Time series analysis

Lecture 2: Forecasting

Alaa Bakhti

Forecasting

Predict the future values of the time series based on previously observed values

Prediction intervals

- Indicates the uncertainty of the forecasts
- The further we forecast in the future, the wider the uncertainty tends to be

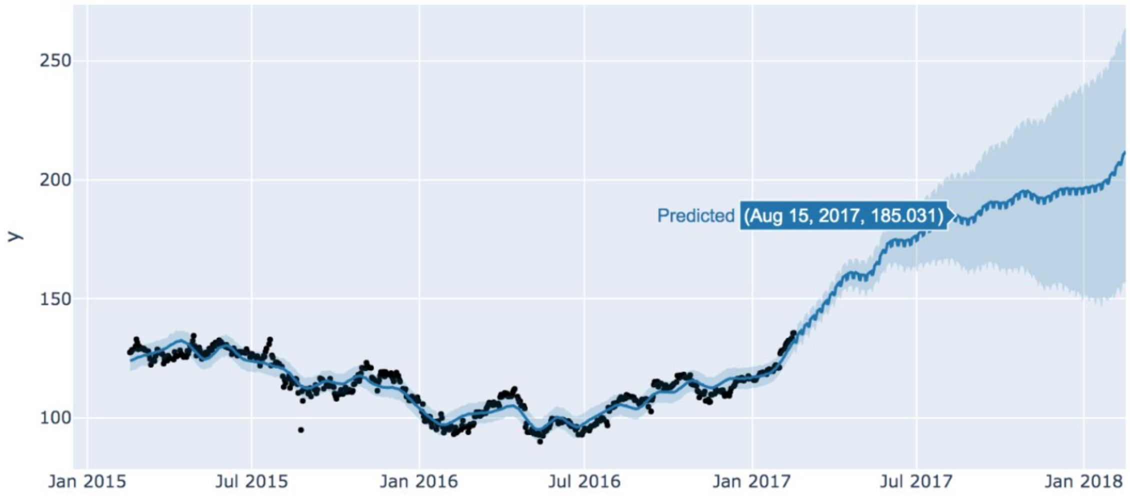


Image source

AR, MA, ARMA, ARIMA, etc models

Autoregressive (AR) model

- Auto regression predicts the values of future time periods as a function of values at previous time periods.
- Regression of a time series against its lagged values
- The last *p* observations are used as predictors in the regression equation

AR(1) model: $$y_t = c + \phi_1 y_{t-1} + e_t$$

AR(2) model: $$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$$

AR(p) model: $$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + e_t$$

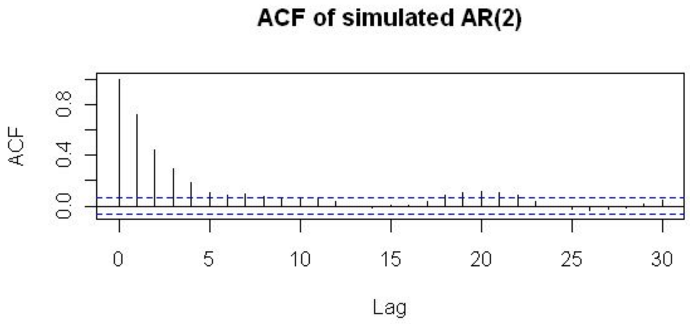
AR model

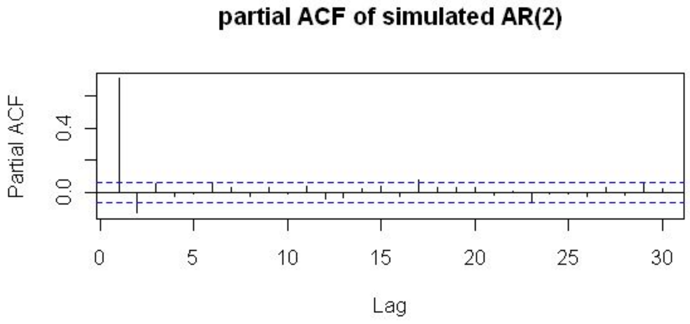
How to determine if the process is AR?

- The ACF tails off: have a gradual decreasing trend
- The PACF cuts off after *p* significant lags

How to determine the AR model order?

