MONASH University

MONASH BUSINESS SCHOOL

ETC3550/ETC5550 Applied forecasting

Week 8: ARIMA models

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.

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.

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.

Transformations help to **stabilize the variance**.

For ARIMA modelling, we also need to **stabilize the mean**.

Non-stationarity in the mean

Identifying non-stationary series

- time plot.
- The ACF of stationary data drops to zero relatively quickly
- The ACF of non-stationary data decreases slowly.
- For non-stationary data, the value of r_1 is often large and positive.

Differencing

Differencing

- Differencing helps to **stabilize the mean**.
- First differencing: *change* between consecutive observations: $y'_t = y_t - y_{t-1}$.
- Seasonal differencing: *change* between years: $y'_t = y_t - y_{t-m}$.
- Sometimes two differences need to be applied (but never more).

Automatic differencing

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.

Seasonal strength

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.

Automatic differencing

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

Seasonal strength

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.

R commands

- Lag 1 difference: `difference(y)`
- Seasonal difference: `difference(y, lag = 4)`
- KPSS test: `unitroot_kpss(y)`
- Seasonal strength: `feat_stl(y, .period = 4)`
- Automatic first differencing: `unitroot_ndiffs(y)`
- Automatic seasonal differencing: `unitroot_nsdiffs(y, .period = 4)`

Relationship to random walks

A random walk is the process:

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

where ε_t is a white noise variable.

So if data did come from such a process, differencing would give white noise:

$$y_t - y_{t-1} = \varepsilon_t$$

Relationship to random walks

A seasonal random walk is the process

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

where ε_t is a white noise variable.

So if data did come from such a process, seasonal differencing would give white noise:

$$y_t - y_{t-m} = \varepsilon_t$$

Relationship to random walk with drift

A random walk with drift is the process:

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

where ε_t is a white noise variable.

So if data did come from such a process, differencing would give white noise with non-zero mean

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

- c is the **average change** between consecutive

Backshift operator notation

- B shifts the data back one period. $By_t = y_{t-1}$
- B^2 shifts the data back two periods: $B(By_t) = B^2 y_t = y_{t-2}$
- A difference can be written as $(1 - B)y_t$
- A dth-order difference can be written as $(1 - B)^d y_t$
- A seasonal difference followed by a first difference can be written as $(1 - B)(1 - B^m)y_t$