171:290 Model Selection
Lecture X: Criteria for Time Series Model Selection (Part II)

Joseph E. Cavanaugh

Department of Biostatistics
Department of Statistics and Actuarial Science
The University of Iowa

October 27, 2009

Introduction

- In this lecture, we continue our review of procedures for model selection in the time series setting.
- We focus on a criterion introduced by Hannan and Quinn, designed to provide a compromise between a consistent and an asymptotically efficient criterion.
- We also discuss model validation in the time series setting.

Introduction

Outline:

- Overview
  - Autoregressive and Moving Average Models
  - Autoregressive Model Selection Framework
- Consistency and asymptotic efficiency in the autoregressive setting.
- The Hannan and Quinn criterion, HQ.
- Simulation study to illustrate consistency and asymptotic efficiency.
- Application
- Model Validation

Overview

- Assume that our data vector y consists of n measurements on a response variable, collected over equally spaced time points indexed by $t = 1, 2, \ldots, n$.
- Denote the response measurements as $y_1, y_2, \ldots, y_n$.

AR(p) and MA(q) Models

- An **autoregressive process** of order p, AR(p), is defined as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + e_t,$$

where $e_t \ \sim \ iid \ N(0, \sigma^2)$.

- An **moving average process** of order q, MA(q), is defined as

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q},$$

where $e_t \ \sim \ iid \ N(0, \sigma^2)$.

AR(p) and MA(q) Models

- A stationary AR(p) process can be represented as an infinite-order moving average, MA(∞), with coefficients θ_i that decay in magnitude as i increases.
- An invertible MA(q) process can be represented as an infinite-order autoregression, AR(∞), with coefficients ϕ_i that decay in magnitude as i increases.

Autoregressive Model Selection Framework

- **True** or **generating model**: $f(y|\,\theta_o)$.
- **Candidate** or **approximating model**: $f(y|\,\theta_k)$.
- **Candidate class:**

$$\mathcal{F}(k) = \{f(y|\,\theta_k) \mid \theta_k \in \Theta(k)\}\,.$$

- Assume $f(y|\,\theta_k)$ corresponds to an autoregressive model of order p. Note that $k = (p+1)$.
- **Candidate family:**

$$\mathcal{F} = \{\mathcal{F}(2), \mathcal{F}(3), \ldots, \mathcal{F}(P+1)\}.$$

- **Parameter vector:** $\theta_k = (\phi_1, \phi_2, \ldots, \phi_p, \sigma^2)'$.
- **True parameter vector:** $\theta_o = (\phi_1^o, \phi_2^o, \ldots, \phi_{p_o}^o, \sigma_o^2)'$.
- **Fitted model:** $f(y|\,\hat{\theta}_k)$.

Popular Criteria for Autoregressive Model Selection

- The Akaike (1973) information criterion:

$$\text{AIC} = -2 \ln f(y|\,\hat{\theta}_k) + 2(p+1).$$

- The corrected Akaike (1973) information criterion (Hurvich and Tsai, 1989):

$$\text{AICc} = -2 \ln f(y|\,\hat{\theta}_k) + \frac{2(p+1)n}{n-p-2}.$$

- The Bayesian information criterion (Schwarz, 1978):

$$\text{BIC} = -2 \ln f(y|\,\hat{\theta}_k) + (p+1) \ln n.$$

Popular Criteria for Autoregressive Model Selection

- Final prediction error (Akaike, 1969):

$$\text{FPE} = \left(\frac{n+p}{n-p} \right) \hat{\sigma}^2.$$

- The Hannan and Quinn (1979) criterion, HQ.

Consistency in the Autoregressive Framework

- Suppose that the true model is contained in the candidate family $\mathcal{F}$: i.e., $p_o \leq P$.
- A criterion is **consistent** if it will asymptotically identify the correct order with probability one.
- BIC and HQ are consistent; AIC, AICc, and FPE are not.

Asymptotic Efficiency in the Autoregressive Framework

- Suppose that the true model is not contained in the candidate family $\mathcal{F}$: i.e., $p_o > P$.
- A criterion is **asymptotically efficient** if it will asymptotically identify a fitted candidate model for which the (one-step) mean squared error of prediction is minimized.
- AIC, AICc, and FPE are asymptotically efficient; BIC and HQ are not.
- In introducing the concept of asymptotic efficiency, Shibata (1980) assumed that the true model was an infinite order autoregression, with coefficients that decay at a certain rate.

