Introduction to Financial Data Analysis

Week 11-Week 12: Classical Properties of Financial Time Series

Financial Time Series

- This chapter discuss the methods and linear models useful in modeling and forecasting financial time series.
- A **time series** is a **series** of data points indexed in **time** order. Most commonly, a **time series** is a sequence taken at successive equally spaced points in **time**. Thus it is a sequence of discrete-**time** data.
- Financial Time Series can exhibit very different patterns thus require different models in prediction and making inference.
- Financial Time Series usually use a series' own past values to forecast its future values. The purpose is to forecast.
- CAPM models, FF-models are using other variables to explain return difference. The purpose is to explain.

Financial Time Series Models

- The models introduced include
  - 1. simple autoregressive (AR) models
  - 2. simple moving average (MA) models
  - 3. unit-root (I) models including unit-root tests
  - 4. seasonal ARIMA models
- For each class of models, we study their fundamental properties, introduce methods for model selection, consider ways to produce prediction, and discuss their applications.
- Financial Time Series exhibit very different patterns as shown in the following examples.

Example: Financial Time Series

- The daily closing price of Apple stock from January 3, 2003 to April 5, 2010.
- The daily prices exhibit certain degrees of variability and show an upward movement during the sample period.

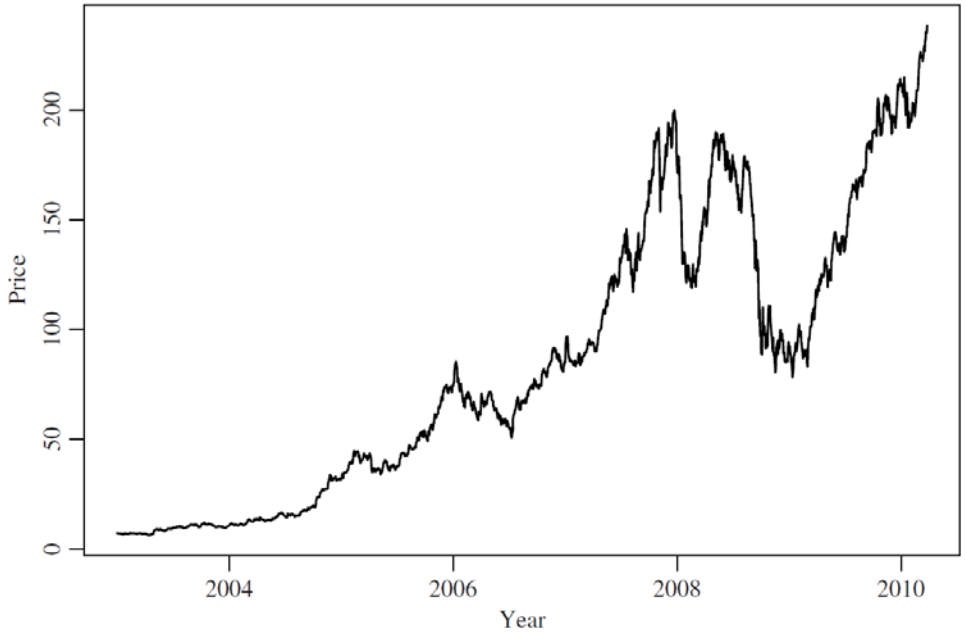


Figure 2.1. Daily closing prices of Apple stock from January 3, 2003 to April 5, 2010.

Example: Financial Time Series

- The quarterly earnings per share of Coca-Cola Company from 1983 to 2009. The quarters are marked in the plot.
- Besides an upward trend, the earnings also exhibit a clear annual pattern, referred to as *seasonality* in the time series analysis. It will be seen later that many economic and financial time series exhibit a clear seasonal pattern.

