Lecture 10

ARMA Models: Estimation, Diagnostics, and Model Selection

Reading: Bowerman, O'Connell, and Koehler (2005): Capter 10.1-10.2; Cryer and Chen (2008): Chapter 7.3-7.5; Chapter 8.1

MATH 4070: Regression and Time-Series Analysis

Whitney Huang
Clemson University

Notes

Agenda

1. Parameter Estimation
2. Model Diagnostics and Selection

Notes

Estimation of the ARMA Process Parameters

Suppose we choose an ARMA(p, q) model for a zero-mean $\{\eta_t\}$

- Need to estimate the $p + q + 1$ parameters:
  - AR component $\{\phi_1, \cdots, \phi_p\}$
  - MA component $\{\theta_1, \cdots, \theta_q\}$
  - $\mathrm{Var}(Z_t) = \sigma^2$
- One strategy:
  - Do some preliminary estimation of the model parameters (e.g., via Yule-Walker estimates)
  - Follow-up with maximum likelihood estimation with Gaussian assumption

Notes

The Yule-Walker Method

Suppose η_t is a causal AR(p) process

$$\eta_t - \phi_1\eta_{t-1} - \cdots - \phi_p\eta_{t-p} = Z_t$$

To estimate the parameters $\{\phi_1, \cdots, \phi_p\}$, we use a method of moments estimation scheme:

- Let $h = 0, 1, \cdots, p$. We multiply η_{t-h} to both sides

$$\eta_t\eta_{t-h} - \phi_1\eta_{t-1}\eta_{t-h} - \cdots - \phi_p\eta_{t-p}\eta_{t-h} = Z_t\eta_{t-h}$$

- Taking expectations:

$$\mathbb{E}(\eta_t\eta_{t-h}) - \phi_1\mathbb{E}(\eta_{t-1}\eta_{t-h}) - \cdots - \phi_p\mathbb{E}(\eta_{t-p}\eta_{t-h}) = \mathbb{E}(Z_t\eta_{t-h}),$$

  we get

$$\boxed{\gamma(h) - \phi_1\gamma(h-1) - \cdots - \phi_p\gamma(h-p) = \mathbb{E}(Z_t\eta_{t-h})}$$

Notes

The Yule-Walker Equations

- When $h = 0$, $\mathbb{E}(Z_t\eta_{t-h}) = \mathrm{Cov}(Z_t, \eta_t) = \sigma^2$ (Why?) Therefore, we have

$$\gamma(0) - \sum_{j=1}^{p}\phi_j\gamma(j) = \sigma^2$$

- When $h > 0$, Z_t is uncorrelated with η_{t-h} (because the assumption of causality), thus $\mathbb{E}(Z_t\eta_{t-h}) = 0$ and we have

$$\gamma(h) - \sum_{j=1}^{p}\phi_j\gamma(h-j) = 0, \quad h = 1, 2, \cdots, p$$

- The Yule-Walker estimates are the solution of these equations when we replace $\gamma(h)$ by $\hat{\gamma}(h)$

Notes

The Yule-Walker Equations in Matrix Form

Let $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \cdots, \hat{\phi}_p)^T$ be an estimate for $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)^T$ and let

$$\hat{\mathbf{\Gamma}} = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(p-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}(p-1) & \hat{\gamma}(p-2) & \cdots & \hat{\gamma}(0) \end{bmatrix}.$$

Then the Yule-Walker estimates of $\boldsymbol{\phi}$ and σ^2 are

$$\hat{\boldsymbol{\phi}} = \hat{\mathbf{\Gamma}}^{-1}\hat{\boldsymbol{\gamma}},$$

and

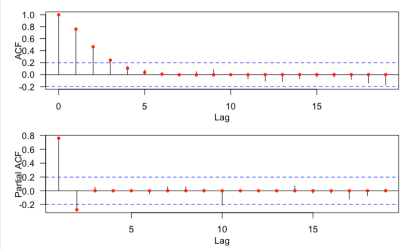
$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{\gamma}},$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}(1), \cdots, \hat{\gamma}(p))^T$

