Lecture 2. Autoregressive Moving Average Model

1. Univariate Time Series Concept

Definition 1.1. An univariate time series x_t with T observations, denoted by $\{x_t\}_{t=1}^{T}$, is a set of repeated observations on the same random variables, ordered in time t.

Definition 1.2. A time series x_t is stationary if both $E[x_t]$ and $Var[x_t]$ are finite and constant over time—namely, $E[x_t] = \mu$ and $Var[x_t] = \sigma^2$ for all t.

- A point of interest in a time series analysis is to estimate unknown parameters governing the stochastic process of the time series data x_t. To do this, we need the stationarity of x_t.

Definition 1.3. For $k = 0, 1, \ldots$, the kth-order autocorrelation of x_t is defined by

$$\rho_k = \frac{\gamma_k}{\gamma_0},$$

where $\gamma_k = Cov[x_t, x_{t-k}]$ is the kth-order autocovariance of x_t and $\gamma_0 = Cov[x_t, x_t]$ is the variance of x_t. The collection of autocorrelations is called the autocorrelation function (ACF).

- The kth-order sample autocorrelation is computed as

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2},$$

  where $\bar{x} = (1/T)\sum_{t=1}^{T} x_t$.

Definition 1.4. A time series x_t is a white noise process, denoted by $x_t \sim WN(0, \sigma^2)$, if $E[x_t] = 0$, $Var[x_t] = \sigma^2$, and $Cov[x_t, x_{t-k}] = 0$ for all t and $k \neq 0$.

- For $x_t \sim WN(0, \sigma^2)$, x_t is often used to represent the new information at time t and referred to as the innovation or shock at time t.
- For $x_t \sim WN(0, \sigma^2)$, it shows $E[x_t^2] = \sigma^2$ and $E[x_t x_{t-k}] = 0$ for $k \neq 0$.

Theorem 1.5. *In testing $H_0 : \rho_1 = \rho_2 = \cdots = \rho_m = 0$ versus $H_1 : \rho_i \neq 0$ for some $i \in \{1, \ldots, m\}$, Ljung and Box (1978) show*

$$Q(m) = T(T+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{T-k} \approx \chi^2_{(m)}$$

under the null hypothesis.

- If $x_t \sim WN(0,\sigma^2)$, ρ_k are zero for all $k > 0$. This means that a test procedure determining whether x_t is a white noise process is equivalent to asking whether x_t has zero autocorrelations. If the null hypothesis $H_0 : \rho_1 = \rho_2 = \cdots = \rho_m = 0$ is rejected in favor of the alternative hypothesis $H_1 : \rho_i \neq 0$ for some $i \in \{1,\ldots,m\}$, we conclude that the time series x_t is not a white noise process. In practice, the order m is selected arbitrarily.

2. Autoregressive Model

2.1. Model Setup

Definition 2.1. An autoregressive (AR) model of order p is defined by

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t, \tag{2.1}$$

where $\varepsilon_t \sim WN(0,\sigma_\varepsilon^2)$.

- The expectation of x_t conditioning on $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$ is

  $$E[x_t | x_{t-1}, x_{t-2}, \ldots, x_{t-p}] = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p},$$

  meaning that the most recent p lagged variables, i.e., $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, jointly determine the conditional expectation of x_t.

Definition 2.2. The lag operator L is defined such that for a time series x_t,

$$Lx_t = x_{t-1}.$$

- It satisfies (a) $L^2 x_t = L \cdot L x_t = L x_{t-1} = x_{t-2}$, (b) $L^j x_t = x_{t-j}$, (c) $L^0 = 1$, (d) $L^{-1} x_t = x_{t+1}$, and (e) $L \cdot a = a$ for any constant a.

Remark 2.3. We write (2.1) as

$$x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \cdots - \phi_p x_{t-p} = \phi_0 + \varepsilon_t$$

or

$$\underbrace{(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)}_{\phi(L)} x_t = \phi_0 + \varepsilon_t.$$

Theorem 2.4. *The stationarity condition for the AR(p) process of x_t is that all the (complex) solutions of the pth-order characteristic equation*

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

are greater than one in modulus.

Example 2.5. We consider the AR(1) process of x_t, i.e.,

$$x_t = \phi_0 + \phi_1 x_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \sim WN(0, \sigma^2)$. The characteristic equation is $\phi(z) = 1 - \phi_1 z = 0$, so the solution is $z = \phi_1^{-1}$. If $|z| = |\phi_1^{-1}| > 1$ or $|\phi_1| < 1$, then the AR(1) process of x_t is stationary.

Suppose that x_t follows a STATIONARY AR(p) process. Based on $E[x_t] = \mu$ for all t, we have

$$\mu = \phi_0 + \phi_1\mu + \phi_2\mu + \cdots + \phi_p\mu$$

or

$$\phi_0 = (1 - \phi_1 - \phi_2 - \cdots - \phi_p)\mu. \tag{2.2}$$

Using (2.2), we write the stationary AR(p) model as

$$x_t = (1 - \phi_1 - \phi_2 - \cdots - \phi_p)\mu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

or

$$(x_t - \mu) = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + \varepsilon_t. \tag{2.3}$$

- Multiplying (2.3) by $x_{t-k} - \mu$ and using $E[(x_{t-k} - \mu)\varepsilon_t] = 0$ (see later), we obtain

$$\begin{aligned} E[(x_t - \mu)(x_{t-k} - \mu)] &= \phi_1 E[(x_{t-1} - \mu)(x_{t-k} - \mu)] + \phi_2 E[(x_{t-2} - \mu)(x_{t-k} - \mu)] + \cdots \\ &\quad + \phi_p E[(x_{t-p} - \mu)(x_{t-k} - \mu)] \end{aligned}$$

  or

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p}.$$

  Consequently, the ACF of the stationary AR(p) process x_t satisfies

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p}$$

for $k \geq p$.

