# Nonstationary time series and forecasting

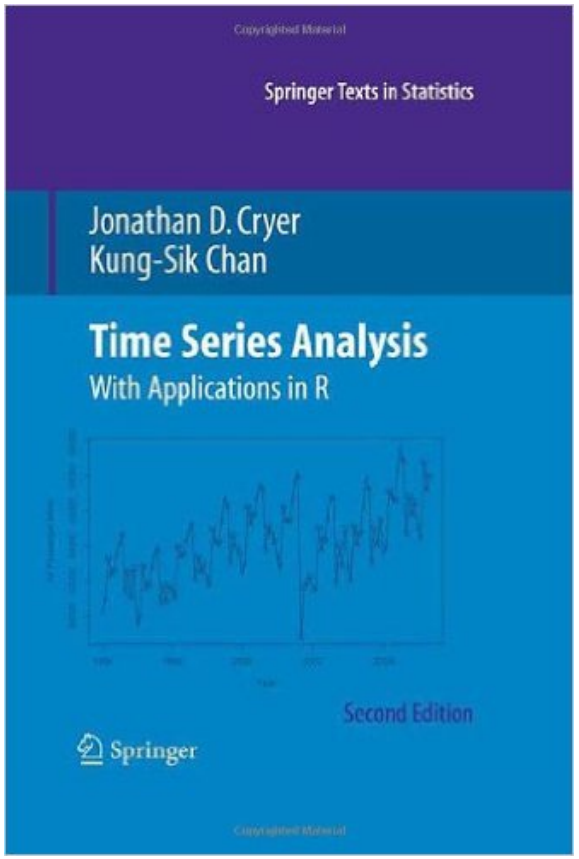
JEN-WEN LIN, PhD, CFA

Oct 11, 2020

# Reading list

- Reference to Wei's book:
  1. Chapter 4.1, 4.2.1, 4.3
  2. Chapter 5.1, 5.2, 5.3
  3. Chapter 8.1, 8.2.1, 8.3
  4. Chapter 9.1

# Time series reference book for R




• Cryer and Chan (2010), *Time Series Analysis: With Applications in R*, Second Edition, Springer.

•Amazon link: http://www.amazon.ca/Time-Analysis-Applications-Jonathan-Cryer

# Classical decomposition of time series

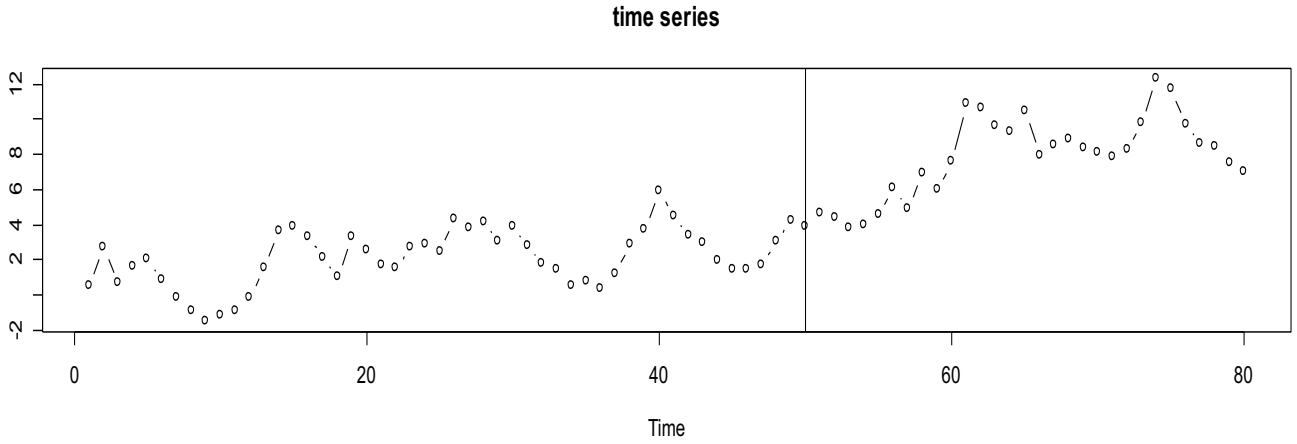
| | |
|---|---|
| **Seasonal variation** | Time series exhibit variation that is annual in period (or every 12 units of time). |
| | For example, the sales of electronic companies in the second quarter are typically the lowest. |
| **Cyclical variation** | Time series exhibit variation at a fixed period due to some other physical cause. |
| | Examples are daily variation in temperature and business cycles. |
| **Trend** | This may be loosely defined as ‘long-term change in the mean level’. |

