MONASH University

MONASH BUSINESS SCHOOL

ETF3231/5231
Business forecasting

Week 7: ARIMA models

https://bf.numbat.space/

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS AMBA ACCREDITED

Monash University CRICOS Provider Number: 00008C

Outline

ARIMA models

AR: autoregressive (lagged observations as inputs)

I: integrated (differencing to make series stationary)

MA: moving average (lagged errors as inputs)

ARIMA models

AR: autoregressive (lagged observations as inputs)
I: integrated (differencing to make series stationary)
MA: moving average (lagged errors as inputs)

An ARIMA model is rarely interpretable in terms of visible data structures like trend and seasonality. But it can capture a huge range of time series patterns.

ARIMA models

AR: autoregressive (lagged observations as inputs)
I: integrated (differencing to make series stationary)
MA: moving average (lagged errors as inputs)

An ARIMA model is rarely interpretable in terms of visible data structures like trend and seasonality. But it can capture a huge range of time series patterns.

Make data stationary (variance & mean), fit model, reverse, forecast.

Outline

Stationarity

Definition

If $\{y_t\}$ is a stationary time series, then for all s, the distribution of $(y_t, \ldots, y_{t+s})$ does not depend on t.

A stationary series is:

- roughly horizontal
- constant variance
- no patterns predictable in the long-term

Stationarity

Definition

If $\{y_t\}$ is a stationary time series, then for all s, the distribution of $(y_t, \dots, y_{t+s})$ does not depend on t.

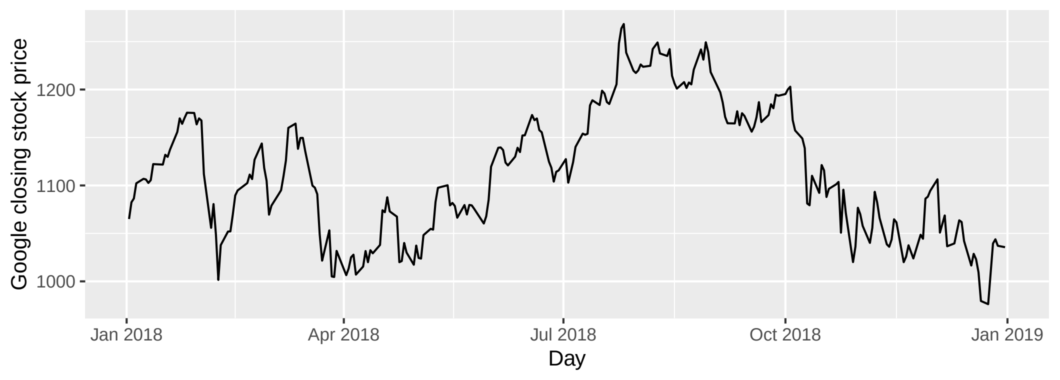
A stationary series is:

- roughly horizontal
- constant variance
- no patterns predictable in the long-term

- Transformations help to stabilize the variance.
- For ARIMA modelling, we also need to stabilize the mean.

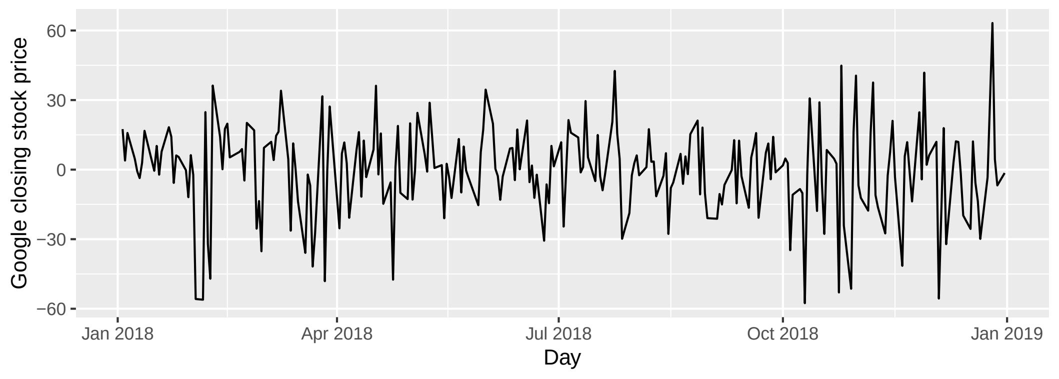
Stationary?

```r
gafa_stock |>
  filter(Symbol == "GOOG", year(Date) == 2018) |>
  autoplot(Close) +
  labs(y = "Google closing stock price", x = "Day")
```



