Building ARIMA Models

The Methodology

- Given a time series X_t, t = 1, 2, …, n, the objective is to select the best fitted ARIMA(p, d, q) model.
- The methodology of building ARIMA models according to Box and Jenkins consists of three stages.
- Stage 1: Identification Stage or Model Selection.
  Autocorrelations and Partial Autocorrelations.
- Stage 2: Estimation Stage or Parameter Estimation.
  Estimate the unknown autoregressive and moving average parameters.
- Stage 3: Checking Stage or Model Checking.
  Check the estimated results.

Stage I: Identification

- The two most useful tools in any attempt at model identification are
  - the sample autocorrelation function and
  - the sample partial autocorrelation function.
- These two functions will provide valuable information with regard to stationarity and to the true generating process.

Sample Autocorrelations

- Given a time series X_t, t =1, 2, …, n, the sample autocorrelation function, which is the plot of the sample autocorrelations

$$r_k = \frac{\sum_{t=k+1}^{n}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

  against k, where k=1, 2, …, provides an obvious estimate of the autocorrelation function ρ_k of the underlying stochastic process.
- Hence the set of r_k's defines the sample autocorrelation function and the graphical presentation of these values defines the correlogram.
- If these estimates do not decay towards zero, the process is not stationary and therefore it needs to be examined in first differences.
- Furthermore, if the process is stationary these estimates will provide useful information with respect to AR or MA components.

Partial Autocorrelations

- The partial autocorrelation coefficient φ_{kk} (population) indicates the autocorrelation (partial or conditional) between X_t and X_{t-k} for given values X_{t-1}, X_{t-2}, …, X_{t-k+1}.
- To determine the value of φ_{kk} we need first to construct the two errors, i.e.,
  - $X_{t.t-1,t-2,..,t-k+1} = X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} - \ldots - \varphi_{k-1} X_{t-k+1}$
  - $X_{t-k.t-k+1,t-k+2,..,t-1} = X_{t-k} - \varphi_1 X_{t-k+1} - \varphi_2 X_{t-k+2} - \ldots - \varphi_{k-1} X_{t-1}$
- Based on these two values the partial autocorrelations are computed as regular autocorrelations.

Determination of PA

- The partial autocorrelation coefficient φ_{kk} are computed as follows:
  - $\varphi_{11} = \text{Corr}(X_t, X_{t-1}) = \text{Cov}(X_t, X_{t-1})/\text{Var}(X_t) = \rho_1$
  - $\varphi_{22} = \text{Corr}[X_t - \varphi_1 X_{t-1}, X_{t-2} - \varphi_1 X_{t-1}) = [\rho_2 - \rho_1^2]/[1 - \rho_1^2]$
  - $\varphi_{33} = \text{Corr}[X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2}, X_{t-3} - \varphi_1 X_{t-1} - \varphi_2 X_{t-2}) =$
  - and so on.
- Moreover, the partial autocorrelation function is related to regression analysis.
- If $X_t \sim$ AR(p), that means that only the first p partial autocorrelations exist.

PA for AR(p) processes

- Consider $X_t \sim$ AR(1), then φ_{kk} are computed as follows:
  - $\varphi_{11} = \rho_1 = \varphi_1$
  - $\varphi_{22} = [\rho_2 - \rho_1^2]/[1 - \rho_1^2] = 0$
  - $\varphi_{kk} = 0$ for all values of $k \geq 3$.
- Consider $X_t \sim$ AR(2), then φ_{kk} are computed as follows:
  - $\varphi_{11} = \rho_1 \neq 0$ [Recall: $\rho_1 = \varphi_1 /(1 - \varphi_2)$ & $\rho_2 = \varphi_1\rho_1 + \varphi_2$]
  - $\varphi_{22} = [\rho_2 - \rho_1^2]/[1 - \rho_1^2] = \varphi_2$
  - $\varphi_{kk} = 0$ for all values of $k \geq 3$.
- Similarly, if $X_t \sim$ AR(p), that means
  - $\varphi_{kk} \neq 0$ for all values of $k \leq p$.
  - $\varphi_{kk} = 0$ for all values of $k \geq p+1$.