Notes

Lake Huron Example in R

```
```{r}
YW_est <- ar(lm$residuals, aic = F, order.max = 2, method = "yw")
# plot sample and estimated acf/pacf
par(las = 1, mgp = c(2.2, 1, 0), mar = c(3.6, 3.6, 0.6, 0.6), mfrow = c(2, 1))
acf(lm$residuals)
acf_YWest <- ARMAacf(ar = YW_est$ar, lag.max = 23)
points(0:23, acf_YWest, col = "red", pch = 16, cex = 0.8)
pacf(lm$residuals)
pacf_YWest <- ARMAacf(ar = YW_est$ar, lag.max = 23, pacf = T)
points(1:23, pacf_YWest, col = "red", pch = 16, cex = 0.8)
```
```



Notes

Remarks on the Yule-Walker Method

- For large sample size, Yule-Walker estimator have (approximately) the same sampling distribution as maximum likelihood estimator (MLE), but with small sample size Yule-Walker estimator can be far less efficient than the MLE
- The Yule-Walker method is a poor procedure for MA(q) and ARMA(p,q) processes with $q > 0$ (see Cryer Chan 2008, p. 150-151)
- We move on the more versatile and popular method for estimating ARMA(p,q) parameters–maximum likelihood estimation[1]

[1] See **Least Squares Estimation** in Chapter 7.2 of Cryer and Chan (2008).

Notes

Maximum Likelihood Estimation

- The setup:
  - Model: $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ has joint probability density function $f(\boldsymbol{x}; \boldsymbol{\omega})$ where $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_p)$ is a vector of p parameters
  - Data: $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$
- The likelihood function is defined as the the "likelihood" of the data, $\boldsymbol{x}$, given the parameters, $\boldsymbol{\omega}$

$$L_n(\boldsymbol{\omega}) = f(\boldsymbol{x}; \boldsymbol{\omega})$$

- The maximum likelihood estimate (MLE) is the value of $\boldsymbol{\omega}$ which maximizes the likelihood, $L_n(\boldsymbol{\omega})$, of the data $\boldsymbol{x}$:

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} L_n(\boldsymbol{\omega}).$$

It is equivalent (and often easier) to maximize the log likelihood,

$$\ell_n(\boldsymbol{\omega}) = \log L_n(\boldsymbol{\omega})$$

Notes

The MLE for an i.i.d. Gaussian Process

Suppose $\{X_t\}$ be a Gaussian i.i.d. process with mean μ and variance σ^2. We observe a time series $\boldsymbol{x} = (x_1, \cdots, x_n)^T$.

- The likelihood function is

$$\begin{aligned} L_n(\mu, \sigma^2) &= f(\boldsymbol{x}|\mu, \sigma^2) \\ &= \prod_{t=1}^{n} f(x_t|\mu, \sigma) \\ &= \prod_{t=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_t - \mu)^2}{2\sigma^2} \right] \right\} \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{t=1}^{n} (x_t - \mu)^2}{2\sigma^2} \right] \end{aligned}$$

- The log-likelihood function is

$$\begin{aligned} \ell_n(\mu, \sigma^2) &= \log L_n(\mu, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{\sum_{t=1}^{n} (x_t - \mu)^2}{2\sigma^2} \end{aligned}$$

$$\Rightarrow \hat{\mu}_{\text{MLE}} = \frac{\sum_{t=1}^{n} X_t}{n} = \bar{X}, \quad \hat{\sigma}^2_{\text{MLE}} = \frac{\sum_{t=1}^{n} (X_t - \bar{X})^2}{n}$$

Notes

Likelihood for Stationary Gaussian Time Series Models

Suppose $\{X_t\}$ be a mean zero stationary Gaussian time series with ACVF $\gamma(h)$. If $\gamma(h)$ depends on p parameters, $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)$

- The likelihood of the data $\boldsymbol{x} = (x_1, \cdots, x_n)$ given the parameters $\boldsymbol{\omega}$ is

$$L_n(\boldsymbol{\omega}) = (2\pi)^{-n/2} |\boldsymbol{\Gamma}|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{x} \right),$$

  where $\boldsymbol{\Gamma}$ is the covariance matrix of $\boldsymbol{X} = (X_1, \cdots, X_n)^T$, $|\boldsymbol{\Gamma}|$ is the determinant of the matrix $\boldsymbol{\Gamma}$, and $\boldsymbol{\Gamma}^{-1}$ is the inverse of the matrix $\boldsymbol{\Gamma}$

- The log-likelihood is

$$\ell_n(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Gamma}| - \frac{1}{2} \boldsymbol{x}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{x}$$

