Model Selection

- Introduction or background

Which variables are important?

What defines a good model?

How to select a model? ← Today's question

With a model selection criteria.

Efficient criteria

A common assumption in both regression and time series is that the generating or true model is of infinite dimension, or that the set of candidate models doesn't contain the true model. The goal is to select one model that best approximates the true model from a set of finite-dimensional candidate models. The candidate model that is closest to the true model is assumed to be the appropriate choice.

In large samples, a model selection criterion that chooses the model with minimum mean squared error distribution is said to be ASYMPTOTICALLY EFFICIENT.

Consistent Criteria

Many researches assume that the true model is of finite dimension, and that it is included in the set of candidate models. Under this assumption the goal of model selection is to correctly choose the true model from the list of candidate models. A model selection criterion that identifies the correct model asymptotically with probability one is said to be CONSISTENT.

Measures of discrepancy or closeness

Let M_T be the true model with density f_T and distribution F_T. Let M_A denote the candidate (approximating) model with density f_A and let Δ denote the discrepancy.

Kullback-Leibler discrepancy

$$\Delta_{K-L}(M_T, M_A) = E_{F_T}\left[\log\left(\frac{f_T(x)}{f_A(x)}\right)\right]$$

Let μ_{M_T} and μ_{M_A} denote the true and candidate model means, respectively. We can define

$$\Delta_{L^2}(M_T, M_A) = \|\mu_{M_T} - \mu_{M_A}\|^2$$

K-L:
- It is a number
- When errors are not normal it must be computed for each distribution

L^2:
- It can be a matrix in multivariate set-ups
- It works independent on the distribution of the errors

Some Foundations of Model Selection Criteria

From minimizing Δ_{L^2} we get

$$FPE_K = \hat{\sigma}_K^2 \frac{n+k}{n-k} = \hat{\sigma}_K^2 \left(1 + \frac{2k}{n-k}\right)$$

where k = # parameters

From minimizing $\Delta_{K\text{-}L}$ we get

$$AIC_K = \log\left(\hat{\sigma}_K^2\right) + \frac{2(k+1)}{n}$$

If now the investigator believes that the true model belongs to the set of candidate models, the objective is to design criteria such we choose the true model with probability equal to one. In this way from a bayesian perspective we get

$$\underset{\text{or}}{BIC_K} = \log\left(\hat{\sigma}_K^2\right) + \frac{\log(n)\,k}{n}$$

$$SIC_K$$

The other consistent criteria (strongly consistent) is

$$HQ_K = \log\left(\hat{\sigma}_K^2\right) + 2\,\frac{\log(\log(n))\,k}{n}$$

- **Model Selection Criteria**

Choose k such that minimizes

$$IC(k) = \ln \hat{\sigma}_k^2 + \boxed{k \frac{C(T)}{T}}$$

penalty term

$C(T) = 2$ for AIC

$C(T) = \ln T$ for BIC, SIC

$C(T) = 2 \ln \ln T$ for HQ

If we are in the world of ARMA (p, q) models, then we choose p and q such that minimizes

$$IC(p, q) = \ln \hat{\sigma}_{p,q}^2 + (p+q) \frac{C(T)}{T}$$

It is easier to think or work in terms of the family of AR(p) models.

$\hat{p}(IC)$ is chosen such that

$$IC(\hat{p}(IC)) = \min \{ IC(p) \mid p = 0, 1, \dots M \}$$

An estimator $\hat{p}$ of the AR order p is called consistent if

$$\text{plim}\ \hat{p}_T = p \quad \text{or} \quad \lim_{T \to \infty} Pr(\hat{p} = p) = 1$$

and strongly consistent if

$$Pr(\lim \hat{p} = p) = 1$$

Proposition

Let y_t be a stationary AR(P) process. Suppose the maximum order $M \geq p$ and $\hat{p}$ is chosen so as to minimize a criterion

$$I((m) = \ln \tilde{\sigma}_m^2 + m \frac{C_T}{T} \quad \text{over } m=0,1,\ldots M$$

where C_T is a nondecreasing sequence of real numbers that depend on the sample size T. Then $\hat{p}$ is consistent iff

$$\boxed{C_T \to \infty} \quad \text{and} \quad \boxed{C_T/T \to 0} \quad \text{as } T \to \infty$$

and $\hat{p}$ is strongly consistent iff

$$\frac{C_T}{2 \ln \ln T} > 1 \quad \text{as } T \to \infty$$

The basic idea of the proof is to show that for $p > m$, the quantity $\ln \tilde{\sigma}_m^2 / \ln \tilde{\sigma}_p^2$ will be greater than one in large samples since $\ln \tilde{\sigma}_m^2$ is essentially the minimum of minus the Gaussian log-likelihood function for an AR(m) model.
Consequently, since the penalty terms

$$m \frac{C_T}{T} \quad \text{and} \quad p \frac{C_T}{T} \to 0 \text{ as } T \to \infty$$

then $IC(m) > IC(p)$ for large T.

Thus the probability of choosing too small an order goes to zero as $T \to \infty$.

Similarly, if $m > p$, $\ln \tilde{\sigma}_m^2 / \ln \tilde{\sigma}_p^2$

approaches one in probability if $T \to \infty$ and the penalty term of the lower order model is smaller than that of a larger order process. Thus the lower order "p" will be chosen if T is large.