Asymptotic Efficiency in the Autoregressive Framework

- For an AR(p) model,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + e_t,$$

the one-step mean square error of prediction (MSEP) for forecasting y_{n+1} is given by

$$d(\theta_k) = \mathrm{E}\left\{ \left(y_{n+1} - (\phi_1 y_n + \phi_2 y_{n-1} + \ldots + \phi_p y_{n-p+1}) \right)^2 \right\}.$$

HQ

- Hannan and Quinn (1979) introduced HQ as a criterion that would serve as a compromise between a consistent criterion (namely BIC) and an asymptotically efficient criterion (namely AIC).
- HQ was introduced specifically for autoregressive model selection, although it is often used in other modeling frameworks as well (McQuarrie and Tsai, 1998).

HQ

- Consider a generalized information criterion, GIC, for AR(p) model selection:

$$\text{GIC} = -2 \ln f(y | \hat{\theta}_k) + a_n (p + 1).$$

- For consistency, a_n must grow with n.
- For asymptotic efficiency, a_n must converge to 2 as $n \rightarrow \infty$.
- Hannan and Quinn (1979) posed the following question: for a GIC in the AR(p) framework, what is the smallest rate of growth for a_n that will ensure consistency?

HQ

- Using the law of the iterated logarithm, they established that consistency can be maintained with a penalty term for which

$$a_n = 2\,c\,\ln\ln n,$$

where c is any constant that exceeds 1.

- Hannan and Quinn chose $c = 1$ to obtain a criterion similar in form to AIC:

$$\text{HQ} = -2\ln f(y|\,\hat{\theta}_k) + 2(p+1)\ln\ln n.$$

- For $n \geq 16$, the term $\ln\ln n$ exceeds 1, making the penalty term of HQ more stringent than that of AIC.

HQ

- However, the term $\ln \ln n$ grows very slowly with n.
  - $\ln \ln 100 = 1.53$
  - $\ln \ln 1\,000 = 1.93$
  - $\ln \ln 10\,000 = 2.22$
  - $\ln \ln 1\,000\,000 = 2.63$
- Thus, the finite sample behavior of HQ is often more typical of an asymptotically efficient criterion than that of a consistent criterion.

Simulation Study to Illustrate Consistency

Study Outline:

- One thousand samples of size $n = 25\,000$ are generated from the AR(1) model
$$y_t = 0.9\, y_{t-1} + e_t,$$
$e_t \sim \textit{iid } N(0, 1)$.
- For every sample, candidate AR(p) models of orders $p = 1, 2, \ldots, 6$ are fit to the data.
- We examine the effectiveness of AIC, AICc, FPE, BIC, and HQ at selecting p.

Simulation Study to Illustrate Consistency

Order selections.

p	AIC	AICc	FPE	HQ	BIC
1	615	615	615	931	994
2	160	160	160	57	6
3	76	76	76	10	0
4	62	62	62	2	0
5	43	43	43	0	0
6	44	44	44	0	0

Simulation Study to Illustrate Asymptotic Efficiency

Study Outline:

- One thousand samples of size $n = 25\,000$ are generated from the MA(1) model

$$y_t = e_t + 0.7\, e_{t-1},$$

$e_t \sim iid\ N(0, 1)$.

- For every sample, candidate AR(p) models of orders $p = 1, 2, \ldots, 15$ are fit to the data.

Simulation Study to Illustrate Asymptotic Efficiency

- Consider an MA(1) process with moving-average coefficient θ_1.
- The process is invertible iff $|\theta_1| < 1$.
- The process can be represented as an infinite-order autoregression, AR(∞), with coefficients ϕ_i defined as $\phi_i = \theta_1^i; \quad i = 1, 2, \ldots$.
- Thus, all models in the candidate family are inherently underspecified.
- However, the "larger order" models might be viewed as more appropriate than the "smaller order" models.

Simulation Study to Illustrate Asymptotic Efficiency

- For each of the criteria AIC, AICc, FPE, BIC, and HQ, we compute the *average observed efficiency*, AOE (McQuarrie and Tsai, 1998).
  - The candidate AR(p) models are fit to every sample.
  - The fitted model which minimizes MSEP is identified, and MSEP is recorded for this model: MSEP_{min}.
  - The fitted model corresponding to the minimum value of the criterion is identified, and MSEP is recorded for this model: MSEP_{crit}.
  - The *observed efficiency* is defined as the ratio $\text{MSEP}_{min}/\text{MSEP}_{crit}$.
  - By definition, this ratio must be less than or equal to one.
  - Over the generated samples, the average of the observed efficiencies is computed, AOE.
- Larger values of AOE are indicative of greater predictive efficiency.