  Typically need to solve it numerically

ARMA Models: Estimation, Diagnostics, and Model Selection
School of Mathematical and Statistical Sciences, Clemson University
Parameter Estimation
Model Diagnostics and Selection

Notes

Decomposing Joint Density into Conditional Densities

A joint distribution can be represented as the product of conditionals and a marginal distribution

- The simple version for $n = 2$ is:

$$f(x_1, x_2) = f(x_2|x_1) f(x_1)$$

- Extending for general n we get the following expression for the likelihood:

$$L_n(\boldsymbol{\theta}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = f(x_1) \prod_{t=2}^{n} f(x_t|x_{t-1}, \cdots, x_1; \boldsymbol{\theta}),$$

  and the log-likelihood is

$$\ell_n(\boldsymbol{\theta}) = \log f(\boldsymbol{x}; \boldsymbol{\theta}) = \log(f(x_1)) + \sum_{t=2}^{n} \log f(x_t|x_{t-1}, \cdots, x_1; \boldsymbol{\theta}).$$

ARMA Models: Estimation, Diagnostics, and Model Selection
School of Mathematical and Statistical Sciences, Clemson University
Parameter Estimation
Model Diagnostics and Selection

Notes

AR(1) Log-likelihood

Let $\{\eta_1, \eta_2, \cdots, \eta_n\}$ be a realization of a zero-mean stationary AR(1) Gaussian time series. Let $\boldsymbol{\theta} = (\phi, \sigma^2)$

$$\ell_n(\boldsymbol{\theta}) = \underbrace{\log(f(\eta_1))}_{\ell_{n,1}} + \underbrace{\sum_{t=2}^{n} \log f(\eta_t | \eta_{t-1}, \cdots, \eta_1; \boldsymbol{\theta})}_{\ell_{n,2}}.$$

Note that for $t \geq 2$, $f(\eta_t|\eta_{t-1}, \cdots, \eta_1) = f(\eta_t|\eta_{t-1})$, where $[\eta_t|\eta_{t-1}] \sim \mathrm{N}(\phi\eta_{t-1}, \sigma^2) \Rightarrow \ell_{n,2} =$

$$-\frac{(n-1)}{2}\log 2\pi - \frac{(n-1)}{2}\log\sigma^2 - \frac{\sum_{t=2}^{n}(\eta_t - \phi\eta_{t-1})^2}{2\sigma^2}$$

Also, we know $[\eta_1] \sim \mathrm{N}\left(0, \frac{\sigma^2}{(1-\phi^2)}\right) \Rightarrow \ell_{1,n} =$

$$\frac{-\log 2\pi}{2} - \frac{\log\sigma^2}{2} + \frac{\log(1-\phi^2)}{2} - \frac{(1-\phi^2)\eta_1^2}{2\sigma^2}$$

$$\Rightarrow \ell_n(\boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{\sum_{t=2}^{n}(\eta_t - \phi\eta_{t-1})^2}{2\sigma^2} + \frac{\log(1-\phi^2)}{2} - \frac{(1-\phi^2)\eta_1^2}{2\sigma^2}$$

Notes

AR(1) Log-likelihood Cont'd

$$\ell_n(\boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 + \frac{\log(1-\phi^2)}{2} - \frac{S(\phi)}{2\sigma^2},$$

where $S(\phi) = \sum_{t=2}^{n}(\eta_t - \phi\eta_{t-1})^2 + (1-\phi^2)\eta_1^2$

- For given value of ϕ, $\ell_n(\phi, \sigma^2)$ can be maximized analytically with respect to σ^2

$$\hat{\sigma}^2 = \frac{S(\hat{\phi})}{n}$$

- Estimation of ϕ can be simplified by maximizing the **conditional sum-of-squares** $(\sum_{t=2}^{n}(\eta_t - \phi\eta_{t-1})^2)$
- **Standard errors** can be obtained by computing the inverse of the **Hessian matrix**: $\mathrm{Var}(\hat{\boldsymbol{\theta}}) = H(\hat{\boldsymbol{\theta}})^{-1}$, where $H(\boldsymbol{\theta}) = \frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$

Notes

`arima` in `R` with the Lake Huron Example

arima: ARIMA Modelling of Time Series

Description

Fit an ARIMA model to a univariate time series.

