ARIMA

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- So far, we have only modelled stationary processes.
- What happens when we have *integrated* or non-stationary time series?

$$Y_t = Y_{t-1} + e_t$$

- We can employ ARIMA models.

Recall: Augmented Dickey-Fuller (ADF) Test

- The null hypothesis of the ADF test is that there is a unit root (not stationary).
- **A rejection of the null hypothesis suggests a stationary process.**
- In R we can use the `adf.test(data)` function from the *tseries* package.

Recall: Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

- The null hypothesis of the KPSS test is that the process is stationary.
- **A failure to reject the null hypothesis suggests a stationary process.**
- In R we can use the `kpss.test(data)` function from the *tseries* package.

Differencing

- Differencing is a method that can be used to create a stationary time series.
- $\nabla Y_t = Y_t - Y_{t-1}$ is the first difference.
- *You may difference multiple times, but the idea is to create a stationary series that is not white noise.*

Autoregressive Integrated Moving Average (ARIMA)

Autoregressive Integrated Moving Average (ARIMA)

- A time series is said to be an **autoregressive integrated moving average** if the differenced series follows an ARMA(p,q) series.
- ARIMA(p,1,q):

$$\begin{aligned} Y_t - Y_{t-1} =& \phi_1(Y_{t-1} - Y_{t-2}) + ... + \phi_p(Y_{t-p} - Y_{t-p-1}) \\ &+ \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} + e_t \end{aligned} \quad (1)$$

- *You may need to difference multiple times.*

Example 1

1. Import the *Germany_Rail.csv* file into R and run the code to convert the object into a time series.

2. Examine the series using the methods that we have covered in this course (no models yet).

3. What did you find?

ARIMA in R

- You can examine the differenced series for ARMA characteristics to determine the p and q values.

- In R:

  `acf(diff(series))` *MA(q)*
  `pacf(diff(series))` *AR(p)*

- We can then use: `model <- arima(data, order = c(p,d,q))` in R (*TSA* package).
- And the `coeftest(model)` function from the *lmtest* package to obtain *p*-values.

- *Do not forget to add the d value if differencing is needed.*

Example 2

- Using the time series from Example 1, do the following:
  1. Examine the ACF and PACF plots of the differenced series.
  2. Estimate an ARIMA model based on your findings.
  3. How did you do?

Recall: `auto.arima()`

- Because we are using MLE to estimate the models, AIC and BIC may be used to pick the *best* model.
- The `auto.arima(data)` function from the forecast package can be used.
- *The function may take many additional arguments.*
- We can still use the `coeftest()` function to test for significance.

Example 3

1. Use the `auto.arima()` function to find a model for the data used in Example 1.

2. How does it compare to your results from Example 2?

Seasonal Autoregressive Integrated Moving Average (SARIMA)

Seasonality I

- Seasonality occurs is many areas where time series are used.
- **Seasonality:** the observed value depends on the season or time of the year.
- *You may use linear regression approaches to deal with seasonality, but it is often too rigid.*
- We can build seasonality naturally into our SARIMA models.

Seasonality II

- Seasonality is generally defined by s, the known seasonal period.
- Examples:
  - $s = 12$ Monthly series
  - $s = 4$ Quarterly series
- Essentially, the correlation structure depends on lags at s in the series.
- *Seasonality may be detected from the ACF and PACF plots.*
- We can try to difference out the seasonality:
  `diff(diff(data),lag=`s`)`

SARIMA(p,d,q)x(P,D,Q)$_s$

- A SARIMA(p,d,q)x(P,D,Q)$_s$ model contains the following components:
    - p: Auto Regressive (AR) order
    - d: The number of ordinary differences
    - q: Moving Average (MA) order
    - P: Seasonal Auto Regressive (AR) order
    - D: Number of seasonal differences
    - Q: Seasonal Moving Average (MA) order
    - s: The seasonal period of the time series

- Values may be obtained from the ACF and PACF plots.

Example 4

1. Examine the seasonally differenced ACF and PACF plots for the data used in Example 1.

2. Do these plots give you any modelling ideas?

SARIMA(p,d,q)x(P,D,Q)$_s$ in R

- Once you have determined the correlation structure, you can use R to estimate a model.
- In R:

```
model <- arima(data,order=c(p,d,q),
seasonal=list(order=c(P,D,Q), period=s))
```

- Use `coeftest(model)` for significance

Example 5

1. Use the information that we have gathered in the first four examples to estimate some SARIMA models.
2. How did we do?

Diagnostics

Recall: White Noise Processes

- A **white noise** process is a random variable indexed in time that has a constant expected value and variance.

- Properties:
  - $E(e_t) = \mu_e$
  - $Var(e_t) = \sigma_e^2$
  - $\rho_{t,s} = 0 \quad \text{if } t \neq s$

- *Similar to the error term we find in regression.*

Residuals

- One way to quantify the relationship between each point and the estimated time series equation is to measure the vertical distance between them.

- The **residuals** (observed errors) are defined as follows:

$$e_t = Y_t - \hat{Y}_t. \tag{2}$$

- *As they residuals relate to the error term, we will use them for model diagnostics.*

- **It is assumed that the error term follows a white noise process where $E[e_t] = 0$.**

Independence

- We can use the ACF and PACF plots to assess the independence of the residuals.

- In R:
  - `acf(residuals(model))`
  - `pacf(residuals(model))`
    Or
  - `acf(rstandard(model))`
  - `pacf(rstandard(model))`

- Ljung-Box Test in R:
  - `Box.test(rstandard(model), lag =` l`, type = "Ljung-Box")`
  - *The null hypothesis is that there is no autocorrelation at the specified number of lags* l.

Example 6

1. Plot the residuals to see if they fluctuate around 0.
2. Do any of your models from Example 5 pass the independence assumption?

Desirable Characteristics

1. Normally distributed residuals
   - We can use the Q-Q plot and the `shapiro.test(rstandard(model))` to assess.
2. Constant Variance
   - Usually visually assessed in the plot of the residuals.
   - Obvious departures may indicate a transformation on the series is required.

Example 7

1. Do any of your models from Example 5 have the desired characteristics?

Exercise 1

- You are now fully equipped to model most time series.
- Use what you have learned to estimate some SARIMA models.
- Do they pass the assumptions?

References & Resources

1. Shumway, R. H., & Stoffer, D. S. (2011). *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Link

2. Jonathan, D. C., & Kung-Sik, C. (2008). *Time series analysis with applications in R*.

- arima.sim()
- TSA
- zoo
- arima()
- auto.arima()
- adf.test()
- kpss.test()
- SARIMA Example