

# 171:290 Model Selection

## Lecture IX: Criteria for Time Series Model Selection (Part I)

Joseph E. Cavanaugh

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

October 20, 2009

## Introduction

- A *time series* refers to a set of measurements on a random variable that are collected over time.
- The following are some examples of time series problems in biostatistics.
  - Using a patient's EEG (electroencephalogram) to determine whether the patient is in a normal state or an epileptic state.
  - Using a patient's EKG (electrocardiogram) to predict life-threatening ventricular cardiac arrhythmia.
  - Utilizing the frequency-to-tidal-volume ( $f/V_t$ ) breathing ratio for a ventilated critical care patient to determine whether the patient can be successfully extubated.
  - Building a forecasting model for weekly influenza-associated mortality by using weekly incidence for influenza-like-illnesses (ILI).

# Introduction

- Much of the groundbreaking work in model selection was developed in time series modeling frameworks.
- In these next two lectures, we review procedures for model selection as well as model validation in the time series setting.

# Introduction

## Outline:

- Brief Introduction to Time Series Analysis
  - Stationarity
  - Time and Frequency Domains
  - Autoregressive Models
  - Moving Average Models
- Autoregressive Model Selection Framework
- Final Prediction Error, FPE
  - Justification of FPE
  - Asymptotic Equivalence of FPE and AIC
- Application

## Introduction to Time Series

- We will assume that the data vector  $y$  consists of  $n$  measurements on a response variable, collected over equally spaced time points indexed by  $t = 1, 2, \dots, n$ .
- We will denote the response measurements as  $y_1, y_2, \dots, y_n$ .
- In time series applications, we assume that the  $y_t$  are temporally correlated.
- Time series methodologies attempt to both characterize and utilize this temporal correlation.
- Most time series methodologies can be classified as belonging to either the *time domain* or the *frequency domain*.

## Introduction to Time Series

- Four important tools in time series analysis are the *mean function*, the *variance function*, the *autocovariance function*, and the *autocorrelation function* (ACF).
- **Mean function:**  $\mu_t \equiv E(y_t)$ .
- **Variance function:**  $\sigma_t^2 \equiv \text{Var}(y_t)$ .
- **Autocovariance function:**  $C(y_{t+m}, y_t) \equiv \text{Cov}(y_{t+m}, y_t)$ .
- **Autocorrelation function** (ACF):

$$R(y_{t+m}, y_t) \equiv \frac{C(y_{t+m}, y_t)}{\sqrt{\sigma_{t+m}^2 \sigma_t^2}}.$$

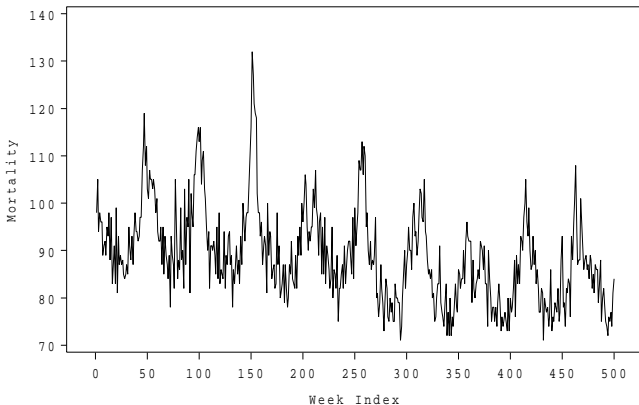
## Stationarity

- A common assumption utilized in time series analysis is that of *(weak) stationarity*.
- The series  $y_t$  is **weakly stationary** if (i) the mean function  $\mu_t$  is constant ( $\mu_t \equiv \mu$ ), and (ii) the autocovariance function  $C(y_{t+m}, y_t)$  depends only on the lag  $m$  ( $C(y_{t+m}, y_t) \equiv C(m)$ ).
- As a consequence of (ii), the variance function  $\sigma_t^2$  is constant:  $\sigma_t^2 = C(0) \equiv \sigma^2$ .
- Under stationarity, the autocorrelation function can be written as  $R(m) = C(m)/\sigma^2$ .

# Stationarity

- The following time series plot displays weekly mortality counts in Los Angeles county during the 1970's (from January 1, 1970 to December 1, 1979).
- Does the series appear stationary?

# Stationarity



## Stationarity

- Non-stationarity in the mean structure is often accommodated by modeling the trend, and treating the detrended series (i.e., the series with the trend extracted) as a stationary series.
- Non-stationarity in the mean structure can also be accommodated by *differencing*.
- A first-order difference of period  $d$  is defined by
$$y_t^* = y_t - y_{t-d}.$$

## Time Domain Methodologies

- Time domain methodologies directly analyze and model the original sample  $y_1, y_2, \dots, y_n$ .
- Popular modeling frameworks: autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), autoregressive conditionally heteroscedastic (ARCH), generalized autoregressive conditionally heteroscedastic (GARCH), state-space or dynamic linear modeling framework.
- Common challenge: To realistically model the autocovariance function.
- Common objective: forecasting or prediction.