Usage

```
arima(x, order = c(0L, 0L, 0L),
      seasonal = list(order = c(0L, 0L, 0L), period = NA),
      xreg = NULL, include.mean = TRUE,
      transform.pars = TRUE,
      fixed = NULL, init = NULL,
      method = c("CSS-ML", "ML", "CSS"), n.cond,
      SSinit = c("Gardner1980", "Rossignol2011"),
      optim.method = "BFGS",
      optim.control = list(), kappa = 1e6)
```

```
```{r}
(MLE_est1 <- arima(lm$residuals, order = c(2, 0, 0), method = "ML"))
```
```

```
Call:
arima(x = lm$residuals, order = c(2, 0, 0), method = "ML")

Coefficients:
         ar1      ar2  intercept
      1.0047  -0.2919     0.0197
s.e.  0.0977   0.1004     0.2350

sigma^2 estimated as 0.4571:  log likelihood = -101.25,  aic = 210.5
```


Notes

Inference for the ARMA Parameters

Motivating example: What is an approximate 95% CI for ϕ_1 in an AR(1) model?

- Let $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)$ and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q)$ denote the ARMA parameters (excluding σ^2), and let $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ be the ML estimates of ϕ and $\boldsymbol{\theta}$. Then for "large" n, $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ have approximately a joint normal distribution:

$$\begin{bmatrix} \hat{\boldsymbol{\phi}} \\ \hat{\boldsymbol{\theta}} \end{bmatrix} \dot{\sim} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\phi} \\ \boldsymbol{\theta} \end{bmatrix}, \frac{V(\boldsymbol{\phi}, \boldsymbol{\theta})}{n} \right)$$

- $V(\boldsymbol{\phi}, \boldsymbol{\theta})$ is a known $(p+q) \times (p+q)$ matrix depending on the ARMA parameters

Notes

$V(\phi, \theta)$ for AR Processes

- For an AR(p) process

$$V(\boldsymbol{\phi}) = \sigma^2 \Gamma^{-1},$$

  where Γ is the $p \times p$ covariance matrix of the series $(\eta_1, \cdots, \eta_p)$

- AR(1) process:

$$V(\phi_1) = 1 - \phi_1^2$$

- AR(2) process:

$$V(\phi_1, \phi_2) = \begin{bmatrix} 1 - \phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1 - \phi_2^2 \end{bmatrix}$$

Notes

Other Examples of $V(\phi, \theta)$

- MA(1) process:

$$V(\theta_1) = 1 - \theta_1^2$$

- MA(2) process:

$$V(\theta_1, \theta_2) = \begin{bmatrix} 1 - \theta_2^2 & \theta_1(1-\theta_2) \\ \theta_1(1-\theta_2) & 1 - \theta_2^2 \end{bmatrix}$$

- Casual and invertible ARMA(1,1) process

$$V(\phi, \theta) = \frac{1 + \phi\theta}{(\phi+\theta)^2} \begin{bmatrix} (1-\phi^2)(1+\phi\theta) & -(1-\phi^2)(1-\theta^2) \\ -(1-\phi^2)(1-\theta^2) & 1 - \theta_2^2 \end{bmatrix}$$

- More generally, for "small" n, the covariance matrix of $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ can be approximated using the second derivatives of the log-likelihood function, known as the Hessian matrix

Notes

MLE for Trend and Temporal Correlation in One Step

~~~~
```{r}
(MLE_est4 <- arima(LakeHuron, order = c(2, 0, 0), xreg = yr))
```
~~~~

```
Call:
arima(x = LakeHuron, order = c(2, 0, 0), xreg = yr)

Coefficients:
         ar1      ar2  intercept     xreg
      1.0048  -0.2913   620.5115  -0.0216
s.e.  0.0976   0.1004    15.5771   0.0081

sigma^2 estimated as 0.4566:  log likelihood = -101.2,  aic = 210.4
```

Fitted model:

$$Y_t = 620.51 - 0.022\text{Year} + \eta_t,$$

where

$$\eta_t = 1.00\eta_{t-1} - 0.29\eta_{t-2} + Z_t, \quad Z_t \sim \text{N}(0, \sigma^2 = 0.46^2).$$

Notes

What About Non-Gaussian Processes?