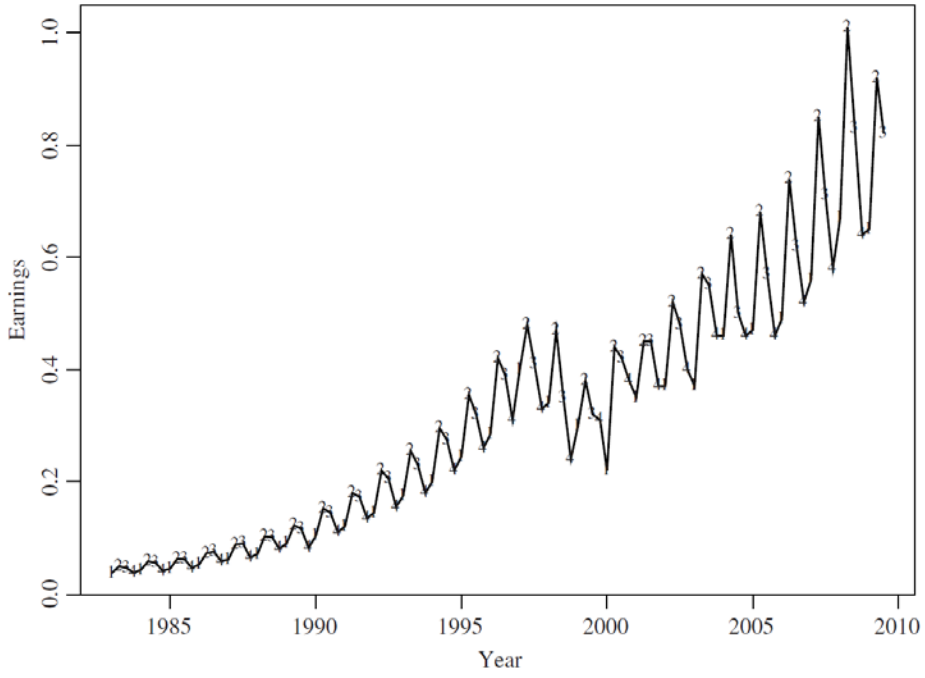


Figure 2.2. Quarterly earnings per share of Coca-Cola Company from the first quarter of 1983 to the third quarter of 2009.

Example: Financial Time Series

- Figure 2.3 gives the monthly log returns of the S&P 500 index from January 1926 to December 2009. From the plot, it is seen that the returns fluctuate around 0 and, except for a few extreme values, are within a fixed range.

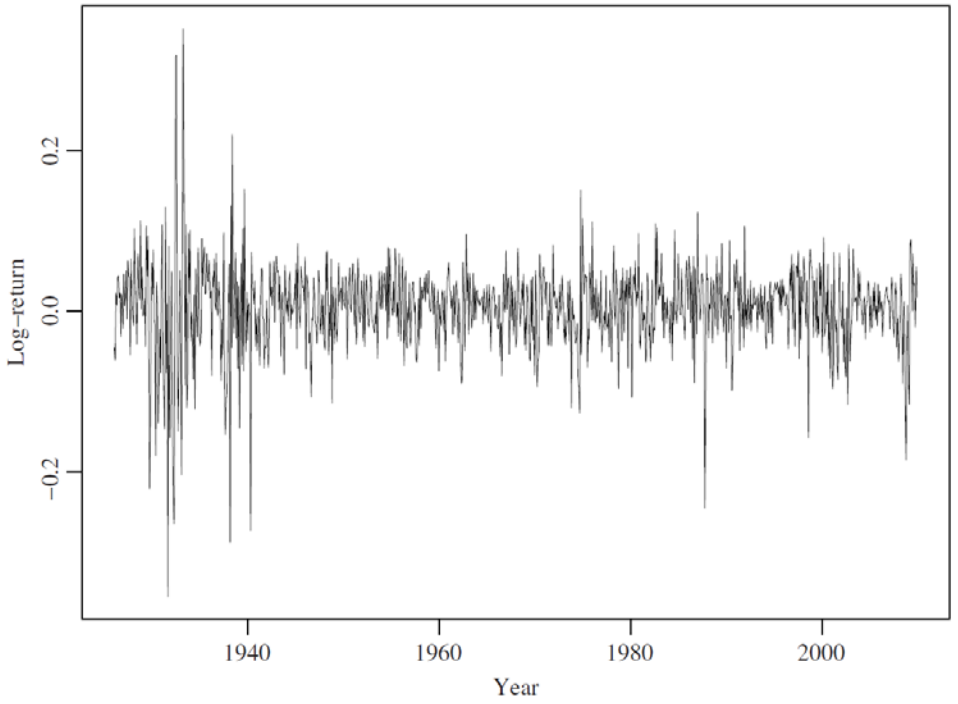


Figure 2.3. Monthly log returns of S&P 500 index from January 1926 to December 2009.

Example: Financial Time Series

- Two time series. They are the weekly US 3-month and 6-month treasury bill rates from January 2, 1959 to April 16, 2010.
- The upper plot is the 6-month rate and the lower plot is the 3-month rate.
- The two series move closely, and also exhibit certain differences.
- The 6-month rate was higher in general, but the 3-month rate appeared to be higher in some periods, for example, the early 1980s. This phenomenon is referred to as an *inverted yield curve* in the term structure of interest rates.