Simulation Study to Illustrate Asymptotic Efficiency

Order selections.

p	AIC	AICc	FPE	HQ	BIC
1-6	0	0	0	0	0
7	0	0	0	1	35
8	7	7	7	52	377
9	72	72	72	324	430
10	181	181	181	329	131
11	249	249	249	199	26
12	175	177	175	60	1
13	121	121	121	22	0
14	109	108	109	8	0
15	86	85	86	5	0

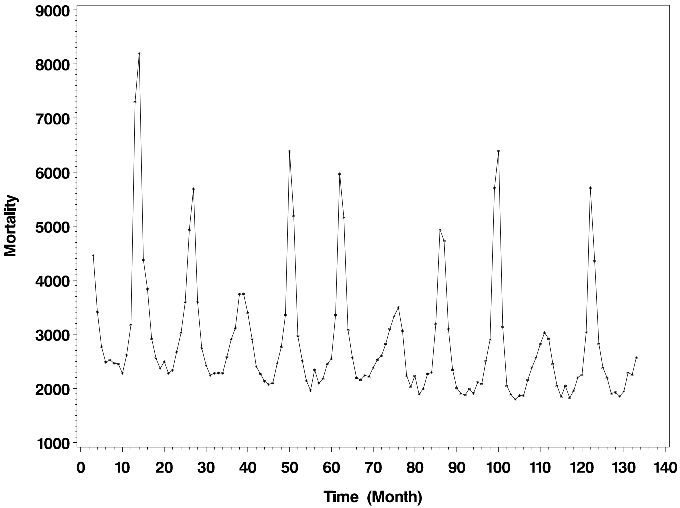
Simulation Study to Illustrate Asymptotic Efficiency

Average observed efficiency, AOE.

	AIC	AICc	FPE	HQ	BIC
AOE	0.830	0.830	0.830	0.742	0.562

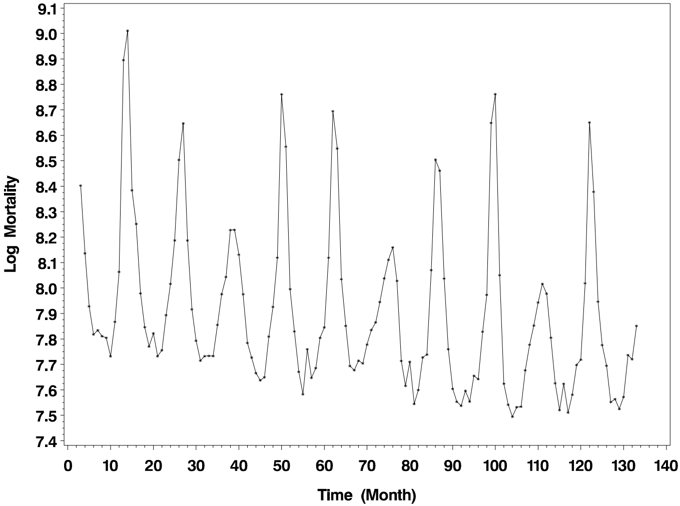
Application

- In this application, we attempt to model a series comprised of monthly U.S. mortality from influenza and pneumonia during an 11-year period extending from 1968 to 1978.
- Incidence and mortality series based on influenza-like-illnesses (ILI) exhibit strong yearly seasonal patterns, typically peaking during the winter months.
- However, such series are difficult to model because of differences in peak amplitudes due to *epidemic* years and *pandemic* years.


9000
8000
7000
6000
5000
4000
3000
2000
1000
Mortality
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Time (Month)

Application

- We begin by applying a log transformation to attenuate the differences in peak amplitudes.