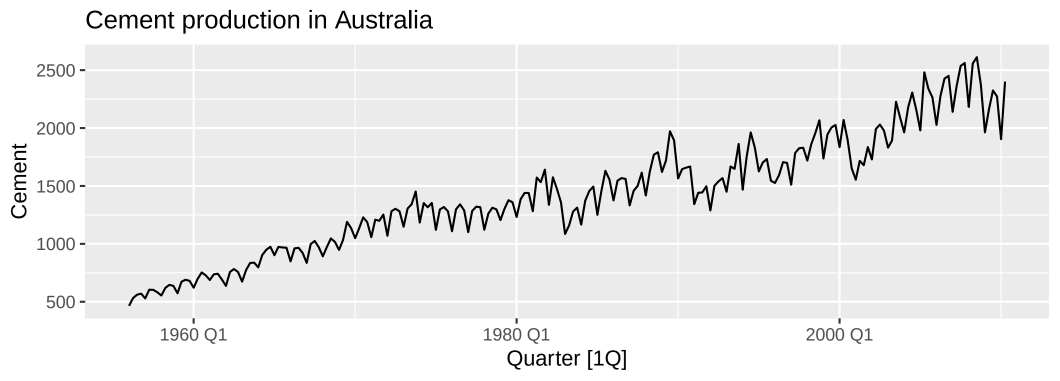
Stationary?

```r
gafa_stock |>
  filter(Symbol == "GOOG", year(Date) == 2018) |>
  autoplot(difference(Close)) +
  labs(y = "Google closing stock price", x = "Day")
```



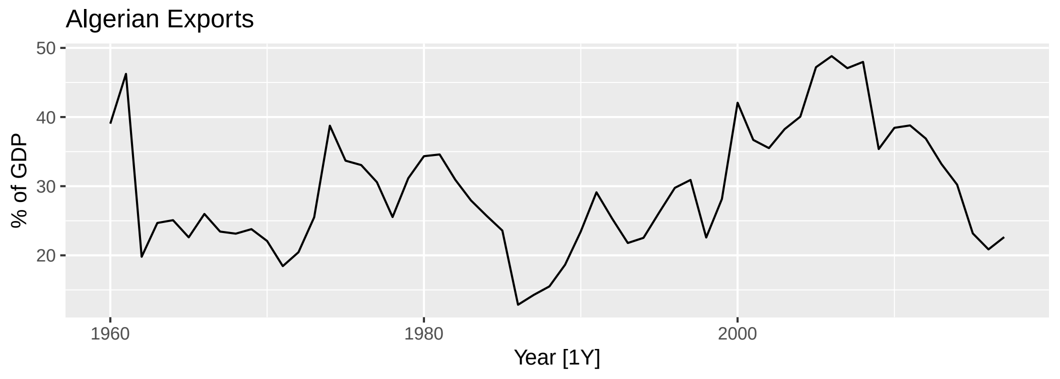
Stationary?

```r
aus_production |>
  autoplot(Cement) +
  labs(title = "Cement production in Australia")
```