PA for MA(q) processes

- If the Moving Average process is invertible that means that it can be written as an AutoRegressive process with infinite parameters.
- Hence, the partial autocorrelations of a MA(q) process will behave very close like the autocorrelations of an AR process.
- Consider $X_t \sim$ MA(1), then φ_{kk} are computed as follows:
  - $\varphi_{11} = \rho_1$ [Recall: $\rho_1 = -\theta/(1 + \theta^2)$]
  - $\varphi_{22} = [\rho_2 - \rho_1^2]/[1 - \rho_1^2] = [-\theta^2 /(1 + \theta^2 + \theta^4)] \neq 0$
  - $\varphi_{kk} \downarrow$ as $k \uparrow$.
- Similarly, if $X_t \sim$ MA(2).

Comments on PA

- The autocorrelations, ρ_k, of a pure autoregressive process decay towards zero with increasing lag length k. By contrast, for a pure moving average process of order q, the autocorrelations cut off abruptly meaning that they are all zero for k bigger than q.
- The partial autocorrelations, φ_{kk}, of a pure moving average process decay towards zero with increasing lag length k. by contrast, for a pure autoregressive process of order p, the partial autocorrelations cut off abruptly meaning that they are all zero for k bigger than p.
- The behavior of ARMA(p, q) models is in contrast to the properties we have already noted of pure AR and pure MA models. The lack of abrupt cuts in either the autocorrelations or in partial autocorrelations makes it more difficult in practice to distinguish among alternative mixed models.

Sample Partial Autocorrelations

- The Sample Partial Autocorrelation function is usually calculated by fitting autoregressive models of increasing order.
- The estimate of the last coefficient of each model is the sample partial autocorrelation coefficient, denoted as φ_{kk} hat.
- Also, estimates of the partial autocorrelations can be obtained through the sample autocorrelations (Yule-Walker equations), i.e.,
  - $k = 1 \rightarrow r_1 = \varphi_{11}^{\wedge}$
  - $k = 2 \rightarrow r_1 = \varphi_{21}^{\wedge} + \varphi_{22}^{\wedge} r_1$ & $r_2 = \varphi_{21}^{\wedge} r_1 + \varphi_{22}^{\wedge} \rightarrow \varphi_{22}^{\wedge} = \ldots$
- Hence, for the identification stage it is important to have the correlogram of the sample autocorrelations and partial ones.

Stage II: Parameter Estimation

- Assuming that a particular ARIMA(p, d, q) model is selected the next step is to estimate the unknown parameters p, d and q.
- If $d \neq 0$, it is necessary to difference the series d times so that to get stationarity.
- The new series $W_t = (1 - B)^d X_t$ will then be used to obtain estimates for the autoregressive, moving average and the mean of the series parameters.
- Frequently, after differencing, it is reasonable to assume that the new series has mean zero, in which case the parameter μ is dropped.
- The estimation procedure for time series is not unique.

Methods of Estimation

- There are three methods of estimation.

  I) Conditional Least Squares (CLS)
- Minimization of the sum of squares (SS) of the white noise error term.
- Computational algorithms for the numerical min SS are widely available.
- These algorithms are based on non-linear regression procedures that provide point estimates and standard errors for statistical inference.
- The parameter estimators have distributions that are close to normal.
- Note that different ARIMA models will use different sample size. Loss of observations depending on the ARIMA model.

Methods of Estimation II

II) Maximum Likelihood Estimation (MLE)

- Maximization of the log likelihood function using standard non-linear algorithm.
- Initial estimates for the parameters are obtained directly from the sample autocorrelations employed in the identification stage.
- Convergence criteria:
  - Change in the values of SSE; ΔSSE
  - Change in the value of parameter estimates: $\Delta\hat{\varphi}$
  - Number of iterations.
- Newbold (1974) and Ansley (1979) in *Biometrika* have provided algorithms.
- Note that all ARMA models will use the same number of observations.

Methods of Estimation III

III) Unconditional Least Squares (ULS)

- Nonlinear estimation and minimization of the sum of squares (SS) of the error term is carried out by numerical iterations.
- The error terms are viewed as "forecast" backward in time.
- For this reason, they frequently called backforecasts or backcasts, whereas the method is known as Backasting Method.

- Ansley and Newbold (1980) *J. of Econometrics*, have suggested to use MLE.