Example 2.6. We consider the stationary AR(1) process of x_t, i.e.,

$$x_t = \phi_0 + \phi_1 x_{t-1} + \varepsilon_t,$$

where $|\phi_1| < 1$ and $\varepsilon_t \sim WN(0, \sigma^2)$. In this case, the ACF satisfies $\rho_k = \phi_1 \rho_{k-1}$ for $k \geq 1$. Since $\rho_0 = 1$, it shows

$$\rho_k = \phi_1(\phi_1 \rho_{k-2}) = \phi_1^2 \rho_{k-2} = \cdots = \phi_1^k \rho_0 = \phi_1^k,$$

which implies that the ACF decays exponentially with rate ϕ_1 since $|\phi_1| < 1$.

2.2. AR(1) Model

Theorem 2.7. *Consider the sum of a finite geometric sequence of the form*

$$S_n = 1 + \phi_1 L + \phi_1^2 L^2 + \cdots + \phi_1^{n-1} L^{n-1}.$$

If $|\phi_1| < 1$, then it shows

$$\lim_{n\to\infty} S_n = \frac{1}{1 - \phi_1 L}.$$

Proof. Notice

$$\begin{aligned} S_n - \phi_1 L S_n &= (1 + \phi_1 L + \phi_1^2 L^2 + \cdots + \phi_1^{n-1} L^{n-1}) - (\phi_1 L + \phi_1^2 L^2 + \phi_1^3 L^3 + \cdots + \phi_1^n L^n) \\ &= 1 - \phi_1^n L^n \end{aligned}$$

and hence

$$S_n = \frac{1 - (\phi_1 L)^n}{1 - \phi_1 L}.$$

If $|\phi_1| < 1$, then $|\phi_1 L| < 1$ because $\phi_1 L = \phi_1$, so $(\phi_1 L)^n \to 0$ as $n \to \infty$. Therefore, we have

$$\lim_{n\to\infty} S_n = 1 + \phi_1 L + \phi_1^2 L^2 + \cdots = \frac{1}{1 - \phi_1 L}.$$

□

Theorem 2.8. *The following AR(1) model*

$$x_t = \phi_0 + \phi_1 x_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$ is stationary if and only if $|\phi_1| < 1$.

Proof. Suppose that x_t is stationary and $E[x_t] = \mu$ for all t. Because $\phi_0 = (1-\phi_1)\mu$, we obtain

$$\begin{aligned} x_t - \mu &= \phi_1(x_{t-1} - \mu) + \varepsilon_t \\ &= \phi_1\left(\phi_1(x_{t-2} - \mu) + \varepsilon_{t-1}\right) + \varepsilon_t \\ &= \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_1^2(x_{t-2} - \mu) \\ &\vdots \\ &= \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_1^2\varepsilon_{t-2} + \cdots, \end{aligned} \tag{2.4}$$

which implies

$$\begin{aligned} E[(x_{t-1} - \mu)\varepsilon_t] &= E[(\varepsilon_{t-1} + \phi_1\varepsilon_{t-2} + \phi_1^2\varepsilon_{t-3} + \cdots)\varepsilon_t] \\ &= E[\varepsilon_t\varepsilon_{t-1}] + \phi_1 E[\varepsilon_t\varepsilon_{t-2}] + \phi_2^2 E[\varepsilon_t\varepsilon_{t-3}] + \cdots \\ &= 0. \end{aligned}$$

Taking the square and the expectation of (2.4), we obtain

$$E[(x_t - \mu)^2] = \phi_1^2 E[(x_{t-1} - \mu)^2] + E[\varepsilon_t^2] + 2\phi_1 E[(x_{t-1} - \mu)\varepsilon_t]$$

or

$$Var[x_t] = \phi_1^2 Var[x_{t-1}] + \sigma_\varepsilon^2.$$

Since the stationarity of x_t implies $Var[x_t] = Var[x_{t-1}]$, we have

$$Var[x_t] = \frac{\sigma_\varepsilon^2}{1 - \phi_1^2},$$

provided $\phi_1^2 < 1$ which ensures that $Var[x_t]$ is bounded and nonnegative. Consequently, the stationarity of the AR(1) process of x_t implies $|\phi_1| < 1$.

Now, suppose $|\phi_1| < 1$. We write the AR(1) model as

$$(1 - \phi_1 L)x_t = \phi_0 + \varepsilon_t$$

or

$$x_t = \frac{\phi_0}{1 - \phi_1 L} + \frac{\varepsilon_t}{1 - \phi_1 L}. \tag{2.5}$$

Since $|\phi_1| < 1$ implies $(1-\phi_1 L)^{-1} = 1 + \phi_1 L + \phi_1^2 L^2 + \cdots$, we write (2.5) as

$$\begin{aligned} x_t &= \phi_0(1+\phi_1 L+\phi_1^2 L^2+\cdots)+\varepsilon_t(1+\phi_1 L+\phi_1^2 L^2+\cdots) \\ &= \phi_0(1+\phi_1+\phi_1^2+\cdots)+(\varepsilon_t+\phi_1\varepsilon_{t-1}+\phi_1^2\varepsilon_{t-2}+\cdots) \\ &= \frac{\phi_0}{1-\phi_1}+\varepsilon_t+\phi_1\varepsilon_{t-1}+\phi_1^2\varepsilon_{t-2}+\cdots, \end{aligned} \tag{2.6}$$

which means that $E[x_t] = \phi_0/(1-\phi_1)$ and thus $E[x_t]$ is time invariant. Based on (2.6) and $Cov[\varepsilon_t, \varepsilon_{t-k}] = 0$ for $k \neq 0$, we show

$$\begin{aligned} Var[x_t] &= Var[\varepsilon_t]+\phi_1^2 Var[\varepsilon_{t-1}]+\phi^4 Var[\varepsilon_{t-2}]+\cdots \\ &= \sigma_\varepsilon^2(1+\phi_1^2+\phi_1^4+\cdots) \\ &= \frac{\sigma_\varepsilon^2}{1-\phi_1^2} \end{aligned}$$

due to $|\phi_1^2| < 1$. Therefore, $Var[x_t]$ is also time invariant. In sum, $E[x_t]$ and $Var[x_t]$ are constant for all t if $|\phi_1| < 1$; equivalently, the AR(1) process of x_t is stationary if $|\phi_1| < 1$. □

2.3. Parameter Estimation

Consider a stationary AR(p) process of x_t of the form

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t, \tag{2.7}$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$. To estimate parameters $\beta = \begin{bmatrix} \phi_0 & \phi_1 & \cdots & \phi_p \end{bmatrix}'$, we take (2.7) as a multiple linear regression model wherein x_t is a dependent variable and $x_{t-1}, \ldots, x_{t-p}$ are a set of independent variables.

