NONSTATIONARY TIME SERIES & FORECASTING

JEN-WEN LIN, PhD

June 21, 2014

Plotting data is the first step of time series analysis

- If there are any apparent discontinuities in the series, such as a sudden change of level, it may be advisable to analyze the series by first breaking it into homogeneous segments.
- If there are outliers, they should be studied carefully to check whether there is any justification for discarding them.
- Then inspection of a graph may also suggest the possibility of representing the data as a realization of the process (the "classical decomposition" model).

Box-Jenkins approach/$ARIMA$ models

An alternative approach by Box-Jenkins (1970) is to apply difference operators repeatedly to the data $\{X_t\}$ until the differenced observations resemble a realization of some stationary process $\{W_t\}$.

A time series $\{X_t\}$ is said to follow an integrated autoregressive moving average model if the dth difference $W_t = (1-B)^d X_t$ is a stationary $ARMA$ model. If $\{W_t\}$ follows an $ARMA(p,q)$ model, we say that $\{X_t\}$ is an $ARIMA(p,d,q)$ process.

- An $ARIMA(p,d,q)$ process is given by $(1-B)^d\phi(B)X_t = \theta(B)a_t$, where $\phi(B) = 1-\phi_1 B - \cdots - \phi_p B^p$, $\theta(B) = 1+\theta_1 B + \cdots + \theta_q B^q$, and $a_t \sim WN(0,\sigma^2)$
- Or written as $\phi(B)W_t = \theta(B)a_t$, where $W_t = (1-B)^d X_t$

Differencing to remove time trend

1−B : differencing

Example:

Let $Y_t = a + bt + ct^2 + X_t$, where X_t is a stationary time series. Consider the following transformation:

- Backward operator B: $By_t = y_{t-1}, Bt = t - 1, Bc = c$.
- Notation: $\nabla^d = (1 - B)^d, \nabla^2 = (1 - B)(1 - B)$
- $(1 - B)^2 a = (1 - 2B + B^2)a = a - 2a + a = 0$
- $(1 - B)^2 bt = \nabla(1 - B)bt = \nabla[bt - b(t - 1)] = \nabla b = 0$
- $(1 - B)^2 ct^2 = \nabla(1 - B)ct^2 = \nabla[ct^2 - c(t - 1)^2] = \nabla[ct^2 - ct^2 + 2ct + c] = \nabla(2ct + c) = 0$
- **Question:** whether $(1 - B)^2 X_t$ is stationary

Differencing at lag d to remove seasonal component

The technique of differencing that we applied to trend-stationary data can be adapted to deal with seasonality of period d by introducing the lag-d difference operator ∇_d defined by $\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t$

- This operator should not be confused with the operator $\nabla^d = (1 - B)^d$ defined earlier.
- Applying the ∇_d operator to the classical decomposition model, $X_t = m_t + s_t + Y_t$, where s_t has period d, we have $\nabla_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d}$.
- $m_t - m_{t-d}$ is a trend component and $Y_t - Y_{t-d}$ is a noise term.

Nonstationarity in variance

Differencing (trend-removing) can be used to reduce a homogeneous nonstationary time series to a stationary (trend-stationary) time series.

- Many nonstationary time series, however, are nonhomogeneous.
- The nonstationarity of these series is not due to their time dependent means but their time-dependent variance and autocovariances.
- To reduce these types of nonstationarity, we need to transformations other than differencing.

Transformation for nonstationarity in variance

(not much about this)

Power transformation by Box and Cox (1964):

$$T(X_t) = (X_t^{\lambda} - 1)/\lambda\,.$$

- One great advantage of using the Box-Cox transformation is that we can treat λ as a transformation parameter and estimate its value from data.
- For example, we can include λ as a parameter in an $ARMA$ model, $\phi(B)\left(X_t^{(\lambda)} - \mu\right) = \theta(B)a_t$ and choose the value of λ that gives the minimum residual mean square error (RMSE).

Remarks

In the preliminary analysis, one can use an AR model to obtain the value of λ through an AR fitting that minimizes the RMSE on a grid of λ values.

A variance stabilizing transformation, if needed, should be performed before any analysis such as differencing.

Frequently, the transformations also improve the approximation of the distribution by a normal distribution.

The variance stabilizing transformations are defined by positive series.

- The definition is not restrictive as it seems because a constant can always be added to the series without affecting the correlation structure of the series.