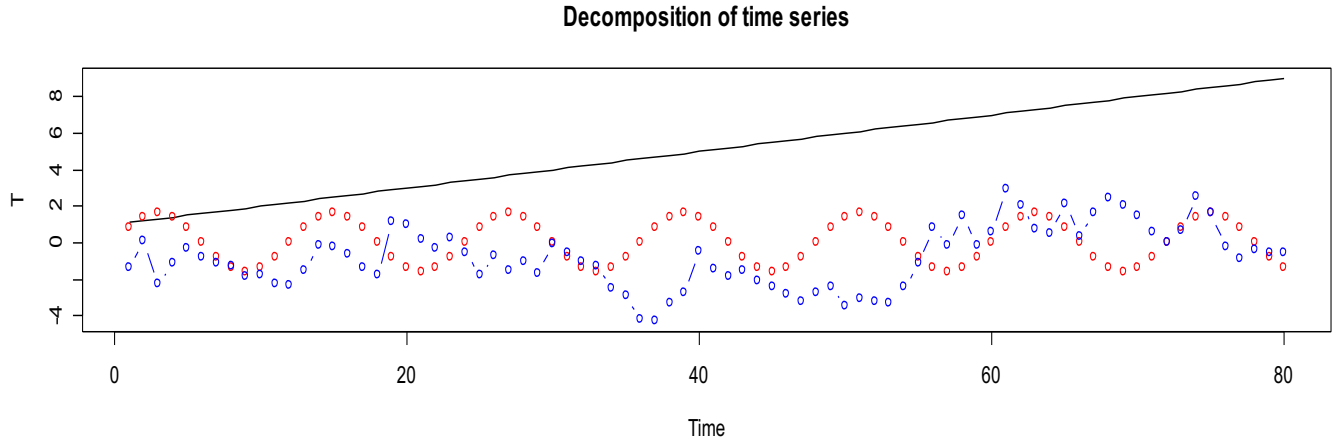
# Hypothetical example

Trend (black): $T_t = 1 + 0.1 \cdot t$

Seasonal (red): $S_t = 1.6 \cdot \sin(\frac{t\pi}{6})$

Irregular (blue): $I_t = 0.7 \cdot I_{t-1} + \varepsilon_t$



# Regression methods to remove time trend

- *Example*: linear regression to remove linear time trend

$$Y_t = \mu_t + X_t,$$

where $\mu_t = \beta_0 + \beta_1 t, \quad t = 1, \ldots, n.$

- Least squared estimation:

$$Q(\beta_0, \beta_1) = \sum_{t=1}^{n} [Y_t - (\beta_0 + \beta_1 t)]^2$$

- Estimator:

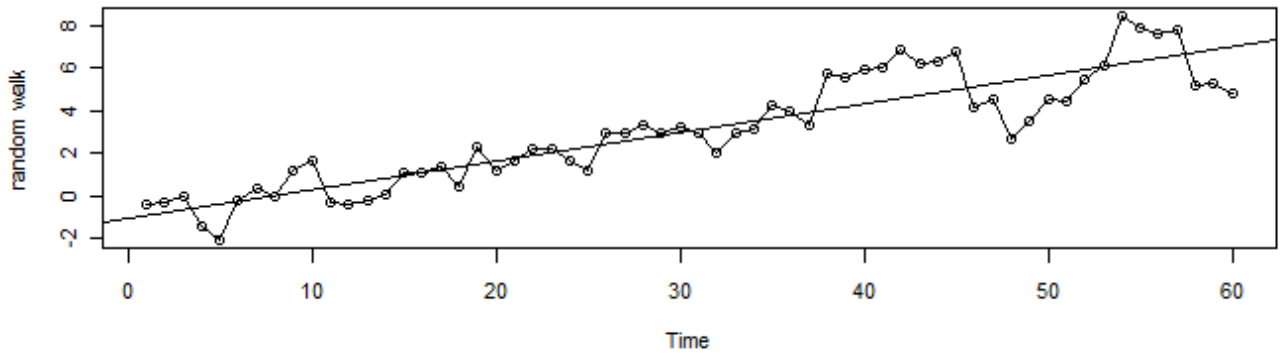
$$\hat{\beta}_1 = \sum_{t=1}^{n} (Y_t - \bar{Y})(t - \bar{t}) / \sum_{1}^{n} (t - \bar{t})$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{t}, \qquad \bar{t} = \frac{n+1}{2}.$$

# Linear and quadratic trends in time

R code:

```
•library(TSA)
•data(rwalk)
•mod_timetr<-lm(rwalk~time(rwalk))
•summary(mod_timetr)
•win.graph(height=2.5, pointsize=8)
•plot(rwalk, type='o', ylab="random walk")
•abline(mod_timetr) # add the fitted regression line
```



# Regression methods to remove seasonality

- *Example*: Monthly mean model:

$$Y_t = \mu_t + X_t, \qquad E(X_t) = 0, \forall t,$$

where $X_t$ denotes the stationary irregular component, and $\mu_t$ is monthly data with 12 constants (parameters) which gives the expected value for each of the 12 months.

- Specifically, we may write

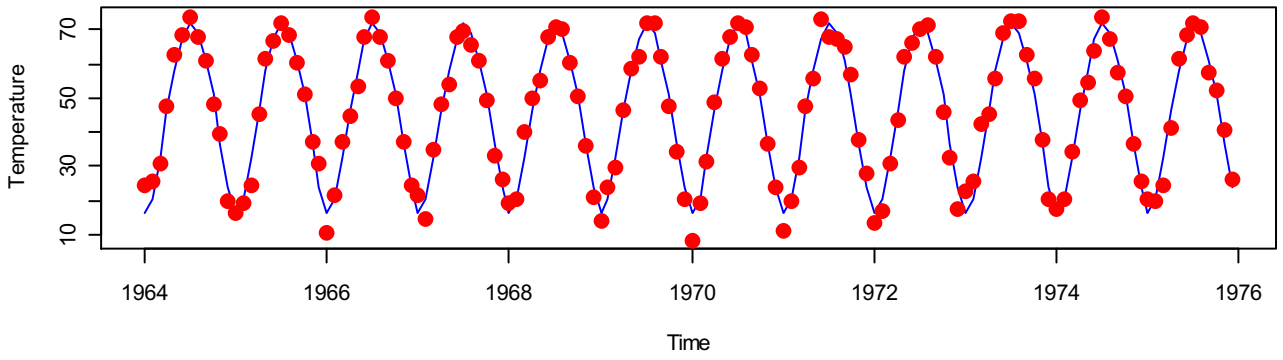
$$\mu_t = \begin{cases} \beta_1, & t = 1,13,25, \dots \\ \beta_2, & t = 2,14,26, \dots \\ \vdots & \\ \beta_{12}, & t = 12,24,36, \dots \end{cases}.$$

# Regression methods to remove seasonality

R code:

```
•data(tempdub)
•month.<-season(tempdub)
•mod_cyctr<-lm(tempdub~month.); temp<-fitted(mod_cyctr)
•win.graph(height=2.5, pointsize=8)
•plot(ts(temp,freq=12,start=c(1964,1)), ylab='Temperature', type="l"
 ,col=4, ylim=range(c(temp,tempdub)))
•points(tempdub,col=2, lwd=4)
```

Monthly average temperature (in degrees Fahrenheit) recorded in Dubuque 1/1964 - 12/1975.