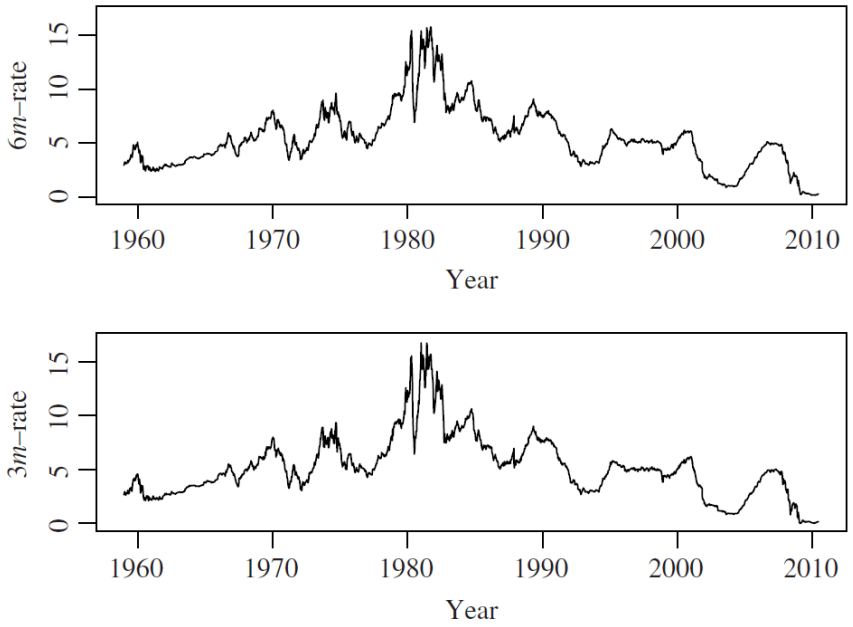


Figure 2.4. Weekly rates, from the secondary market, of the US 3-month and 6-month treasury bills from January 2, 1959 to April 16, 2010.

Example: Financial Time Series

- In these four examples, the series x_t is observed at (roughly) equally spaced time intervals. They are the examples of financial time series that we analyze in this chapter.
- Our goal is to study the dynamic dependence of the series (on its own past values) so that proper inference of the series can be made.

Stationary

- The foundation of statistical inference in time series analysis is the concept of weak stationarity.
- If a time series is stationary, that means, the properties of the time series is stable over time and thus, predictable.
- Statistically, stationary requires the distribution of future values are exactly the same as in the past.
- This is almost impossible in real financial time series.
- Weak Stationary: mean and variance are stable over time.

Stationary Example

- The mean is stable:
  - The monthly log returns of the S&P 500 index vary around 0 over time.
  - One can divide the time span into several subperiods, and the resulting sample means of the subperiods would all be close to 0.
  - In statistics, the mean of the returns is constant over time or simply the expected return is time invariant.
- The variance is stable:
  - Furthermore, except for the Great Depression era, the range of the monthly log returns is approximately $[-0.2, 0.2]$ throughout the sample span.
  - In statistics, this fact indicates that the variance of the returns is constant over time.

Non-stationary

- The quarterly earnings per share of Coca-Cola Company is non-stationary:
- 1. If one divides the time span into few subperiods, the resulting sample means differ substantially from one subperiod to the other. Therefore, the earnings are not weakly stationary. This is not surprising because one would expect that the quarterly earnings of a good company increase over time. The time
- 2. Figure also shows that the variability of the earnings increased over time. Therefore, the variance of quarterly earnings is also time varying.

Weak Stationary

- A time series is weakly stationary: if its first two moments (mean and variance) are time invariant.
- The weak stationarity is important because they provide the basic framework for prediction.
  - For the monthly log returns of the S&P 500 index, we can predict with reasonable confidence that the future monthly returns will be around 0 and vary between −0.2 and 0.2.
- For a given integer k, define the lag-k autocovariance of x_t *as* $\Gamma_k = \mathrm{Cov}(x_t, x_{t-k})$, Γ_k is also time-invariant for a stationary time series.
  - The linear dependency of x_t on its previous value is stable.

Autocorrelation Function (ACF)

- The ACF is the time series counter party of the correlation between two random variables
- The ACF evaluations the correlation between x_t and x_{t-k}.

$$\rho_k = \frac{\text{Cov}(x_t, x_{t-k})}{\sqrt{\text{Var}(x_t)\text{Var}(x_{t-k})}} = \frac{\text{Cov}(x_t, x_{t-k})}{\text{Var}(x_t)} = \frac{\gamma_k}{\gamma_0},$$

- In calculation, the lag k ACF is

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \quad 0 \le k < T - 1.$$

- If x_t weakly stationary, then $\hat{\rho}_k$ is asymptotically normal with mean 0 and variance $1/T$ *for* $k>0$.

Test for Significance in Autocorrelation

- Testing Individual ACF: $H_0 : \rho_k = 0 \text{ versus } H_a : \rho_k \neq 0.$
- The t statistics (t-ratio) is $\sqrt{T}\hat{\rho}_k$ and is standard normal distributed
- Reject H0 if $|t\text{-ratio}| > Z_{\alpha/2}$ and p is small.
- Where $Z_{\alpha/2}$ is the $100(1-\alpha/2)\text{th}$ percentile of the standard normal distribution.

Test for Significance in Autocorrelation

- Joint Test: $H_0 : \rho_1 = \cdots = \rho_m = 0$ $H_a : \rho_i \neq 0$

- Ljung and Box (1978)

$$Q(m) = T(T+2)\sum_{\ell=1}^{m} \frac{\hat{\rho}_\ell^2}{T-\ell}.$$

- In R, Box.test(ts, lag, type="Ljung")
- Reject the H0 if Q(m) is large and p-value is small.

Example: Autocorrelation in Crude Oil

- Example: Is there serial autocorrelation in crude oil returns?
- WTI crude oil price is highly driven by demand and supply of oil and is reflecting the economic growth status.
- The demand and supply growth are usually long term and tend to have trend.
- These fundamentals changes will be reflected in oil price.
- We examine whether there is autocorrelations in oil returns from month to month.