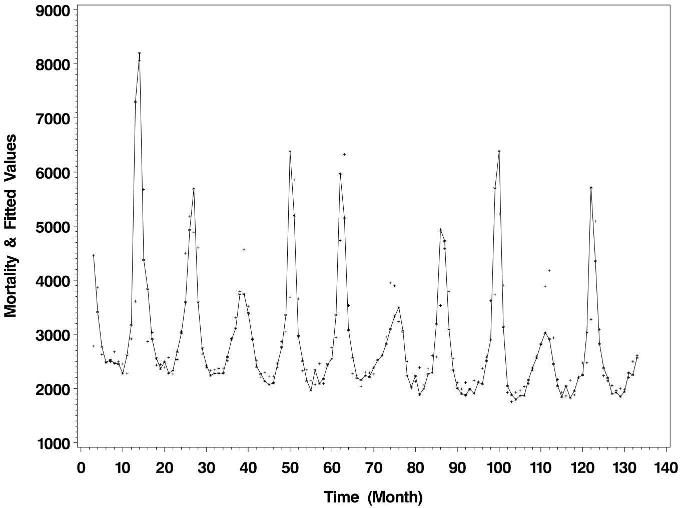

Log Mortality
9.1
9.0
8.9
8.8
8.7
8.6
8.5
8.4
8.3
8.2
8.1
8.0
7.9
7.8
7.7
7.6
7.5
7.4
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Time (Month)

Application

- The log-transformed series exhibits a subtle decreasing trend.
- However, it does not appear that this trend induces non-stationarity that needs to be accommodated by modeling the trend or by taking a first-order difference.
- We consider various autoregressive models for characterizing the log-transformed series.

Application

Model	AIC Value	BIC Value
AR lag 1	-26.8	-21.1
AR lag 1, 2	-71.6	-63.0
AR lag 1, 2, 3	-69.8	-58.3
AR lag 1, 2, 12	-114.1	-102.6
AR lag 1, 2, 12, 13	-112.1	-97.8


9000
8000
7000
6000
5000
4000
3000
2000
1000
Mortality & Fitted Values
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Time (Month)

Procedures for Model Validation

- As emphasized in Lecture VIII, a model selection criterion attempts to find the "best" fitted model among those models in a candidate collection.
- However, there is no guarantee that the selected model will be an adequate model, since all of the models in the candidate collection could be inappropriate.
- Recall (from Lecture I) that an optimal statistical model is characterized by three fundamental attributes.
  - Parsimony: model simplicity.
  - Goodness-of-fit: conformity of the fitted model to the data at hand.
  - Generalizability: applicability of the fitted model to describe or predict new data.

Procedures for Model Validation

- From Lecture VIII, recall that *model validation* refers to the process of ensuring that the selected fitted model provides an adequate fit to the data used in its own construction, and is capable of adequately describing and predicting new data.

Procedures for Model Validation

- Goodness-of-fit can be investigated by checking whether the residuals mimic the assumed distributional characteristics of the model errors: $e_t \sim \text{ iid } N(0, \sigma^2)$.
  - A stochastic process which consists of a sequence of uncorrelated, mean zero, homoscedastic random variables is called *white noise*.
  - A stochastic process which consists of a sequence of uncorrelated, mean zero, homoscedastic normal random variables is called *Gaussian white noise*.

Procedures for Model Validation

- For checking whether the residuals of a fitted time series model exhibit the properties of a Gaussian white noise process, the techniques discussed in Lecture VIII (for the traditional regression framework) can be applied.
- Temporal correlation in the residuals is a particular concern in modeling time series.
- The temporal correlation of the residuals can be investigated by inspecting the autocorrelation function of the residuals (ACF).

Procedures for Model Validation

- The *Ljung-Box test* checks for autocorrelation in the residuals by assessing whether the sample ACF exhibits the properties of the theoretical ACF for a white noise process.
  - The test is conducted at various lags, say $m, 2m, 3m, \ldots, Km$, where m and K are determined by the user.
  - The null distribution of each test statistic is chi-square, with degrees of freedom determined by the lag.
- The *Durbin-Watson test* provides an overall check for temporal correlation in the residuals, focusing on the lag-one autocorrelation.
  - The test is based on the statistic

$$\frac{\sum_{t=2}^{n}(\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{n}\hat{e}_t^2},$$

where $\hat{e}_t$, $t = 1, 2, \ldots, n$, are the residuals.

Procedures for Model Validation

- Since forecasting if often a key inferential objective in time series modeling, assessing the predictive effectiveness of a model may be crucial.
- A split-sample validation procedure is often used, where the *training sample* is comprised of the initial part of the series, and the *test sample* is comprised of a non-overlapping set of recent observations in the series.
- The model is fit based on the training sample.
- The predictive efficacy of the fitted model is then assessed based on the ability of the model to forecast the observations in the test sample.

Reference

- Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society (Series B)* **41**, 190–195.