## Frequency Domain Methodologies

- Frequency domain methodologies transform the original sample  $y_1, y_2, \dots, y_n$  using a discrete Fourier transform (DFT).
- The transformed data is indexed by a frequency  $\nu$  as opposed to a time  $t$ .
- The central objective is to characterize the frequencies and periodicities in the series.
- Nonstationary series can be analyzed in the frequency domain using either a localized DFT or a discrete wavelet transform (DWT).

## Autoregressive Model

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

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

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

- The autoregressive coefficients  $\phi_1, \phi_2, \dots, \phi_p$  must satisfy certain conditions for the  $AR(p)$  process to be stationary.
- The ACF for an  $AR(p)$  process decays quickly, but is nonzero for all lags.
- Common problem in AR modeling: the determination of the order  $p$ .

## Moving Average Model

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

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

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

- For any values of the moving average coefficients  $\theta_1, \theta_2, \dots, \theta_q$ , the  $MA(q)$  process is stationary.
- The ACF for an  $MA(q)$  process decays until lag  $q$ , and is zero for all lags beyond  $q$ .
- Common problem in MA modeling: the determination of the order  $q$ .

## AR and MA Relationships

- An *invertible*  $MA(q)$  process can be represented as an infinite-order autoregression,  $AR(\infty)$ , with coefficients  $\phi_i$  that decay in magnitude as  $i$  increases.
- The moving average coefficients  $\theta_1, \theta_2, \dots, \theta_q$  must satisfy certain conditions for the  $MA(q)$  process to be invertible.
- A stationary  $AR(p)$  process can be represented as an infinite-order moving average,  $MA(\infty)$ , with coefficients  $\theta_i$  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  $k = (p + 1)$ .
- **Parameter vector:**  $\theta_k = (\phi_1, \phi_2, \dots, \phi_p, \sigma^2)'$ .
- **True parameter vector:**  $\theta_o = (\phi_1^o, \phi_2^o, \dots, \phi_p^o, \sigma_o^2)'$ .
- **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):

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

- Other popular criteria for autoregressive model selection: final prediction error, FPE (Akaike, 1969); the Hannan and Quinn (1979) criterion, HQ.

## Final Prediction Error, FPE

- Final prediction error, FPE, was proposed by Akaike (1969) for the selection of the order of an autoregression.
- In the large-sample justification of FPE, we assume that  $f(y|\theta_o) \in \mathcal{F}(k)$  (as in the large-sample justification of AIC).
- In the autoregressive setting, this assumption amounts to requiring that the order of the true autoregressive model,  $p_o$ , is less than or equal to  $p$ , the order of the candidate autoregressive model.

## Final Prediction Error, FPE

- Suppose we wish to choose a fitted AR(p) model that will yield an accurate predictor of  $y_{n+1}$ .
- For an AR(p) model,

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

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

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

## Final Prediction Error, FPE

- The mean square error of prediction (MSEP) can be viewed as a *discrepancy*; i.e., a measure that reflects the disparity between the true model  $f(y|\theta_o)$  and the candidate model  $f(y|\theta_k)$ .
- MSEP depends upon both the parameters of the true model,  $\theta_o$ , and the parameters of the candidate model,  $\theta_k$ .
- Let  $R_o(m)$  denote the true ACF. One can show that

$$d(\theta_k) = \sigma_o^2 + \sigma_o^2 \sum_{i=1}^p \sum_{j=1}^p (\phi_i - \phi_i^o) R_o(i-j) (\phi_j - \phi_j^o),$$

where

$$\phi_{p_o+1}^o = \phi_{p_o+2}^o = \dots = \phi_p^o = 0.$$

## Final Prediction Error, FPE

- The corresponding *expected discrepancy* is given by

$$\begin{aligned}\Delta(k) &= E\{d(\hat{\theta}_k)\} \\ &= \sigma_o^2 + \sigma_o^2 E \left\{ \sum_{i=1}^p \sum_{j=1}^p (\hat{\phi}_i - \phi_i^o) R_o(i-j) (\hat{\phi}_j - \phi_j^o) \right\}.\end{aligned}$$

- In large-sample settings,

$$n \sum_{i=1}^p \sum_{j=1}^p (\hat{\phi}_i - \phi_i^o) R_o(i-j) (\hat{\phi}_j - \phi_j^o)$$

has an approximate chi-squared distribution with  $p$  degrees of freedom.