Example: Autocorrelation in Crude Oil

- Example: Can we use past crude oil returns to predict future oil returns?
- We examine the autocorrelation in the time series of crude oil returns.
- Step 1. Download data from St. Louis Fed Website
- https://fred.stlouisfed.org/series/DCcrudeWTICO
- Step 2. Visualize the returns and ACF function
- Step 3. Calculate the t stats that is used to test individual autocorrelation is not zero.
- Step 4. Joint test for the serial autocorrelation.

Autoregression Models

- When X_t has a statistically significant lag-1 autocorrelation, the lagged value X_{t-1} might be useful in predicting X_t. A simple model that makes use of such predictive power is
- $$x_t = \phi_0 + \phi_1 x_{t-1} + a_t, \qquad (1)$$
- where a_t is assumed to be a white noise series with mean 0 and variance σ_a^2
- This is the **AR model** of order 1 or simply an AR(1)
- The expected return and conditional variance conditional on past return is

$$E(x_t|x_{t-1}) = \phi_0 + \phi_1 x_{t-1}, \qquad \text{Var}(x_t|x_{t-1}) = \text{Var}(a_t) = \sigma_a^2.$$

Autoregression Model

- AR(p) model
- $$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t, \qquad (2)$$

- When is the AR(p) model weak stationary?

$$E(x_t) = \mu, \ \mathrm{Var}(x_t) = \gamma_0, \ \text{and} \ \mathrm{Cov}(x_t, x_{t-j}) = \gamma_j$$

- Mathematically, define $1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$ as the characteristic equation of model (2), if all the solutions of this equation are greater than 1 in modulus, then series x_t is stationary.

Autoregression Model ACF Examples

- AR(1) Model:

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t,$$

- AR(1) ACF:
- Rho(l) = gamma(l)/gamma(0)

$$\rho_\ell = \phi_1 \rho_{\ell-1}, \quad \text{for} \quad \ell > 0.$$

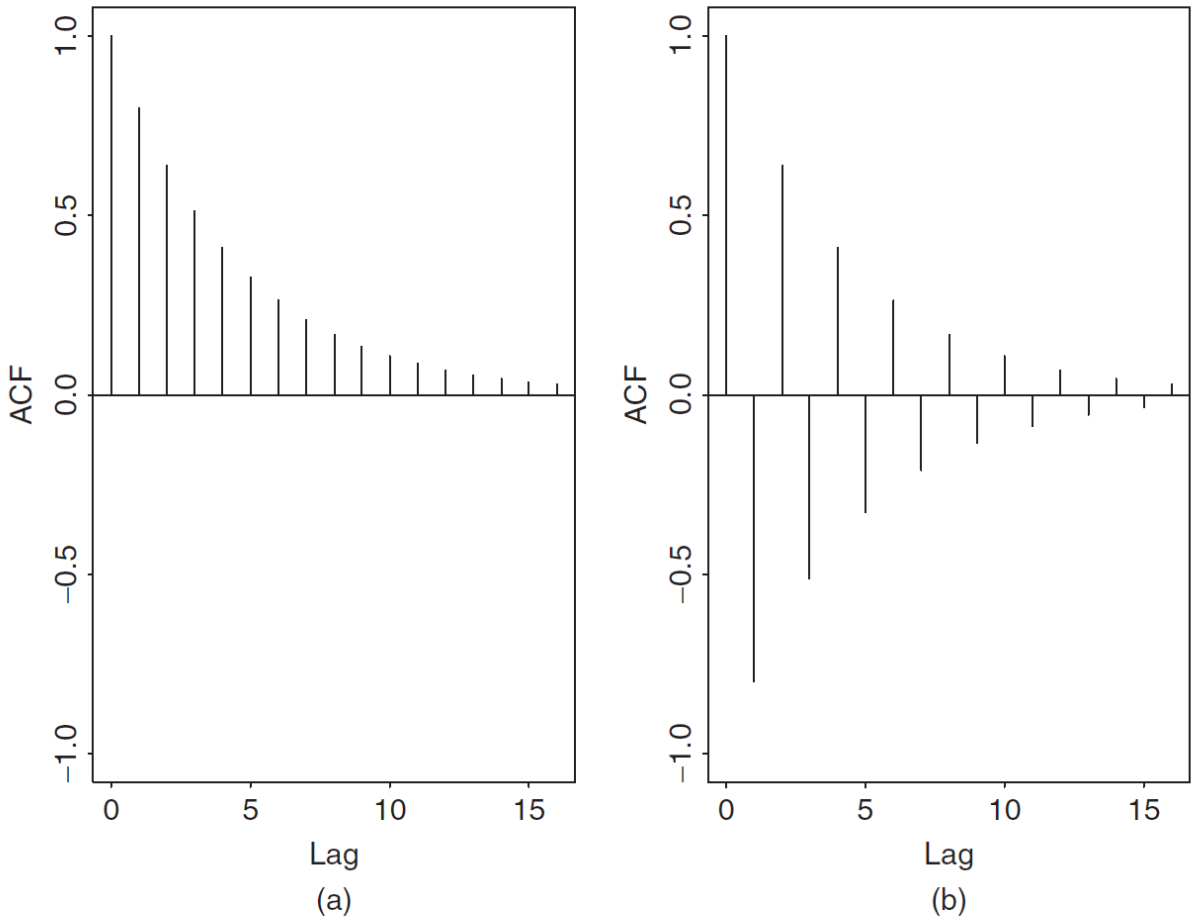


Figure 2.3 Autocorrelation function of an AR(1) model: (a) for $\phi_1 = 0.8$ and (b) for $\phi_1 = -0.8$.