$I(d)$ process (usually in final, test def)

Definition: A series with no deterministic component that has a stationary, invertible $ARMA$ representation after differencing d times is said to be integrated of order d, which is denoted as $I(d)$.

- A random walk model $y_t = y_{t-1} + a_t = y_0 + \sum_1^t a_s$, $a_t \sim NID(0, \sigma_a^2)$. This process is not stationary;
- If a_t is not uncorrelated over time, such as $a_t = \phi a_{t-1} + \eta_t$ with $|\phi| > 1$ and $\eta_t \sim NID(0, \sigma_\eta^2)$, then y_t is not a random walk process but an $I(1)$ process.

Dickey-Fuller unit root test

The Dickey-Fuller test was developed by Dickey and Fuller (1979). The use of the test contains two steps: remove the deterministic time trend (if any) and conduct statistical inference using the above test.

- Consider a regression test on the following model $X_t = \phi X_{t-1} + a_t$, $a_t \sim NID(0, \sigma^2)$.
- $\Delta X_t = (\phi - 1)X_{t-1} + a_t = \pi X_{t-1} + a_t$
- $H_0: \pi = 0$; H_a: a trend stationary process
- Under H_0, $X_t \sim I(1)$ and $\Delta X_t \sim I(0)$ so the OLS estimate of π does not follow not a Student-t distribution.

One-step unit root test

$\Delta X_t = \tau^T DR_t + \pi X_{t-1} + a_t$, where DR_t are deterministic independent variables, τ is the corresponding coefficient vector, and $a_t \sim NID(0, \sigma^2)$.

For the one-step procedure, the distribution of the t-ratio of π depends on the nuisance parameters.

A weakness of the aforementioned tests is that they do not take a possible serial correlation of the error process into account.

Augmented Dickey Fuller test

ADF

Dickey and Fuller (1981) have suggested replacing the AR(1) process for $\{X_t\}$ with an $ARMA(p, q)$ process, or an $AR(p)$ process. It can be shown that the following test regression ensure the serial correlation in error terms is removed.

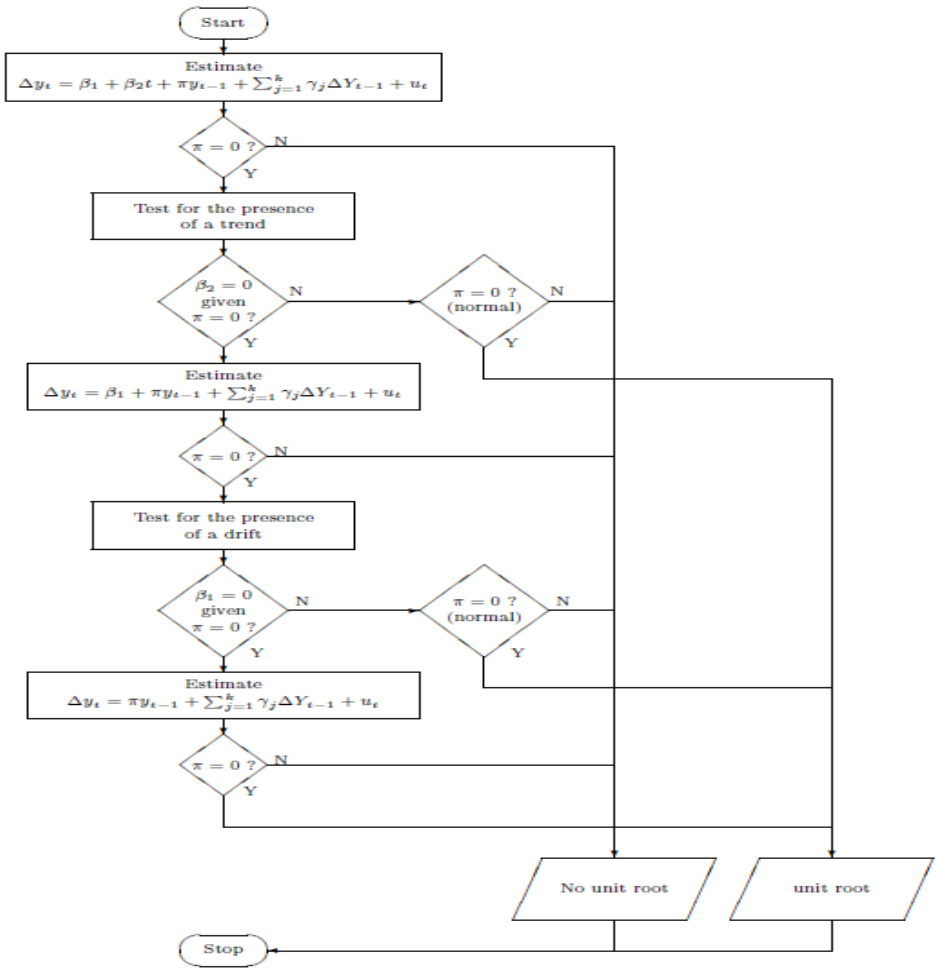
Idea: $\begin{cases} \Delta X_t = \Pi X_{t-1} + a_t \\ \Gamma(B)\Delta X_t = \Pi X_{t-1} + \eta_t \end{cases}$

