Time Series Analysis 3
DS351

AutoRegressive Integrated Moving Average (ARIMA)

Stationarity

A time series is **stationary** if **its statistical properties do not change over time**.

Stationarity

A time series is **stationary** if **its statistical properties do not change over time**.

What are statistical properties?

- mean
- variance
- covariance
- etc.

Stationarity

More precise definition:

A time series is stationary if
the distribution of $(Y_t, Y_{t+1}, \ldots, Y_{t+s})$ does not depend on t.

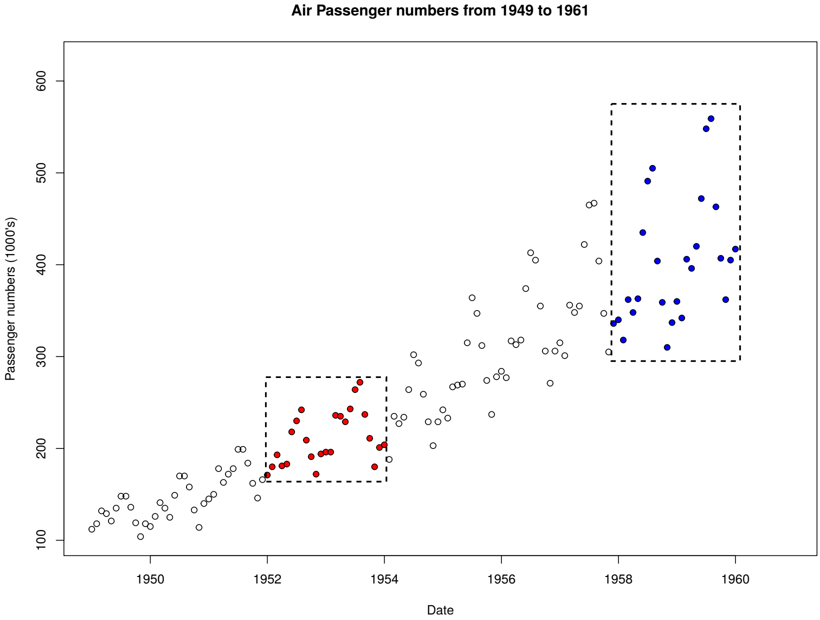
Stationarity

More precise definition:

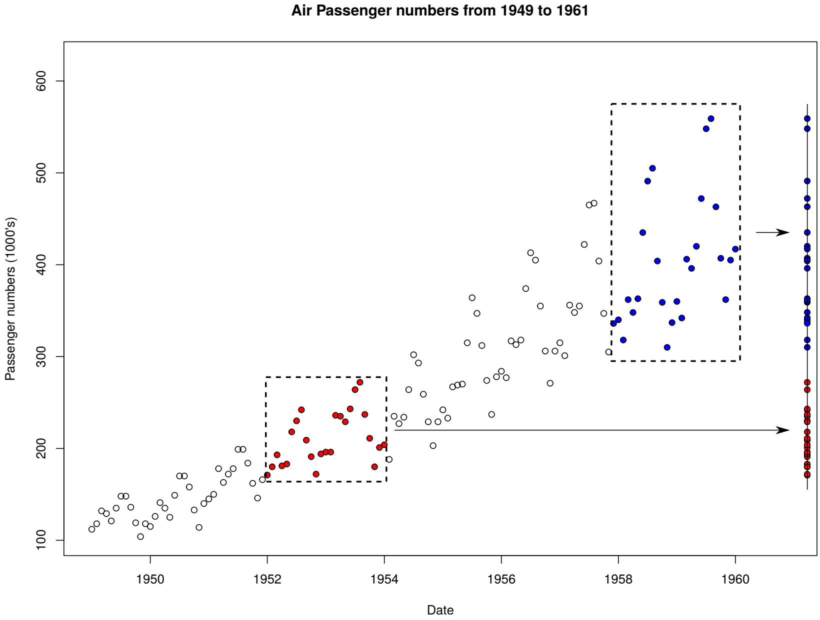
A time series is stationary if
the distribution of $(Y_t, Y_{t+1}, \ldots, Y_{t+s})$ does not depend on t.

- we usually don't know the distribution of these variables.
- It is usually easier to detect that a time series is not stationary by looking at its plot.