Autoregression Model ACF Examples

- AR(2) Model:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + a_t.$$

- AR(2) ACF:

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}$$

$$\rho_\ell = \phi_1 \rho_{\ell-1} + \phi_2 \rho_{\ell-2}, \quad \ell \geq 2.$$

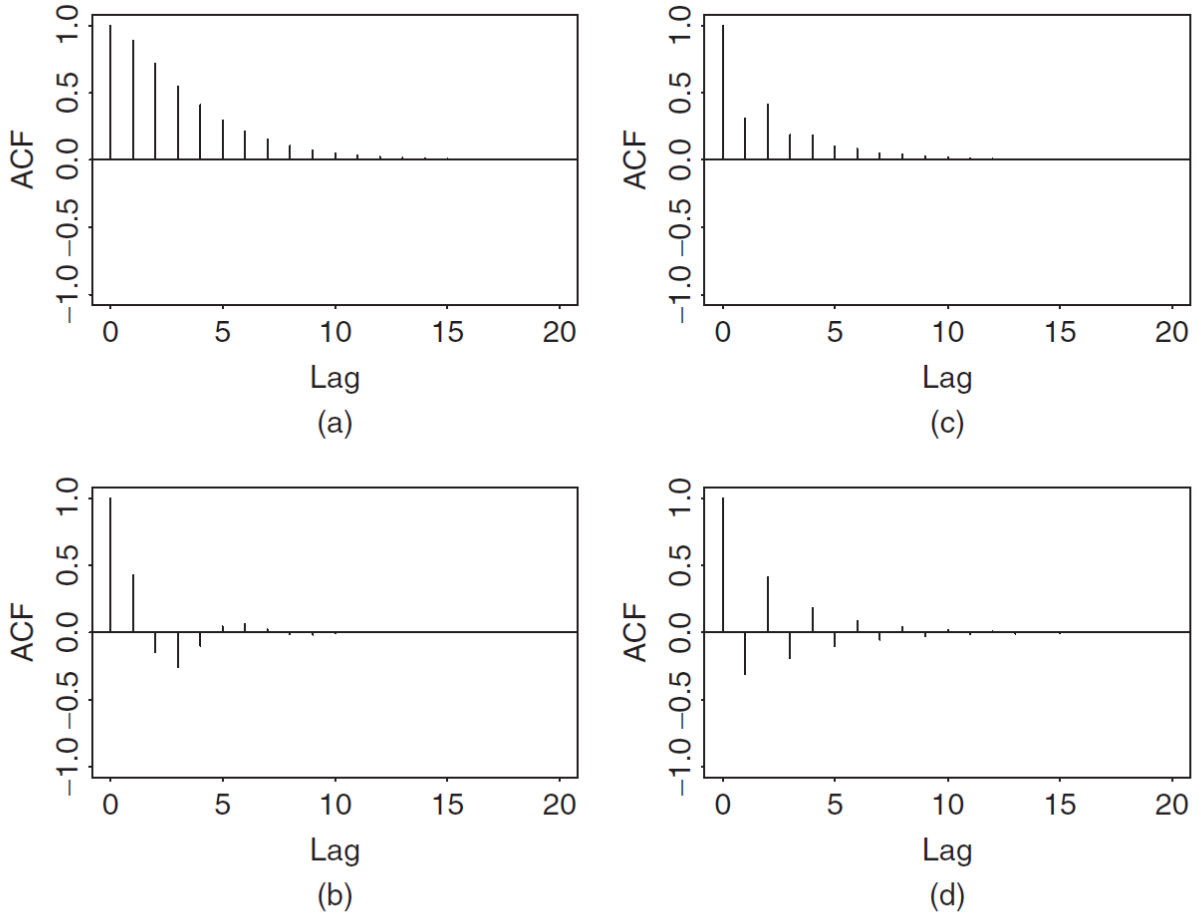


Figure 2.4 Autocorrelation function of an AR(2) model: (a) $\phi_1 = 1.2$ and $\phi_2 = -0.35$, (b) $\phi_1 = 0.6$ and $\phi_2 = -0.4$, (c) $\phi_1 = 0.2$ and $\phi_2 = 0.35$, and (d) $\phi_1 = -0.2$ and $\phi_2 = 0.35$.