# Regression methods to remove seasonality

- *Example*: Monthly mean model:

$$Y_t = \mu_t + X_t, \qquad E(X_t) = 0, \forall t,$$

$$\mu_t = \beta_1 \cos(2\pi f t) + \beta_2 \sin(2\pi f t),$$

Where $X_t$ denotes the stationary irregular component, and $1/f$ is called the period.

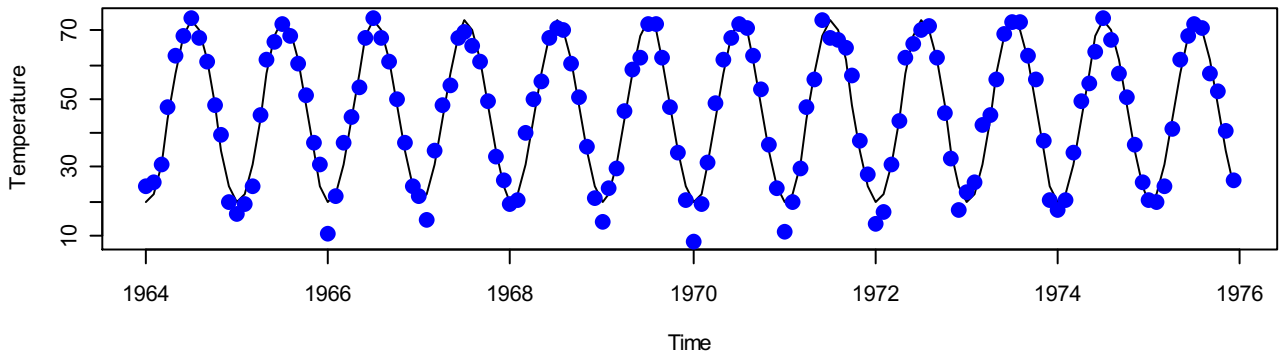
- For example, monthly data with time index as $1,2,\ldots$, has $f = 1/12$ because such sinusodial function will repeat itself every $12$ months. In this case, the period is $12$.

- Least square estimation: use $\cos(2\pi f t)$ and $\sin(2\pi f t)$ as predictor variables.

# Regression methods to remove seasonality

R code:

```
• har.<-harmonic(tempdub,1)
• mod_costr<-lm(tempdub~har.); temp_<-fitted(mod_costr)
• win.graph(height=2.5, pointsize=8)
• plot(ts(temp_,freq=12,start=c(1964,1)),
• ylab='Temperature', type="l" ,ylim=range(c(temp_,tempdub)))
• points(tempdub,col=2, lwd=4)
```



# Autoregressive integrated moving average (ARIMA) models

- For nonstationary time series, Box-Jenkins (1970) suggested applying difference operators repeatedly to the data $\{X_t\}$ until the differenced observations resemble a realization of some stationary process $\{W_t\}$.

- $\{X_t\}$ is said to follow an ARIMA model of order $(p, d, q)$ if $W_t = (1-B)^d X_t$ is a stationary $ARMA$ model. Mathematically, we have

$$(1-B)^d \phi(B) X_t = \theta(B) a_t, \qquad a_t \sim N(0, \sigma^2)$$

where

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p,$$
$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q.$$

# Differencing to remove time trend

- 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: $\nabla^d = (1 - B)^d$, $\nabla^2 = (1 - B)(1 - B)$
  - $(1 - B)^2 a = (1 - 2B + B^2)a = a - 2a + a = 0$
  - $(1 - B)^2 bt = \nabla(1 - B)bt = \nabla[bt - b(t - 1)] = \nabla b = 0$
  - $(1 - B)^2 ct^2 = \nabla(1 - B)ct^2 = \nabla[ct^2 - c(t - 1)^2] = \nabla[ct^2 - ct^2 + 2ct + c] = \nabla(2ct + c) = 2c$
- *Question*: whether $(1 - B)^2 X_t$ is stationary