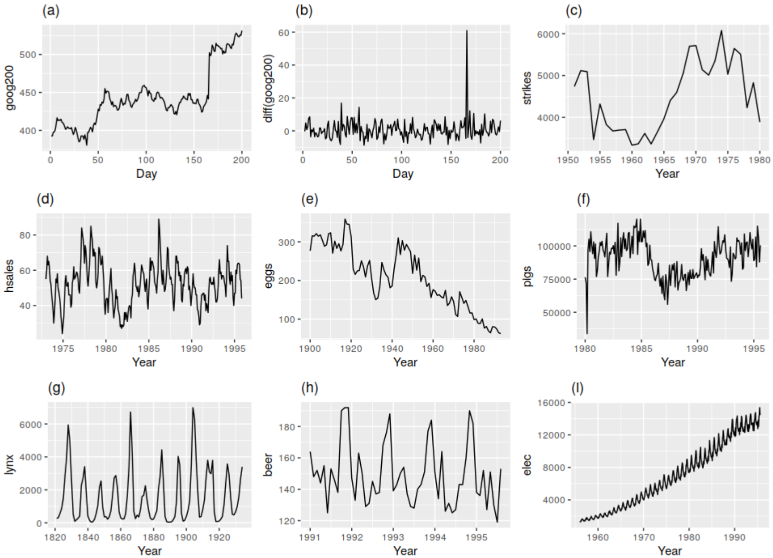
Example


Air Passenger numbers from 1949 to 1961
Passenger numbers (1000's)
100
200
300
400
500
600
1950
1952
1954
1956
1958
1960
Date

Example


Air Passenger numbers from 1949 to 1961
Passenger numbers (1000's)
100
200
300
400
500
600
1950
1952
1954
1956
1958
1960
Date

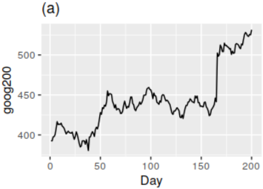
Examples


(a)
goog200
500
450
400
0
50
100
150
200
Day
(b)
diff(goog200)
60
40
20
0
0
50
100
150
200
Day
(c)
strikes
6000
5000
4000
1950
1955
1960
1965
1970
1975
1980
Year
(d)
hsales
80
60
40
1975
1980
1985
1990
1995
Year
(e)
eggs
300
200
100
1900
1920
1940
1960
1980
Year
(f)
pigs
100000
75000
50000
1980
1985
1990
1995
Year
(g)
lynx
6000
4000
2000
0
1820
1840
1860
1880
1900
1920
Year
(h)
beer
180
160
140
120
1991
1992
1993
1994
1995
Year
(I)
elec
16000
12000
8000
4000
1960
1970
1980
1990
Year

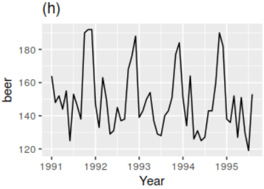
Find stationarity from the plot

In summary, a time series is **not** stationary if there is

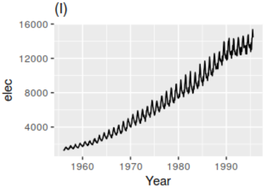
- trend
- seasonality
- increase/decrease in variance



has trend



has seasonality



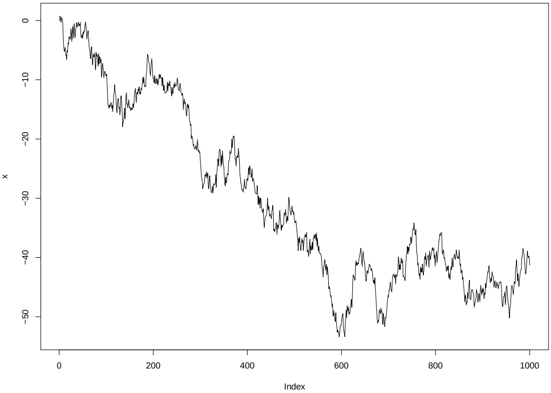
has both

Random walk

Random walk is a simple non-stationary process.

$$y_t = y_{t-1} + \epsilon_t.$$

where ϵ_t is a *white noise* with zero mean and variance σ^2 e.g. $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.



Differencing

From the random walk

$$y_t = y_{t-1} + \epsilon_t,$$

which is not stationary, we can transform it into

$$z_t = y_t - y_{t-1} = \epsilon_t.$$

Now, z_t is stationary.