- With T observations, we obtain

$$\begin{bmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \\ x_T \end{bmatrix} = \begin{bmatrix} 1 & x_p & \cdots & x_1 \\ 1 & x_{p+1} & \cdots & x_2 \\ \vdots & \vdots & & \vdots \\ 1 & x_{T-1} & \cdots & x_{T-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \vdots \\ \varepsilon_T \end{bmatrix}$$

  or simply

$$y = X\beta + \varepsilon.$$

Given the observed sample of x_t, the ordinary least squares (OLS) estimate of β, denoted by b, minimizes the sum of squared errors with respect to β, i.e.,

$$\begin{aligned} b &= \underset{\beta}{\text{argmin}} \varepsilon' \varepsilon \\ &= \underset{\beta}{\text{argmin}} (y - X\beta)'(y - X\beta). \end{aligned}$$

Theorem 2.9. *Provided that $X'X$ is nonsingular, the OLS estimate of β is given by*

$$b = (X'X)^{-1}X'y.$$

Proof. We write the sum of squared errors, $\varepsilon'\varepsilon$, as

$$(y - X\beta)'(y - X\beta) = y'y - y'X\beta - \beta'X'y + \beta'X'X\beta.$$

The first order condition is

$$\frac{\partial \varepsilon'\varepsilon}{\partial \beta} = -2X'y + 2X'X\beta = 0,$$

leading to

$$X'Xb = X'y.$$

Since there exists $(X'X)^{-1}$, the solution is given by

$$b = (X'X)^{-1}X'y.$$

□

Remark 2.10. The order p of an AR model is unknown, so we must specify it in practice. Consider the following AR models

$$\begin{aligned} x_t &= \phi_1 + \phi_{11}x_{t-1} + \varepsilon_t \\ x_t &= \phi_2 + \phi_{21}x_{t-1} + \phi_{22}x_{t-2} + \varepsilon_t \\ x_t &= \phi_3 + \phi_{31}x_{t-1} + \phi_{32}x_{t-2} + \phi_{33}x_{t-3} + \varepsilon_t \\ &\vdots \end{aligned}$$

The coefficient ϕ_{jj} of the jth equation is called the jth partial autocorrelation function (PACF) of x_t. For the AR(p) process of x_t, the pth sample PACF, denoted by $\hat{\phi}_{jj}$, should not be zero for

$j = p$ but $\hat{\phi}_{jj}$ should be close to zero for all $j > p$. That is, the sample PACF cuts off at lag p if x_t follows an AR(p) process.

Remark 2.11. In `R`, an AR(p) model is in the form

$$(x_t - \mu) = \phi_1(x_{t-1} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + \varepsilon_t,$$

where $\mu = \phi_0(1 - \phi_1 - \cdots - \phi_p)^{-1}$ is referred to as `intercept`.

2.4. Forecasting

Suppose that we know $\phi_0, \phi_1, \ldots, \phi_p$. The s-step forecast of x_T, denoted by $x_T[s]$, is the conditional expectation of x_{T+s} given the information set available at time T, denoted by I_T, i.e.,

$$x_T[s] = E[x_{T+s}|I_T]$$

for $s \geq 1$.

- For $x_t \sim$AR(p), the 1-step forecast is

$$\begin{aligned} x_T[1] &= E[x_{T+1}|I_T] \\ &= E[\phi_0 + \phi_1 x_T + \cdots + \phi_p x_{T+1-p} + \varepsilon_{T+1}|I_T] \\ &= \phi_0 + \phi_1 x_T + \cdots + \phi_p x_{T+1-p} \end{aligned}$$

  and the 2-step forecast is

$$\begin{aligned} x_T[2] &= E[x_{T+2}|I_T] \\ &= E[\phi_0 + \phi_1 x_{T+1} + \phi_2 x_T + \cdots + \phi_p x_{T+2-p} + \varepsilon_{T+2}|I_T] \\ &= \phi_0 + \phi_1 E[x_{T+1}|I_T] + \phi_2 x_T + \cdots + \phi_p x_{T+2-p} \\ &= \phi_0 + \phi_1 x_T[1] + \phi_2 x_T + \cdots + \phi_p x_{T+2-p} \end{aligned}$$

  The s-step forecasts can be computed recursively.

Remark 2.12. With the s-step forecasts, we have

$$\begin{aligned} x_{T+1} - x_T[1] &= (\phi_0 + \phi_1 x_T + \cdots + \phi_p x_{T+1-p} + \varepsilon_{T+1}) - (\phi_0 + \phi_1 x_T + \cdots + \phi_p x_{T+1-p}) \\ &= \varepsilon_{T+1}, \end{aligned}$$

$$\begin{aligned}
x_{T+2} - x_T[2] &= (\phi_0 + \phi_1 x_{T+1} + \phi_2 x_T + \cdots + \phi_p x_{T+2-p} + \varepsilon_{T+2}) \\
&\quad -(\phi_0 + \phi_1 x_T[1] + \phi_2 x_T + \cdots + \phi_p x_{T+2-p}) \\
&= \phi_1 (x_{T+1} - x_T[1]) + \varepsilon_{T+2} \\
&= \varepsilon_{T+2} + \phi_1 \varepsilon_{T+1},
\end{aligned}$$

and so on. Assume $\varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ and treat x_{T+s} as unknown parameters for $s = 1, 2, \ldots$. Then we have

$$x_T[1] \sim N(x_{T+1}, \sigma_\varepsilon^2)$$

which means

$$0.95 = \Pr\left(-1.96 \leq \frac{x_T[1] - x_{T+1}}{\sigma_\varepsilon} \leq 1.96\right).$$

This suggests that a 95% confidence interval for x_{T+1} is computed as

$$x_T[1] \pm 1.96\sigma_\varepsilon.$$

Since $x_T[2] \sim N(x_{T+2}, (1+\phi_1^2)\sigma_\varepsilon^2)$, we compute a 95% confidence interval for x_{T+2} as

$$x_T[2] \pm 1.96\sqrt{1+\phi_1^2}\sigma_\varepsilon.$$

The 95% confidence intervals for x_{T+s} become wider as s increases, so the uncertainty of the s-step forecast increases with the forecast horizon s.

