Lecture 5

Autoregressive-Moving Average Model I

Readings: CC08 Chapter 4.4-4.6, 7.1, 7.3, 7.4; BD16 Chapter 2.3, 3.1-3.2; SS17 Chapter 3.1-3.3, 3.5

MATH 8090 Time Series Analysis
Week 5

Whitney Huang
Clemson University

Agenda

1. Autoregressive-Moving Average Model: Stationarity, Causality, and Invertibility
2. Partial Autocorrelation Functions
3. Parameter Estimation

ARMA(p, q) Processes

$\{\eta_t\}$ is an ARMA(p, q) process if it satisfies

$$\eta_t - \sum_{i=1}^{p} \phi_i \eta_{t-i} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j},$$

where $\{Z_t\}$ is a $\mathrm{WN}(0, \sigma^2)$ process.

- Let $\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$ and $\theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$. Then we can write it as

$$\phi(B)\eta_t = \theta(B)Z_t$$

- An ARMA(p, q) process $\{\tilde{\eta}_t\}$ with mean μ can be written as

$$\phi(B)(\tilde{\eta}_t - \mu) = \theta(B)Z_t$$

A Stationary Solution to the ARMA Equation

A zero-mean ARMA process is stationary if it can be written as a linear process, i.e., $\eta_t = \psi(B)Z_t$, where $\psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j$ for an absolutely summable sequence $\{\psi_j\}$

- This only happens if one can "divide" by $\phi(B)$, i.e., it is stationary only if the following makes senese:

$$(\phi(B))^{-1}\,\phi(B)\eta_t = (\phi(B))^{-1}\,\theta(B)Z_t$$

- Let's forget about B is the backshift operator and replace it with z. Now consider whether we can divide $\theta(z)$ by $\phi(z)$

The Roots of AR Characteristic Polynomial and Stationarity

- A root of the polynomial $f(z) = \sum_{j=0}^{p} a_j z^j$ is a value ξ such that $f(\xi) = 0 \Rightarrow$ it can be real-valued $\mathbb{R}$ or complex-valued $\mathbb{C}$

- For example, a root can take the form $\xi = a + b\,i$ for real number a and b. The modulus of a complex number $|\xi|$ is defined by

$$|\xi| = \sqrt{a^2 + b^2}$$

- For any ARMA(p,q) process, a stationary and unique solution exists if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0,$$

for all $|z| = 1$.

Note: Stationarity of the ARMA process has nothing to do with the MA polynomial!

AR(4) Example

Consider the following AR(4) process

$$\eta_t = 2.7607\eta_{t-1} - 3.8106\eta_{t-2} + 2.6535\eta_{t-3} - 0.9238\eta_{t-4} + Z_t,$$

the AR characteristic polynomial is

$$\phi(z) = 1 - 2.7607z + 3.8106z^2 - 2.6535z^3 + 0.9238z^4$$

- Hard to find the roots of $\phi(z)$ –we use the `polyroot` function in `R`:
- Use `Mod` in `R` to calculate the modulus of the roots
- **Conclusion:**

Causal ARMA Processes

An ARMA process is causal if there exists constants $\{\psi_j\}$ with $\sum_{j=0}^{\infty} |\psi_j| < 0$ and $\eta_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, that is, we can write $\{\eta_t\}$ as an MA(∞) process depending only on the current and past values of $\{Z_t\}$

- Equivalently, an ARMA process is causal if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0,$$

  for all $|z| \leq 1$

- The previous AR(4) example is causal since each zero, ξ, of $\phi(\cdot)$ is such that $|\xi| > 1$

Invertible ARMA Processes

An ARMA process is invertible if there exists constants $\{\pi_j\}$ with $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$Z_t = \sum_{j=0}^{\infty} \pi_j \eta_{t-j},$$

that is, we can write $\{Z_t\}$ as an AR(∞) process depending only on the current and past values of $\{\eta_t\}$