Differencing

From the random walk

$$y_t = y_{t-1} + \epsilon_t,$$

which is not stationary, we can transform it into

$$z_t = y_t - y_{t-1} = \epsilon_t.$$

Now, z_t is stationary.

In general, we try to make a stationary time series by transforming $z_t = y_t - y_{t-1}$. This is known as **differencing**.

Second-order differencing

Second-order differencing
If differencing is not enough to make a time series stationary, we do it twice.

$$z'_t = z_t - z_{t-1} = z_t - 2z_{t-1} + z_{t-2}.$$

Second-order differencing

Second-order differencing

If differencing is not enough to make a time series stationary, we do it twice.

$$z'_t = z_t - z_{t-1} = z_t - 2z_{t-1} + z_{t-2}.$$

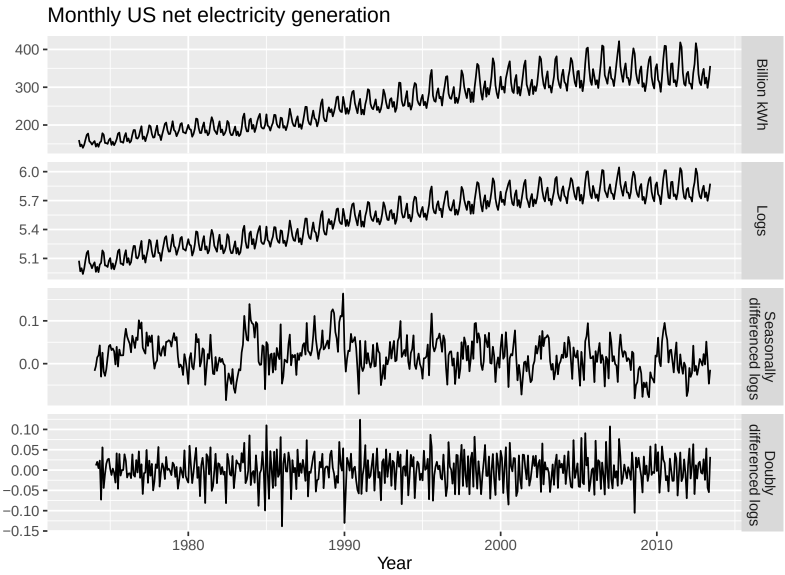
Seasonal differencing

To remove seasonality, we take the difference between an observation and the previous observation from the same season.

$$z_t = y_t - y_{t-m},$$

where m is the length of seasonality e.g. $m = 12$ for annual seasonality.

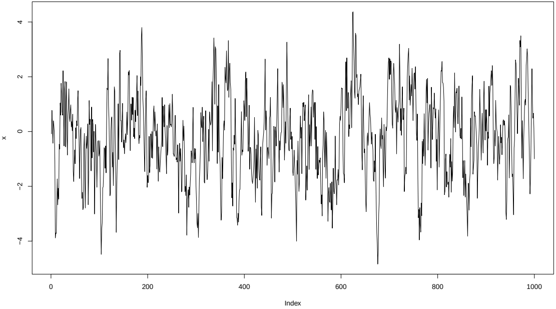
Example


Monthly US net electricity generation
400
300
200
Billion kWh
6.0
5.7
5.4
5.1
Logs
0.1
0.0
Seasonally differenced logs
0.10
0.05
0.00
−0.05
−0.10
−0.15
Doubly differenced logs
1980
1990
2000
2010
Year

Unit root

The random walk is an example of a time series that has a **unit root**: the value of α in $y_t = \alpha y_{t-1} + \epsilon_t$ is $\alpha = 1$.

In general, **a time series is non-stationary if it has a unit root**.