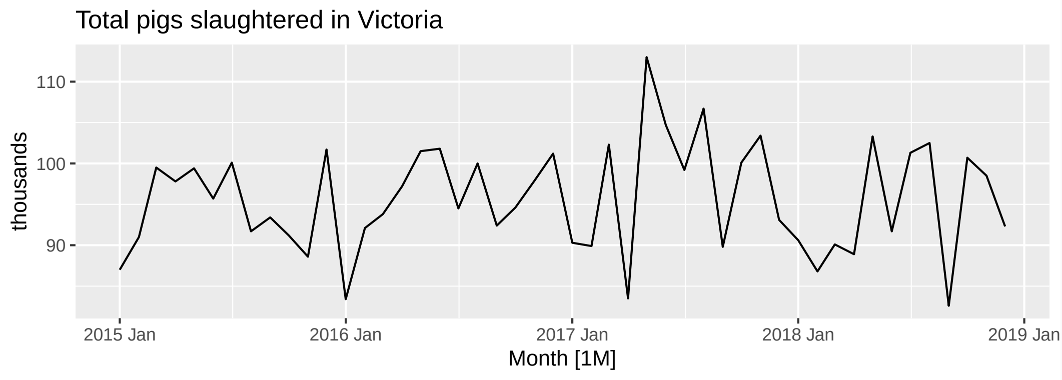
Stationary?

```
global_economy |>
  filter(Country == "Algeria") |>
  autoplot(Exports) +
  labs(y = "% of GDP", title = "Algerian Exports")
```



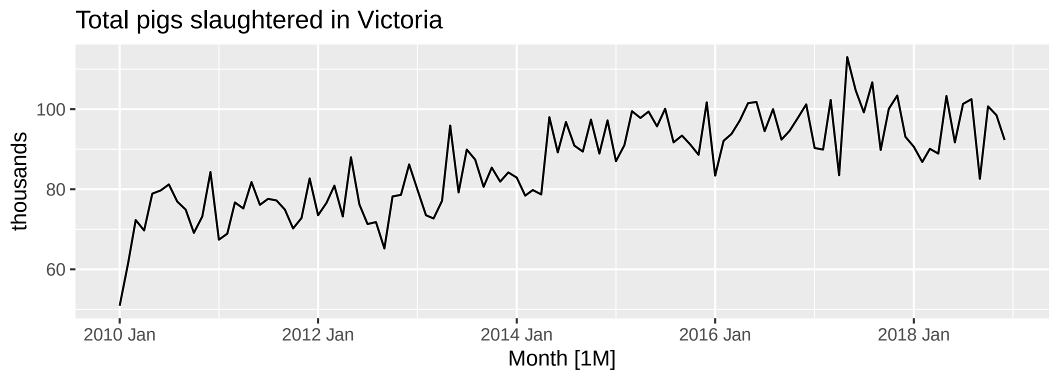
Stationary?

```r
aus_livestock |>
  filter(Animal == "Pigs", State == "Victoria", year(Month) >= 2015) |>
  autoplot(Count/1e3) +
  labs(y = "thousands", title = "Total pigs slaughtered in Victoria")
```



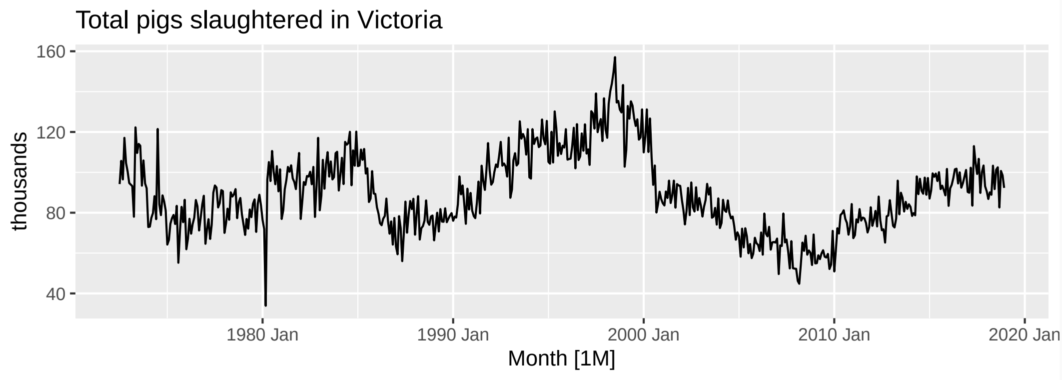
Stationary?

```
aus_livestock |>
  filter(Animal == "Pigs", State == "Victoria", year(Month) >= 2010) |>
  autoplot(Count/1e3) +
  labs(y = "thousands", title = "Total pigs slaughtered in Victoria")
```