Estimations of Autoregression Model

- For a stationary AR(p) model, we have

$$E(x_t) = \frac{\phi_0}{1 - \phi_1 - \cdots - \phi_p}$$

- In R, use arima(data,c(p,0,0)) to estimate the AR(p) model
- In the estimation, the fitted model is

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-1} - \mu) + \ldots + \phi_p(x_{t-p} - \mu)$$

- The intercept coefficient estimated is

$$\mu = \frac{\phi_0}{1 - \phi_1 - \phi_2 - \ldots - \phi_p} = E(x_t)$$

Estimations of Autoregression Model (Optional)

- Solution of unconditional variance of a stationary AR(p) is mathematical tedious and is not required for this course.
- Define the lag-k autocovariance of x_t as $\Gamma_k = \mathrm{Cov}(x_t, x_{t-k})$
- For AR(1):

$$\mathrm{Var}(r_t) = \gamma_0 = \frac{\sigma^2}{1-\phi_1^2} \qquad \text{and} \qquad \gamma_\ell = \phi_1 \gamma_{\ell-1}, \quad \text{for} \quad \ell > 0.$$

Estimations of Autoregression Model (Optional)

- For AR(2) process y(t), for simplicity, we remove the mean from the y(t):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t.$$

$$\begin{aligned}
\gamma(\tau) &= \mathrm{E}((y_t - \mu)(y_{t-\tau} - \mu)) = \mathrm{E}(y_t y_{t-\tau}) \\
&= \mathrm{E}\Big((\phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t) y_{t-\tau}\Big) \\
&= \phi_1 \mathrm{E}(y_{t-1} y_{t-\tau}) + \phi_2 \mathrm{E}(y_{t-2} y_{t-\tau}) + \mathrm{E}(\epsilon_t y_{t-\tau}) \\
&= \begin{cases} \phi_1 \gamma(\tau - 1) + \phi_2 \gamma(\tau - 2), & \text{for } \tau \neq 0, \\ \phi_1 \gamma(\tau - 1) + \phi_2 \gamma(\tau - 2) + \sigma_\epsilon^2, & \text{for } \tau = 0. \end{cases}
\end{aligned}$$

Estimations of Autoregression Model (Optional)

- The initial conditions can be solved by the following three equations:

$$\gamma(0) = \phi_1\gamma(1) + \phi_2\gamma(2) + \sigma_\epsilon^2,$$

$$\gamma(1) = \phi_1\gamma(0) + \phi_2\gamma(1),$$

$$\gamma(2) = \phi_1\gamma(1) + \phi_2\gamma(0).$$

Estimations of Autoregression Model (Optional)

- The initial conditions are given by

$$\gamma(0) = \left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_\epsilon^2}{(1-\phi_2)^2-\phi_1^2},$$
$$\gamma(1) = \frac{\phi_1}{1-\phi_2}\gamma(0) = \left(\frac{\phi_1}{1-\phi_2}\right)\left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_\epsilon^2}{(1-\phi_2)^2-\phi_1^2}.$$

- Furthermore, we have

$$\gamma(\tau) = \phi_1\gamma(\tau-1) + \phi_2\gamma(\tau-2), \qquad \text{for } \tau = 2, 3, \cdots.$$

Identify AR Models in Practice

- In application, the order p of an AR time series is unknown.
- Two general approaches are available for determining the value of p.
  - 1. Use the partial autocorrelation function (PACF)
  - 2. Use some information criteria.

Partial Autocorrelation Function (PACF)

- Consider the following AR models in consecutive orders

$$
\begin{aligned}
x_t &= \phi_{0,1} + \phi_{1,1}x_{t-1} + e_{1t}, \\
x_t &= \phi_{0,2} + \phi_{1,2}x_{t-1} + \phi_{2,2}x_{t-2} + e_{2t}, \\
x_t &= \phi_{0,3} + \phi_{1,3}x_{t-1} + \phi_{2,3}x_{t-2} + \phi_{3,3}x_{t-3} + e_{3t}, \\
x_t &= \phi_{0,4} + \phi_{1,4}x_{t-1} + \phi_{2,4}x_{t-2} + \phi_{3,4}x_{t-3} + \phi_{4,4}x_{t-4} + e_{4t}, \\
&\vdots
\end{aligned}
$$

- These models are in the form of a multiple linear regression and can be estimated by the least squares (LS) method.

Partial Autocorrelation Function (PACF)

- The estimate of the $\hat{\phi}_{1,1}$ from the first equation is called the lag-1 sample PACF.
- The estimate of $\hat{\phi}_{2,2}$ from the second equation is called the lag-2 sample PACF, and so on.
- The lag-2 sample PACF $\hat{\phi}_{2,2}$ shows the added contribution of x_{t-2} to the AR(1) model. The lag-3 sample PACF shows the added contribution of x_{t-3} over the AR(2) model.
- Therefore, for an AR(p) model, the lag-p sample PACF should not be 0, but the PACF of lag-j with j>p should be 0.