3. Moving Average Model

3.1. Model Setup

Definition 3.1. A moving average (MA) model of order q is defined by

$$\begin{aligned}
x_t &= \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \\
&= \mu + (1 + \theta_1 L + \cdots + \theta_q L^q)\varepsilon_t \\
&= \mu + \theta(L)\varepsilon_t,
\end{aligned}$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$.

- The MA(q) process of x_t is always stationary because

$$\begin{aligned}
E[x_t] &= \mu + E[\varepsilon_t] + \theta_1 E[\varepsilon_{t-1}] + \cdots + \theta_q E[\varepsilon_{t-q}] \\
&= \mu
\end{aligned}$$

and

$$\begin{aligned} Var[x_t] &= Var[\varepsilon_t] + \theta_1^2 Var[\varepsilon_{t-1}] + \cdots + \theta_q^2 Var[\varepsilon_{t-q}] \\ &= (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma_\varepsilon^2, \end{aligned}$$

both of which are time invariant.

- For $x_t \sim$MA(q), it shows that the kth-order autocovariance is

$$\begin{aligned} \gamma_k &= E[(x_t - E[x_t])(x_{t-k} - E[x_{t-k}])] \\ &= E[(\varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q})(\varepsilon_{t-k} + \theta_1 \varepsilon_{t-k-1} + \cdots + \theta_q \varepsilon_{t-k-q})] \\ &= \begin{cases} (\theta_k + \theta_{k+1}\theta_1 + \theta_{k+2}\theta_2 \cdots + \theta_q \theta_{q-k})\sigma_\varepsilon^2 & \text{for } k = 1, 2, \ldots, q \\ 0 & \text{for } k > q, \end{cases} \end{aligned}$$

meaning that the ACF cuts off at lag q.

Example 3.2. Consider an MA(1) process of the form

$$x_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1},$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$. The kth-order autocovariance is

$$\begin{aligned} \gamma_k &= E[(x_t - E[x_t])(x_{t-k} - E[x_{t-k}])] \\ &= E[(\varepsilon_t + \theta_1 \varepsilon_{t-1})(\varepsilon_{t-k} + \theta_1 \varepsilon_{t-1-k})] \\ &= E[\varepsilon_t \varepsilon_{t-k} + \theta_1 \varepsilon_t \varepsilon_{t-1-k} + \theta_1 \varepsilon_{t-1} \varepsilon_{t-k} + \theta_1^2 \varepsilon_{t-1} \varepsilon_{t-1-k}] \\ &= \begin{cases} \theta_1 \sigma_\varepsilon^2 & \text{for } k = 1 \\ 0 & \text{for } k > 1, \end{cases} \end{aligned}$$

which implies that the ACF cuts off at lag 1.

Theorem 3.3. *The Wold theorem states that any stationary process x_t with $E[x_t] = \mu$ can be represented as*

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i},$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, $\psi_0 = 1$, and $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.

- The theorem implies that any stationary AR series x_t has the corresponding MA representation.

Example 3.4. Consider a stationary AR(1) process x_t of the form

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \varepsilon_t,$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$ and $|\phi_1| < 1$. We already showed

$$\begin{aligned} x_t &= \mu + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_1^2\varepsilon_{t-2} + \cdots \\ &= \mu + (1 + \phi_1 L + \phi_1^2 L^2 + \cdots)\varepsilon_t \\ &= \mu + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}, \end{aligned}$$

which is the MA representation of the AR(1) process of x_t.

Theorem 3.5. *The MA(q) process of x_t is invertible if all of the solutions of the qth-order characteristic equation*

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

are greater than one in modulus. If the MA(q) process of x_t is invertible, there exists the corresponding AR process of x_t.

Example 3.6. Consider an MA(1) process x_t of the form

$$x_t = \varepsilon_t - \theta_1\varepsilon_{t-1}, \tag{3.1}$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$. We write (3.1) as

$$\varepsilon_t = \frac{x_t}{1 - \theta_1 L}. \tag{3.2}$$

From (3.1), the characteristic equation is $\theta(z) = 1 - \theta_1 z = 0$ and, hence, the solution is $z = \theta_1^{-1}$. If $|z| > 1$ or $|\theta_1| < 1$, then we know

$$\frac{1}{1 - \theta_1 L} = 1 + \theta_1 L + \theta_1^2 L^2 + \cdots$$

and write (3.2) as

$$\begin{aligned} \varepsilon_t &= (1 + \theta_1 L + \theta_1^2 L^2 + \cdots)x_t \\ &= x_t + \theta_1 x_{t-1} + \theta_1^2 x_{t-2} + \cdots \end{aligned}$$

or

$$x_t = -\theta_1 x_{t-1} - \theta_1^2 x_{t-2} - \cdots + \varepsilon_t.$$

This is the corresponding AR process to the MA(1) process of x_t.

3.2. Parameter Estimation

Consider an MA(q) model of the form

$$x_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$. Let $\beta = (\mu, \theta_1, \ldots, \theta_q, \sigma_\varepsilon^2)$ be all parameters of the MA(q) model. The technique of maximum likelihood estimation (MLE) is used to estimate β.

Definition 3.7. For the observed sample $\tilde{x} = (x_1, \ldots, x_T)$ from a population with joint pdf $f(x_1, x_2, \cdots, x_T; \theta)$, the likelihood function (LF) is defined by

$$L(\theta|\tilde{x}) = f(\tilde{x}; \theta).$$

That is, the LF is the joint pdf evaluated at the observed sample point $\tilde{x}$.