$\Delta X_t = \frac{\Pi}{\Gamma(B)} X_t + \frac{\eta_t}{\Gamma(B)} \Sigma \psi_j B^j \eta_t$

- The encompassing ADF-test equation:
- $\Delta X_t = \beta_1 + \beta_2 t + \pi X_{t-1} + \sum_1^k \gamma_j \cdot \Delta X_{t-j} + \eta_t$
- General form:
- $\Delta X_t = \tau^T DR_t + \pi X_{t-1} + \sum_1^k \gamma_j \cdot \Delta X_{t-j} + \eta_t$, with $k = p - 1$
- How to determined p or k?

Testing sequences

- The encompassing ADF-test equation is estimated first.
- Then steps on RHS are taken until one can conclude that the series is
  - stationary around mean,
  - stationary around a nonzero mean,
  - stationary around a linear trend, containing a unit root with zero drift,
  - and containing a unit root with non-zero drift.



Remarks

If the null hypothesis π cannot be rejected, then the series may be integrated of order one or above. Hence, the test procedure does not end at the end of the aforementioned procedure. One has to test whether the series is $I(2)$, or even integrated to a higher degree.

- The derivation of the asymptotic distribution of the DF test may be taught in class if time allowed.

Objective of Forecasting

An $ARIMA(p, d, q)$ models may be defined as

$$\varphi(B)X_t = \theta(B)a_t, \quad (1)$$

where $\varphi(B) = \phi(B)(1 - B)^d$,

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p,$$

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q.$$

- In what follows, we shall learn how to forecast a value $X_{t+l}, l \geq 1$ when we are currently standing at time t.
- This forecast is said to be made at origin t for lead time l.

Three explicit forms of $ARIMA$ models

- Any observation X_{t+l} generated by the process (1) may expressed as the following explicit forms:

1. **Difference equation form:** directly in terms of the difference equation

$$X_{t+l} = \varphi_1 X_{t+l-1} + \cdots + \varphi_{p+d} X_{t+l-p-d}$$

$$+ a_{t+l} + \theta_1 a_{t+l-1} + \cdots + \theta_q a_{t+1-q}$$

Three explicit forms of *ARIMA* models

2. **Integrated form:** as an infinite weighted sum of current and previous shocks,

$$X_{t+l} = \sum_0^\infty \psi_i a_{t+l-j} \quad (2),$$

where $\{\psi_j\}$ may be calculated using the techniques in §5.2.1 and §5.2.2 of Wei (2006). We could rewrite eqn. (2) as

$$X_{t+l} = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} + C_t(l) \quad (3),$$

where $C_t(l)$ is associated with the truncated infinite sum

$$C_t(l) = \sum_l^\infty \psi_j a_{t+l-j} \quad (4).$$

Causal: $X_t = \sum_{j=0}^\infty \psi_j a_{t-j}$, $\psi_0 = 1$

want this: $X_{t+l} = \left(\sum_{j=0}^\infty \psi_j a_{t+l-j}\right) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} + \sum_{j=l}^\infty \psi_j a_{t+l-j}$

2 parts information

Three explicit forms of *ARIMA* models

We don't use this in this class

3. **Weighted average of previous observations:**

$$X_{t+l} = a_{t+l} + \sum_{j=1}^{\infty} \pi_j X_{t+l-j}$$

$$= a_{t+l} + \pi_1 X_{t+l-1} + \cdots + \pi_{l-1} X_{t+1} + \sum_{j=l}^{\infty} \pi_j X_{t+l-j} \quad (5)$$

where $\{\pi_j\}$ can be obtained from equating the coefficients of the order of the backward operators B using

$$\varphi(B) = (1 - \pi_1 B - \pi_2 B^2 - \cdots)\theta(B)$$

Minimum mean square error forecast

- Suppose the best forecast for X_{t+l} standing at origin t using innovations $a_t, a_{t-1}, \ldots$ is written as

$$\hat{X}_t(l) = \psi_l^* a_t + \psi_{l+1}^* a_{t-1} + \psi_{l+2}^* a_{t-2} + \cdots \quad (6),$$

  where $\psi_l^*, \psi_{l+1}^*, \ldots$ are to be determined.