Comment on Parameter Estimation

- In Regression Analysis we will the same estimates and standard errors using any computer package that does OLS.
- This is not true in Time Series Analysis.
- The standard ARIMA computer packages will yield somewhat different parameter estimates using identical data set.
- Very often these differences will be of no great practical significance, in the sence that quit similar forecast will result.
- This is true only for simple models, not for complicated ones.
- See Newbold, Agiakloglou and Miller (1994), *J. of Forecasting*, 10, 573-581, "Adventures with ARIMA software".
- **Rule:** Before you estimate an ARIMA model, make sure that you know what the program does for you.

Stage III: Model Checking

- Two major approaches to model checking have traditionally been used.
- **I) Fit a more elaborate model**
- Fit a model that contains additional parameters.
- If an ARMA(p, q) model is fitted, then fit an alternative model with one or two autoregressive parameters and then with one or two moving average parameters.
- Note: It is not a good strategy to add both extra autoregressive and extra moving average terms, because if the model is adequately specified, the additional parameters will in effect cancel out (common roots).
- The resulting parameter estimators of the augmented model will have very large variances.
- For example: if X_t is a white noise process and we fit ARMA(1, 1) model, then $\varphi = \theta$.

An example

- Newbold, Agiakloglou and Miller (1993) using the Nelson and Plosser (1982) log U.S. unemployment rate series found that the best fitted model was an ARIMA(1, 1, 2), that is:

$$(1 - 0.57B)(1 - B)X_t = (1 - 0.45B - 0.55B^2)\varepsilon_t$$

which is $(1 - 0.57B)(1 - B)X_t = (1 - B)(1 + 0.55B)\varepsilon_t$

- This suggests over-differencing.
- The best fitted model for d = 0 is the ARIMA(1, 0, 1) , that is:

$$(1 - 0.50B)(X_t - \mu) = (1 + 0.62B)\varepsilon_t$$

- A model that is very close to the one above after the cancellation of the common root.

The Portmanteau test – I

- The second approach to model checking is the portmanteau test.
- This approach is based on the fact that is the model is correctly specified, the error term will be white noise, which means that all autocorrelations of these errors will be zero.
- In practice the true errors will be unknown, but they can be estimated by the residuals of the fitted model.
- The autocorrelations of the residuals are then calculated:

$$\hat{r}_k = \frac{\sum_{t=k+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}$$

- The test is carried out based on these autocorrelations.

The Portmanteau test – II

- The portmanteau test is based on the squares of the first m residual autocorrelations where m is a moderate large number - at least ten.
- The test introduced by Ljung and Box (1978), *Biometrika*, is based on the Q statistic:

$$Q = n(n+2)\sum_{j=1}^{m}\frac{\hat{r}_j^2}{n-j}$$

- This is modification of the Box and Pierce (1970), *JASA*, whereas similar result can be found in Davies, Triggs and Newbold (1977), *Biometrika*.
- The Q statistic follow a X^2 distribution with (m – p – q) degrees of freedom regardless of the presence of the mean of the series or constant.
- If $Q < X^2_{m-p-q,\alpha}$ the null hypothesis will be accepted and the model will be correctly specified.

Identification in Practice – I

- "The principle of parsimony"
- "A small model is always preferable to a large"
- **Information Criteria**
  - I) Akaike Information Criterion: AIC = ln(SSE/n) + 2k
  - II) Schwarz Bayesian Criterion: SBC = ln(SSE/n) + (k/n)ln(n)

where

  - SSE = Residual Sum of Squares
  - n = number of observations
  - k = number of parameters estimated (p+q+possible constant term)
- Select the best fitted model such that the value of AIC and SBC is the smallest among alternatives.
- Both criteria min the same function with different penalty function.
- To adequately compare the alternatively models, the number of observations should be kept fixed.

Identification in Practice – II

- AIC has the tendency to select over-parameterized models.
- SBC selects small models and it is asymptotically consistent.
- If d = 0 fit ARMA(p, q) models with constant for all $p + q \leq 5$ and select the best fitted model according to the min value of AIC or SBC.
- If the value of d cannot be determined then:
  - Fit ARIMA(p, 0, q) with constant and without the first observation for all $p + q \leq 5$.
  - Fit ARIMA(p, 1, q) with no constant for all $p + q \leq 5$.
- Use AIC and SBC to select the best fitted model.
- The omission of the first observation for the undifferenced series is mandatory to allow comparability.