It is more challenging to express the joint distribution of X_t for non-Gaussian processes. Instead, we often rely on the Gaussian likelihood as an approximate likelihood

- In practice:
  - Transform the data to make the series as close to Gaussian as possible (e.g., using a log, square-root, or Box-Cox transformation)
  - Then use the Gaussian likelihood to estimate parameters, assuming the transformed series follows a near-Gaussian structure
  - For many real-world applications, this approximation works well and simplifies estimation. However, residual diagnostics are needed to ensure the model fits the data adequately

Notes

Assessing Fit / Comparing Different Time Series Models

- We can use diagnostic plots for the "residuals" of the fitted time series, along with Box tests to assess whether an i.i.d. process is reasonable

```
> Box.test(YW_est$resid[-(1:2)], type = "Ljung-Box")

	Box-Ljung test

data:  YW_est$resid[-(1:2)]
X-squared = 0.56352, df = 1, p-value = 0.4528
```

- Use confidence intervals for the parameters. Intervals that contain zero may indicate that we can simplify the model
- We can also use model selection criteria, such as AIC, to compare between different models

Notes

Diagnostics via the Time Series Residuals

- Recall the innovations are given by

$$U_t = X_t - \hat{X}_t$$

- Under a Gaussian model, $\{U_t : t = 1, \cdots, T\}$ is an independent set of RVs with

$$U_t \sim \mathrm{N}(0, \nu_{t-1}) \overset{d}{=} \sigma \mathrm{N}(0, r_{t-1}).$$

- Define the residuals $\{R_t\}$ by

$$R_t = \frac{U_t}{\sqrt{r_{t-1}}} = \frac{X_t - \hat{X}_t}{\sqrt{r_{t-1}}}$$

Under Gaussian model $R_t \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma^2)$

Notes

ARMA Order Selection

- We would prefer to use models that compromise between a small residual error $\hat{\sigma}^2$ and a small number of parameters $(p + q + 1)$
- To choose the order (p and q) of ARMA model it makes sense to penalize models with a large number of parameters
- Here we consider an information based criteria to compare models

ARMA Models: Estimation, Diagnostics, and Model Selection
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Parameter Estimation
Model Diagnostics and Selection

Notes

Akaike Information Criterion (AIC)

- The Akaike information criterion (AIC) is defined by

$$\mathrm{AIC} = -2\ell_n(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) + 2(p + q + 1)$$

- We choose the values of p and q that minimizes the AIC value
- For AR(p) models, AIC tends to overestimate p. The bias corrected version is

$$\mathrm{AICc} = 2\ell_n(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) + \frac{2n(p + q + 1)}{(n-1) - (p + q + 1)}$$

ARMA Models: Estimation, Diagnostics, and Model Selection
School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University
Parameter Estimation
Model Diagnostics and Selection

Notes

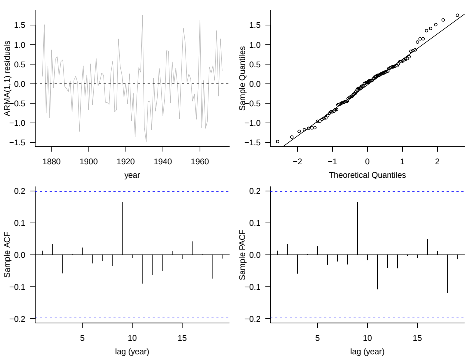
Lake Huron Example: AIC and AICc

```
m1 <- arima(LakeHuron, order = c(1, 0, 0), xreg = yr)
m2 <- arima(LakeHuron, order = c(1, 0, 1), xreg = yr)
m3 <- arima(LakeHuron, order = c(2, 0, 0), xreg = yr)
m4 <- arima(LakeHuron, order = c(2, 0, 1), xreg = yr)
AIC(m1); AIC(m2); AIC(m3); AIC(m4)
library(MuMIn)
AICc(m1); AICc(m2); AICc(m3); AICc(m4)
```

```
[1] 218.4501
[1] 212.3954
[1] 212.3965
[1] 214.0638
[1] 218.8803
[1] 213.0476
[1] 213.0487
[1] 214.9868
```


Notes

Lake Huron Model Diagnostics




```
> Box.test(resids, lag = 10, type = "Ljung-Box")

        Box-Ljung test

data:  resids
X-squared = 3.7882, df = 10, p-value = 0.9564
```


Notes

Notes