## Final Prediction Error, FPE

- Thus, if  $n$  is large, we have

$$\begin{aligned}\Delta(k) &= \sigma_o^2 + \left(\frac{\sigma_o^2}{n}\right) \mathbb{E} \left\{ n \sum_{i=1}^p \sum_{j=1}^p (\hat{\phi}_i - \phi_i^o) R_o(i-j) (\hat{\phi}_j - \phi_j^o) \right\} \\ &\approx \sigma_o^2 + \left(\frac{\sigma_o^2}{n}\right) \mathbb{E} \{ \chi_p^2 \} \\ &= \sigma_o^2 \left(1 + \frac{p}{n}\right).\end{aligned}$$

## Final Prediction Error, FPE

- Define the statistic

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

where  $\hat{\sigma}^2$  is the MLE of  $\sigma^2$ .

- If  $n$  is large, we have  $E\{\hat{\sigma}^2\} \approx ((n-p)/n)\sigma_o^2$ .
- Thus, for large  $n$ , we have

$$\begin{aligned} E\{\text{FPE}\} &= \left( \frac{n+p}{n-p} \right) E\{\hat{\sigma}^2\} \\ &\approx \left( \frac{n+p}{n-p} \right) \left( \left( \frac{n-p}{n} \right) \sigma_o^2 \right) \\ &= \sigma_o^2 \left( 1 + \frac{p}{n} \right) \\ &\approx \Delta(k). \end{aligned}$$

## Final Prediction Error, FPE

- In large-sample settings, FPE therefore provides an approximately unbiased estimator of  $\Delta(k)$ .
- How does the penalization imposed by FPE compare to that imposed by AIC?
- We will show that choosing the fitted model corresponding to the minimum value of FPE is asymptotically equivalent to choosing the fitted model corresponding to the minimum value of AIC.
- In the autoregressive setting,

$$-2 \ln f(y | \hat{\theta}_k) \approx n \ln \hat{\sigma}^2 + n(\ln 2\pi + 1).$$

## Final Prediction Error, FPE

- Note that choosing the fitted model corresponding to minimum value of FPE is equivalent to choosing the fitted model corresponding to the minimum value of

$$\begin{aligned} & n \ln\{\text{FPE}\} + n(2\pi + 1) + 2 \\ &= n \ln \left\{ \left( \frac{n+p}{n-p} \right) \hat{\sigma}^2 \right\} + n(2\pi + 1) + 2 \\ &= \{n \ln \hat{\sigma}^2 + n(2\pi + 1)\} + n \ln \left( \frac{n+p}{n-p} \right) + 2 \\ &\approx -2 \ln f(y|\hat{\theta}_k) + n \ln \left( \frac{n+p}{n-p} \right) + 2. \end{aligned}$$

## Final Prediction Error, FPE

- Consider a first-order Taylor series expansion of  $n \ln \{(n+p)/(n-p)\}$  in the argument  $\{(n+p)/(n-p)\}$  about the point 1.
- We have

$$\begin{aligned} n \ln \left( \frac{n+p}{n-p} \right) &\approx n \ln(1) + \left( \left( \frac{n+p}{n-p} \right) - 1 \right) n \\ &= \frac{2np}{(n-p)}. \end{aligned}$$

## Final Prediction Error, FPE

- Thus, choosing the fitted model corresponding to the minimum value of FPE is asymptotically equivalent to choosing the fitted model corresponding to the minimum value of

$$\begin{aligned} & -2 \ln f(y | \hat{\theta}_k) + n \ln \left( \frac{n+p}{n-p} \right) + 2 \\ & \approx -2 \ln f(y | \hat{\theta}_k) + \frac{2np}{(n-p)} + 2 \\ & = -2 \ln f(y | \hat{\theta}_k) + a_n(p+1), \end{aligned}$$

where

$$a_n = 2 \left\{ \frac{n}{(n-p)} - \frac{p}{(n-p)(p+1)} \right\}.$$

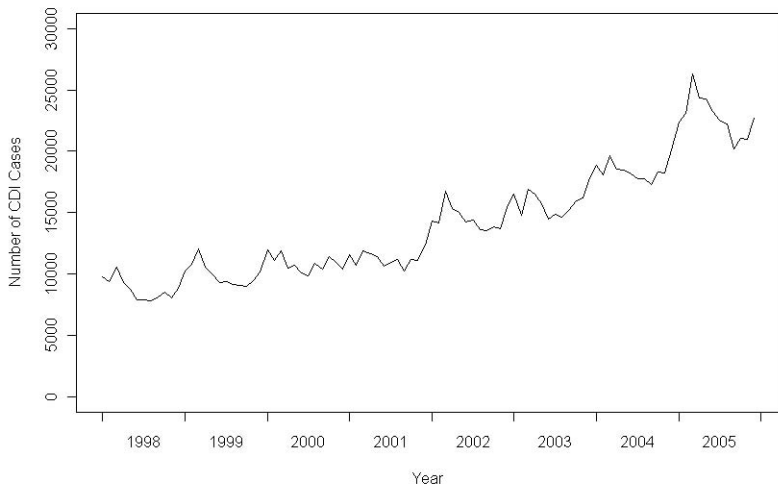
- Note that  $a_n \rightarrow 2$  as  $n \rightarrow \infty$ .