- A process is invertible if and only if

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0,$$

for all $|z| \leq 1$

- An ARMA process

$$\phi(B)\eta_t = \theta(B)Z_t,$$

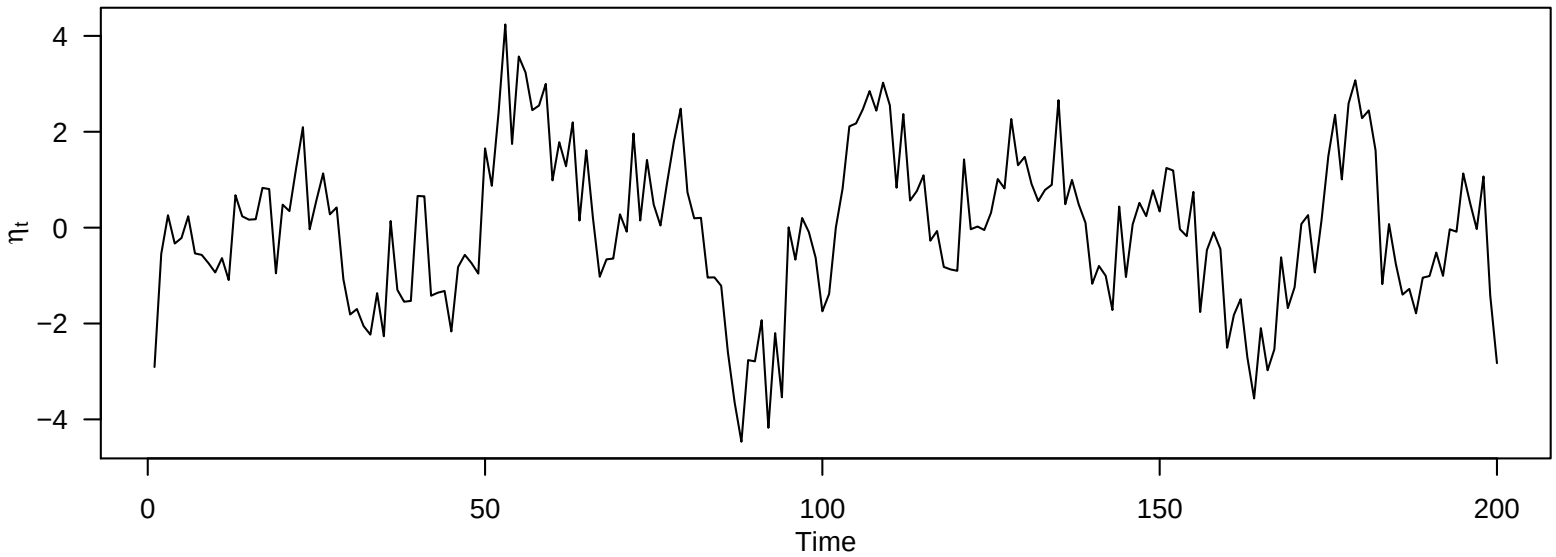
with $\phi(z) = 1 - 0.5z$ and $\theta(z) = 1 + 0.4z$ has a root of the MA characteristic polynomial at $z = \frac{-1}{0.4} = -2.5$

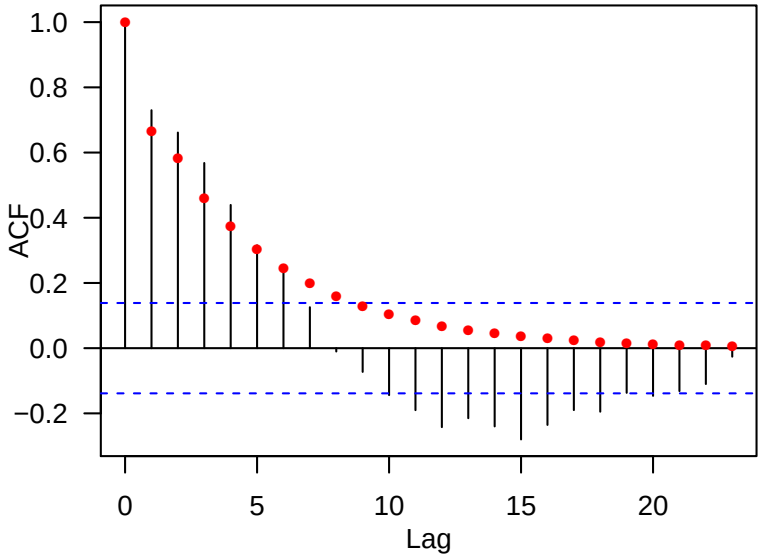
Partial Autocorrelation Functions (PACF)

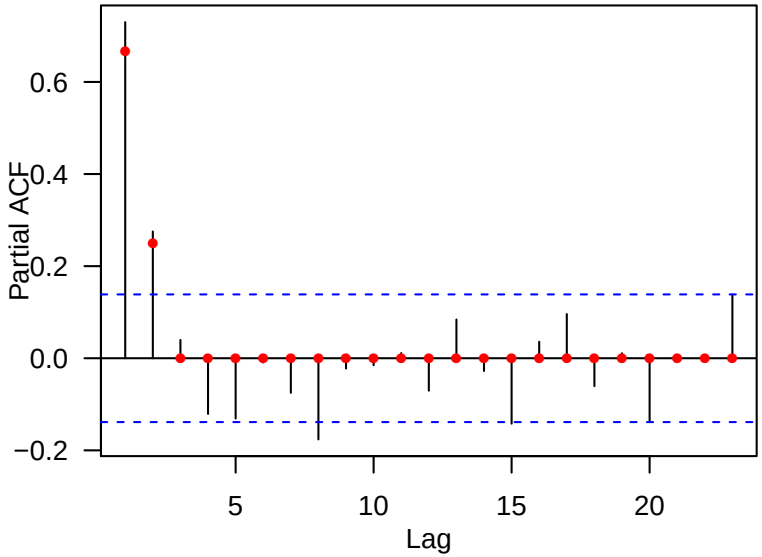
The partial autocorrelation function (PACF) represents the partial correlation of a stationary time series $\{\eta_t\}$ with its own lagged values, while regressing out the effects of the time series at all shorter lags

