STAT302: Time Series Analysis

Chapter 5. Nonstationary Processes and ARIMA Models

Sangbum Choi, Ph.D

Department of Statistics, Korea University

Outline

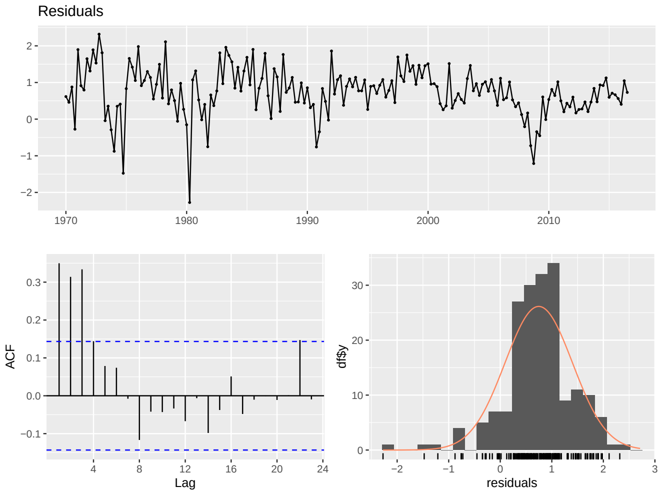
Building ARIMA models

There are a few basic steps to fitting ARIMA models to time series data. These steps involve

- plotting the data,
- possibly transforming the data,
- identifying the dependence orders of the model,
- parameter estimation,
- diagnostics, and
- model choice.

US comsumption data

```
library(fpp2); library(forecast)
y = uschange[,c("Consumption")]
checkresiduals(y)
```

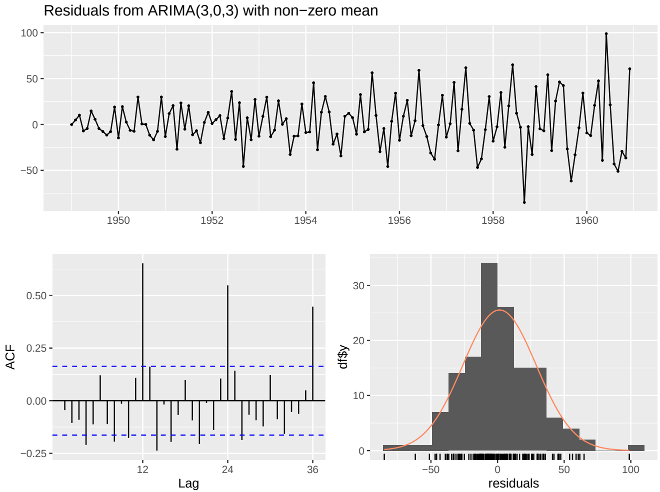


US comsumption data

```r
auto.arima(y)
## Series: y
## ARIMA(1,0,3)(1,0,1)[4] with non-zero mean
##
## Coefficients:
##           ar1     ma1     ma2     ma3     sar1    sma1    mean
##       -0.3548  0.5958  0.3437  0.4111  -0.1376  0.3834  0.7460
## s.e.   0.1592  0.1496  0.0960  0.0825   0.2117  0.1780  0.0886
##
## sigma^2 = 0.3481:  log likelihood = -163.34
## AIC=342.67   AICc=343.48   BIC=368.52
```

AirPassengers data

```r
AP = Arima(AirPassengers, order=c(3,0,3)) #ARMA
checkresiduals(AP)
```



