_1 &\approx \pi(\hat{\theta}) \int \exp\left\{ \frac{1}{2} \left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \left(\theta - \hat{\theta}\right) \right\} d\theta \\ I_2 &\approx \frac{\partial \pi(\hat{\theta})}{\partial\theta'} \int \left(\theta - \hat{\theta}\right) \exp\left\{ \frac{1}{2} \left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \left(\theta - \hat{\theta}\right) \right\} d\theta \\ I_3 &\approx 0 \end{aligned}$$

Because $\hat{\theta}$ is the MLE,

$$\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}$$

must be negative definite, so

$$-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}$$

is positive definite. It follows that

$$\exp\left\{ \frac{1}{2} \left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \left(\theta - \hat{\theta}\right) \right\} = \exp\left\{ -\frac{1}{2} \left(\theta - \hat{\theta}\right)' \left[\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1} \right]^{-1} \left(\theta - \hat{\theta}\right) \right\}$$

can be viewed as the kernel of a Normal distribution with mean $\hat{\theta}$ and variance

matrix

$$\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}$$

Thus,

$$\int \exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2\ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\}d\theta = (2\pi)^{p/2}\left|\left(-\frac{\partial^2\ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right|^{1/2}$$

and

$$\int \left(\theta-\hat{\theta}\right)\exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2\ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\}d\theta = 0$$

Therefore,

$$\begin{aligned}
\int \pi(\theta)f(\mathbf{y}|\theta)d\theta &\approx \exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}\left|\left(-\frac{\partial^2\ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right|^{1/2} \\
&= \exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}\left|n\left(-\frac{1}{n}\frac{\partial^2\ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)\right|^{-1/2} \\
&= \frac{\exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}}{n^{p/2}\left|J(\hat{\theta})\right|^{1/2}}
\end{aligned}$$

□

5.3 Finally the BIC

Now we re-introduce the dependence on the model M_i. Taking logs of the Laplace Approximation and multiplying by two (again, this is traditional but has no effect on model comparisons)

$$\begin{aligned}
2\log f(y|M_i) &= 2\log\left\{\int f_i(y|\theta_i)\pi(\theta_i)d\theta_i\right\} \\
&\approx 2\ell(\hat{\theta}_i) - p\log(n) + p\log(2\pi) - \pi(\hat{\theta}_i) - \log\left|J(\hat{\theta}_i)\right|
\end{aligned}$$

The first two terms are $O_p(n)$ and $O_p(\log n)$, while the last three are $O_p(1)$, hence negligible as $n \to \infty$. This gives us Schwarz's BIC

$$BIC(M_i) = 2 \log f_i(\mathbf{y}|\hat{\theta}_i) - p \log n$$

We choose the model M_i for which $BIC(M_i)$ is largest. Notice that the prior on the parameter, $\pi(\theta)$, drops out in the limit, and recall that we began by putting a uniform prior on the *models* under consideration.