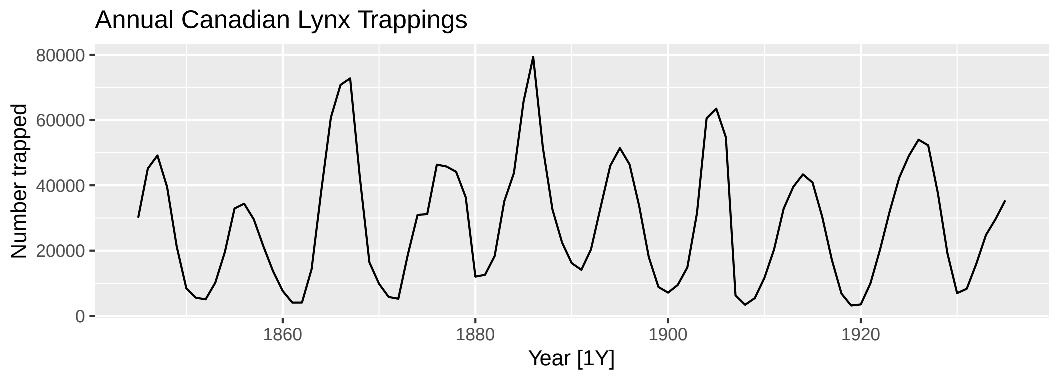
Stationary?

```
aus_livestock |>
  filter(Animal == "Pigs", State == "Victoria") |>
  autoplot(Count/1e3) +
  labs(y = "thousands", title = "Total pigs slaughtered in Victoria")
```



Stationary?

```r
pelt |>
  autoplot(Lynx) +
  labs(y = "Number trapped",
       title = "Annual Canadian Lynx Trappings")
```

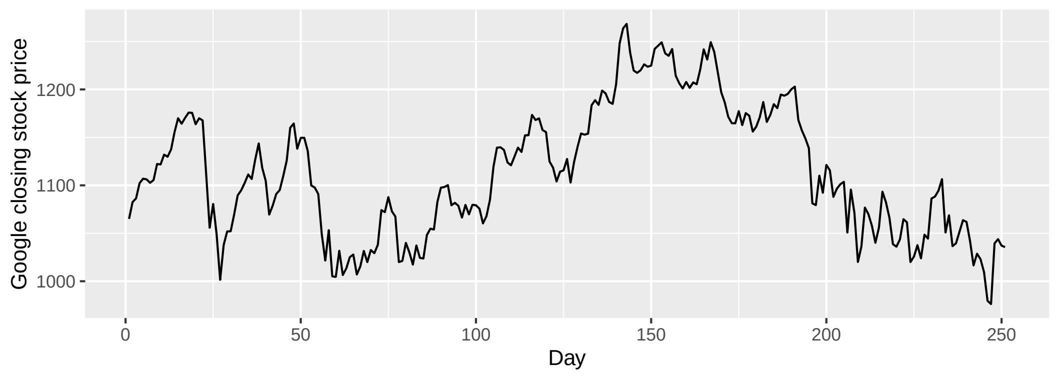


Example: Google stock price

```
google_2018 <- gafa_stock |>
  filter(Symbol == "GOOG", year(Date) == 2018) |>
  mutate(trading_day = row_number()) |>
  update_tsibble(index = trading_day, regular = TRUE)
```