## Application

- The following application is taken from the paper “A Time Series Analysis of *Clostridium difficile* and its Seasonal Association with Influenza,” by Polgreen PM, Yang M, Bohnett LC, and Cavanaugh JE. (To appear in *Infection Control and Hospital Epidemiology*.)
- *Clostridium difficile* infection (CDI) is a nosocomial infectious disease of the digestive track that is frequently acquired by hospitalized patients.
- The use of antibiotics increases the risk of CDI, since antibiotics may adversely affect the natural flora of the digestive track.

## Application

- In this paper, it was hypothesized that CDI is a seasonal disease with increasing incidence.
  - “Because antimicrobial use is the major risk factor for CDI, and because the use of antimicrobials varies by season, peaking in the winter months, the epidemiology of CDI may also be seasonal.”
  - “Furthermore, because antibiotic use increases during influenza seasons, we hypothesize that the seasonal variation of CDI may be related to influenza activity.”
  - “The purpose of this study is to characterize the temporal progression of the monthly incidence of CDI, and to determine if the incidence of CDI is associated with the seasonal variation of influenza.”
- The following time series plot illustrates national CDI incidence by month from 1998 to 2005.



## Application

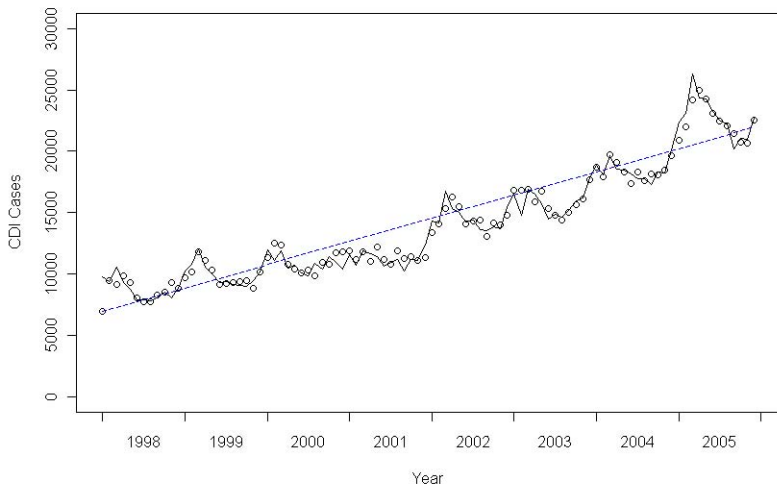
- To establish on a national basis that CDI is a seasonal disease with increasing incidence, we fit a time series autoregressive model with a linear trend to the incidence series.
- The final model was selected using backwards elimination (in the exploratory phase) and AIC.
  - “The final autoregressive model for national CDI incidence includes a positive linear trend ( $p$ -value  $< 0.0001$ ) to account for temporal case escalation, lags of 1, 2, and 3 months to account for recent activity (all  $p$ -values  $< 0.10$ ), and lags of 12 and 13 months to account for yearly seasonal variation (both  $p$ -values  $< 0.0001$ ).”
- We compare the AIC value for this final fitted model to that of other candidate fitted models.

## Application

Model	AIC Value
Trend	1706.9
Trend + AR lag 1	1602.1
Trend + AR lag 1, 2	1603.8
Trend + AR lag 1, 2, 3	1591.1
Trend + AR lag 1, 2, 3, 12	1580.2
Trend + AR lag 1, 2, 3, 12, 13	1541.3
Trend + AR lag 1, 2, 3, 12, 13, 14	1540.2

## Application

- The preceding AIC values clearly show that the seasonal cyclic behavior of the CDI incidence series is an important component of the series.
- The following time series plot illustrates the trend and the fitted values for the final model superimposed against the CDI incidence series.



## Reference

- Akaike, H. (1969). Fitting autoregressive models for prediction. *The Annals of the Institute of Statistical Mathematics* **21**, 243–247.

## Upcoming Topics

### Topics for Lecture X:

- Consistency and asymptotic efficiency in the autoregressive setting.
- The Hannan and Quinn criterion, HQ.
- A simulation study to illustrate consistency and asymptotic efficiency.
- A second time series modeling application.