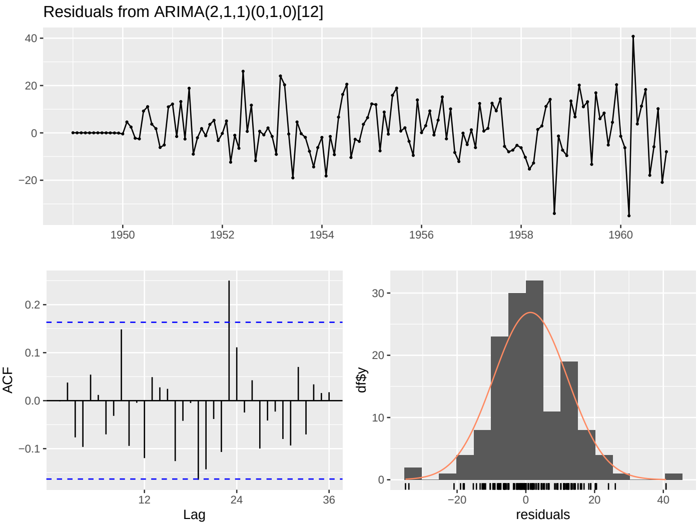

```
##
```

AirPassengers data

```r
print(AP2 <- auto.arima(AirPassengers))
## Series: AirPassengers
## ARIMA(2,1,1)(0,1,0)[12]
##
## Coefficients:
##          ar1     ar2      ma1
##       0.5960  0.2143  -0.9819
## s.e.  0.0888  0.0880   0.0292
##
## sigma^2 = 132.3:  log likelihood = -504.92
## AIC=1017.85   AICc=1018.17   BIC=1029.35
```

AirPassengers data

```r
checkresiduals(AP2)
```



##

Outline

Building ARIMA Models

Nonstationary Processes and Differencing

Autoregressive Integrated Moving Average (ARIMA)

Multiplicative Seasonal ARIMA Models

Linear models for nonstationary processess

- Often time series $\{X_t\}$ is nonstationary in a particular way known as "homogeneous stationarity".
- That is, apart from differences in local mean level and perhaps local linear trend, different portions of the time series have similar statistical characteristics.
- A class of time series we consider are nonstationary series, but the first-order difference

$$W_t = \nabla X_t = X_t - X_{t-1} = (1 - B)X_t$$

may form a stationary series.

Linear models for nonstationary processess

- Or even first difference is nonstationary, but the second-order difference

$$
\begin{aligned}
\nabla^2 X_t &= \nabla W_t \\
&= W_t - W_{t-1} \\
&= (1-B)X_t - (1-B)X_{t-1} \\
&= (1-B)X_t - (1-B)BX_t \\
&= (1-B)^2 X_t
\end{aligned}
$$

may form a stationary series.

Linear models for nonstationary processess

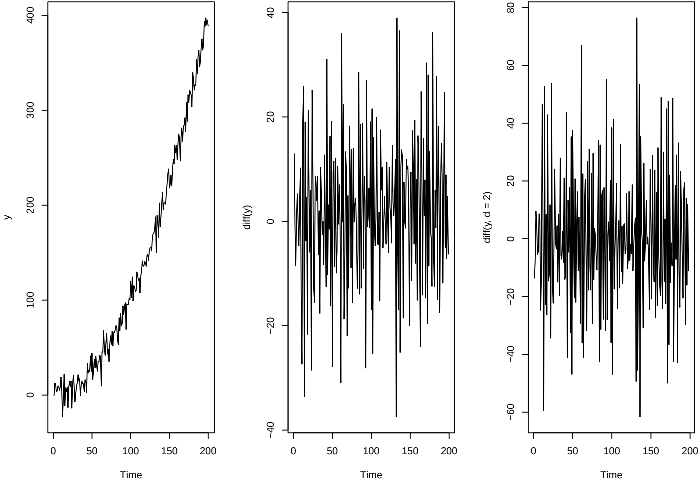
- In general, the d^{th} differences

$$\nabla^d X_t = (1-B)^d X_t, \quad d = 1, 2, \ldots$$

  form a stationary series.
- In practice, usually $d = 1$ and occasionally $d = 2$.
- We say that the differencing operator **reduces the nonstationary series to stationarity**.

Nonstationary processes and differencing

```r
y = (1:200)^2/100 + rnorm(200)*10
par(mfrow=c(1,3))
ts.plot(y); ts.plot(diff(y)); ts.plot(diff(y,d=2))
```



Unit-root nonstationarity

- Consider an AR(1) model

$$X_t = \phi X_{t-1} + \epsilon_t$$

  we see that

  - $|\phi| < 1 \iff$ AR(1) is stationary
  - $|\phi| = 1 \iff$ AR(1) is nonstationary
  - $|\phi| > 1 \iff$ AR(1) is explosive.