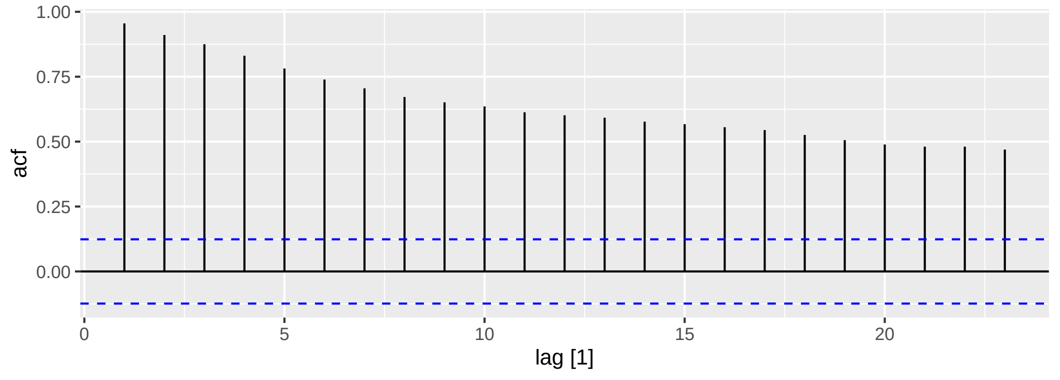
Example: Google stock price

```
google_2018 |>
  autoplot(Close) + labs(y = "Google closing stock price", x = "Day")
```



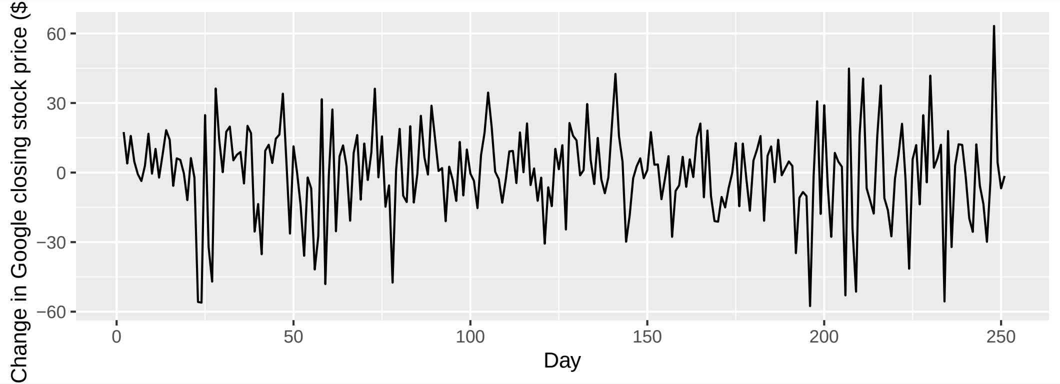
Example: Google stock price

```r
google_2018 |>
  ACF(Close) |> autoplot()
```



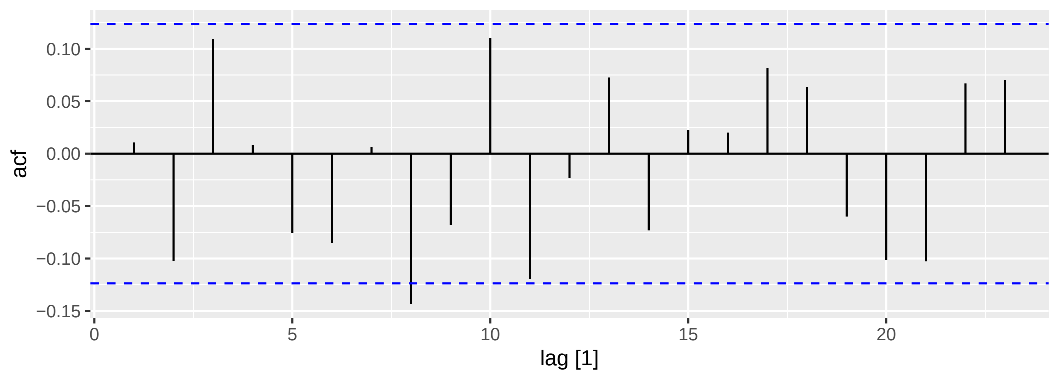
Example: Google stock price

```
google_2018 |>
  autoplot(difference(Close)) +
  labs(y ="Change in Google closing stock price ($USD)", x="Day")
```



Example: Google stock price

```
google_2018 |> ACF(difference(Close)) |> autoplot()
```



Differencing

- Differencing helps to stabilize the mean.
- The differenced series is the *change* between each observation in the original series: $y'_t = y_t - y_{t-1}$.
- The differenced series will have only $T - 1$ values since it is not possible to calculate a difference y'_1 for the first observation.

