Lectures 22-23

1 A primer on Model Selection

Broadly speaking, model selection refers to the use of data to select a particular statistical model within a certain class. In the context of autoregressive models, model selection is tantamount to deciding the number of lags (p) that should be included in the autoregressive specification.

A popular approach to select the number of lags of an AR(p) model is the Bayes Information Criterion. Let T be the sample size. For any integer $p < T$ define the function:

$$\text{BIC}(p) \equiv \ln\left(\frac{1}{T}\sum_{t=p+1}^{T}(y_t - \widehat{\mu}(p) - \sum_{j=1}^{p}\widehat{\phi}_j(p)y_{t-j})^2\right) + (p+1)\frac{\ln(T)}{T}. \quad (1)$$

where $\widehat{\mu}(p)$ and $\widehat{\phi}_j(p)$ are the OLS estimators of the constant term and the autoregressive coefficients in an AR(p) model.

Definition: Let $\overline{p}$ be a user-specified upper bound on the number of lags. Given a time series $(Y_1, Y_2, \ldots, Y_T)$ the number of lags selected by the Bayes Information criterion is given by:

$$\widehat{p} \equiv \arg\min_{p\in\{0,1,2,\ldots,\overline{p}\}} \text{BIC}(p). \quad (2)$$

How do we know if the Bayes Information Criterion above "works"? Let $\theta^* \equiv (p^*, \mu, \phi_1, \ldots, \phi_{p^*}, \sigma^2)$ denote the parameters of an AR(p^*) model. We say that a model selection criterion is consistent for the true p^* if:

$$\mathbb{P}_{\theta^*}(\widehat{p} = p^*) \to 1$$

That is, if the model selection criterion selects the true lag order with probability close to 1 (regardless of what the true order is).

2 Consistency of the BIC

We know provide a sketch of the proof that shows that the BIC "works".

2.1 Lemma: Wald tests and sums of squared residuals

The first thing to note is that the BIC is based on a comparison of the sums of squared residuals for different models. Thus, we start by analyzing the sum of squared residuals that corresponds to two different autoregressive models.

Consider two AR models with lags p and $\tilde{p}$, with $p > \tilde{p}$. We of course now that the sum of squared residuals for the "larger" model will be smaller than the sum of squared residuals for the "smaller" model. We want to use standard regression algebra to quantify this difference.

Let y be the $T \times 1$ vector that collects the T realizations of y_t; that is

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}$$

and let X denote the matrix that collects its p lags and the constant term:

$$X = \begin{pmatrix} 1 & y_{1-1} & y_{1-2} & \cdots & y_{1-p} \\ 1 & y_{2-1} & y_{2-2} & \cdots & y_{2-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_{T-1} & y_{T-2} & \cdots & y_{T-p} \end{pmatrix}$$

(I have assumed that I have access to the lags before period 1, which is just a normalization).

I we let:

$$\beta = \begin{pmatrix} \mu \\ \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix},$$

the sum of squared residuals of a model with p lags is simply:

$$SSR(p) \equiv \min_{\beta}(y - X\beta)'(y - X\beta),$$

which is the unconstrained Ordinary Least Squares problem. Interestingly, the sum of squared residuals of a model with $\tilde{p}$ lags $\tilde{p} < p$ can be written as a constrained OLS problem:

$$SSR(\tilde{p}) \equiv \min_{\beta}(y - X\beta)'(y - X\beta) \quad \text{s.t.} \quad R(\tilde{p})\beta = \mathbf{0},$$

where R is a matrix that enforces the restriction that the autoregressive coefficients with index larger than $\tilde{p}$ are zero (for example, if $p = 2$ and $\tilde{p} = 1$, then R=[0, 0,1]). In the appendix I show (painful, but standard matrix algebra) that:

$$SSR(\tilde{p}) = SSR(p) + T(R(\tilde{p})\widehat{\beta}_u)' \left[R(X'X/T)^{-1}R'\right]^{-1} (R(\tilde{p})\widehat{\beta}_u), \tag{3}$$

which implies that

$$\widehat{\sigma}^2(\tilde{p}) - \widehat{\sigma}^2(p) = \frac{\widehat{\sigma}^2(p)}{T} \text{ WaldStat}(R(\tilde{p})\beta) \tag{4}$$

where $\widehat{\sigma}^2(p)$ to denote $SSR(p)/T$ and

$$\text{WaldStat}(R(\tilde{p})\beta) \equiv \left[T(R(\tilde{p})\widehat{\beta}_u)'(R(X'X/T)^{-1}R')^{-1}(R(\tilde{p})\widehat{\beta}_u)/\widehat{\sigma}^2(p)\right].$$

The expression above is the Wald statistic for the null hypothesis that the smaller model is correct—that is $R(\widetilde{\rho}) = 0$—based on the unconstrained OLS estimator of β.