# Differencing 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 $\nabla_d$ defined by $\nabla_d X_t = X_t - X_{t-d} = \left(1 - B^d\right)X_t$
- This operator should not be confused with the operator $\nabla^d = (1 - B)^d$ defined earlier.
  - Applying the $\nabla_d$ 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 can be used to transform a nonstationary time series due to the unstable mean level over time to a stationary (trend-stationary) time series.
- Many nonstationary time series, however, are not due to their time dependent means but their time-dependent variance and autocovariances.
  - We refer to time-dependent unconditional second moments rather than conditional second moments.
- To reduce these types of nonstationarity, we need to different transformations other than differencing.

# Nonstationarity in variance

- Power transformation by Box and Cox (1964):

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

- We can incorporate the Box-Cox transformation into model estimation. For example, we can include $\lambda$ as one of the parameters

$$\phi(B)\left(X_t^{(\lambda)} - \mu\right) = \theta(B)a_t, \qquad a_t \sim NID(0, \sigma^2),$$

  and choose the values of $\lambda$ as well as $\{\phi_i\}_{i=1}^{p}$ and $\{\theta_i\}_{i=1}^{q}$ that give the minimum residual mean square error (RMSE).

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

# Some remarks

- In the preliminary analysis, one can use an $AR$ model to obtain the value of $\lambda$ through an AR fitting that minimizes the RMSE on a grid of $\lambda$ values.
- Frequently, the transformations also improve the approximation of the distribution by a normal distribution.
- Finally, it is worth noting that 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 and Dickey-Fuller unit root test

- A series follows a stationary *ARMA* model after differencing $d$ times is said to be integrated of order $d$, or $I(d)$ process.
- The Dickey-Fuller test is used to test $I(1)$ processes. Consider

$$H_0: \pi = 0 \ (or \ \phi = 1) \ ; \ H_a: \pi > 0 \ (or \ \phi > 0).$$

$$X_t = \phi X_{t-1} + a_t, \qquad a_t \sim NID(0, \sigma^2).$$

$$\Delta X_t = (\phi - 1) X_{t-1} + a_t = \pi X_{t-1} + a_t$$

  - Remark: Under $H_0: X_t \sim I(1)$, the OLS estimate of $\pi$ does not follow a Student-t distribution.

# More on Dickey Fuller test

- The general Dickey-Fuller test may contain an intercept and a deterministic time trend as

$$\Delta X_t = a + \tau^T DR_t + \pi X_{t-1} + a_t,$$

- where $a$ denotes the regression intercept and $DR_t$ are deterministic independent variables, $\tau$ is the corresponding coefficient vector, and $a_t \sim NID(0, \sigma^2)$

# Issues on Dickey-Fuller test

- The Dickey-Fuller test considers only a single unit root.
- Correct model specification
  - Correct specification of time trend and intercept
  - The DGP may contain both autoregressive and moving average terms
  - There might be structural breaks in the data

# Augmented Dickey-Fuller test

- Dickey and Fuller (1981) have suggested the encompassing Augmented Dickey-Fuller test equation:

$$\Delta X_t = \tau^T D R_t + \pi X_{t-1} + \sum_{j=1}^{k} \gamma_j \cdot \Delta X_{t-j} + a_t \,,$$

where $k = p - 1$. The above equation use the autoregression to take into account the presence of serial correlated errors.

# Selection of the lag length

- **Autoregression approximation:**

Said and Dickey (1984) later on show that an unknown $ARIMA(p, 1, q)$ process can often be approximated by an $ARIMA(n, 1,0)$ autoregression of order $n$ where $n \leq T^{\frac{1}{3}}$.

- **General-to-specific methodology:**

Start with a relatively long lag length and pare down the model by the usual t-test or F-test.