- Note that if $\phi = 1$, the AR(1) model,

$$X_t = X_{t-1} + \epsilon_t,$$

  is in fact a **random walk** and called a **unit-root nonstationary time series.**

Unit-root nonstationarity

- Clearly,

$$\epsilon_t = X_t - X_{t-1} \sim \mathrm{WN}(0, \sigma_\epsilon^2)$$

- By adding all ϵ_t terms, we obtain

$$X_t = X_0 + \epsilon_1 + \epsilon_2 + \cdots + \epsilon_t$$

- In the random walk, we can show that

$$EX_t = 0,\ \mathrm{Var}(X_t) = \mathrm{Var}\left(\sum_{j=1}^{t} \epsilon_j\right) = t\sigma_\epsilon^2 \to \infty \text{ as } t \to \infty.$$

- Note X_t is not weakly stationary, but ∇X_t is stationary.

Dickey–Fuller (DF) test

- In statistics, the **Dickey–Fuller test** tests the null hypothesis that a unit root is present in an autoregressive (AR) time series model.
- It tests $H_0 : \xi = 0$ in the AR model:

$$X_t - X_{t-1} = \xi X_{t-1} + \epsilon_t$$

- If the null hypothesis is accepted, the model is a random walk and a differencing is required to achieve stationarity.
- The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity.
- Also, can use KPSS test, Phillips–Perron (PP) test.

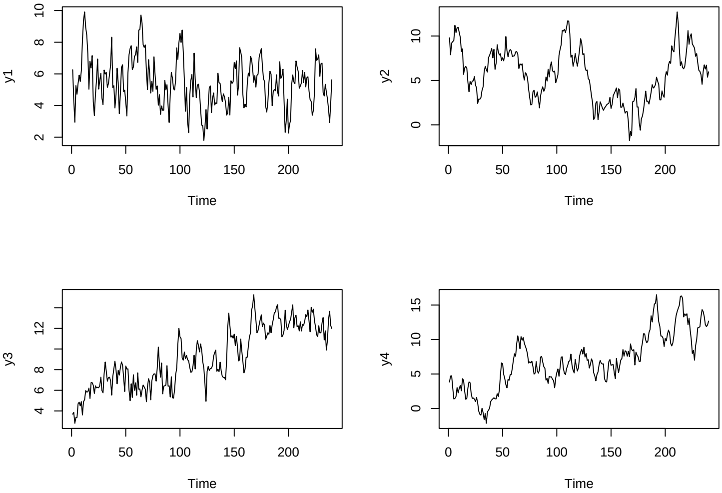
Dickey–Fuller (DF) test

- Note that the "near" unit root is present in time series `y2` and `y4`, and thus we hope that the Dickey-Fuller test suggests unit roots in `y2` and `y4`.
- One may use `adf.test()`, `kpss.test()`, and `pp.test()` in R for hypothesis testing of unit roots.

```
library(tseries)
library(forecast)
set.seed(12345)
n <- 240
time <- (1:n)/10
y1 <- arima.sim(n=n, model=list(ar=0.75)) + 5
y2 <- arima.sim(n=n, model=list(ar=0.92)) + 5
y3 <- arima.sim(n=n, model=list(ar=0.80)) + 5 + 0.3*time
y4 <- arima.sim(n=n, model=list(ar=0.92)) + 5 + 0.3*time
```

Dickey–Fuller (DF) test

```
par(mfrow=c(2,2))
plot(y1); plot(y2); plot(y3); plot(y4)
```



Dickey–Fuller (DF) test

```
adf.test(y1)
##
##  Augmented Dickey-Fuller Test
##
## data:  y1
## Dickey-Fuller = -4.9139, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
adf.test(y2)
##
##  Augmented Dickey-Fuller Test
##
## data:  y2
## Dickey-Fuller = -2.7009, Lag order = 6, p-value = 0.2811
## alternative hypothesis: stationary
```