- The mean square error of the forecast in eqn. (6) is

$$E\left(X_{t+l} - \hat{X}_t(l)\right)^2 = E\left\{(a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}) + (C_t(l) - \hat{X}_{t+l}(l))\right\}$$

$$= E\left(\Delta^2 + E\left(\sum_0^\infty (\psi_{l+j} - \psi_{l+j}^*) a_t\right)^2 \infty\right)$$

$$= \left(1 + \psi_1^2 + \cdots \psi_{l-1}^2\right)\sigma^2 + \sum_0 \left(\psi_{l+j} - \psi_{l+j}^*\right)^2 \sigma^2$$

derivation

  which is minimized by setting $\psi_{l+j}^* = \psi_{l+j}$.

Minimum mean square error forecast cont'd

- Using eqn. (2), and (6), we have

$$X_{t+l} =$$
$$a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} + (\psi_l a_t + \psi_{l+1} a_{t-1} + \cdots)$$

$$= e_t(l) + \hat{X}_t(l),$$

where $\hat{X}_t(l)$ is also called the forecast function for origin t, and $e_t(l)$ is the corresponding "error of forecast".

Important facts

The minimum mean square error forecast at origin t, for lead time l, $\hat{X}_t(l)$ is the conditional expectation of X_{t+l} at time t.

Conditional expectation

- $E\left(a_{t+j}|X_t, X_{t-1}, \dots\right) = E\left(a_{t+j}\right) = 0, \forall\, j > 0$
- $E_t(X_{t+l}) = \hat{X}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots,$
- where $E(X_{t+l}|X_t, X_{t-1}, \dots) = E_t(X_{t+l})$ is the conditional expectation of X_{t+l} given knowledge of all X's up to time t.
- The minimum requirement on the random shocks $\{a_t\}$ in eqn. (1) in order for the minimum mean square error forecast $E_t(X_{t+l})$ coincide with the mean square error "linear" forecast is that $E_t\left(a_{t+j}\right) = 0, \forall\, j > 0$. This may not hold for certain types of intrinsically nonlinear processes.

Important facts

The forecast error for lead time l is
$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}$

- $E_t(e_t(l)) = 0$ so the forecast is unbiased
- $Var(e_t(l)) = \left(1 + \psi_1^2 + \cdots + \psi_{l-1}^2\right)\sigma^2$ is the variance of forecast error

$\hat{X}\_t(l)$ is not only the minimum mean square error forecast of X_{t+l} but any linear function $\sum_1^L w_l \hat{X}_t(l)$ of the forecasts is a minimum mean square error forecast of the corresponding linear function $\sum_1^L w_l X_{t+l}$.

- For example, $\hat{X}_t(1) + \hat{X}_t(2) + \hat{X}_t(3)$ is the minimum mean square error forecast for $X_{t+1} + X_{t+2} + X_{t+3}$.

Important facts

The shock is the one-step-ahead forecast error. Specifically, $e_t(1) = X_{t+1} - \hat{X}_t(1) = a_{t+1}$. It follows that for a minimum mean square error forecast, the one-step-ahead forecast errors must be uncorrelated.

Correlation between the forecast errors: Although the optimal forecast errors at lead time 1 will be uncorrelated, the forecast errors for longer lead time in general will be correlated.

Autocorrelation function of forecast errors at different origins

skip

Consider forecasts for lead time l, made at origins t and $t-j$ respectively, where j is a positive integer.

If $j = l, l+1, l+2, \ldots,$ the forecast errors will contain no common component so the forecasts are uncorrelated

But for $j = 1,2, \ldots, l-1$, certain of shocks will be included in both forecast errors.