$$y_t = 0.8 y_{t-1} + \epsilon_t$$

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

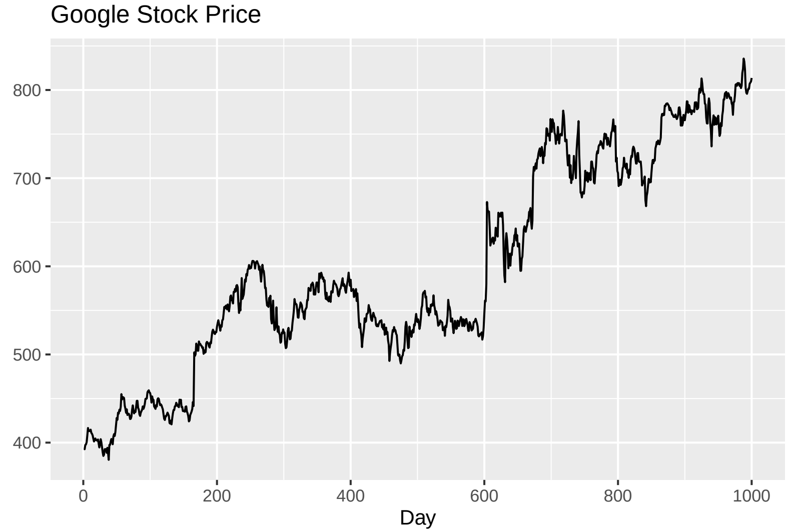
This is a test to see if we want differencing.

KPSS test

$$\begin{aligned} H_0 &: y_t \text{ is stationary} \\ H_a &: y_t \text{ is \textbf{not} stationary} \end{aligned}$$

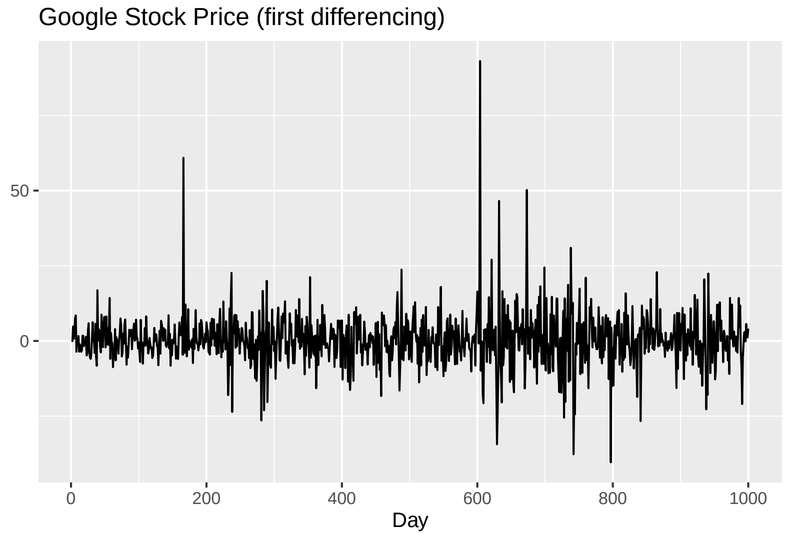
If H_0 is rejected ($p-$value < 0.05) then differencing y_t is required.

Example



$p-$value < 0.01

Example



$p\text{−value} > 0.1$

Autocorrelation

Autocorrelation measures the linear relationship between a time series and its *lagged values*.

$\boldsymbol{Y}_t$	y_1	$\dots$	y_{k+1}	$\dots$	y_t	$\dots$	y_T		
$\boldsymbol{Y}_{t-k}$		$\rightarrow$	y_1	$\dots$	y_{t-k}	$\dots$	y_{T-k}	$\dots$	$\rightarrow$

Autocorrelation

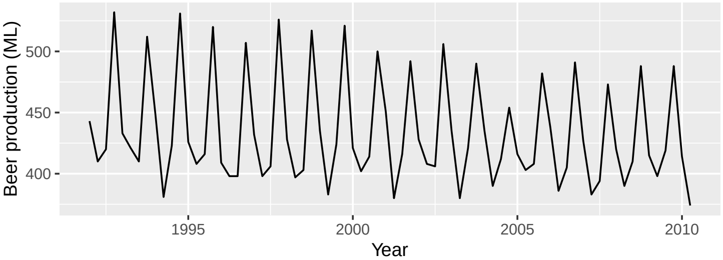
Autocorrelation measures the linear relationship between a time series and its *lagged values*.