- Using the PACF plot
- The number of lags after which the PACF cuts off





Moving average (MA) model

- Use the previous time periods errors to make a better estimate of the current time period
- The observations are always centered on the average c

MA(1) model:

$$y_t = c + \theta_1 e_{t-1} + e_t$$

MA(2) model:

$$y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$$

MA(q) model:

$$y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots + \theta_q e_{t-q} + e_t$$

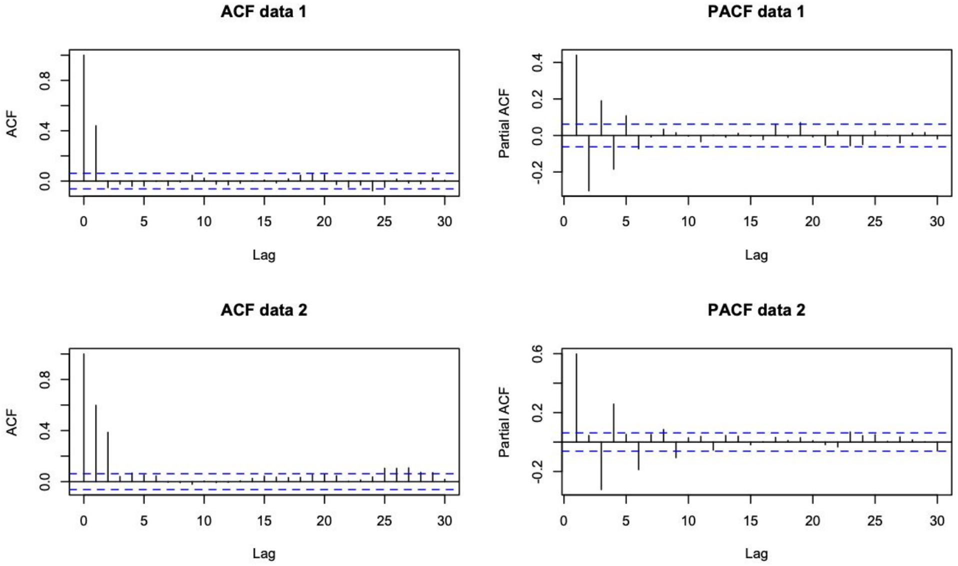
MA model

How to determine if the process is MA?

- The ACF cuts off after *q* lags
- The PACF tails off: have a gradual decreasing trend

How to determine the MA model order?

- Using the ACF plot
- The number of lags after which the ACF cuts off



Autoregressive moving-average (ARMA) model

- ARMA = AR + MA
- An auto-regressive moving-average models the value of a variable as a linear function of previous values and residual errors at previous time steps of a stationary time series
- Use the time series lagged values and errors as predictors

ARMA(1, 1):

$$y_t = c + \phi_1 y_{t-1} + \theta_1 e_{t-1} + e_t$$

ARMA(p, q):

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q} + e_t$$

Where

- *p: order of the* AR *model*
- *q*: order of the *MA* model

ARIMA model

- Autoregressive integrated moving average model
- Used when the time series have a linear trend
- The time series is differentiated d times to become stationary and then an ARMA model is applied

ARIMA(p, d, q): d = order of differencing

- The forecasted time series is the differentiated one time series and not the original one
- How to define the d parameter?
    - Differentiate the time series multiple times to make stationary
    - Use the number of differentiation for the d parameter

Training and evaluation of time series

Training and test sets

Training set

- composed of the observations that will be used to build the forecasts

Test set

- composed of the observations that will be used to evaluate the forecasting model
- The observations in this set should be future observations compared to the training set