- If $\tilde{x} = (x_1, \ldots, x_T)$ are iid, the LF is simplified to

$$\begin{aligned} L(\theta|\tilde{x}) &= f(x_1, \ldots, x_T; \theta) \\ &= f(x_1; \theta) \times \cdots \times f(x_T; \theta) \\ &= \prod_{t=1}^{T} f(x_t; \theta). \end{aligned}$$

- If $\tilde{x} = (x_1, \ldots, x_T)$ are not iid (which is true in most time-series data), we factor the LF by using $f(x, y) = f(x|y)f(y)$ as

$$\begin{aligned} L(\theta|\tilde{x}) &= f(x_T, \ldots, x_1; \theta) \\ &= f(x_T|x_{T-1}, \ldots, x_1; \theta) \times f(x_{T-1}, \ldots, x_1; \theta) \\ &= f(x_T|x_{T-1}, \ldots, x_1; \theta) \times f(x_{T-1}|x_{T-2}, \ldots, x_1; \theta) \times f(x_{T-2}, \ldots, x_1; \theta) \\ &\vdots \\ &= f(x_T|x_{T-1}, \ldots, x_1; \theta) \times f(x_{T-1}|x_{T-2}, \ldots, x_1; \theta) \times \cdots \times f(x_2|x_1; \theta) \\ &\quad \times f(x_1; \theta). \end{aligned} \tag{3.3}$$

Definition 3.8. For each sample point $\tilde{x}$, let $\hat{\theta}$ be a parameter value at which $L(\theta|\tilde{x})$ attains its maximum as a function of θ, with $\tilde{x}$ held fixed. Then the maximum likelihood estimator of the parameter θ based on a sample $\tilde{x} = (x_1, \ldots, x_T)$ is $\hat{\theta}$.

- Intuitively, the MLE is the parameter point for which the observed sample is most likely. If the likelihood function is differentiable in θ, the MLE is the value that solves
$$\frac{\partial L(\theta|\tilde{x})}{\partial \theta} = 0$$
and satisfies
$$\frac{\partial^2 L(\theta|\tilde{x})}{\partial \theta^2} < 0.$$

Remark 3.9. It is easier to work with the natural logarithm of $L(\theta|\tilde{x})$, which is known as the log likelihood function, than it it is to work with $L(\theta|\tilde{x})$ directly—that is to say, the MLE is the value that solves
$$\frac{\partial \ln L(\theta|\tilde{x})}{\partial \theta} = 0$$
and satisfies
$$\frac{\partial^2 \ln L(\theta|\tilde{x})}{\partial \theta^2} < 0.$$
This is because the log function is strictly increasing on $(0, \infty)$, which implies that the extrema of $L(\theta|\tilde{x})$ and $\ln L(\theta|\tilde{x})$ coincide.

Example 3.10. Let $\tilde{x} = (x_1, \ldots, x_T)$ be an iid sample from $N(\mu, 1)$. The LF is

$$\begin{aligned}
L(\mu|\tilde{x}) &= \prod_{t=1}^{T} f(x_t|\mu) \\
&= \prod_{t=1}^{T} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{(x_t-\mu)^2}{2}\right) \\
&= \frac{1}{(2\pi)^{T/2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T}(x_t-\mu)^2\right).
\end{aligned}$$

The log LF is

$$\begin{aligned}
\ln L(\mu|\tilde{x}) &= \ln\left((2\pi)^{-T/2}\right) + \ln\left(\exp\left(-\frac{1}{2}\sum_{t=1}^{T}(x_t-\mu)^2\right)\right) \\
&= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(x_t-\mu)^2.
\end{aligned}$$

The first-order derivative is

$$\frac{\partial \ln L(\mu|\tilde{x})}{\partial \mu} = -\sum_{t=1}^{T}(x_t - \mu),$$

so that letting it equal to zero, we obtain

$$\sum_{t=1}^{T} x_t - T\hat{\mu} = 0$$

or

$$\hat{\mu} = \frac{\sum_{t=1}^{T} x_t}{T}.$$

Remark 3.11. The observed sample $\tilde{x}$ from the MA model is not iid, so we obtain from (3.3) the following log LF

$$\begin{aligned}\ln L(\theta|\tilde{x}) &= \ln f(x_T|x_{T-1},\ldots,x_1;\theta) + \ln f(x_{T-1}|x_{T-2},\ldots,x_1;\theta) + \cdots \\ &\quad + \ln f(x_2|x_1;\theta) + \ln f(x_1;\theta).\end{aligned} \tag{3.4}$$

While we want to find $\hat{\theta}$ by maximizing (3.4), an analytical solution is not available. To find the MLE, we have to use numerical optimization methods.

Remark 3.12. Recall that the ACF of the MA(q) series of x_t cuts off at lag q. If $\rho_k \neq 0$ for $k = 1,\ldots,q$ but $\rho_k = 0$ for $k > q$, therefore, x_t follows an MA(q) process.

Remark 3.13. In `R`, an MA(q) model is in the form

$$x_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q},$$

where $\mu = E[x_t]$ is referred to as `intercept`.

3.3. Forecasting

Suppose that we know $\mu, \theta_1, \ldots, \theta_q$. The s-step forecast is the conditional expectation of x_{T+s} given I_T, i.e., $x_T[s] = E[x_{T+s}|I_T]$ for $s \geq 1$. Notice that the forecasts go to the mean μ after the first q periods in the MA(q) process of x_t.

- For $x_t \sim$MA(q), we have the followings:

$$\begin{aligned}x_T[1] &= E[x_{T+1}|I_T] \\ &= E[\mu + \varepsilon_{T+1} + \theta_1\varepsilon_T + \cdots + \theta_q\varepsilon_{T+1-q}|I_T] \\ &= \mu + \theta_1\varepsilon_T + \cdots + \theta_q\varepsilon_{T+1-q},\end{aligned}$$

$$\begin{aligned}
x_T[2] &= E[x_{T+2}|I_T] \\
&= E[\mu + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \varepsilon_T + \cdots + \theta_q \varepsilon_{T+2-q}|I_T] \\
&= \mu + \theta_2 \varepsilon_T + \cdots + \theta_q \varepsilon_{T+2-q},
\end{aligned}$$

$$\vdots$$

$$\begin{aligned}
x_T[q] &= E[x_{T+q}|I_T] \\
&= E[\mu + \varepsilon_{T+q} + \theta_1 \varepsilon_{T+q-1} + \cdots + \theta_q \varepsilon_T|I_T] \\
&= \mu + \theta_q \varepsilon_T,
\end{aligned}$$

and $x_T[s] = \mu$ for $s > q$.

