Model Selection

Jesus Gonzalo

UC3M

November 3, 2020

Introduction

Important QUESTIONS are:

- *Which variables have to be included in a model?*
- *What defines a good model?*

Introduction

Important QUESTIONS are:

- *Which variables have to be included in a model?*
- *What defines a good model?*

The Purpose of today's lectures is to answer the following question:

Introduction

Important QUESTIONS are:

- *Which variables have to be included in a model?*
- *What defines a good model?*

The Purpose of today's lectures is to answer the following question:

- *How to select a model?*
- The answer is: with an information *model selection* criteria.

Efficient Criteria

Model selection criteria have two main properties: **efficiency** and **consistency**.
A common assumption is both regression and time series is that the generating or true model is of infinite dimension: thus the set of candidate madoels does not contain the true model.
The goal is to select one model that best approximate the true modle from a set of finite-dimensional candidate model. The candidate model that is closest to the true model is assumed to be the appropriate choice.

Definition

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 researchers 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.

Definition

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. M_A denote the candidate (approximating) model with density f_A and let Δ denote the discrepancy.
Two different discrepancy types are presented here:

- The 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 the candidate model means, respectively. We can define:

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

Measures of Discrepancy or Closeness

Two comments are important:

- For Δ_{K-L}:
  - it is a *number*;
  - when errors are **not** normal it must be computed for each distribution
- For Δ_{L^2} :
  - it can be a matrix in multivariate set-ups;
  - it works *independently* of the errors' distribution.

Foundations of Model Selection Criteria

From minimizing Δ_{L^2} we get the Final Prediction Error criterion :

$$FPE_k = \hat{\sigma}_k^2 \frac{n+k}{n-k} = \hat{\sigma}_k^2 (1 + \frac{2k}{n-k})$$

where k is the number of parameters.
From minimizing Δ_{K-L} we get the Akaike information criterion:

$$AIC_k = \ln(\hat{\sigma}_k^2) + \frac{2(k+1)}{n}.$$

Foundations of Model Selection Criteria

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

$$BIC_k = \ln(\hat{\sigma}_k^2) + \frac{\ln(n)k}{n}.$$

The other consistent criterion (strongly consistent) is the Hannan and Quinn criterion:

$$HQ_k = \ln(\hat{\sigma}_k^2) + 2\frac{\ln(\ln(n))k}{n}.$$

Model Selection Criteria

The different way to select models choose k so as to minimize the following information criterion:

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

where T is the sample size (it can be replaced by n); and $\frac{kC(T)}{T}$ is the *penalty term*:

- for AIC we have $C(T) = 2$
- for BIC we have $C(T) = \ln T$
- for HQ we have $C(T) = 2\ln(\ln(T))$.

Model Selection Criteria

In the case of $ARMA(p,q)$ models, p and q have to be chosen in order to minimize the following criterion:

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

However, It is easier to think and work in terms of the family of $AR(p)$ models. In this case, $\hat{p}_{(IC)}$ has to be chosen such that:

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

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

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

and strongly consistent if

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

Model Selection Criteria

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 the following criterion over $m = 0, 1, ...M$

$$IC(m) = \ln \hat{\sigma}_m^2 + m\frac{C(T)}{T}$$

where $C(T)$ is a nondecreasing sequence of real numbers that depend on the sample size T. Then, $\hat{p}$ is consistent iff, as $T \to \infty$

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

and $\hat{p}$ is strongly consistent iff, as $T \to \infty$

$$\frac{C(T)}{2\ln(\ln T)} > 1$$

Model selection

Proof

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

$$m\frac{C(T)}{T}, p\frac{C(T)}{T} \to 0$$

then $IC(m) > IC(p)$ for large T.
Thus the probability of choosing too small order goes to zero as $T \to \infty$.
Similarly, if $m > p$, $\frac{\ln \hat{\sigma}_m^2}{\ln \hat{\sigma}_p^2}$ approaches one in probability as $T \to \infty$ 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 too large.

Model selection

Proof

Consider $I(p) = S_p^2 = \ln \hat{\sigma}_p^2 + p\frac{C(T)}{T}$ and $\hat{p} = \min S_p^2$ with $p \in \{1, ..., \bar{p}\}$ where $\bar{p}$ is fixed and known.
Then, recalling that

$$\begin{cases} C(T) \to \infty & \text{as } T \to \infty \\ \\ \dfrac{C(T)}{T} \to 0 & \text{as } T \to \infty \end{cases}$$

We have to show the following facts:

1. $\Pr\left[S_1^2, ..., S_{p_0-1}^2 > S_{p_0}^2\right] \to 1$: the probability of "no-underfit" $(S_{p_0}^2)$
2. $\Pr\left[S_{p_0-1}^2, ..., S_{\bar{p}}^2 > S_{p_0}^2\right] \to 1$: the probability of "no-overfit" $(S_{p_0}^2)$.

Model selection

Proof

For the first point let us consider the difference $S_p^2 - S_{p_0}^2$ assuming $p < p_0$:

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

Recalling that $\sigma_p^2 = V(y_t \mid y_{t-1}, ..., y_{t-r})$, we have that under some conditions (ergodicity):

$$\ln\left(\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{p_0}^2}\right) \xrightarrow{p} \ln\left(\frac{\sigma_p^2}{\sigma_{p_0}^2}\right) > 0 \implies S_p^2 - S_{p_0}^2 \xrightarrow{p} \ln\left(\frac{\sigma_p^2}{\sigma_{p_0}^2}\right) > 0.$$

In this way we show that the probability of "no-underfit" the true model tends to one.