Information Criteria Method

- Information Criteria Method: There are several information criteria available to determine the order p of an AR process. All of them are likelihood based.
- Likelihood: Statistically, the likelihood is the probability of observing a sample of data given a set of model parameter values. By definition, the higher the likelihood, the better the model is.
- More complicated models, eg, with a lot of control variable x, usually give higher likelihood. Therefore, information criteria statistics are deviated from the likelihood to punish the complexity of the model, this part usually called the penalty function.
- Different IC use different penalty function.
- Information Criteria can be used broadly to compare different regression models efficiency, even if models are very different in their settings. (Linear/Non-linear, different independent variables, etc.)

Information Criteria Method

- Two popular information criteria: AIC, BIC
- AIC (Akaike Information Criterion) (Akaike, 1973)

$$\text{AIC} = \frac{-2}{T}\ln(\text{likelihood}) + \frac{2}{T} \times (\text{number of parameters}),$$

- For a Gaussian $\text{AR}(\ell)$ model,

$$\text{AIC}(\ell) = \ln(\tilde{\sigma}_\ell^2) + \frac{2\ell}{T},$$

- The lower the AIC, the better the model.

Information Criteria Method

- BIC (Bayesian information criterion), for a Gaussian AR(l) model

$$\mathrm{BIC}(\ell) = \ln(\tilde{\sigma}_\ell^2) + \frac{\ell \ln(T)}{T}.$$

- The lower the BIC, the better the model.
- AIC vs. BIC: The penalty for each parameter used is 2 for AIC and ln(T) for BIC.
- Thus, compared with AIC, BIC tends to select a lower AR model when the sample size is moderate or large.

Information Criteria Method

- Selection Rule: To use AIC to select an AR model in practice, one computes AIC(l) for $\ell = 0, \ldots, P$, where P is a prespecified positive integer and selects the order k that has the minimum AIC value. The same rule applies to BIC.

Model Checking

- A fitted model must be examined carefully to check for possible model inadequacy. If the model is adequate, then the residual series should behave as a white noise.
- Use Ljung–Box statistics to check the residual series is purely random or not.

$$H_0 : \rho_1 = \cdots = \rho_m = 0$$

- The function in R is Box.test()
- The null hypothesis is no serial correlation.

Example: Selecting Models

- Identify the order of the AR model for the GNP Growth.
- See “GNP_AR.html”

  - Step 1. Load the GNP data, and visualize the GNP and its growth.
  - Step 2. Look into the PACF to select the AR model order
  - Step 3. Look into the IC to select the AR(p) model
  - Step 4. Examine the coefficients of the AR(3) model
  - Step 5. Modify the AR(9) model and restrict some coefficients to 0
  - Step 6. Model checking for residual serial correlations

Forecast

- Forecasting is an important application of time series analysis.
- For the AR(p) model, suppose we are at time h and is interested in forecasting x_{h+l}
- For an AR(p) Model:

$$x_{h+1} = \phi_0 + \phi_1 x_h + \cdots + \phi_p x_{h+1-p} + a_{h+1}.$$

- One step forecast: The point forecast of x_{h+1} given F_h is

$$\hat{x}_h(1) = E(x_{h+1}|F_h) = \phi_0 + \sum_{i=1}^{p} \phi_i x_{h+1-i}$$

- Where F_h is the all the information available at h

Forecast

- The forecast error is

$$e_h(1) = x_{h+1} - \hat{x}_h(1) = a_{h+1}.$$

- The variance of the forecast error is

$$\text{Var}[e_h(1)] = \text{Var}(a_{h+1}) = \sigma_a^2$$

- The 95% confidence interval for your forecast x_{h+1} is
-

$$\hat{x}_h(1) \pm 1.96 \times \sigma_a$$

Two Step Forecast

- Two Step Forecast:

$$x_{h+2} = \phi_0 + \phi_1 x_{h+1} + \cdots + \phi_p x_{h+2-p} + a_{h+2}$$

- The conditional expectation is

$$\hat{x}_h(2) = E(x_{h+2}|F_h) = \phi_0 + \phi_1 \hat{x}_h(1) + \phi_2 x_h + \cdots + \phi_p x_{h+2-p}$$

- The associated forecast error is

$$e_h(2) = x_{h+2} - \hat{x}_h(2) = \phi_1[x_{h+1} - \hat{x}_h(1)] + a_{h+2} = a_{h+2} + \phi_1 a_{h+1}.$$

- The variance of the forecast error is

$$\text{Var}[e_h(2)] = (1 + \phi_1^2)\sigma_a^2$$

- Obviously, the forecast uncertainty increases as the forecast horizon increases

$$\text{Var}[e_h(2)] \geq \text{Var}[e_h(1)]$$

Multiple Step Forecast

- The l-step ahead forecast is

$$\hat{x}_h(\ell) = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{x}_h(\ell - i),$$

- This forecast can be computed recursively using forecasts

$$\hat{x}_h(i) \text{ for } i = 1, \ldots, \ell - 1$$

- Mathematically, it can be shown that $\hat{x}_h(\ell)$ approaches $E(x_t)$ as $\ell \to \infty$
- This means, such a series long-term point forecast approaches its unconditional mean. This property is referred to as the **mean reversion** in the finance literature.
- The variance of the forecast error then approaches the unconditional variance of x_t