Dickey–Fuller (DF) test

```
adf.test(y3)
##
##  Augmented Dickey-Fuller Test
##
## data:  y3
## Dickey-Fuller = -4.3518, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
adf.test(y4)
##
##  Augmented Dickey-Fuller Test
##
## data:  y4
## Dickey-Fuller = -3.4632, Lag order = 6, p-value = 0.04696
## alternative hypothesis: stationary
```

Linear regression with ARIMA errors

For time series `y3` and `y4`, we can fit a linear regression model with ARIMA errors. For example,

```
Arima(y4, order=c(1,0,0), xreg=time)
## Series: y4
## Regression with ARIMA(1,0,0) errors
##
## Coefficients:
##          ar1  intercept    xreg
##       0.9013     1.7010  0.4423
## s.e.  0.0273     1.1824  0.0834
##
## sigma^2 = 0.9815:  log likelihood = -337.64
## AIC=683.27   AICc=683.44   BIC=697.2
```

Outline

Autoregressive Integrated Moving Average (ARIMA)

- A (nonstationary) process $\{X_t\}$ is **autoregressive integrated moving average of orders** (p, d, q), denoted as ARIMA(p, d, q), if the d^{th} differences

$$W_t = (1 - B)^d X_t$$

form a stationary invertible ARMA(p, q).

- That is, the process $\{W_t\}$ satisfies the ARMA format:

$$\phi(B) W_t = \delta + \theta(B)\epsilon_t$$

- The ARIMA(p, d, q) process $\{X_t\}$ satisfies the model equation

$$\phi(B)(1 - B)^d X_t = \delta + \theta(B)\epsilon_t$$

ARIMA processes

- An ARIMA process $\{X_t\}$ is called an "integrated" process of order d, denoted by $I(d)$.
- Consider the case $d = 1$, then X_t is $I(1)$ as

$$(1 - B)X_t = W_t$$

  is stationary.
- From $X_t - X_{t-1} = W_t$, by successive back substituting, we have

$$\begin{aligned} X_t &= W_t + X_{t-1} \\ &= W_t + W_{t-1} + X_{t-2} \\ &= \cdots \\ &= W_t + W_{t-1} + \cdots + W_2 + W_1 + X_0 \end{aligned}$$

- Thus, $\{X_t\}$ is an "integrated" form of stationary $\{W_t\}$.

ARIMA processes

- In ARIMA(p, d, q), the order of differencing d is understood as the smallest integer that produces a stationary series for $W_t = (1 - B)^d X_t$.
- A random walk defined as

$$X_t = X_{t-1} + \delta + \epsilon_t$$

  is the most elementary example of the ARIMA process with order $(0, 1, 0)$.
- An ARIMA process is also called a "unit-root" process.

Three forms of ARIMA models

The ARIMA(p, d, q) process$\{X_t\}$ satisfying the model equation

$$\phi(B)(1 - B)^d X_t = \delta + \theta(B)\epsilon_t$$

can be represented in three forms.

1. Difference equation or ARIMA equation form.
2. Infinite MA representation.
3. Infinite AR representation.

ARIMA equation form

- Define the generalized AR operator $\varphi(B)$ as

$$\begin{aligned}\phi(B)(1-B)^d &= (1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)(1-B)^d \\ &= 1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_{p+d}B^{p+d} \\ &= \varphi(B).\end{aligned}$$

- Then an ARIMA model can be represented as

$$\varphi(B)X_t = \delta + \theta(B)\epsilon_t$$

- Equivalently,

$$X_t - \sum_{i=1}^{p+d} \varphi_i X_{t-i} = \delta + \epsilon_t - \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

ARIMA equation form

- In this form, the model equation for X_t is of the same form as an ARMA$(p+d, q)$ model.
- However this ARMA$(p+d, q)$ model is nonstationary, because the generalize AR operator

  $$\varphi(B) = \phi(B)(1-B)^d$$

  has a corresponding polynomial function

  $$m^{p+d} - \varphi_1 m^{p+d-1} - \cdots - \varphi_{p+d} = (m^p - \phi_1 m^{p-1} - \cdots - \phi_p)(m-1)^d$$

  which has d unit roots (i.e. equal to 1).
- Hence this ARIMA model is called a "unit-root process".