Example: Google stock price

- The differences are the day-to-day changes.
- Now the series looks just like a white noise series:
  - No autocorrelations outside the 95% limits.
  - Large Ljung-Box p-value.
- Conclusion: The daily change in the Google stock price is essentially a random amount uncorrelated with previous days.

Random walk model

- Graph of differenced data suggests the following model:

$$y_t - y_{t-1} = \varepsilon_t \qquad \text{or} \qquad y_t = y_{t-1} + \varepsilon_t$$

where $\varepsilon_t \sim NID(0, \sigma^2)$.

- Very widely used for non-stationary data.
- This is the model behind the naïve method.
- Random walks typically have:
  - long periods of apparent trends up or down.
  - Sudden/unpredictable changes in direction - stochastic trend.
- Forecast are equal to the last observation (Naive)
  - future movements are unpredictable - movements up or down are equally likely.

Random walk with drift model

- If the differenced series has a non-zero mean then:

$$y_t - y_{t-1} = c + \varepsilon_t \qquad \text{or} \qquad y_t = c + y_{t-1} + \varepsilon_t$$

where $\varepsilon_t \sim NID(0, \sigma^2)$.

- c is the non-zero average change between consecutive observations.
- If $c > 0$, y_t will tend to drift upwards and vice versa.
  - Stochastic and deterministic trend.
- This is the model behind the drift method.

Further differencing

- Occasionally you need to difference non-seasonal data twice.

Further differencing

- Occasionally you need to difference non-seasonal data twice.
- We seasonally difference seasonal data.

$$y'_t = y_t - y_{t-m}$$

  where m = number of seasons.
  - For monthly data $m = 12$, for quarterly data $m = 4$.
  - Seasonally differenced series will have $T - m$ obs.

Seasonal random walk

If seasonally differenced data is white noise it implies:

$$y_t - y_{t-m} = \varepsilon_t \qquad \text{or} \qquad y_t = y_{t-m} + \varepsilon_t$$

- The model behind the seasonal naïve method.

Seasonal differencing

Common to take both seasonal and first differences. When both are applied…

Seasonal differencing

Common to take both seasonal and first differences. When both are applied…

- it makes no difference which is done first—the result will be the same.
- If seasonality is strong, we recommend that seasonal differencing be done first because sometimes the resulting series will be stationary and there will be no need for further first difference.

Unit root tests

Statistical tests to determine the required order of differencing.

1. Augmented Dickey Fuller test: null hypothesis is that the data are non-stationary and non-seasonal.
2. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test: null hypothesis is that the data are stationary and non-seasonal.
3. Other tests available for seasonal data.

Unit root tests

Statistical tests to determine the required order of differencing.

1. Augmented Dickey Fuller test: null hypothesis is that the data are non-stationary and non-seasonal. H_0: non-stationary
2. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test: null hypothesis is that the data are stationary and non-seasonal. H_0: stationary
3. Other tests available for seasonal data.

Seasonal differencing

STL decomposition: $y_t = T_t + S_t + R_t$

Seasonal strength $F_s = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t+R_t)}\right)$

If $F_s > 0.64$, do one seasonal difference.

Outline

1. Stationarity and differencing
2. Backshift notation

Backshift notation

- First-order difference is denoted as $(1 - B)y_t$;
- Second-order difference is denoted as $(1 - B)^2 y_t$;
- Second-order difference is not the same as a second difference, which would be denoted $(1 - B^2)y_t$;
- In general, a dth-order difference can be written as $(1 - B)^d y_t$

Backshift notation

- First-order difference is denoted as $(1 - B)y_t$;
- Second-order difference is denoted as $(1 - B)^2y_t$;
- Second-order difference is not the same as a second difference, which would be denoted $(1 - B^2)y_t$;
- In general, a dth-order difference can be written as $(1 - B)^d y_t$
- A seasonal difference is denoted as $(1 - B^m)y_t$;
- A seasonal difference followed by a first difference can be written as

$$(1 - B^m)(1 - B)y_t$$