- PACF of lag h is the autocorrelation between η_t and η_{t+h} with the linear dependence between η_t and $\eta_{t+1}, \cdots, \eta_{t+h-1}$ removed
- PACF plots are a commonly used tool for identifying the order of an AR model, as the theoretical PACF "shuts off" past the order of the model
- One can use the function `pacf` in `R` to plot the PACF plots

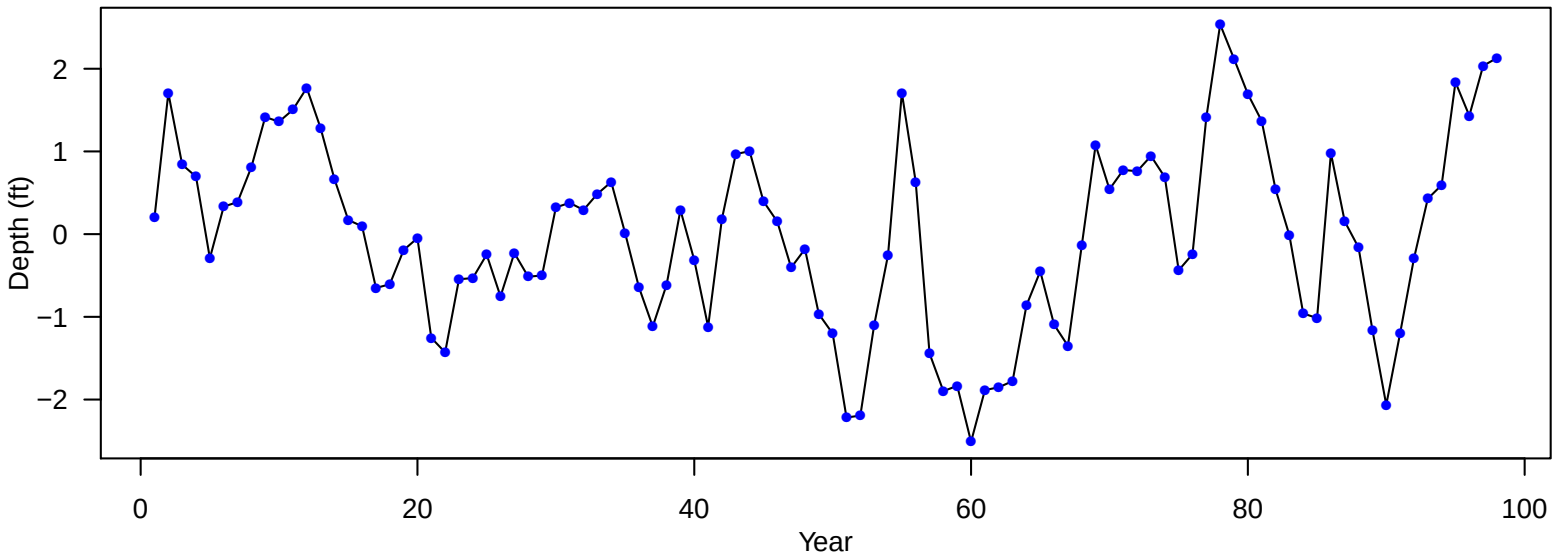
An Example of PACF Plot

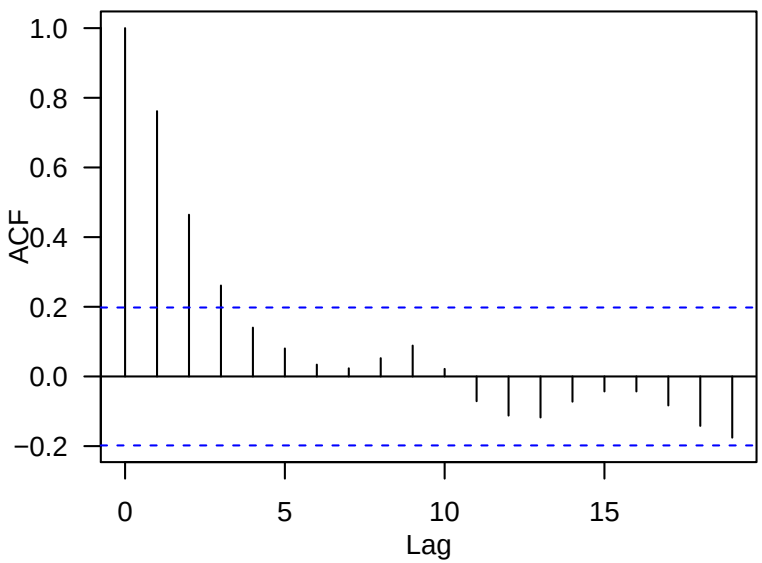


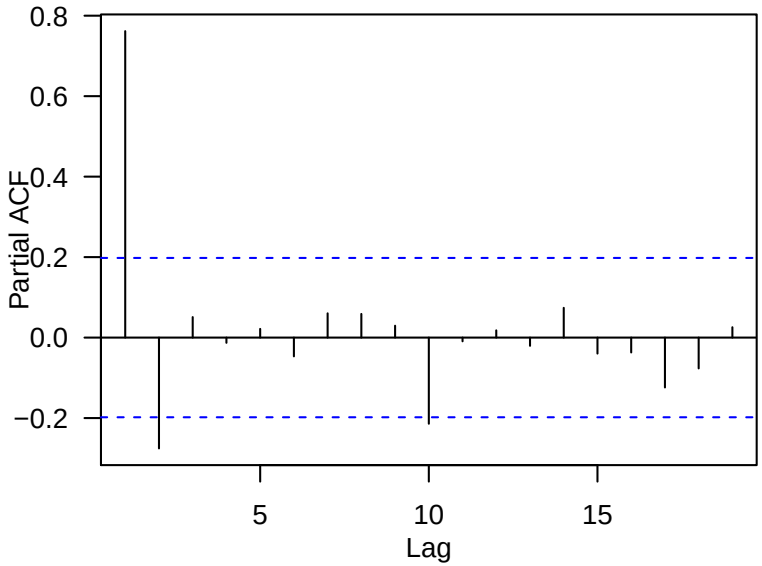




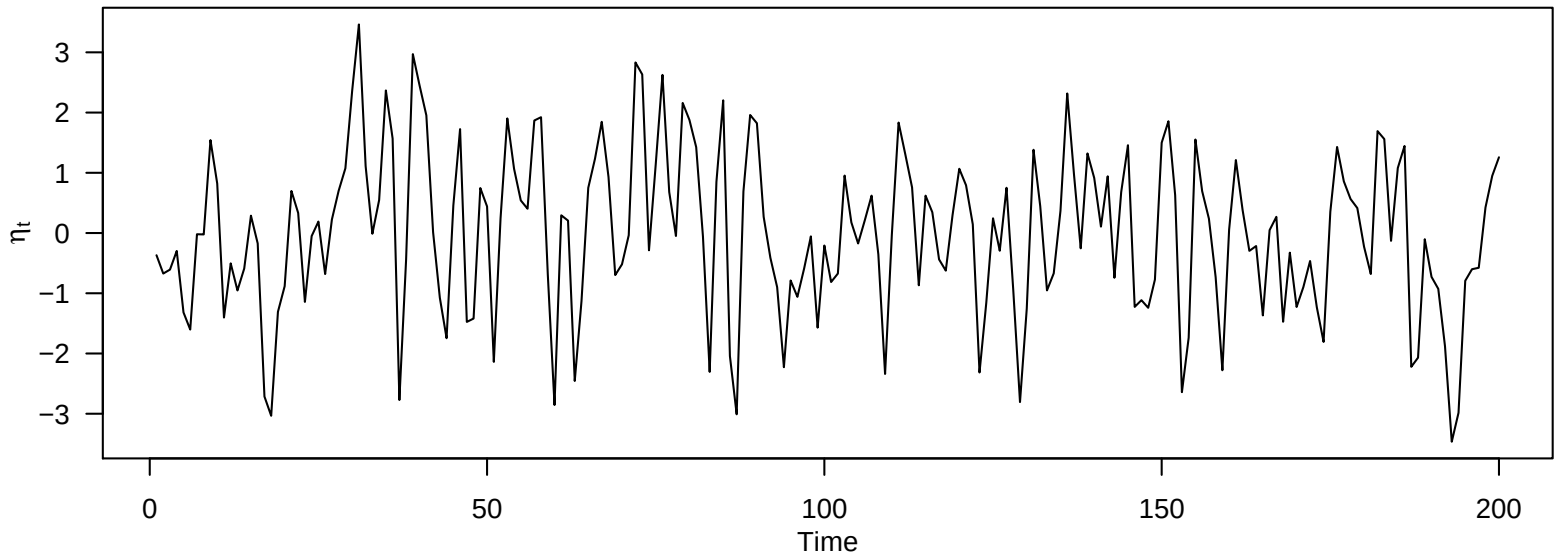
Lake Huron Series PACF Plot

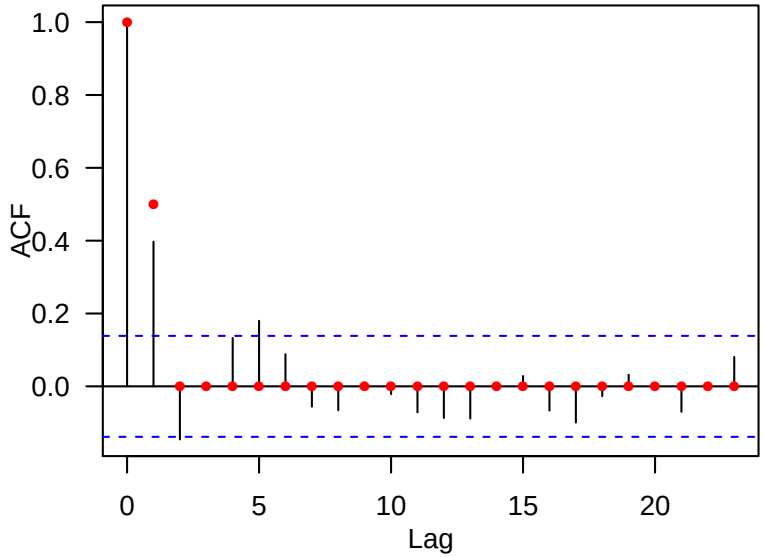


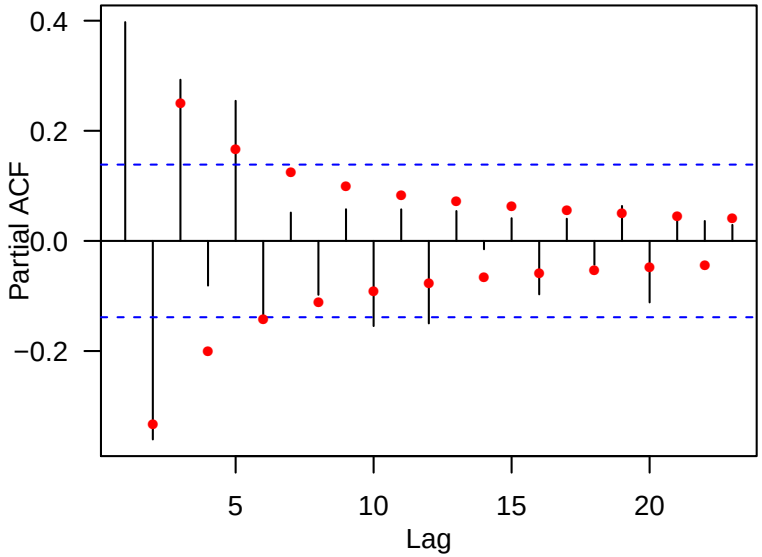




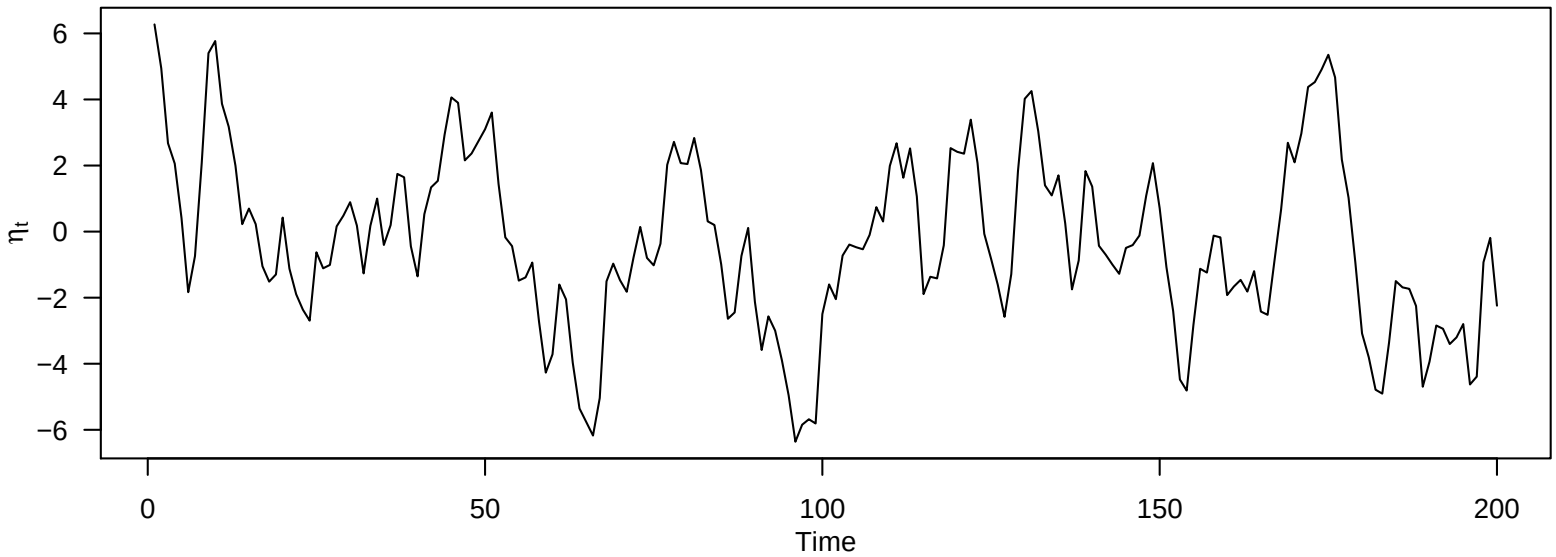
PACF Plot for a MA Process