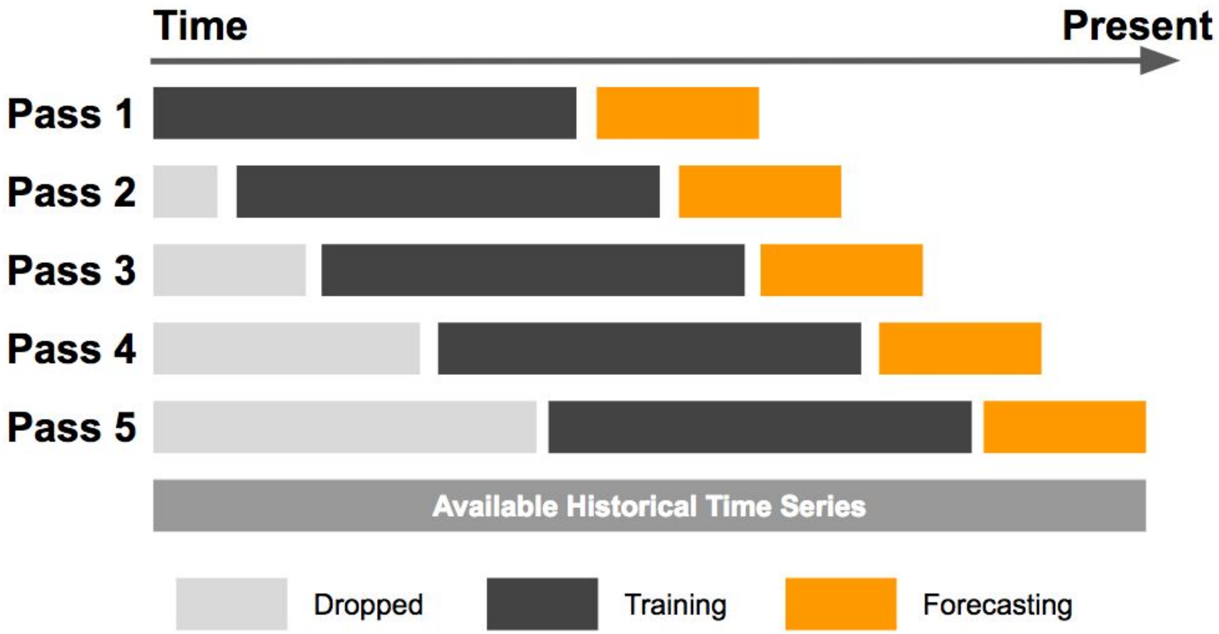
⚠️

- Observations from the future should not be used to build the forecasting model, otherwise you will have a Data Leakage (introduction of future value that cannot be available at prediction time)

Residuals and forecast errors

- **Fitted value**
  - the forecast of an observation using all previous observations (one-step forecast)
- **Residual**
  - Is the difference between an observation and its fitted value in the training set (what is left and the model did not predict)
  - Plotting the residuals is helpful to find if more modeling power is required or not, if the residuals do not look like white noise, that means that there may be some informations in the residuals that were not captured by the model
  - 💡 Always check the residuals after building the forecast model
- **Forecast errors**
  - The difference between the observed value and its forecast in the test set
  - Based on multi-step forecasts

Cross validation methods: sliding window



For more informations: https://eng.uber.com/omphalos/

Cross validation methods: expanding window

For more informations: https://eng.uber.com/omphalos/

Practical work

Practical work

In this practical work, we will build some forecasting models using the statsmodel package. We continue using the same air passengers dataset.

1. Load the last practical work data (stationary time series)
2. Train an AR model
    a. Determine if the process is AR or not using the ACF and PACF plots
    b. Choose an AR model order and train the model
        i. Split the data on training and test sets (test period = [1958-01-01, 1960-12-01])
        ii. Build a AR forecasting model using the statsmodel ARIMA class
        iii. Print and interpret the model results summary
        iv. Plot the true observations and the fitted values of the model using the fitted model *fittedvalues* attribute
        v. Compute and plot the residuals

Practical work

3. Evaluate your AR model
    a. Forecast the test data using the *predict* method with a start and end date with multi-step forecasting
    b. Compute the MAPE score
    c. Plot the test data and the forecasts
    d. Implement the expanding window cross validation method with one step and multi step forecasting. Each time plot the predictions against correct values and compute the MAPE score
        i. One step forecasting: window = 1 (1957-12-01 as a start date for the test set)
        ii. And multi step forecasting: window = 3
4. Repeat the previous steps for the MA and ARMA model
5. Benchmark the results for the 3 models using the MAPE score, AIC and BIC criteria and the fit-time

Practical work

6. Forecast with the ARMA model
    a. Forecast using the *get_forecast* method for a single step and multiple steps
    b. Transform and plot the forecasted back to the time series original form (before applying the transformations to make it stationary
    c. **Bonus**: Plot the prediction intervals for:
        i. Future forecasts (multiple steps)
        ii. The test data using multiple steps forecasting
        iii. The test data using one step forecasting

Ressources

- Forecasting: Principles and Practice - Rob J Hyndman and George Athanasopoulos (2nd ed)
- Machine Learning Why Good Forecasts Treat Human Input as Part of the Model