THE SCHWARZ CRITERION, OR BIC

1. Taylor expansion of log likelihood

$$LLH = \sum_{t=1}^{T} \log p(y_t;\beta) \doteq \sum_{t=1}^{T} \log p(y_t;\hat\beta) + \tfrac{1}{2}(\beta-\hat\beta)' \sum_{t=1}^{T} \frac{\partial^2 \log p(y_t;\hat\beta)}{\partial\beta\partial\beta'}(\beta-\hat\beta) + R(Y_T;\beta-\hat\beta)$$

- The first order term in the expansion drops out because $\hat\beta$ is the MLE and we are assuming differentiability of at least second order and an interior maximum.
- This form is satisfied both by i.i.d. y_t and by dynamic models in which the y_t vector includes lagged conditioning variables and the p's are conditional densities for current data.
- y_t can contain strongly exogenous variables, which would contribute to the pdf of the observed (exogenous plus endogenous) data only via a factor that does not depend on β.

2. Integrating it

Exponentiating this (ignoring the error term) produces an object proportional to a $N(\hat\beta, \Omega_T^{-1})$ pdf, where

$$\Omega_T = \sum_{t=1}^{T} \frac{\partial^2 \log p(y_t;\hat\beta)}{\partial\beta\partial\beta'}\,.$$

The leading constant, though, is $\exp(\widehat{LLH})$, the likelihood maximum, whereas the normalizing constant for the corresponding normal pdf would be $(2\pi)^{-k/2}\,|\Omega_T|^{\frac{1}{2}}$. Therefore the log of the integral of this object over β is

$$\widehat{LLH} + \frac{k}{2}\log(2\pi) - \tfrac{1}{2}\log|\Omega_T|\,.$$

If the y_t process is ergodic,

$$\frac{1}{T}\Omega_T \xrightarrow[t\to\infty]{} \Omega = E\left[\frac{\partial^2 \log p(y_t;\hat\beta)}{\partial\beta\partial\beta'}\right]$$

Date: November 28, 2006.

Plugging in this approximation, and recalling that for a $k \times k$ matrix A, $|cA| = c^k |A|$, we get

$$\widehat{LLH} + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log T - \tfrac{1}{2}\log|\Omega| \ .$$

The maximized log likelihood LLH also converges to a limit when scaled by T, so that

$$\widehat{LLH} \doteq TE[\log p(y_t;\beta)] = T\int \log p(y_t;\beta)q(y_t)dy_t \ ,$$

where q is the true pdf of y_t. This is minus the **Kullback-Leibler distance** from q to $p(y_t;\beta)$, which is known to be minimized when $q = p$.

3. The Schwarz criterion or BIC

- When comparing a true model to a false model, the $\widehat{LLH}$ term will grow at the rate T, dominating the other terms.
- When comparing two true models, one of which has more parameters, the $\widehat{LLH}$ terms will converge to a bounded limiting distribution. (This is the standard result that in this case twice the difference in maximized likelihoods is asymptotically χ^2.) So in this case the $\frac{k}{2}\log T$ term dominates.
- the **Schwarz criterion** or **BIC** (Bayesian information criterion) ranks models using only these two terms of the expansion. It will in a large enough sample pick the same model as optimal as does a full calculation of Bayesian posterior odds, if regularity conditions are satisfied, yet does not depend on the prior or on any aspect of the data other than the maximized likelihood values for the two models.

4. Limitations

- The Taylor series approximation
- In a given sample, the $k\log(2\pi)/2$ term and the $\frac{1}{2}\log|\Omega|$ terms will differ across models, and may dominate the comparison.
- Ergodicity. Models with unit roots do not satisfy the ergodicity assumption. The penalty term becomes larger, possibly, but the $\widehat{L}LH$ term also behaves differently. How much difference this makes depends on how the models are related, and there is no standard, simple correction to the BIC for this case.

5. interpretations

In the case of a linear regression, Ω_T is $X'X/\sigma^2$. So bigger models are penalized according to how potentially powerful their additional explanatory variables are, relative to the model's residual variance.