# General-to-specific methodology

- For example, let's start with a lag length $p^*$. If the t-statistic of lag $p^*$ is insignificant at some specified critical value, re-estimate the regression using the length $p^* - 1$.
- Repeat the process until the last lag is significant different from zero.
- In the pure autoregressive case, such a procedure will yield the true lag length with an asymptotic probability of unity, provided the initial choice of lag length include the true length.

# More on selection of lag Length

- Once a tentative lag length has been determined, diagnostic checking should be conducted.
  - Residual autocorrelation plot
  - Portmanteau tests on regression residuals
- If the regression equation does not omit a deterministic regressor in the data-generating process, it is possible to perform lag-length test using t-tests or F-tests. (Sims, Stock, and Watson, 1990)

# Spurious regression

- Consider a simple regression on two random walks

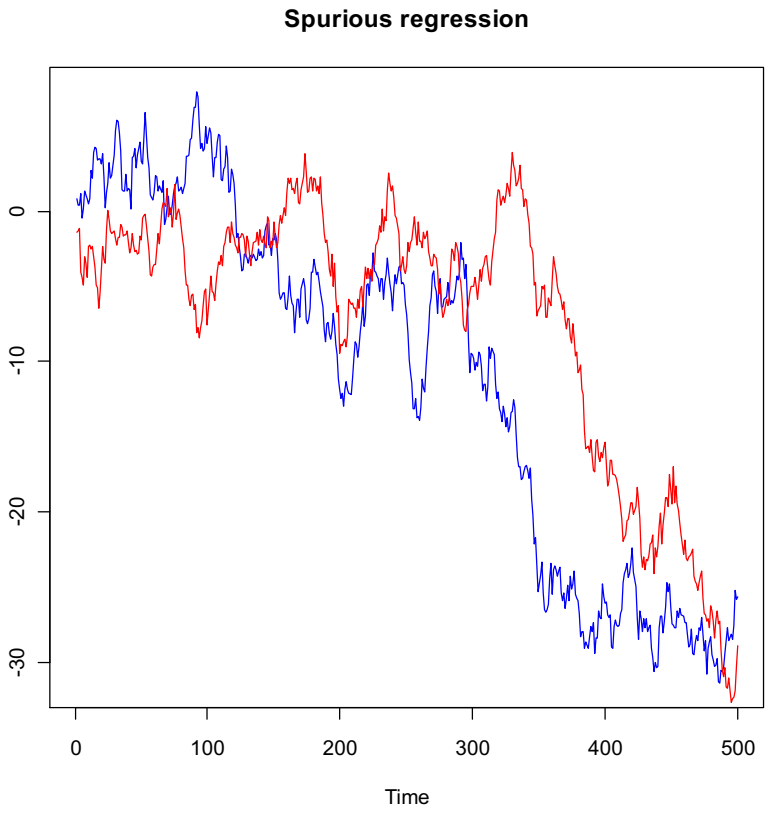
$$y_t = \alpha + \beta x_t + \epsilon_t,$$

  where $x_t = x_{t-1} + a_t$ and $y_t = y_{t-1} + e_t$ with $a_t$ and $e_t$ are mutually independent. For simplicity, let's assume that all error terms $\{\epsilon_t, a_t, e_t\}$ are IID random variables.

- What statistical inference can we know about a conventional simple regression?
  - $\hat{\beta} \to 0$ in probability
  - $R^2 \to 0$ in probability
  - $t_\beta = \frac{\hat{\beta} - 0}{se(\hat{\beta})}$ converges to Student $t$-distribution

# Simulation example

- library(lmtest)
- set.seed(1112)
- e1 <- rnorm(500)
- e2 <- rnorm(500)
- y1 <-cumsum(e1)
- y2 <-cumsum(e2)
- sr.reg<- lm(y1 ~ y2)
- sr.dw <- dwtest(sr.reg1)$statistic
- R-square is 0.58 and the Durbin-Watson statistic 0.0507 is close to zero, as expected.