Model selection

Proof

For the second point let us consider again the difference $S_p^2 - S_{p_0}^2$:

$$S_p^2 - S_{p_0}^2 = \ln\left(\frac{\sigma_p^2}{\sigma_{p_0}^2}\right) + (p - p_0)\frac{C(T)}{T}.$$

Argue that $-2T \ln\left(\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{p_0}^2}\right)$ is a likelihood ratio (LR) test with the following two hypotheses:

$$H_0 : p = p_0, H_a : p = p_0 + 1.$$

Under $H_0 : p = p_0$, we have that $LR \xrightarrow{d} \chi^2$.

Model selection

Proof

Accordingly, we have that:

$$\begin{bmatrix} -2T \ln\left(\hat{\sigma}^2_{p_0+1}/\hat{\sigma}^2_{p_0}\right) \\ \vdots \\ -2T \ln\left(\hat{\sigma}^2_{\bar{p}}/\hat{\sigma}^2_{p_0}\right) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \chi^2_1 \\ \vdots \\ \chi^2_{\bar{p}-p_0} \end{bmatrix}$$

Therefore:

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

Model selection

Proof

Finally, we have that, with $p > p_0$:

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

and

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

and since $\ln\left(\hat{\sigma}_p^2/\hat{\sigma}_{p_0}^2\right) \to -\chi^2_{p-p_0}$ and $(p - p_0)C(T) \to \infty$, we have:

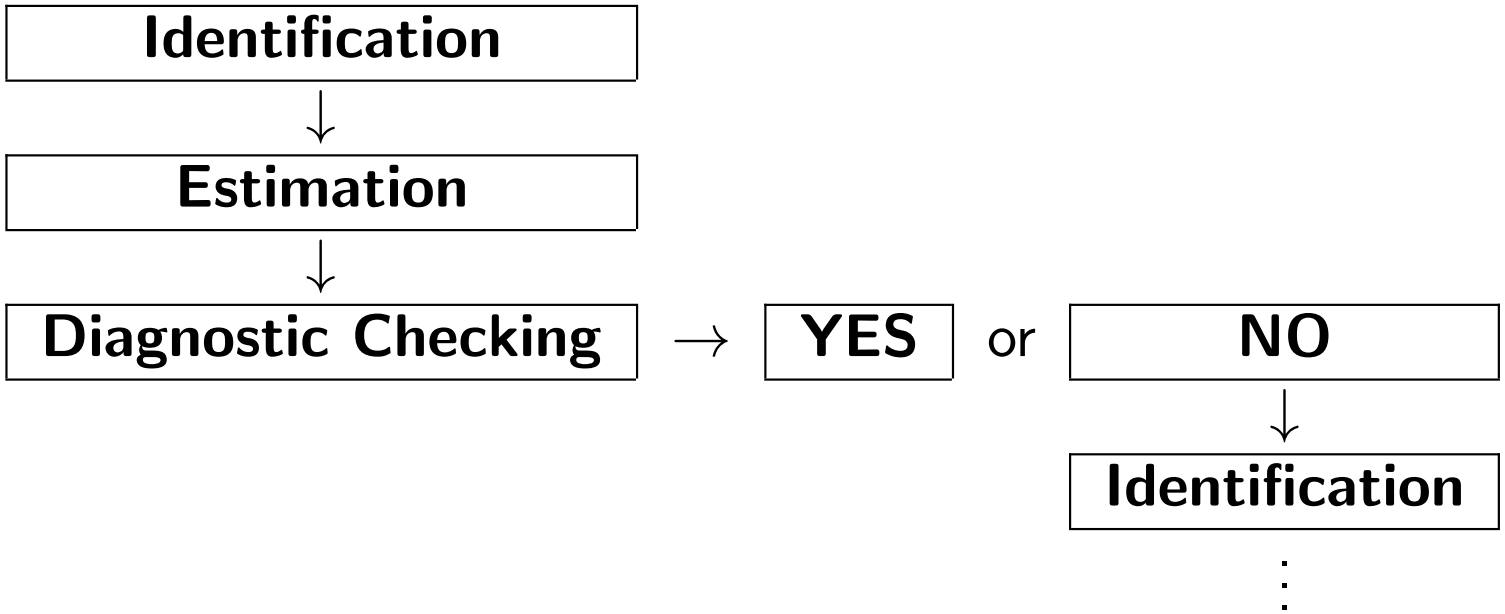
$$\Pr\left[S_p^2 > S_{p_0}^2\right] = \Pr\left[\chi^2_{p-p_0} < \infty\right] \to 1.$$

Model selection

Akaike criterion results to be *not consistent*, while BIC is consistent; still, we have to be careful preferring BIC to Akaike since **Akaike has only problems in overfitting**.

Three Philosophies to Select a Model

- *From Particular to General*: Box-Jenkins methodology



- *From General to Particular*: Testing
- *Information Crtiteria*: no testing

Form your own philosophy to select a model based on what do you need your model for (Focused information criterion is an example).