ARIMA(1,1,1)

ARIMA(1,1,1) has the equation, with $|\phi| < 1$ and $|\theta| < 1$,

$$(1 - \phi B)(1 - B)X_t = \delta + (1 - \theta B)\epsilon_t$$

$$\Rightarrow (1 - \varphi_1 B - \varphi_2 B^2)X_t = \delta + (1 - \theta B)\epsilon_t$$

where

$$\varphi_1 = 1 + \phi, \quad \varphi_2 = -\phi$$

ARIMA(1,1,1)

- Suppose $\phi = 0.8$, then we have

$$(1 - 0.8B)(1 - B)X_t = \delta + (1 - \theta B)\epsilon_t$$

$$\Rightarrow (1 - 1.8B - 0.8B^2)X_t = \delta + (1 - \theta B)\epsilon_t$$

which has the same form as ARMA(2,1).

- However,

$$\varphi(B) = 1 - 1.8B + 0.8B^2$$

has the associated equation

$$m^2 - 1.8m + 0.8 = (m - 0.8)(m - 1) = 0$$

which has roots 0.8 and 1.

- Hence X_t is not stationary, but so is ∇X_t.

ARIMA(1,2,1)

- ARIMA(1,2,1) has the equation with $|\phi| < 1$ and $|\theta| < 1$:

$$(1 - \phi B)(1 - B)^2 X_t = \delta + (1 - \theta B)\epsilon_t$$

- Notice that ARIMA(1,2,1) = Nonstationary ARMA(3,1). Clearly, it is not invertible, because it has two unit roots.
- On the other hand, $\nabla^2 X_t = (1 - B)^2 X_t \sim \text{ARMA}(1, 1)$ is invertible.

ARIMA(0,1,1) $\equiv$ IMA(1,1)

- A special case of ARIMA(1,1,1) is when $\phi = 0$.
- Then the model equation is

$$(1 - B)X_t = \delta + (1 - \theta B)\epsilon_t$$

  where $\varphi_1 = 1, \varphi_2 = 0$.
- This model is called IMA(1,1), which is ARIMA(0,1,1).

Outline

Modeling seasonal effects

- Now, we introduce several modifications made to the ARIMA model to account for seasonal and nonstationary behavior.
- Often, the dependence on the past tends to occur most strongly at multiples of some underlying seasonal lag s.
- For example, with monthly data, there is a strong yearly component occurring at lags that are multiples of $s = 12$. Data taken quarterly will exhibit the yearly repetitive period at $s = 4$ quarters.
- Let S_t denote the seasonal effect of X_t with $s = 12$. Often,

$$S_t \approx S_{t-12} \approx S_{t-24} \approx \cdots \approx S_{t-ks}$$

- Perhaps, S_t is stationary and follows an ARMA model.

Seasonal ARMA processes

- The pure seasonal autoregressive moving average model, say, ARMA$(P, Q)_s$, takes the form

$$\Phi_P(B^s)S_t = \Theta_Q(B^s)\epsilon_t,$$

where the operators

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}$$

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs}$$

are the seasonal autoregressive operator and the seasonal moving average operator of orders P and Q, respectively, with seasonal period s.

Seasonal AR(1) series

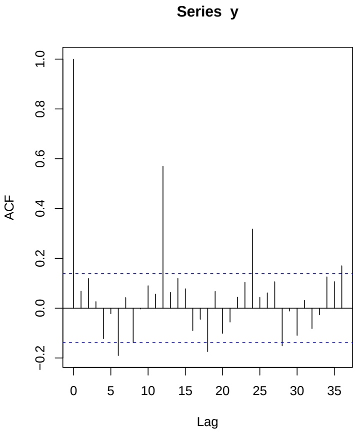
- A first-order seasonal autoregressive series that might run over months could be written as

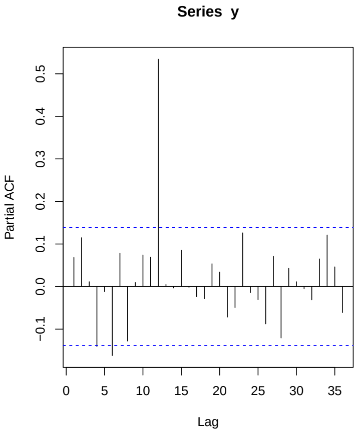
$$(1 - \Phi B^{12})S_t = \epsilon_t \quad \Leftrightarrow \quad S_t = \Phi S_{t-12} + \epsilon_t$$

- This model exhibits the series S_t in terms of past lags at the multiple of the yearly seasonal period $s = 12$ months.
- It is clear from the above form that estimation and forecasting for such a process involves only straightforward modifications of the unit lag case already treated.
- In particular, the causal condition requires $|\Phi| < 1$.