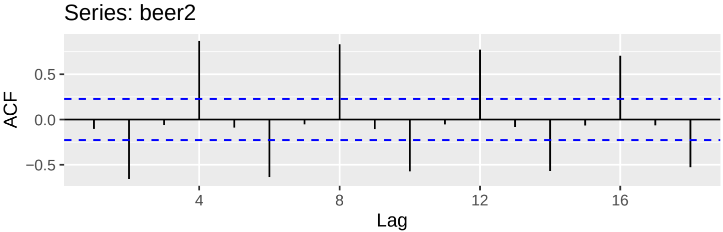
$\boldsymbol{Y}_t$	y_1	$\dots$	y_{k+1}	$\dots$	y_t	$\dots$	y_T		
$\boldsymbol{Y}_{t-k}$		$\rightarrow$	y_1	$\dots$	y_{t-k}	$\dots$	y_{T-k}	$\dots$	$\rightarrow$

The autocorrelation at lag k, r_k can be written as

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y - \bar{y})^2}$$

Beer production



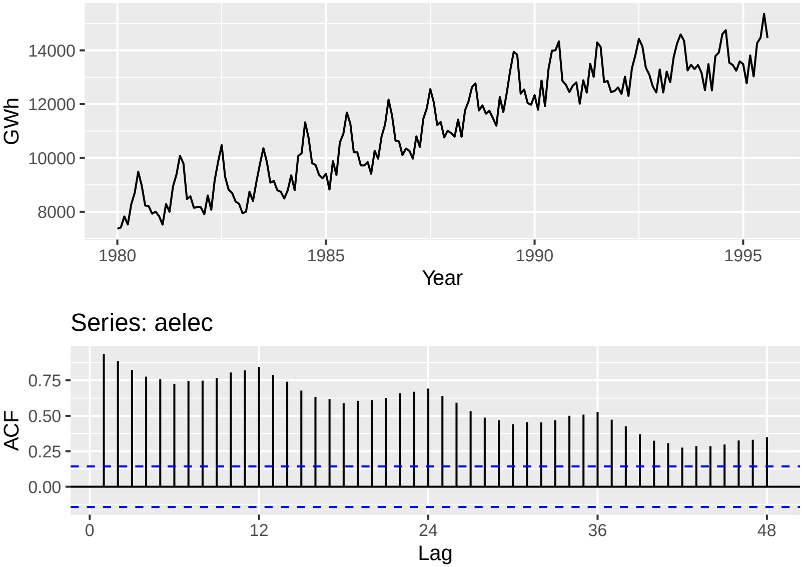


The values beyond blue lines are significantly different than zero.

Trend and seasonality in ACF plots

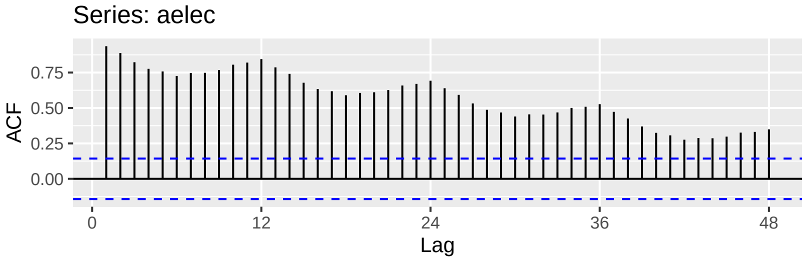
- When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.
- When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags.

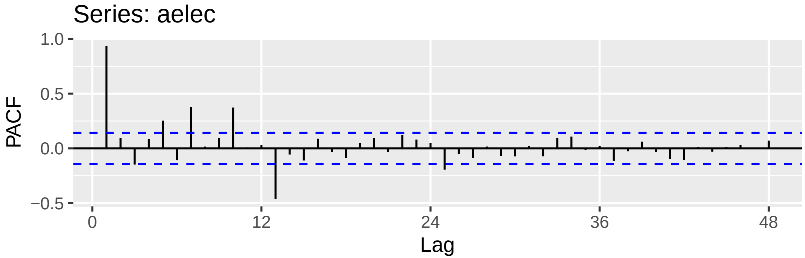
Australian electricity demand



Partial autocorrelation function (PACF)

Partial autocorrelation function measures the part of the correlation that has not been explained by the earlier lags.