# False statistical inference

- What if $x_t$ and $y_t$ are both random walks?
  - The absolute value of $t_\beta$ tends to become larger and larger as the series length $T$ increases;
  - Therefore, we will eventually rejects the null hypothesis that $\beta = 0$ with probability one as $T \to \infty$.
  - Additionally, $R^2$ does not converge to zero but to a random, positive number that varies from sample to sample.
- When a regression model appears to find relationship that do not really exist, it is called spurious regression.
- We have discuss in class that spurious regression can occur even when all variables are stationary. The risk can be far from negligible with stationary series that exhibit substantial series correlation.

# Forecasting

- The objective of forecasting is 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 $l$.
- According to Box and Jenkins (1976), $ARIMA$ models can be expressed into three explicit forms:

1. *Difference equation form*: in terms of an 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*: an infinite weighted sum of current and previous error terms,

$$X_{t+l} = \sum_0^\infty \psi_i a_{t+l-j} \quad (1),$$

where $\{\psi_j\}$ may be calculated using the techniques in §5.2.1 and §5.2.2 of Wei (2005) if $X_t$ is not stationary.

- Alternatively, we could rewrite eqn. $(1)$ as

$$X_{t+l} = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} + C_t(l) \quad (2),$$

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

$$C_t(l) = \sum_l^\infty \psi_j a_{t+l-j} \quad (3).$$

# Three explicit forms of $ARIMA$ models

3. *Weighted average of previous observations*: **(optional)**

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

$$= a_{t+l} + \pi_1 X_{t+l-1} + \cdots + \pi_{l-1} X_{t+1} + \sum_{j=l}^{\infty} \pi_j X_{t+l-j} \quad (4)$$

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

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

- We require the invertible $ARIMA(p, d, q)$ model to ensured that the $\{\pi_j\}$ coefficients form a convergent series.

# Minimum mean square error forecast

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

$$\hat{X}_t(l) = \psi_l^* a_t + \psi_{l+1}^* a_{t-1} + \psi_{l+2}^* a_{t-2} + \cdots \quad (5),$$

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

- The mean square error of the forecast in eqn. (5) is

$$E\left(X_{t+l} - \hat{X}_t(l)\right)^2 = \left(1 + \psi_1^2 + \cdots \psi_{l-1}^2\right)\sigma^2 + \sum_0^\infty \left(\psi_{l+j} - \psi_{l+j}^*\right)^2 \sigma^2$$

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

# Minimum mean square error forecast cont'd

- Using eqn. (1) and (5), 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 of MSE (1)

- The forecast error for lead time $l$ is

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

  - $E_t\big(e_t(l)\big) = 0$ so the forecast is unbiased
  - $Var\big(e_t(l)\big) = \big(1 + \psi_1^2 + \cdots + \psi_{l-1}^2\big)\sigma^2$ 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 $X_{t+1} + X_{t+2} + X_{t+3}$.

# Application
## Probability limits of the forecast

- The variance of $l$-step-ahead forecast error for any origin $t$ is the expected value of $e_t^2(l)$

$$E[e_t^2(l)] = E\{X_{t+l} - \hat{X}_t(l)\}^2 = \left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)\sigma^2$$

- *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 $\sqrt{\left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)\sigma^2}$

# Important facts of MSE (2)

- The minimum mean square error forecast $\hat{X}_t(l)$ is the conditional expectation of $X_{t+l}$ at time $t$.
  - $E\left(a_{t+j}\middle|X_t, X_{t-1}, \ldots\right) = E\left(a_{t+j}\right) = 0, \forall\, j > 0$
  - $E_t(X_{t+l}) = \hat{X}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \psi_{l+2} a_{t-2} + \cdots,$

    where $E(X_{t+l}|X_t, X_{t-1}, \ldots) = E_t(X_{t+l})$ is the conditional expectation of $X_{t+l}$ given knowledge of all $X$'s up to time $t$.

# Rules for Calculating the conditional expectations

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

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

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

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

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

# 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 $\psi_l$.

# 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.