Autocorrelation function of forecast errors at different origins

skip

Consider forecast errors at different origins:

- $e_t(l) = X_{t+l} - \hat{X}_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}$
- $e_{t-j}(l) = X_{t+l-j} - \hat{X}_{t-j}(l) = a_{t-j+l} + \psi_1 a_{t-j+l-1} + \cdots + \psi_{l-1} a_{t-j+1}$

For $j < l$, the lag j auto-covariance of forecast errors for lead time l is

- $\mathrm{Cov}\big(e_t(l), e_{t-j}(l)\big) = \sigma^2 \sum_{i=j}^{l-1} \psi_i \psi_{i-j}$

The corresponding autocorrelations are

- $Corr\big(e_t(l), e_{t-j}(l)\big) = \begin{cases} \sum_{i=1}^{l-1} \psi_i \psi_{i-j} / \sum_{i=0}^{l-1} \psi_i^2 , 0 \le j \le l \\ 0, j \ge 0 \end{cases}$

Correlation between forecast errors at the same origin with different lead times skip

Suppose that we make a series of forecasts for different lead times from the same fixed origin t. Then the errors for these forecasts will be correlated. For $j = 1,2,3, \ldots$. the forecast errors may be expressed as

- $e_t(l) = X_{t+l} - \hat{X}_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}$
- $e_t(l+j) = X_{t+l-j} - \hat{X}_t(l+j) = a_{t+j+l} + \psi_1 a_{t+j+l-1} + \cdots + \psi_j a_{t+1} + \psi_{j+1} a_{t+l-1} + \cdots + \psi_{l+j-1} a_{t+1}$

Therefore, the covariance between the t-origin forecasts at lead times l and $l+j$ is

- $Corr(e_t(l), e_t(l+j)) = \sum_{i=0}^{l-1} \psi_i \psi_{i+j} / \sqrt{\sum_{h=0}^{l-1} \psi_h^2 \cdot \sum_{g=0}^{l+j-1} \psi_g^2}$

Three form of forecast (skip)

Notations: $[a_{t+l}] = E_t(a_{t+l})$ and $[X_{t+l}] = E_t(X_{t+l}) = \hat{X}_t(l)$.

Forecasts from difference equations:

- $[X_{t+l}] = \varphi_1[X_{t+l-1}] + \cdots + \varphi_{p+d}[X_{t+l-p-d}] + [a_{t+l}] + \theta_1[a_{t+l-1}] + \cdots + \theta_q[a_{t+l-q}]$

Forecasts in integrated form:

- $[X_{t+l}] = \sum_0^\infty \psi_i[a_{t+l-j}]$

Forecasts as a weighted average of previous observations and forecasts made at previous lead times from the same origin

- $[X_{t+l}] = \sum_{j=1}^\infty \pi_j[X_{t+l-j}] + [a_{t+l}]$

Remarks (skip)

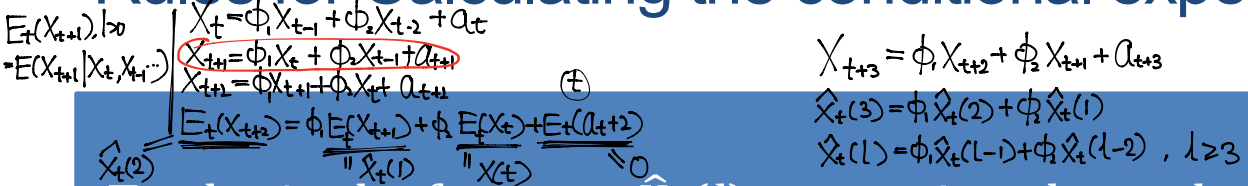
This approach theoretically requires knowledge of the X's stretching back into the infinite past.

However, the requirement of invertibility, which we have imposed on the $ARIMA(p,d,q)$ model, ensures that the $\{\pi_j\}$ coefficients form a convergent series.

Hence, for the computation of a forecast to a given degree of accuracy, for some k, the dependence on $X_{t-j}, \forall\, j > k$ can be ignored.

In practice, the $\{\pi_j\}$ coefficients usually decay rather quickly, only a moderate length of series $\{X_t\}$ is needed to calculate the forecasts to sufficient accuracy.

Rules for Calculating the conditional expectations



$$X_{t+3} = \phi_1 X_{t+2} + \phi_2 X_{t+1} + a_{t+3}$$

$$\hat{X}_t(3) = \phi_1 \hat{X}_t(2) + \phi_2 \hat{X}_t(1)$$

$$\hat{X}_t(l) = \phi_1 \hat{X}_t(l-1) + \phi_2 \hat{X}_t(l-2), \quad l \geq 3$$

To obtain the forecast $\hat{X}_t(l)$, one writes down the model for X_{t+l} in any one of the three explicit forms and treats the terms on the right according to the following rules:

- The X_{t-j} $(j = 0,1,2, \ldots)$, which have already happened at origin t, are left unchanged.
- The X_{t+j} $(j = 1,2, \ldots)$, which have not yet happened, are replaced by their forecasts at origin t, i.e., $\hat{X}_t(j)$.
- The a_{t-j} $(j = 0,1,2, \ldots)$, which have already happened, are available from $X_{t-j} - \hat{X}_{t-j-1}(1)$.
- The a_{t+j} $(j = 1,2, \ldots)$, which have not yet happened, are replaced by zeros.

