ECON 4115/5115 Outline of Lecture Notes

Chapter 9. ARIMA Models

- Stationarity and differencing
  - <u>Stationary Time Series</u>: Properties do not depend on time which they are observed
    - Trends and seasonality make a series non-stationary
  - Figure 9.1: Which ones are stationary?
  - Differencing can eliminate trends and seasonality; logarithms can stabilize variances
  - ACF can help identify stationarity (see Figure 9.2)
  - Random walk model can be used to represent non-stationary data: $y_t = y_{t-1} + \varepsilon_t$
  - Random walk + drift: $y_t = c + y_{t-1} + \varepsilon_t$
  - First differences of random walk are stationary: $y_t - y_{t-1} = c + \varepsilon_t$
  - Occasionally you need to take second differences (i.e., differences of differences)
  - Seasonal differencing: $y_t - y_{t-m}$
  - You may need to take both first differences and seasonal differences
  - Unit root tests: statistical test for stationarity
  - KPSS test: H_0: data are stationary vs. H_A: data are non-stationary
  - Small *p* values indicate differencing is necessary
  - KPSS test can be computed using *urca* package in R
    - *unitroot_kpss*() option will execute the KPSS test
    - *unit_ndiffs*() will identify the number of first differences necessary
    - *unitroot_nsdiffs*() will identify the number of necessary seasonal differences

- Backshift notation
  - The backshift (or lag) operator is: $B^j y_t = y_{t-j}$
  - The first difference operator is: $(1-B)y_t = y_t - y_{t-1}$
    - The term "unit root" comes from solving the equation: $(1-z)=0$, where z replaces B.
  - The seasonal difference operator is: $(1-B^m)y_t = y_t - y_{t-m}$
  - Difference operators can be mixed: $(1-B)(1-B^m)y_t = (1-B-B^m+B^{m+1})y_t$
- Autoregressive (AR) models
  - An AR model regresses y_t on its own past values: $y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t$
  - This is referred to as an AR(*p*) model, an autoregressive model of order p
  - AR(*p*) models are capable of fitting a wide-range of time series patterns
  - If $p=1$ and $\phi_1 = 1$, then we have a random walk with drift (i.e., non-stationary series)
  - Usually, AR(*p*) models are restricted to be stationary
  - If $p \geq 2$, then it is possible to generate cycles if the roots are complex
  - See Excel spreadsheet for examples of AR(1) and AR(2) processes
- Moving average (MA) models
  - An MA model regresses y_t on current and past ε_t: $y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$
  - This is referred to as an MA(*q*) model, a moving-average model of order q
  - MA(*q*) models are different than earlier MA-smoothing methods for fitting a trend
  - Stationary AR(*p*) models can be expressed as an MA(∞) model
    - For example, an AR(1) model, through repeated substitutions, is: $y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \cdots$

- Invertible MA(q) models can be expressed as AR(∞) model
  - For example, an MA(1) model can be expressed as: $y_t = \theta_1 y_{t-1} - \theta_1^2 y_{t-2} + \theta_1^3 y_{t-3} - \theta_1^4 y_{t-4} + \cdots + \varepsilon_t$
- See Excel spreadsheet for examples of MA (1) and MA(2) processes

- Non-seasonal ARIMA models
  - A non-seasonal ARIMA(p, d, q) model is:

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

    where y'_t indicates the differenced version of y_t.

  - Using the backshift operator, the ARIMA(p, d, q) model can be written as:

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B^d) y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\, \varepsilon_t$$

  - The *ARIMA*() function in R will automatically choose the best model.
  - It's hard to determine p and q from a time series plot.
  - However, you can use the ACF and PACF to determine p and q.
    - The ACF measures $corr(y_t, y_{t-k})$ for $k = 1, \ldots, K$.
    - A statistically significant $corr(y_t, y_{t-k})$ could be caused by an AR(1), AR(2), … , or AR(k) process. So the ACF alone cannot determine p.
    - The partial autocorrelation function (PACF) solves this by removing the impact of $y_{t-1}, \ldots, y_{t-(k-1)}$ and then calculating the correlation between y_t and y_{t-k}.
    - The differenced data may be generated by an ARMA(p,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.
    - The differenced data may be generated by an ARMA(0,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.

- Estimation and order selection
    - The parameters (ϕ's and θ's) are estimated with maximum likelihood methods.
    - The order of the ARMA(p,q) model can also be found with the AIC, AICc and BIC.
- ARIMA modelling in R
    - The *ARIMA*() function in R will choose the model using the H-K algorithm.
    - In general, here are the steps you should follow in ARIMA modelling:
        - Plot the data and identify any unusual observations.
        - If necessary, transform the data (using a Box-Cox transformation) to stabilize the variance.
        - If the data are non-stationary, take first differences until the data are stationary.
        - Examine the ACF/PACF: Is an ARIMA($p, d, 0$) or ARIMA($0, d, q$) model appropriate?
        - Try your chosen model(s), and use the AICc to search for a better model.
        - Check the residuals from your chosen model by plotting the ACF and do a test of the residuals. If they are not white noise, try a modified model.
        - Once the residuals look like white noise, calculate forecasts.
- Forecasting
    - Calculating point forecasts is done automatically with the *forecast*() function in R.
    - If you need to do it manually, you follow these three steps:
        - Step #1. Move all terms other than y_t to the right side of the equation.
        - Step #2. Rewrite the equation replacing t with $T + h$.
        - Step #3. Replace future observations with forecasts, future errors with zero, and past errors with their corresponding residuals.
    - Calculating the prediction intervals is done automatically with the *forecast*() function.
    - If you need to do it manually, you first re-write as an MA(∞) process and then calculate the variance.

- The prediction intervals from the *forecast*() function are overly optimistic because they ignore the uncertainty with estimating the parameters and choosing the ARIMA order.

- Seasonal ARIMA models

  - The seasonal ARIMA model is written as: ARIMA$(p, d, q)(P, D, Q)_m$

  - Using the backshift operator, an ARIMA$(1,1,1)(1,1,1)_4$ model can be written as:

$$(1 - \phi_1 B)(1 - \Phi_1 B^4)(1 - B)(1 - B^4)y_t = (1 + \theta_1 B)(1 + \Theta_1 B^4)\varepsilon_t$$

  - You can use the ACF & PACF to manually select the model or use the *ARIMA*() function.

  - See R script #15 for an example of a seasonal ARIMA model.

- ARIMA vs. ETS

  - See R script #15 for an example of comparing ARIMA vs. ETS models.