Auto Regressive Integrated Moving Average (ARIMA)

Auto-Regressive (AR) Models

- Auto-regression means regression of a variable on itself.
- One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process.
- When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR models

Concept of Stationarity

- A Time Series is considered to be stationary whose statistical properties such as the variance and (auto) correlation are all constant over time. The properties of a stationary time series do not depend on time. The (auto) correlation observations only depend on how far apart these observations are and not where they are.
- Note: Correlation = $\frac{Covariance(xt, xt_{+k})}{S_t * St_{+k}}$

Here, the standard deviation is constant and is represented by S. The x_t is the Time Series random variable.

The Importance of Stationarity

- Stationary Time Series allows us to essentially have “copies” of things which enables us to do build appropriate statistical models for forecasting.
- A Time Series with a pronounced trend is not going to have a similar mean everywhere across the observed the time frame. This means that the entire Time Series is not part of the *same sample* which is a need for a regression model.
- One more intuitive understanding of the importance of stationarity is that the coefficients of the AR model should not be biased because the Time Series has a pronounced trend or seasonality.

How to Check for Stationarity?

- To check whether the series is stationary, we use the Augmented Dickey Fuller (ADF) test whose null and alternate hypothesis can be simplified to
  - Null Hypothesis H_0 : Time Series is non-stationary
  - Alternate Hypothesis H_a : Time Series is stationary
- At our desired level of significance (chosen alpha value), we can test for stationarity using the ADF test.

How does the ADF test work?

- The ADF test works on the principle of finding the probability that a unit-root is present in the AR model.
- If an unit root is present in a Time Series, the Time Series shows a systematic pattern which is unpredictable thereby violating the idea of a stationary Time Series.
- At a very basic level, a process can be written as a series of monomials and each of these monomials corresponds to a root. If one of these roots is 1, then that can be said as a unit root. This is a very intuitive definition of an unit root.

How does the ADF test work? cont'd….

- The ARIMA model can be modelled as a autoregressive polynomial (of order 'p') which has 'd' roots on the unit circle.
- Correspondingly, we calculate a (modified) version of the t-statistic with appropriate degrees of freedom and compare it with the empirical values to conclude whether an unit root is present (and subsequently the Time Series can be said to be non-stationary).

How to make a non-stationary Time Series stationary?

- We can take appropriate levels of differencing to make a Time Series stationary.
- We can try various mathematical transformations to make the series stationary.
  - Apply transformation and/or differencing.
  - Check for stationarity.
  - If the time series is not stationary repeat the process of differencing.
  - Remember, complicated transformations might give us a stationary series very easily but after the forecast values are obtained we need to get back to the original series by tracing back the transformation steps.

- An AR(p) model (Auto-Regressive model of order p) can be written as:
  $y_t(\text{hat}) = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + e_t$
- Here, *e* is an error term which is an independent and identically distributed random variable (or in other words, a white noise) with the parameters mean = 0 and variance = σ^2.

How to choose the order of an AR(p) model? – Partial Autocorrelation

- We look at the Partial Auto-Correlations of a stationary Time Series to understand the order of a Auto-Regressive models.
- For an AR model, the PACF (Partial Auto Correlation Function) values cuts-off after a certain lag. The PACF values closes to 0 (at appropriate confidence intervals for the PACF plots) beyond that order (or lag).
- Partial auto-correlation of lag k, $\boldsymbol{\rho_{pk}}$, is the correlation between $\boldsymbol{Y_t}$ and $\boldsymbol{Y_{t-k}}$when the influence of all intermediate values ($\boldsymbol{Y_{t-1}}$,$\boldsymbol{Y_{t-2}}$,$\cdots$,$\boldsymbol{Y_{t-k+1}}$)is removed from both $\boldsymbol{Y_t}$ $\boldsymbol{and}$ $\boldsymbol{Y_{t-k}}$
- For building AR models, we look at the PACF plots and determine the order of the AR model.

Partial Autocorrelation cont'd…

- A more formal definition of the Partial Autocorrelation of order 2 is Covariance of $(x_t, x_{t-2} \mid x_{t-1})$/ square-root (Variance of $(x_t|x_{t-1})$ * Variance of $(x_{t-2}|x_{t-1})$)
- Since the series is stationary, the variance terms should be equal.
- Similarly for the Partial Autocorrelation of order p is the corresponding covariance values taking into account the intermediate values divided by the subsequent variance values.