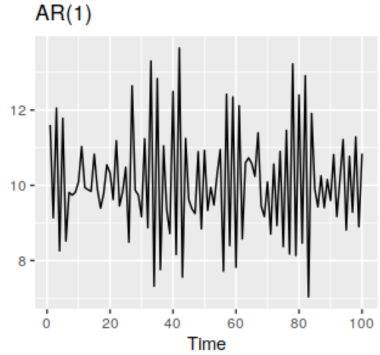


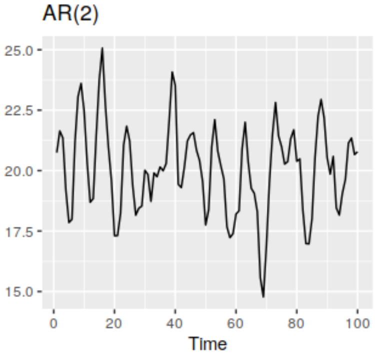
Autoregressive model

An autoregressive model of order p, $\mathbf{AR}(\boldsymbol{p})$, is

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

which is a multiple linear regression with $y_{t-p}, \ldots, y_{t-1}$ as predictors and y_t as the responses variable.

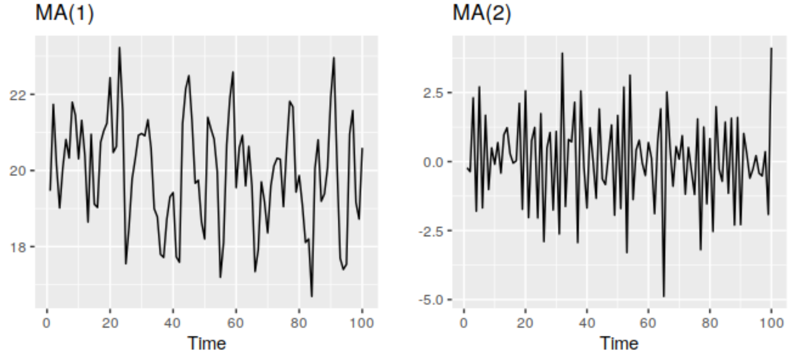




Moving average model

A moving average model, **MA($\boldsymbol{q}$)**, uses **past errors** to forecast.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}.$$



ARIMA model

AutoRegressive Integrated Moving Average (ARIMA) is a combination of both AR and MA models:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (1)$$

where y'_t is the differenced series. Thus there are a total of $p + q$ predictors. We call this an **ARIMA**$(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q})$ model.

ARIMA model

AutoRegressive Integrated Moving Average (ARIMA) is a combination of both AR and MA models:

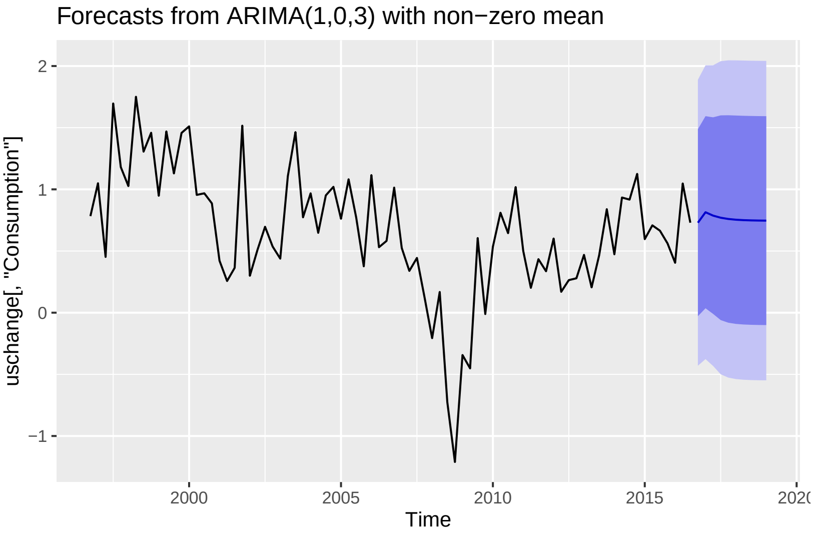
$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (1)$$

where y'_t is the differenced series. Thus there are a total of $p + q$ predictors. We call this an **ARIMA($\boldsymbol{p, d, q}$)** model.

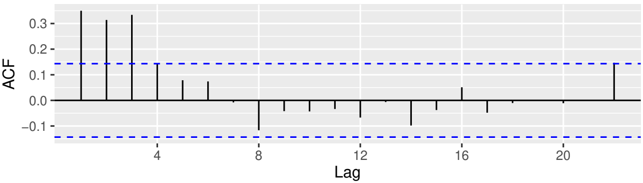
The value of c and d have the following effects on the long-term forecast.