Rules for Calculating the conditional expectations

$$[X_{t-j}] = E_t[X_{t-j}] = X_{t-j}, \quad j = 0,1,2,\ldots$$

$$[X_{t+j}] = E_t[X_{t+j}] = \hat{X}_t(j), \quad j = 1,2,\ldots$$

$$[a_{t-j}] = E_t[a_{t-j}] = a_{t-j} = X_{t-j} - \hat{X}_{t-j-1}(1), \quad j = 0,1,2,\ldots$$

$$[a_{t+j}] = E_t[a_{t+j}] = 0, \quad j = 1,2,\ldots$$

Remarks on the third rule

- The current and past errors may be calculated as follows: $a_t = X_t - \hat{X}_{t-1}(1)$ and $a_{t-1} = X_{t-1} - \hat{X}_{t-2}(1)$, where The forecasting process may be started off initially by setting unknown a's equal to their unconditional expected values of zeros.
- Assuming that the data are available starting from time $s = 1$, the necessary a's are computed recursively from $a_s = X_s - \hat{X}_{s-1}(1) = X_s - (\sum_{j=1}^{p+d} \varphi_j X_{s-j} + \sum_{j=1}^{q} \theta_j a_{s-j})$, where $s = p + d + 1, \ldots, t$ and we may set a's equal to zero for $s < p + d + 1$.

Use of the $\{\psi_j\}$ weights in updating forecasts

Using the result

$$\hat{X}_{t+1}(l) = \psi_l a_{t+1} + \psi_{l+1} a_t + \psi_{l+2} a_{t-1} + \dots$$

$$\hat{X}_t(l+1) = \psi_{l+1} a_t + \psi_{l+2} a_{t-1} + \psi_{l+3} a_{t-2} + \dots$$

- On subtraction, it follows that $\hat{X}_{t+1}(l) = \hat{X}_t(l+1) + \psi_l \cdot a_{t+1}$.

That is, the t-origin forecast for X_{t+l+1} can be updated to become the $t+1$ origin forecast for X_{t+l+1}, by adding a constant multiple of the one-step-ahead forecast error a_{t+1}, with multiplier ψ_l.

Probability limits of the forecast

The variance of l-step-ahead forecast error for any origin t is the expected value of $e_t^2(l) = \{X_{t+l} - \hat{X}_t(l)\}^2$ and is given by $\left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)\sigma^2$.

The probability limits of the forecasts at any lead times: Assuming that the a's are Gaussian, it follows that given information up to time t, the conditional probability distribution $p(X_{t+l}|X_t, X_{t-1}, \dots)$ of a future values X_{t+l} of the process will be Gaussian with mean $\hat{X}_t(l)$ and standard deviation $\sqrt{\left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)\sigma^2}$

Exercises

- Calculate the following conditional expectation:
  1. $(1-0.8B)(1-B)X_t = a_t$
  2. $(1-B)^2 X_t = (1-0.9B+0.5B^2)a_t$
- Consider an $AR(1)$ model $(1-0.6B)(X_t-9) = a_t$, where $a_t \sim NID(0,1)$. Suppose that we have the observations $(X_{97}, X_{98}, X_{99}, X_{100}) = (9.6,9,9,8.9)$.
  1. Forecast $\{X_t\}, t = 101, 102, 103$ and 104 and their associated 95% forecast limits.
  2. Calculate $\hat{X}_{101}(1), \hat{X}_{101}(2), \hat{X}_{101}(3)$ using "updating forecast".

References

- Brockwell and Davis (1980), *Time Series: Theory and Methods*, Springer-Verlag.
- Box and Jenkins (1976), *Time series analysis: Forecasting and control*, Holden-Day.
- Wei (2006), *Time series analysis: Univariate and mltivariate methods*, Addison Wesley.
- *Pfaff (2008), Analysis of Integrated and Cointegrated Time Series with R,* Springer.