Seasonal AR(1) series

```
phi = c(rep(0,11),.6)
y = arima.sim(list(order=c(12,0,0), ar=phi), n=200)
par(mfrow=c(1,2)); acf(y,lag.max=36); pacf(y,lag.max=36)
```





A mixed seasonal model

- Consider an ARMA$(0,1) \times (1,0)_{12}$ model

$$X_t = \Phi X_{t-12} + \epsilon_t + \theta \epsilon_{t-1},$$

  where $|\Phi| < 1$ and $|\theta| < 1$.

- Because X_{t-12}, ϵ_t, and ϵ_{t-1} are uncorrelated, and X_t is stationary,

$$\gamma(0) = \frac{1+\theta^2}{1-\Phi^2}\sigma_\epsilon^2$$

A mixed seasonal model

- In addition, multiplying the model by $X_{t-h}, h > 0$, and taking expectations, we have $\gamma(1) = \Phi\gamma(11) + \theta\sigma_\epsilon^2$, and $\gamma(h) = \Phi\gamma(h-12)$, for $h \geq 2$.
- Thus, the ACF for this model is

$$\rho(12h) = \Phi^h, \quad h = 1, 2, \ldots$$

$$\rho(12h-1) = \rho(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h, \quad h = 0, 1, 2, \ldots$$

$$\rho(h) = 0, \quad \text{otherwise.}$$

Behavior of ACF and PACF for SARMA

Table 3.3. Behavior of the ACF and PACF for pure SARMA models

	$AR(P)_s$	$MA(Q)_s$	$ARMA(P, Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots,$ are zero

Strategy for seasonal ARIMA processes

For general nonstationary and seasonal time series X_t,

(1) take a difference $\nabla^d X_t = (1 - B)^d X_t$ for removing a trend and making it stationary;

(2) fit an ARMA(p, q) for $\nabla^d X_t$.

Combining (1) and (2) yields an ARIMA(p, d, q) process, say Z_t. If seasonal effects are present, do

(3) take a seasonal difference $\nabla_s^D Z_t = (1 - B^s)^D Z_t$ for seasonal detrending;

(4) fit an ARMA$(P, Q)_s$ for $\nabla_s^D Z_t$.

Combining (1)-(4) yields a ARIMA$(p, d, q) \times (P, D, Q)_s$ process.

Seasonal ARIMA processes

- The multiplicative **seasonal autoregressive integrated moving average model** or **SARIMA** model is given by

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d Xt = \delta + \Theta_Q(B^s)\theta(B)\epsilon_t,$$

  where ϵ_t is the usual (Gaussian) white noise process. The general model is denoted as **ARIMA**$(p, d, q) \times (P, D, Q)_s$.

- The ordinary autoregressive and moving average components are represented by polynomials $\phi(B)$ and $\theta(B)$ of orders p and q, respectively, and the seasonal autoregressive and moving average components by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ of orders P and Q and ordinary and seasonal difference components by $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$.

Seasonal ARIMA processes

- Consider the following model, which often provides a reasonable representation for seasonal, nonstationary, economic time series.
- We exhibit the equations for the model, denoted by ARIMA$(0,1,1) \times (0,1,1)_{12}$, where the seasonal fluctuations occur every 12 months.
- Then, with $\delta = 0$, the SARIMA model becomes

$$\nabla_{12}\nabla X_t = \Theta(B^{12})\theta(B)\epsilon_t$$

- Equilvalently,

$$(1 - B^{12})(1 - B)X_t = (1 + \Theta B^{12})(1 + \theta B)\epsilon_t.$$

Seasonal ARIMA processes

- Expanding both sides leads to

$$(1 - B - B^{12} + B^{13})X_t = (1 + \theta B + \Theta B^{12} + \theta\Theta B^{13})\epsilon_t,$$

  or in difference equation form

$$X_t = X_{t-1} + X_{t-12} - X_{t-13} + \epsilon_t + \theta\epsilon_{t-1} + \Theta\epsilon_{t-12} + \theta\Theta\epsilon_{t-13}.$$