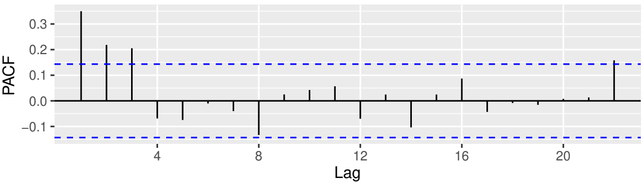
$c = 0$	$d = 0$	forecasts go to zero.
$c = 0$	$d = 1$	forecasts go to a non-zero constant.
$c = 0$	$d = 2$	forecasts follow a straight line.
$c \neq 0$	$d = 0$	forecasts go to the mean of the data.
$c \neq 0$	$d = 1$	forecasts follow a straight line.
$c \neq 0$	$d = 2$	forecasts follow a quadratic trend.

Example: Quarterly US consumption



Find p and q from ACF and PACF plots

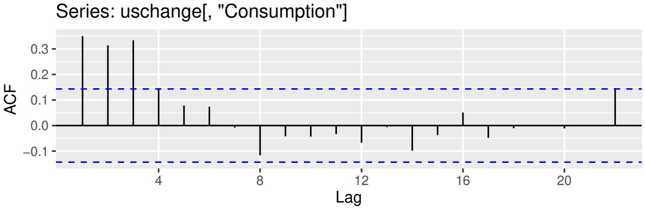

Series: uschange[, "Consumption"]

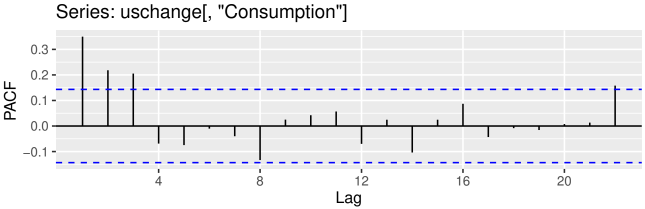

Series: uschange[, "Consumption"]

ACF and PACF plots of **differenced data** can help us find

- p in ARIMA$(p, d, 0)$
- q in ARIMA$(0, d, q)$
- Plots do not help for ARIMA(p, d, q) when $p, q > 0$.

Find p and q from ACF and PACF plots

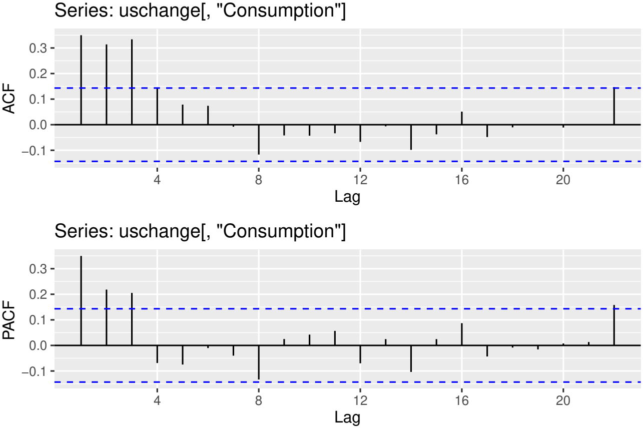




The data may follow an ARIMA$(p, d, 0)$ model if

- the ACF is exponentially decaying or sinusoidal
- there is a significant spike at lag p in the PACF, but none beyond lag p.

Find p and q from ACF and PACF plots

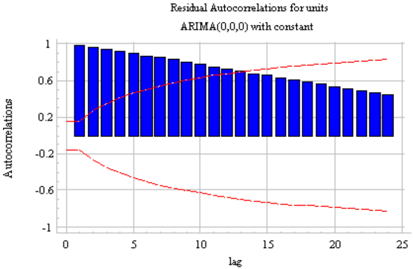

Series: uschange[, "Consumption"]

Series: uschange[, "Consumption"]

The data may follow an ARIMA$(0, d, q)$ model if

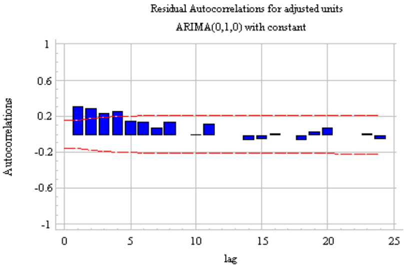
- the PACF is exponentially decaying or sinusoidal
- there is a significant spike at lag q in the ACF, but none beyond lag q.