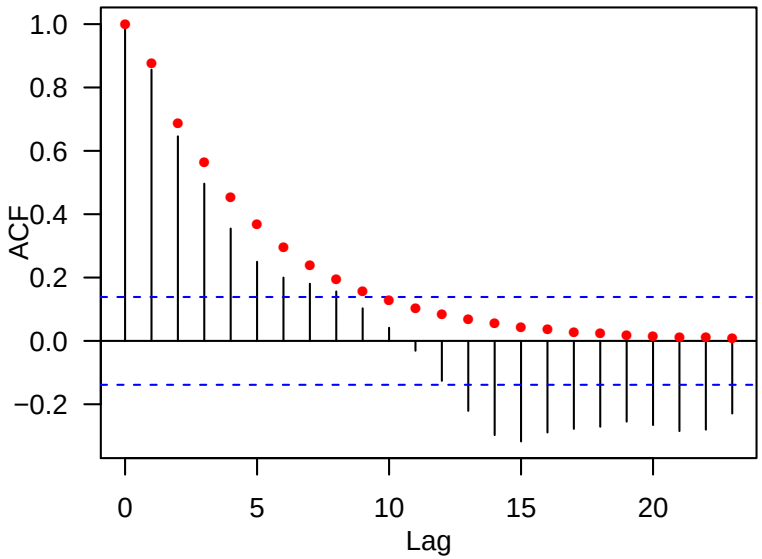


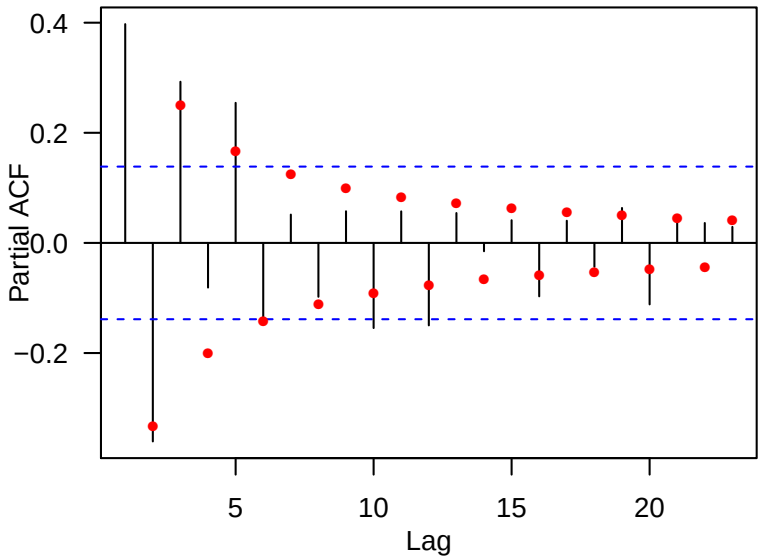




PACF Plot for a ARMA Process





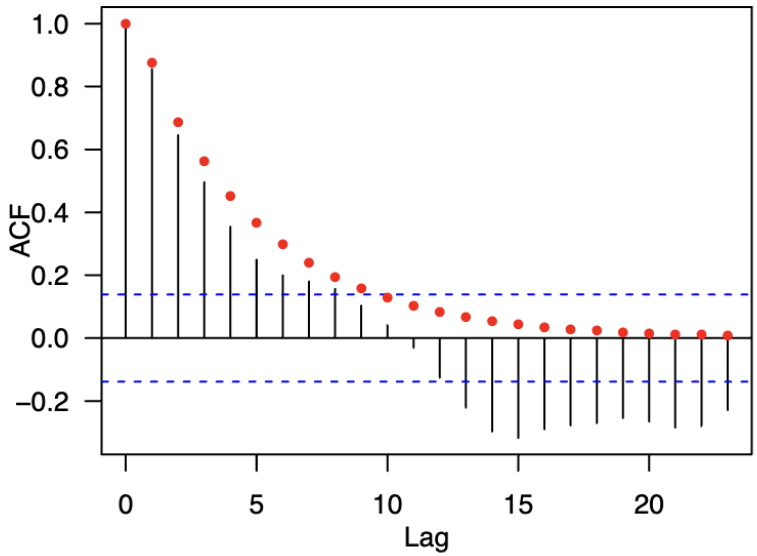


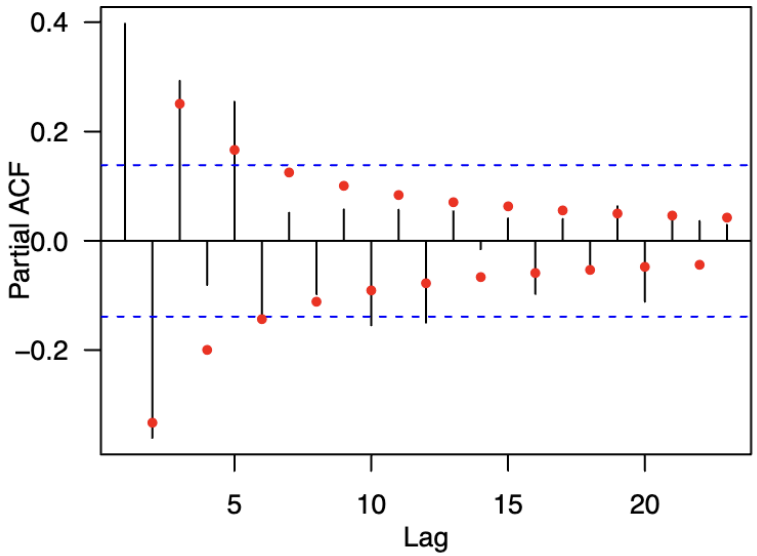
Identifying Plausible Stationary ARMA Models

We can use the sample ACF and PACF to help identify plausible models:

Model	ACF	PACF
MA(q)	cuts off after lag q	tails off exponentially
AR(p)	tails off exponentially	cuts off after lag p

For ARMA(p, q) we will see a combination of the above





Estimation of the ARMA Process Parameters

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

- Need to estimate the $p + q + 1$ parameters:
  - AR component $\{\phi_1, \cdots, \phi_p\}$
  - MA component $\{\theta_1, \cdots, \theta_q\}$
  - $\mathbb{V}\mathrm{ar}(Z_t) = \sigma^2$
- 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

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

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

The Yule-Walker Equations

- When $h = 0$, $\mathbb{E}(Z_t\eta_{t-h}) = \mathbb{C}\text{ov}(Z_t, \eta_t) = \sigma^2$ (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 $\hat{\gamma}(h)$