4. Autoregressive Moving Average Model

The autoregressive moving average (ARMA) model of Box and Jenkins (1976) is popular for modeling and estimating stationary processes. This is because significantly fewer parameters are used to describe a time-series data with the ARMA model, compared to the AR and MA models.

Definition 4.1. The ARMA(p, q) model is defined by

$$x_t = \phi_0 + \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \tag{4.1}$$

where p is the order of the AR part, q is the order of the MA part, and $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$.

- We simplify the ARMA(p, q) model to

$$x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} = \phi_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

or

$$\underbrace{(1 - \phi_1 L - \cdots - \phi_p L^p)}_{\phi(L)} x_t = \phi_0 + \underbrace{(1 + \theta_1 L + \cdots + \theta_q L^q)}_{\theta(L)} \varepsilon_t.$$

Theorem 4.2. *If all of the solutions of the AR characteristic equation $\phi(z) = 0$ are greater than one in modulus, then the ARMA process x_t is stationary. A sufficient condition for invertibility is that all the solutions of the MA characteristic equation $\theta(z) = 0$ are greater than one in modulus.*

Remark 4.3. If x_t follows a stationary ARMA(p, q) process with $E[x_t] = \mu$ for all t, we obtain from (4.1)

$$\mu = \phi_0 + \phi_1\mu + \cdots + \phi_p\mu.$$

So, the stationary ARMA(p, q) model is given by

$$x_t = (\mu - \phi_1\mu - \cdots - \phi_p\mu) + \sum_{i=1}^{p}\phi_i x_{t-i} + \varepsilon_t + \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}$$

or

$$(x_t - \mu) = \sum_{i=1}^{p}\phi_i(x_{t-i} - \mu) + \varepsilon_t + \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}. \tag{4.2}$$

Remark 4.4. In estimating an ARMA(p, q) model, we must decide p and q in practice. A rule of thumb is to use some information criteria. Two common information criteria are the Akaike Information Criterion (AIC) and Schwartz-Bayesian Criterion (BIC):

$$\begin{aligned}\text{AIC}(p,q) &= \ln\left(\tilde{\sigma}^2(p,q)\right) + \frac{2}{T}(p+q)\\ \text{BIC}(p,q) &= \ln\left(\tilde{\sigma}^2(p,q)\right) + \frac{\ln T}{T}(p+q),\end{aligned}$$

where $\tilde{\sigma}^2(p,q)$ is the estimate of σ^2, computed from the ARMA(p, q) model. We compute AIC(p,q) for $p = 0,1,\ldots,p_{max}$ and $q = 0,1,\ldots,q_{max}$, where p_{max} and q_{max} are pre-specified positive integers (usually less than 3) and select the orders p and q for minimizing the AIC value. The same rule applies to BIC.

Remark 4.5. We estimate parameters of the ARMA model using the MLE method. In `R`, the ARMA(p, q) model is in the form of (4.2) and μ is referred to as `intercept`.

Remark 4.6. For $x_t \sim$ARMA(p,q), the s-step forecasts are computed as

$$x_T[1] = \phi_0 + \phi_1 x_T + \phi_2 x_{T-1} + \cdots + \phi_p x_{T+1-p} + \theta_1\varepsilon_T + \cdots + \theta_q\varepsilon_{T+1-q},$$

$$x_T[2] = \phi_0 + \phi_1 x_T[1] + \phi_2 x_T + \cdots + \phi_p x_{T+2-p} + \theta_2\varepsilon_T + \cdots + \theta_q\varepsilon_{T+2-q},$$

and so on. In general, it shows

$$x_T[s] = \phi_0 + \sum_{i=1}^{p}\phi_i x_T[s-i] + \sum_{j=1}^{q}\theta_j\varepsilon_T[s-j],$$

where $x_T[s-i] = x_{T+s-i}$ if $s \le i$; and $\varepsilon_T[s-j] = \varepsilon_{T+s-j}$ if $s \le j$ and $\varepsilon_T[s-j] = 0$ if $s > j$.

5. R Code

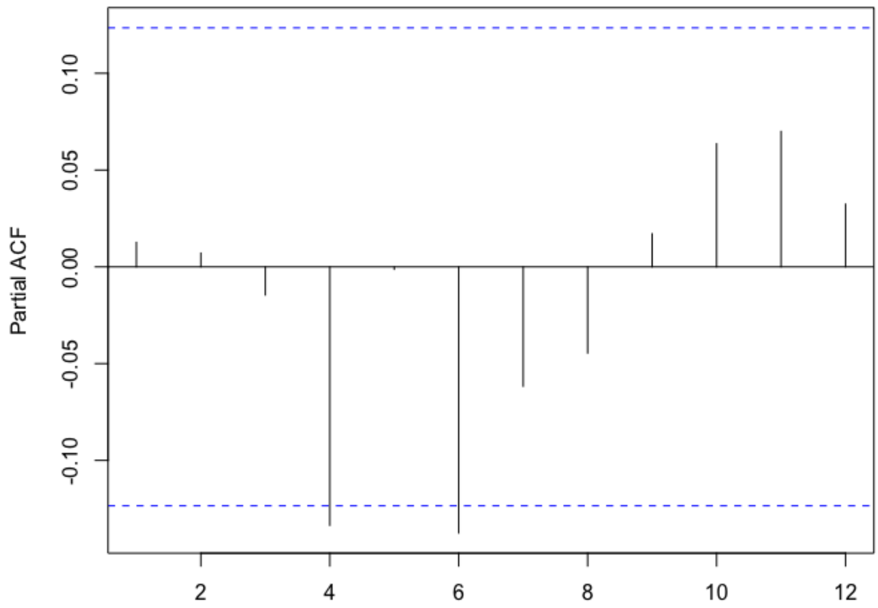
Example 5.1. We test whether the monthly returns on the KOSPI index follow a white noise process over the period from February 1980 to December 2021.

```
> mydat <- read.csv("data2_1.csv", header = T)
> head(mydat)
    TRD_DD    PRC     RTN
1 19800229 103.74 -0.0203
2 19800331 105.00  0.0121
3 19800430 116.09  0.1056
4 19800531 116.43  0.0029
5 19800630 112.62 -0.0327
6 19800731 112.71  0.0008
> Box.test(mydat$RTN, lag = 12, type = "Ljung")
Box-Ljung test
data:  mydat$RTN
X-squared = 6.6766, df = 12, p-value = 0.8782
```

- The Ljung-Box statistics with $m = 12$ cannot reject the null hypothesis of no serial correlations in the KOSPI returns, which concludes that the KOSPI return follows a white noise process.