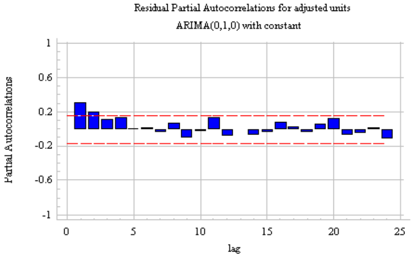
Example



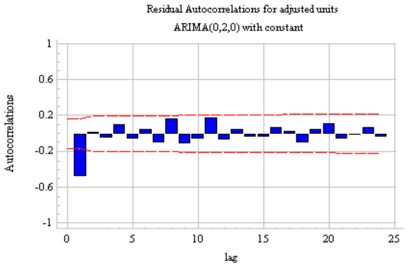
Residual Autocorrelations for units
ARIMA(0,0,0) with constant
Autocorrelations
1
0.6
0.2
-0.2
-0.6
-1
0
5
10
15
20
25
lag

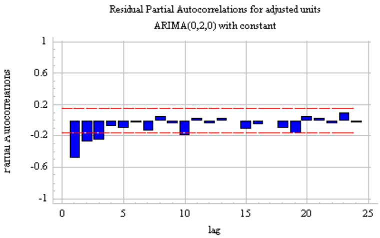
Example


Residual Autocorrelations for adjusted units
ARIMA(0,1,0) with constant
Autocorrelations
1
0.6
0.2
-0.2
-0.6
-1
0
5
10
15
20
25
lag


Residual Partial Autocorrelations for adjusted units
ARIMA(0,1,0) with constant
Partial Autocorrelations
1
0.6
0.2
-0.2
-0.6
-1
0
5
10
15
20
25
lag

Example


Residual Autocorrelations for adjusted units
ARIMA(0,2,0) with constant
Autocorrelations
1
0.6
0.2
-0.2
-0.6
-1
0
5
10
15
20
25
lag


Residual Partial Autocorrelations for adjusted units
ARIMA(0,2,0) with constant
Partial Autocorrelations
1
0.6
0.2
-0.2
-0.6
-1
0
5
10
15
20
25
lag

Model selection

We can also choose p, q, d by comparing the models' "score", which measures how likely it is that the observed data is generated from ARIMA(p, q, d)

Model selection

We can also choose p, q, d by comparing the models' "score", which measures how likely it is that the observed data is generated from ARIMA(p, q, d)

Our "score" of ARIMA model is based on the **likelihood**

$$\text{Likelihood} = L = \mathbb{P}(\text{data}|\text{model})$$

Model selection

We can also choose p, q, d by comparing the models' "score", which measures how likely it is that the observed data is generated from ARIMA(p, q, d)

Our "score" of ARIMA model is based on the **likelihood**

$$\text{Likelihood} = L = \mathbb{P}(\text{data}|\text{model})$$

The likelihood depends on the model. For time-series, the likelihood can be very complicated

Model selection

Choose p and q, find parameters and compute $L = L(\text{data}|\text{model})$.

There are three “scores” that we can use

- Akaike's Information Criterion (AIC)

$$\text{AIC} = -2\log(L) + 2(p + q + k + 1),$$

where $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$.

Model selection

Choose p and q, find parameters and compute $L = L(\text{data}|\text{model})$.

There are three “scores” that we can use

- Akaike’s Information Criterion (AIC)

  $$\text{AIC} = -2\log(L) + 2(p+q+k+1),$$

  where $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$.
- corrected AIC

  $$\text{AICc} = \text{AIC} + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2},$$

  where T is the number of observations in the data.

Model selection

Choose p and q, find parameters and compute $L = L(\text{data}|\text{model})$.

There are three "scores" that we can use

- Akaike's Information Criterion (AIC)

  $$\text{AIC} = -2\log(L) + 2(p+q+k+1),$$

  where $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$.
- corrected AIC

  $$\text{AICc} = \text{AIC} + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2},$$

  where T is the number of observations in the data.
- Bayesian Information Criterion

  $$\text{BIC} = \text{AIC} + [\log(T) - 2](p+q+k+1).$$

Model selection

These scores follow the same concepts.

- Better model has higher likelihood.
- But model with too many parameters (high $p + q$) tends to overfit and should be penalized.

We prefer AICc for ARIMA. The lower the score, the better.

Also check out seasonal ARIMA (this week's lab).