- Note that the multiplicative nature of the model implies that the coefficient of ϵ_{t-13} is the product of the coefficients of ϵ_{t-1} and ϵ_{t-12} rather than a free parameter.
- The multiplicative model assumption seems to work well with many seasonal time series data sets while reducing the number of parameters that must be estimated.

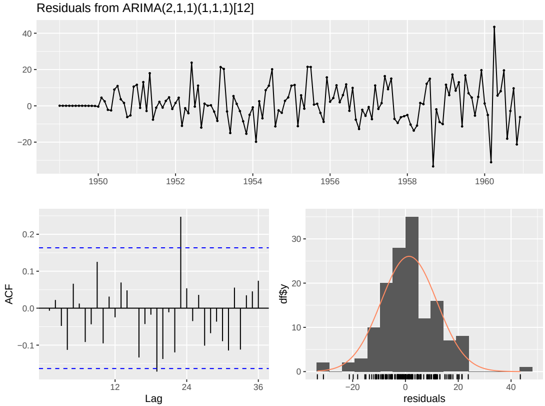
SARIMA for AirPassengers data

Seasonal ARIMA for AirPassengers data

```
AP1 = Arima(AirPassengers, order=c(2,1,1), seasonal=c(1,1,1))
AP1
## Series: AirPassengers
## ARIMA(2,1,1)(1,1,1)[12]
##
## Coefficients:
##          ar1     ar2      ma1      sar1    sma1
##       0.5800  0.2287  -0.9782  -0.9010  0.8095
## s.e.  0.0892  0.0880   0.0289   0.2516  0.3462
##
## sigma^2 = 129.4:  log likelihood = -503.12
## AIC=1018.25   AICc=1018.93   BIC=1035.5
```

SARIMA for AirPassengers data

```
checkresiduals(AP1)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,1)(1,1,1)[12]
## Q* = 34.929, df = 19, p-value = 0.01424
```

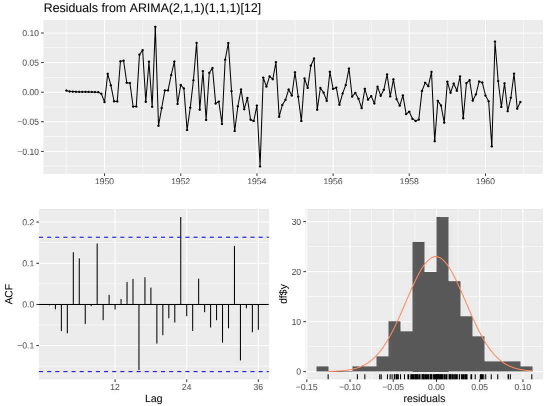
SARIMA for AirPassengers data

Seasonal ARIMA for AirPassengers data

```
AP2 = Arima(AirPassengers, order=c(2,1,1),
                seasonal=c(1,1,1), lambda=0)
AP2
## Series: AirPassengers
## ARIMA(2,1,1)(1,1,1)[12]
## Box Cox transformation: lambda= 0
##
## Coefficients:
##          ar1     ar2      ma1     sar1     sma1
##       0.5552  0.2530  -0.9653  -0.0598  -0.5168
## s.e.  0.0956  0.0949   0.0466   0.1551   0.1367
##
## sigma^2 = 0.001359:  log likelihood = 246.21
## AIC=-480.42   AICc=-479.74   BIC=-463.17
```

SARIMA for `AirPassengers` data

```r
checkresiduals(AP2)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,1)(1,1,1)[12]
## Q* = 27.162, df = 19, p-value = 0.1009
```

SARIMA for AirPassengers data

```
forecast(AP2, h=24) %>% plot()
```