Forecast

- Once an ARIMA model has been fitted to a time series, the projection forward of that model to derive forecast values is quit straight forward.
- Consider: ARIMA(p, d, q):

$$(1 - \varphi_1 B - \ldots - \varphi_p B^p)(1 - B)^d X_t = (1 - \theta_1 B - \ldots - \theta_q B^q)\varepsilon_t$$

which is

$$(1 - \Phi_1 B - \ldots - \Phi_{p+d} B^{p+d})X_t = c + (1 - \theta_1 B - \ldots - \theta_q B^q)\varepsilon_t$$

where

$$(1 - \Phi_1 B - \ldots - \Phi_{p+d} B^{p+d}) = (1 - \varphi_1 B - \ldots - \varphi_p B^p)(1 - B)^d$$

and $c = (1 - \varphi_1 - \ldots - \varphi_p)\mu$

- The model can be written as:

$$X_t = c + \Phi_1 X_{t-1} + \ldots + \Phi_{p+d} X_{t-p-d} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

- The above model can be used to generate forecasts.

Prediction

- Given the fitted model:

$$X_t = c + \Phi_1 X_{t-1} + \ldots + \Phi_{p+d} X_{t-p-d} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

- And starting at time n, we can predict future values of X_{n+h}.
- In particular, setting t=n+h we have:

$$X_{t+n} = c + \Phi_1 X_{n+h-1} + \ldots + \Phi_{p+d} X_{p+h-p-d} + \varepsilon_{n+h} - \theta_1 \varepsilon_{n+h-1} - \ldots - \theta_q \varepsilon_{n+h-q}$$

- I) The parameters c, $\Phi_1, \ldots, \Phi_{p+d}, \theta_1, \ldots, \theta_q$ are replaced by their estimates derived from the parameter estimation stage.
- II) For $t \leq n$, X_t will be a known observation. For $t > n$ X_t is replaced by its forecast made at t.
- III) For $t \leq n$, the error term ε_t is replaced by its estimate, that is, the residual from the fitted model. For $t > n$, the unknown ε_t is replaced by its beast forecast, which is zero.

Example of an AR(1) Process

- Consider an AR(1) process: $X_t = \varphi X_{t-1} + \varepsilon_t$
  where $|\varphi| < 1$, t = 1, 2, …, n and ε_t is a white noise process.
- Forecast: $\hat{X}_n(h) = \hat{\phi}\hat{X}_n(h-1) + \hat{\varepsilon}_{n+h}$
- h = 1 $\hat{X}_n(1) = \hat{\phi}\hat{X}_n(0) + \hat{\varepsilon}_{n+1} = \hat{X}_n(1) = \hat{\phi}X_n$
- h = 2 $\hat{X}_n(2) = \hat{\phi}\hat{X}_n(1) = \hat{\phi}^2 X_n$
- In general: $\hat{X}_n(h) = \hat{\phi}^h X_n$
- With mean: $\hat{X}_n(h) = \mu + \hat{\phi}^h [X_n - \mu]$
- Note: 1) since $|\varphi| < 1$, the forecast will go to the mean of the series and 2) the forecast depends on X_n.

Forecast error of AR(1)

- One step-ahead forecast error:

$\varepsilon_n(1) = X_{n+1} - E[X_{n+1}|X_n, X_{n-1}, \ldots, X_1] = \varphi X_n + \varepsilon_{n+1} - \varphi X_n = \varepsilon_{n+1}$

- Thus, $Var[\varepsilon_n(1)] = \sigma^2$ (replace by its estimate).
- Forecast error for longer leads:

$\varepsilon_n(h) = X_{n+h} - E[X_{n+h}|X_n, X_{n-1}, \ldots, X_1] = \varepsilon_{n+h} + \varphi\varepsilon_{n+h-1} + \varphi^2\varepsilon_{n+h-2} + \ldots + \varphi^{h-1}\varepsilon_{n+1}$

- We can write: $\varepsilon_n(h) = \sum_{j=0}^{h-1} \varphi^j \varepsilon_{n+h-j}$
- The Variance is: $Var[\varepsilon_n(h)] = \sigma^2 \sum_{j=0}^{h-1} \varphi^{2j}$
- For summing finite geometric series: $Var[\varepsilon_n(h)] = \sigma^2 \frac{1-\varphi^{2h}}{1-\varphi^2}$
- For long lead times: $Var[\varepsilon_n(h)] \approx \frac{\sigma^2}{1-\varphi^2} \approx Var(X_t) = \gamma_0$