Example 5.2. We apply AR, MA, and ARMA models to describe the monthly returns on the Samsung Electronics stock over the period from January 2000 to December 2020 and compute the s-step forecasts for $s = 1, \ldots, 12$.

```
> mydat <- read.csv("data2_2.csv", header = T)
> tail(mydat)
      TRD_DD TRD_PRC TRD_RTN
259 20210731   78500 -0.0273
260 20210831   76700 -0.0229
(...)
264 20211231   78300  0.0982
> rtn <- ts(mydat$TRD_RTN, start = c(2000, 1), freq = 12)
> rtn1 <- window(rtn, end = c(2020, 12))
# AR Model Fitting
> pacf(rtn1, lag = 12, main = "")
```



- The sample PACF $\hat{\phi}_{jj}$ is significantly different from 0 at the 5% level for $j = 6$, which identifies that x_t may follow the AR(6) model.

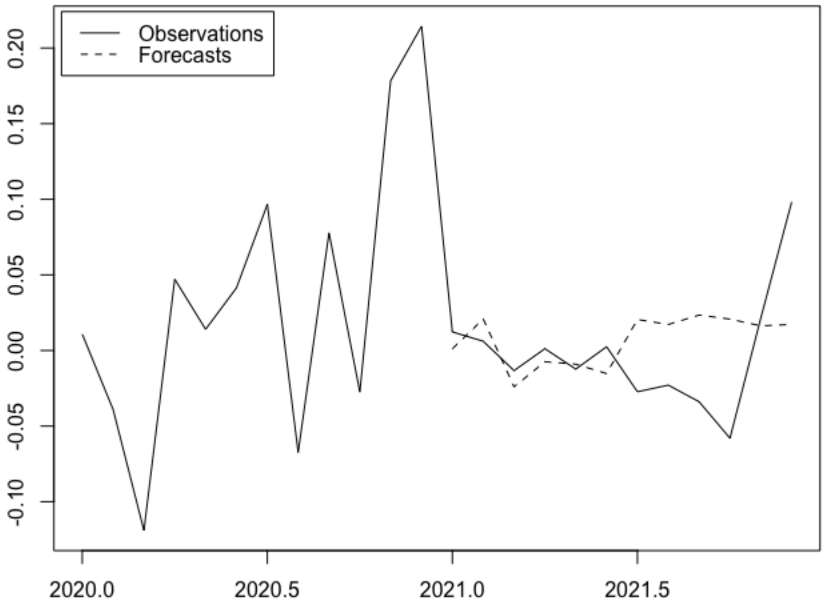
```
> fit <- arima(rtn1, order = c(6, 0, 0), method = "CSS")
> fit
Coefficients:
        ar1      ar2    ar3      ar4      ar5      ar6 intercept
     0.0419  -0.0153 -0.039  -0.1293  -0.0007  -0.1446    0.0138
s.e. 0.0624   0.0629  0.062   0.0608   0.0613   0.0614    0.0044
```

- The AR(6) model is estimated as

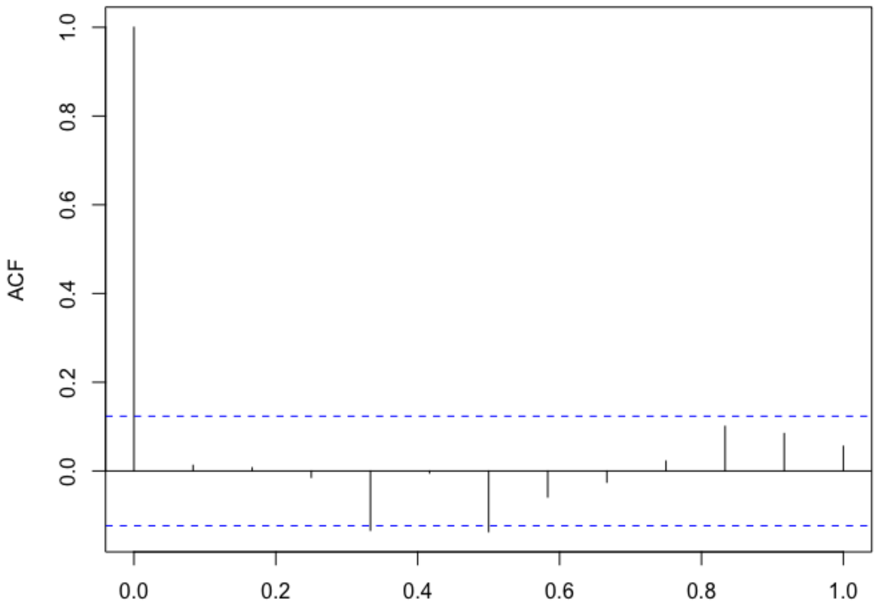
$$
\begin{aligned}
(x_t - 0.0138) &= 0.0419(x_{t-1} - 0.0138) - 0.0153(x_{t-2} - 0.0138) \\
&\quad -0.039(x_{t-3} - 0.0138) - 0.1293(x_{t-4} - 0.0138) \\
&\quad -0.0007(x_{t-4} - 0.0138) - 0.1446(x_{t-4} - 0.0138) + \varepsilon_t.
\end{aligned}
$$

```
> fcst.x <- predict(fit, 12)$pred
> new.x <- window(rtn, start = c(2020, 1))
> plot(new.x, xlab = "", ylab = "")
> lines(fcst.x, lty = 2)
> legend("topleft", c("Observations", "Forecasts"), lty = c(1, 2),
```

```
inset = 0.01)
```




```
# MA Model Fitting
> acf(rt1, lag = 12, main = "")
```