Proof :

$$I(p) : S_p^2 = \ln \hat{\sigma}_p^2 + p \frac{C(T)}{T}$$

$$\begin{cases} C(T) \to \infty \\ T \to \infty \\ \frac{C(T)}{T} \to 0 \end{cases}$$

$$\hat{p} = \min S_p^2 \qquad p = 1 \dots \bar{p}$$

$\bar{p}$ fixed and known

Then $\hat{p}_T \not\to p_0$ true lag value

We have to show

(i) $Pr\left[S_1^2, \dots, S_{p_0-1}^2 > S_{p_0}^2\right] \to 1$ "undefit"

(ii) $Pr\left[S_{p_0+1}^2, \dots, S_{\bar{p}}^2 > S_{p_0}^2\right] \to 1$ "overfit"

(i)

$$S_p^2 - S_{p_0}^2 = \ln\left(\hat{\sigma}_p^2 / \hat{\sigma}_{p_0}^2\right) + \underbrace{(p - p_0)}_{\text{fixed}} \frac{C(T)}{T} \quad \Big| \quad p < p_0$$

$\frac{C(T)}{T} \to 0$

Recall that $\sigma_p^2 = V(y_t \mid y_{t-1}, \ldots y_{t-p})$

Then under some condition

$$\ln\left(\hat{\sigma}_p^2 / \hat{\sigma}_{p_0}^2\right) \xrightarrow{p} \ln\left(\sigma_p^2 / \sigma_{p_0}^2\right) > 0$$

so $S_p^2 - S_{p_0}^2 \xrightarrow{p} \ln\left(\sigma_p^2 / \sigma_{p_0}^2\right) > 0$

and we "underfit" with probability one

(ii) Now

$$S_p^2 - S_{p_0}^2 = \ln\left(\hat{\sigma}_p^2 / \hat{\sigma}_{p_0}^2\right) + (p - p_0)\frac{C(T)}{T}$$

but

$$-2T \ln\left(\hat{\sigma}_{p_0+1}^2 / \hat{\sigma}_{p_0}^2\right) = \text{LR test of}$$

$$H_0: p = p_0$$

versus

$$H_a: p = p_0 + 1$$

under $H_0: p = p_0$, $LR \xrightarrow{d} \xi_1^2$ $(\xi_1 \sim N(0,1))$

so

$$\begin{bmatrix} -2T \ln\left(\hat{\sigma}_{p_0+1}^2 / \hat{\sigma}_{p_0}^2\right) \\ \vdots \\ -2T \ln\left(\hat{\sigma}_p^2 / \hat{\sigma}_{p_0}^2\right) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \chi_1^2 \\ \vdots \\ \chi_{p-p_0}^2 \end{bmatrix}$$

Therefore

$$T\left(S_p^2 - S_{p_0}^2\right) = \underbrace{T\left(\ln\left(\hat{\sigma}_p^2 / \hat{\sigma}_{p_0}^2\right)\right)}_{O_p(1)} + \underbrace{(p - p_0)\frac{C(T)}{T}}_{\to \infty}$$

Thus $T\left(S_p^2 - S_{p_0}^2\right) \xrightarrow{p} \infty$

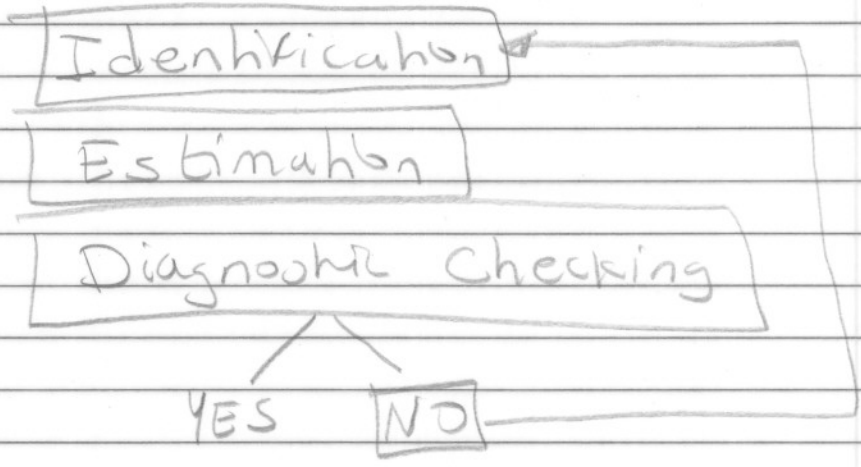
Then $Pr\left[S_p^2 > S_{p_0}^2\right] = Pr\left[T\left(S_p^2 - S_{p_0}^2\right) > 0\right]$

$$= Pr\left[\underbrace{T \ln\left(\hat{\sigma}_p^2 / \hat{\sigma}_{p_0}^2\right)}_{-\chi_{p-p_0}^2} + \underbrace{(p - p_0)\frac{C(T)}{T}}_{\to \infty} > 0\right]$$

$$= Pr\left[\chi_{p-p_0}^2 < \infty\right] \longrightarrow 1$$

Two phibsophies to select a model

- From particular to general
  - * Box-Jenkins Methodology



- From General to particular
  - * Testing

Form your own phibsophy to select a model, based on what do you need your model for?