Example of a MA(1) Process

- Consider a MA(1) process: $X_t = \varepsilon_t - \theta\varepsilon_{t-1}$
  where $|\theta| < 1$, t = 1, 2, …, n and ε_t is a white noise process.
- Forecast: $\hat{X}_n(h) = \hat{\varepsilon}_{n+h} - \hat{\theta}\hat{\varepsilon}_{n+h-1}$
- h = 1 $\hat{X}_n(1) = \hat{\varepsilon}_{n+1} - \hat{\theta}\hat{\varepsilon}_n = -\hat{\theta}\hat{\varepsilon}_n$
- h ≥ 2 $\hat{X}_n(h) = 0$
- With mean: $\hat{X}_n(h) = \hat{\mu} + \hat{\varepsilon}_{n+h} - \hat{\theta}\hat{\varepsilon}_{n+h-1}$
- h = 1 $\hat{X}_n(1) = \hat{\mu} + \hat{\varepsilon}_{n+1} - \hat{\theta}\hat{\varepsilon}_n = \hat{\mu} - \hat{\theta}\hat{\varepsilon}_n$
- h ≥ 2 $\hat{X}_n(h) = \hat{\mu}$
- Note: The forecast depends on recent values of residual.

Forecast error of MA(1)

- One step-ahead forecast error:

$$\varepsilon_n(1) = X_{n+1} - E[X_{n+1}|X_n, X_{n-1}, \ldots, X_1] = \varepsilon_{n+1} - \theta\varepsilon_n - (-\theta\varepsilon_n) = \varepsilon_{n+1}$$

- Thus, $Var[\varepsilon_n(1)] = \sigma^2$ (replace by its estimate).
- Note: $\varepsilon_n(1) \sim N(0, \sigma^2)$.
- Then Construct Confidence Interval for $X_n(1)$.

General Comments for Prediction

- Pure Moving Average Processes:
  - Have actual forecasts up to q periods ahead only that is up yo the order of the moving average model.
  - After the order of the moving average model the forecast is the mean of the series.
  - The forecasts depend on the most recent values of the error term (residuals).
- Pure Autoregressive Processes:
  - Have forecasts for large periods ahead.
  - The forecasts depend on the most recent observations of the series.
  - For very large periods ahead the forecast is the mean of the series.
  - Also for large periods ahead the variance of the forecast converges to the variance of the series.
- Mixed comments are valid for mixed models.

Example of a R.W. Process

- Consider Random Walk with drift process: $X_t = X_{t-1} + \mu + \varepsilon_t$
- Forecast: $\hat{X}_n(h) = \hat{X}_n(h-1) + \hat{\mu} + \hat{\varepsilon}_{n+h}$
- h = 1 $\hat{X}_n(1) = \hat{X}_n(0) + \hat{\mu} + \hat{\varepsilon}_{n+1} = X_n + \hat{\mu}$
- h = 2 $\hat{X}_n(2) = X_n + 2\hat{\mu}$
- In general: $\hat{X}_n(h) = X_n + h\hat{\mu}$
- If $\mu \neq 0$ the forecast does not converge for long leads h but follows a straight line with slope μ for all h periods.

Forecast error of R.W.

- One step-ahead forecast error:

$\varepsilon_n(1) = X_{n+1} - E[X_{n+1}|X_n, X_{n-1}, \ldots, X_1] = X_n + \mu + \varepsilon_{n+1} - (X_n + \mu) = \varepsilon_{n+1}$

- Thus, $Var[\varepsilon_n(1)] = \sigma^2$ (replace by its estimate).
- Forecast error for longer leads:

$\varepsilon_n(h) = X_{n+h} - E[X_{n+h}|X_n, X_{n-1}, \ldots, X_1] = (X_n + h\mu + \varepsilon_{n+1} + \ldots + \varepsilon_{n+h}) - (X_n + h\mu)$

- We can write: $\varepsilon_n(h) = \sum_{j=0}^{h-1} \varepsilon_{n+h-j}$
- The Variance is: $Var[\varepsilon_n(h)] = h\sigma^2$

- In contrast to the stationary case, here the $Var[\varepsilon_n(h)]$ grows without limit as the forecast lead time h increases.
- This property is characteristic of the forecast error variance for all non-stationary processes.