Moving Average (MA) Models

- For a MA model, the error component is modelled.
- A MA(q) model (Moving Average model of order q) can be written as:
  y_t (hat) $= e_t + \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + \ldots + \alpha_q e_{t-q}$

How to choose the order of a MA(q) model? – Autocorrelation

- We look at the Auto-Correlations of a stationary Time Series to understand the order of a Moving-Average models.
- For a MA model, the ACF (Auto Correlation Function) values cut-off at a certain lag. The ACF values closes to 0 (at appropriate confidence intervals for the ACF plots) beyond that order (or lag).
- Auto-correlation of lag k, $\boldsymbol{\rho_{qk}}$, is the correlation between $\boldsymbol{Y_t}$ and $\boldsymbol{Y_{t-k}}$. This particular function does not depend on 't' since the Time Series is stationary.
- For building MA models, we look at the ACF plots and determine the order of the MA model.

Autocorrelation cont'd...

- The ACF(0) = 1 since this is the correlation of series with itself (without any lags).
- The PACF(1) = ACF (1) as for the PACF of order 1 we do not need to factor out the effect of any other lags in between.

ARIMA (p,d,q) Model

- An ARIMA model consists of the Auto-Regressive (AR) part and the Moving Average (MA) part after we have made the Time Series stationary by taking the correct degree/order of differencing.
- The AR order is selected by looking at where the PACF plot cuts-off (for appropriate confidence interval bands) and the MA order is selected by looking at where the ACF plots cuts-off (for appropriate confidence interval bands)
- The correct degree or order of difference gives us the value of ‘d’ while the ‘p’ value is for the order of the AR model and the ‘q’ value is for the order of the MA model.
- This is the Box-Jenkins methodology for building the ARIMA models.

ARIMA (p,d,q) Model cont'd...

- ARIMA models can be built keeping the Akaike Information Criterion (AIC) in mind as well. In this case, we choose the 'p' and 'q' values to determine the AR and MA orders respectively which gives us the lowest AIC value. Lower the AIC better is the model.
- Coding languages tries different orders of 'p' and 'q' to arrive to this conclusion. Remember, even for such a way of choosing the 'p' and 'q' values, we must make sure that the series is stationary.
- The formula for calculating the AIC is $2k - 2\ln(L)$, where k is the number of parameters to be estimated and L is the likelihood estimation

Seasonal ARIMA (p,d,q)(P,D,Q)F Model

- For a Seasonal Auto-Regressive Integrated Moving Average we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the 'F' parameter indicates the seasonality/seasonal effects over a particular period.
- We can follow the Box-Jenkins method over here as well to decide the 'p', 'q', 'P' and 'Q' values.
- For deciding the 'P' and 'Q' values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of 'F' and see where these cut-off (for appropriate confidence interval bands)

- For the SARIMA models, we can also estimate ‘p’, ‘q’ , ‘P’ and ‘Q’ by looking at the lowest AIC values.
- The seasonal parameter ‘F’ can be determined by looking at the ACF plots. The ACF plot is expected to show a spike at multiples of ‘F’ thereby indicating a presence of seasonality.
- Also, for Seasonal models, the ACF and the PACF plots are going to behave a bit different and they will not always continue to decay as the number of lags increase.

Conclusion

1. Split the whole data into a training and test sets with the most recent observations in the test set.
2. Check for stationarity of the training data and if the data is non-stationary take appropriate measures to make the data stationary.
3. Follow either the Box-Jenkins methodology of estimating the ‘p’ , ‘q’ , ‘P’ (if Time Series has a seasonality) and ‘Q’ (if Time Series has a seasonality) by looking at the PACF and the ACF plots or estimate these parameters by looking at the lowest Akaike Information Criterion.
4. Use the model built on the training data to forecast on the test data and calculate the necessary model evaluation parameters like the Root Mean Squared Error (RMSE) or the Mean Absolute Percentage Error (MAPE).
5. If you are satisfied with the model thus built, check for stationarity on the whole data and forecast for the desired future time points using this model.

Note: Complicated models by definition would work better on the training data but make sure to select the optimum model to predict on the test data.

Further Study: (S)ARIMAX and TVLM Models

- For (S)ARIMAX (X stands for exogenous variables) models, we can include Exogenous variables as well. These are variables which affect the target stationary Time Series. These variables can also be included in our models to aid in the decrease of forecast errors.
- There are Time Varying Linear Models as well in which the exogenous variables are also allowed to look at their past observations for forecasting into the future.

Reference Books

1. Time Series Analysis, Forecasting and Control by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung
2. Introduction to Time Series and Forecasting by Peter J. Brockwell, Richard A. Davis
3. Time Series Analysis by James D.Hamilton