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

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

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

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

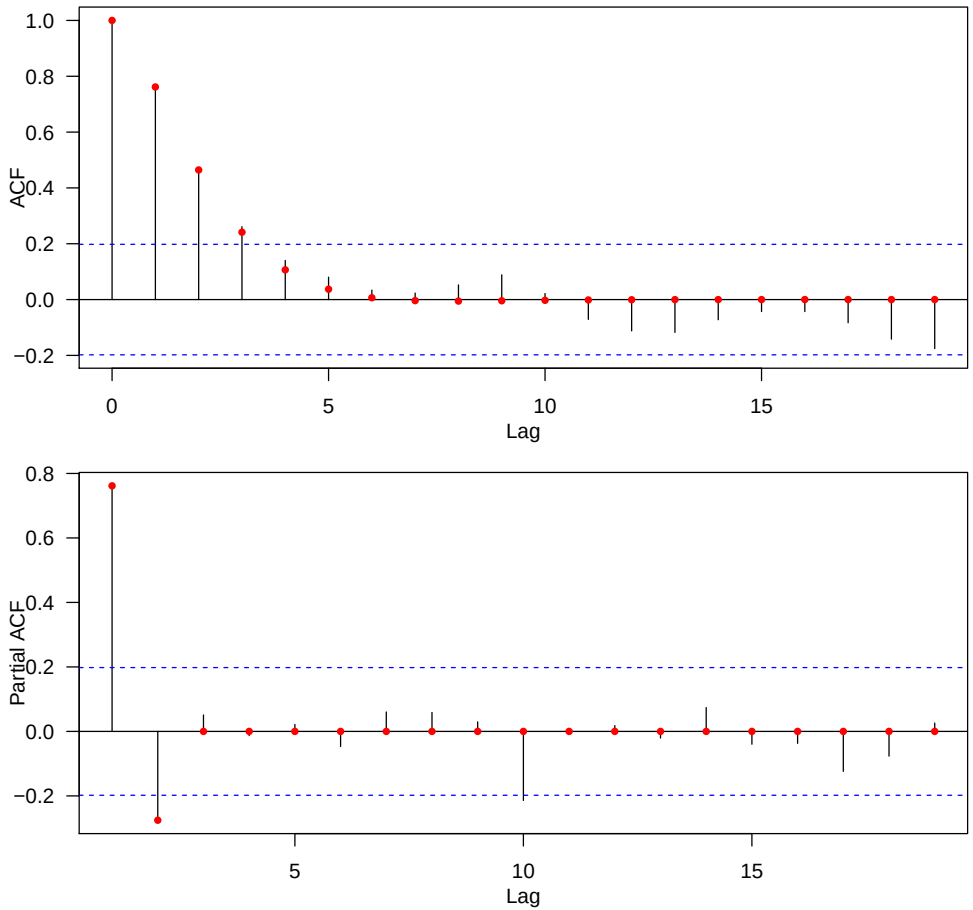
and

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

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

Lake Huron Example in R

```r
YW_est <- ar(lm$residuals, aic = FALSE, order.max = 2, method = "yw")
# plot sample and estimated acf/pacf
par(las = 1, mgp = c(2, 1, 0), mar = c(3.6, 3.6, 0.6, 0.6), mfrow = c(2, 1))
acf(lm$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)
```



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 ARMA(p,q) processes with $q > 0$

- We move on the more versatile and popular method for estimating ARMA(p,q) parameters–maximum likelihood estimation

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,

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

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

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

- The log-likelihood function is

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

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

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

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:

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

and the log-likelihood is

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

Simplifying the Likelihood Calculation

- Let the best linear one-step predictor of X_t be

$$\hat{X}_t = \begin{cases} 0, & t = 1; \\ P_{t-1}X_t, & t = 2, \cdots, n \end{cases}$$

- The one-step prediction errors or innovations are defined

$$U_t = X_t - \hat{X}_t, \quad t = 1, \cdots, n,$$

  and the associated mean squared error is

$$\nu_{t-1} = \mathbb{E}\left[(X_t - \hat{X}_t)^2\right] = \mathbb{E}(U_t^2), \quad t = 1, \cdots, n.$$

- For a causal ARMA process we can write $\nu_{t-1} = \sigma^2 r_{t-1}$, where r_t and U_t only depends on the AR and MA parameters ϕ and θ, but not σ^2

Working with the Innovations

- Result I: $\{U_t\}$ is an independent set of RVs with

$$U_t \sim \mathrm{N}(0, \nu_{t-1}), t = 1, \cdots, n$$

$\Rightarrow$ the one-step prediction errors are uncorrelated with one another, and each each a normal distribution

- Result II: The likelihoods are the same if we use a model based on realizations of $\{X_t\}$ or a model based on realizations of $\{U_t\}$

- Therefore

$$\ell_n(\boldsymbol{\omega}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log(\nu_{t-1}) - \frac{1}{2}\sum_{t=1}^{n}\left(\frac{u_t^2}{\nu_{t-1}}\right).$$

For a causal ARMA process this becomes

$$\ell_n(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{t=1}^{n}\log(r_{t-1}) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left(\frac{u_t^2}{r_{t-1}}\right)$$

The MLEs of σ^2, ϕ, and θ

- Now take the derivative of ℓ_n with respect to σ^2, setting the derivative equal to zero and solving for $\sigma^2 \Rightarrow$

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

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \left(\frac{u_t^2}{r_{t-1}} \right).$$

- Substituting $\hat{\sigma}^2$ into ℓ_n, the MLE estimates of ϕ and $\boldsymbol{\theta}$, denoted by $\hat{\phi}$ and $\hat{\boldsymbol{\theta}}$, respectively, are those values which **maximize**

$$\tilde{\ell}_n(\boldsymbol{\phi}, \boldsymbol{\theta}, \hat{\sigma}^2) = -\frac{n}{2} \log \left(\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n} \right) - \frac{1}{2} \sum_{t=1}^{n} \log(r_{t-1})$$

What About Non-Gaussian Processes?

- Not as easy to express the joint distribution of $\{X_t\}$ if the process is not Gaussian, instead consider the Gaussian likelihood as an approximate likelihood

- In practice:

  - Transform the data to make the series "as Gaussian" as possible

  - Then use the Gaussian likelihood to estimate the parameters of interest

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 $\boldsymbol{\phi}$ 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

$V(\phi, \theta)$ 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 $(X_1, \cdots, X_p)$

- AR(1) process:

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

- AR(2) process:

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

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 $(\hat{\phi}, \hat{\theta})$ can be approximated by using the second derivatives of the log-likelihood function

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

Diagnostics via the Time Series Residuals

- Recall the innovations are given by

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

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

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

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

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

Akaike Information Criterion (AIC)

- The Akaike information criterion (AIC) is defined by

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

- 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

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