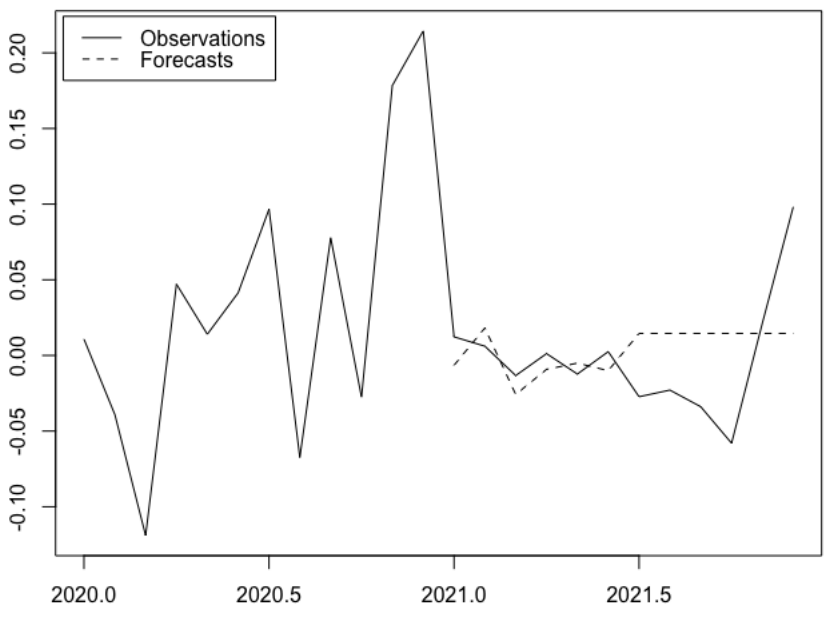
- The autocorrelation is significant at lag 6, thereby meaning that the MA order q could be 6.

```
> fit <- arima(rtn1, order = c(0, 0, 6), method = "ML"))
> fit
Coefficients:
         ma1     ma2     ma3     ma4    ma5     ma6 intercept
     -0.0104 -0.0229 -0.0426 -0.1519 0.0102 -0.1273    0.0146
s.e.  0.0649  0.0640  0.0641  0.0687 0.0654  0.0643    0.0038
```

- The MA(6) model is estimated as

$$\begin{aligned} x_t &= 0.0146 + \varepsilon_t - 0.0104\varepsilon_{t-1} - 0.0229\varepsilon_{t-2} - 0.0426\varepsilon_{t-3} - 0.1519\varepsilon_{t-4} \\ &\quad +0.0102\varepsilon_{t-5} - 0.1273\varepsilon_{t-6}. \end{aligned}$$

```
> fcst.x <- predict(fit, 12)$pred
> plot(new.x, xlab = "", ylab = "")
> lines(fcst.x, lty = 2)
> legend("topleft", c("Observations", "Forecasts"), lty = c(1, 2),
         inset = 0.01)
```




```
# ARMA Model Fitting
> computeAIC <- function(p, q){
  fit <- arima(rtn1, order = c(p, 0, q), method = "ML")
```

```
  fit$aic
}
> computeAIC(0, 0)
[1] -476.6945
> computeAIC(0, 1)
[1] -474.7348
(...)
> computeAIC(3, 3)
[1] -476.0351
```

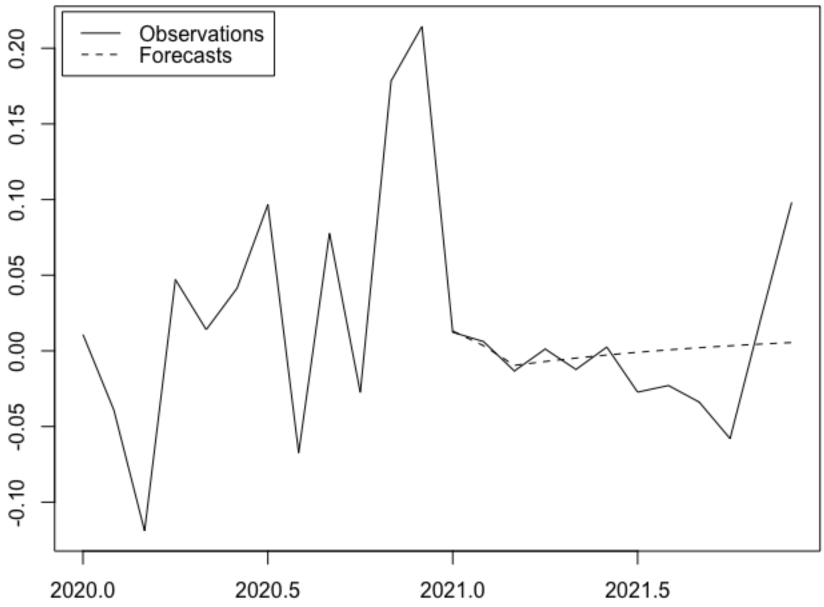
- By setting $p_{max} = q_{max} = 3$, we find that the minimum AIC arises when $p = 1$ and $q = 3$.

```
> fit <- arima(rtn1, order = c(1, 0, 3), method = "ML")
> fit
Coefficients:
        ar1     ma1    ma2     ma3 intercept
     0.8903 -0.9329 0.0046 -0.0717    0.0137
s.e. 0.0367  0.0712 0.0936  0.0763    0.0007
```

- The ARMA(1, 3) model is estimated as

$$\begin{aligned}(x_t - 0.0137) &= 0.8903(x_{t-1} - 0.0137) + \varepsilon_t - 0.9329\varepsilon_{t-1} + 0.0046\varepsilon_{t-2} \\ &\quad -0.0717\varepsilon_{t-3}.\end{aligned}$$

```
> fcst.x <- predict(fit, 12)$pred
> plot(new.x, xlab = "", ylab = "")
> lines(fcst.x, lty = 2)
> legend("topleft", c("Observations", "Forecasts"), lty = c(1, 2),
         inset = 0.01)
```


Observations
Forecasts
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
2020.0
2020.5
2021.0
2021.5