2.2 BIC selects a smaller model with probability zero

Let p^* denote the true model and let $\tilde{p} < p^*$. Assume that the coefficients of the true model are such that $R(\tilde{p})\beta \neq 0$ (we need this assumption, otherwise we would not be able to distinguish between the smaller and the larger model). We show that:

$$\mathbb{P}_{p^*}(\widehat{p} = \tilde{p}) \to 0 \tag{5}$$

Note first that the probability of selecting the smaller model $\tilde{p}$ is smaller than the probability that $\text{BIC}(\tilde{p}) < \text{BIC}(p^*)$. This happens because whenever $\widehat{\tilde{p}} = \tilde{p}$, it follows that

$$\text{BIC}(\widehat{p}) = \min_{p \in \{0,1,\ldots,\bar{p}\}} \text{BIC}(p).$$

In particular, whenever $\widehat{\tilde{p}} = \tilde{p}$ we must have that $\text{BIC}(\tilde{p}) \leq \text{BIC}(p^*)$. Therefore:

$$\begin{aligned}
\mathbb{P}_{p^*}(\widehat{p} = \tilde{p}) &\leq \mathbb{P}_{p^*}(\text{BIC}(\tilde{p}) < \text{BIC}(p^*)) \\
&= \mathbb{P}_{p^*}(\ln(\widehat{\sigma}^2(\tilde{p})) + (\tilde{p}+1)\ln(T)/T < \ln(\widehat{\sigma}^2(p^*)) + (p^*+1)\ln(T)/T) \\
&= \mathbb{P}_{p^*}(\ln(\widehat{\sigma}^2(\tilde{p})/\widehat{\sigma}^2(p^*)) < (p^* - \tilde{p})\ln(T)/T) \\
&= \mathbb{P}_{p^*}(\ln(1 + (\widehat{\sigma}^2(\tilde{p}) - \widehat{\sigma}^2(p^*)/\widehat{\sigma}^2(p^*)) < (p^* - \tilde{p})\ln(T)/T) \\
&= \mathbb{P}_{p^*}(\ln(1 + \text{WaldStat}(R(\tilde{p})\beta)/T) < (p^* - \tilde{p})\ln(T)/T) \\
&\quad \text{(by equation (3))}
\end{aligned}$$

Now, note that the if the true coefficients of the model are such that $R(\tilde{p})\beta \neq 0$ (as we have assumed), the Wald statistic (divided by T) will converge in probability to a nonnegative constant (call it w). Since, for large enough T:

$$\ln(1+w) \leq (p^* - \tilde{p})\ln(T)/T)$$

for any nonnegative constant w:

$$\mathbb{P}_{p^*}(\widehat{p} = \tilde{p}) \to 0$$

2.3 BIC selects a larger model with probability zero

Once again, let now p^* denote the true model and let $\tilde{p}$ be some larger model. We want to show that:

$$\mathbb{P}_{p^*}(\widehat{p} = \tilde{p}) \to 0 \tag{6}$$

Note that:

$$\begin{aligned}
\mathbb{P}_{\tilde{p}}(\widehat{p} = p^*) &\leq \mathbb{P}_{\tilde{p}}(\text{BIC}(\tilde{p}) > \text{BIC}(p^*)) \\
&= \mathbb{P}_{p^*}(\ln(1 + \text{WaldStat}(R(p^*)\beta)/T) > (\tilde{p} - p^*)\ln(T)/T) \\
&\quad \text{(by equation (3))} \\
&\approx \mathbb{P}_{p^*}(\ln(1 + \chi^2_{\tilde{p}-p^*}/T) > (\tilde{p} - p^*)\ln(T)/T) \\
&\quad \text{(since under } p^* \text{ the Wald stat converges to a } \chi^2 \text{ distribution)} \\
&\approx \mathbb{P}_{p^*}(\chi^2_{\tilde{p}-p^*}/T > (\tilde{p} - p^*)\ln(T)/T),
\end{aligned}$$

where the last line used the fact that $\ln(1 + a) \approx a$. This shows that:

$\mathbb{P}_{p^*}(\widehat{p} = \tilde{p}) \to 0$

as $n \to \infty$.

3 Appendix

Let:

$$SSR = \min_{\beta}(y - X\beta)'(y - X\beta),$$

and

$$SSR(R) \equiv \min_{\beta}(y - X\beta)'(y - X\beta) \quad \text{s.t.} \quad R\beta = \mathbf{0}_{k\times 1},$$

where k is the number of restrictions imposed by R. Note that the first order conditions of the constrained problem are:

$$2X'(y - X\widehat{\beta}_c) - R'\lambda = \mathbf{0}_{k\times 1}, \tag{7}$$

where λ is the vector of Lagrange multipliers. Note that left multiplying equation (6) by $R(X'X)^{-1}$ we get:

$$2R(X'X)^{-1}X'y = R(X'X)^{-1}R'\lambda,$$

which implies that the vector of lagrange multipliers λ equals:

$$\frac{\lambda}{2} = \left[R(X'X)^{-1}R'\right]^{-1} R(X'X)^{-1}X'y = \left[R(X'X)^{-1}R'\right]^{-1} R\widehat{\beta}_u \tag{8}$$

and also we get that:

$$\widehat{\beta}_c = \widehat{\beta}_u - (X'X)^{-1}R'\frac{\lambda}{2} \tag{9}$$

($\widehat{\beta}_u$ is the unconstrained OLS estimator and $\widehat{\beta}_c$ is the constrained). Note that (8) implies that:

$$y - X\widehat{\beta}_c \quad = \quad y - X\widehat{\beta}_u + X(X'X)^{-1}R'\lambda/2$$

Note that by definition of the OLS residuals, $(y - X\widehat{\beta}_u)'X = 0$, therefore:

$$\begin{aligned} SSR(R) &= SSR + (\lambda/2)'R(X'X)^{-1}X'X(X'X)^{-1}R'\lambda/2 \\ &= SSR + (\lambda/2)'R(X'X)^{-1}R(\lambda/2) \\ &= SSR + (R\widehat{\beta}_u)'\left[R(X'X)^{-1}R'\right]^{-1}(R\widehat{\beta}_u) \end{aligned}$$