Example: Forecast GNP Growth (Continued)

- See “GNP_AR.html”

- Step 7. Generate 3 Years (12 Quarters) forecast of GNP Growth
- Use predict(model, n.ahead) to generate forecast in AR model

Simple Moving Average Model

- MA model
- MA(1): $x_t = c_0 + a_t - \theta_1 a_{t-1}$
- MA(q): $x_t = c_0 + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$

  Where c_0 is constant and $\{a_t\}$ are white noise series
- White Noise: $\{a_t\}$ are iid normal distributed with mean 0 and variance σ^2

MA Model

- MA models are always weakly stationary because they are finite linear combinations of a white noise sequence for which the first two moments are time invariant.
- The Mean is $E(x_t) = c_0$
- The Variance is $\text{Var}(x_t) = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_a^2$
- ACF: MA(q) has non-zero ACF only up to q'th order. Therefore, we call the MA model a "finite memory" model.
- In R, use arima(data, c(0,0,q)) to estimate a MA(q) model.

Identify MA Model

- Method 1: The ACF is useful in identifying the order of an MA model.
  - Detecting the order of MA(q) by looking into the non-zero ACF order.
- Method 2: Use IC to select correct order of the model.

Integrated Model

- The most simple Integrated model is the unit root model:
- Unit Root Model: $y_t = y_{t-1} + c + e_t$
- Unit Root process is non-stationary.
- Non-stationary data cannot be modeled or forecasted.
- For c=0, the process is called a random walk

$$y_t = y_{t-1} + \epsilon_t$$

$$y_{t-1} = y_{t-2} + \epsilon_{t-1}$$

$$\Rightarrow y_t = \sum_{s=0}^{t} \epsilon_s$$

where $y_0 = \epsilon_0$

Transformation of The Random Walk

- Take the first order difference:

$$\Delta y_t = y_t - y_{t-1} = \epsilon_t$$

- The difference process is stationary.
- In time series literature, if y_t has a unit root and while Δy_t is stationary, then y_t is called the integrated of first order, I(1).
- If y_t is non-stationary but the kth difference of y_t is stationary, then y_t is called the integrated of order k, I(k).
- I(k) model are less frequently used in finance literatures .
- I(k) indicated the increments of the time series have a trend.

Example of Unit-Root: Level of SPY

Figure: SPY ETF - Violation of Weak Stationarity

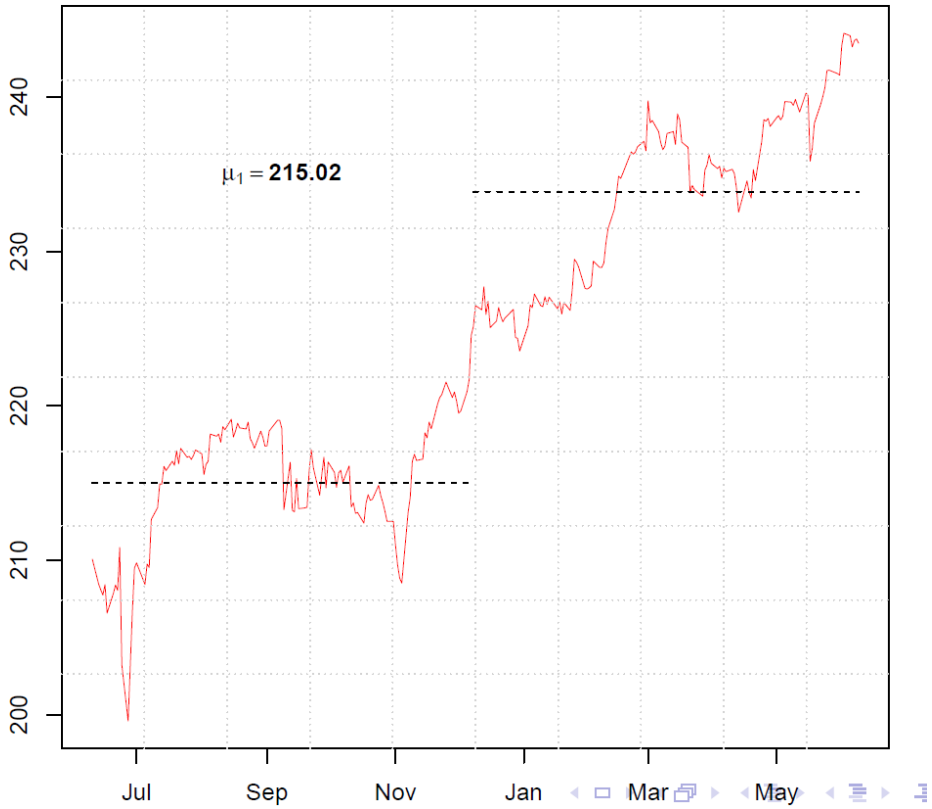
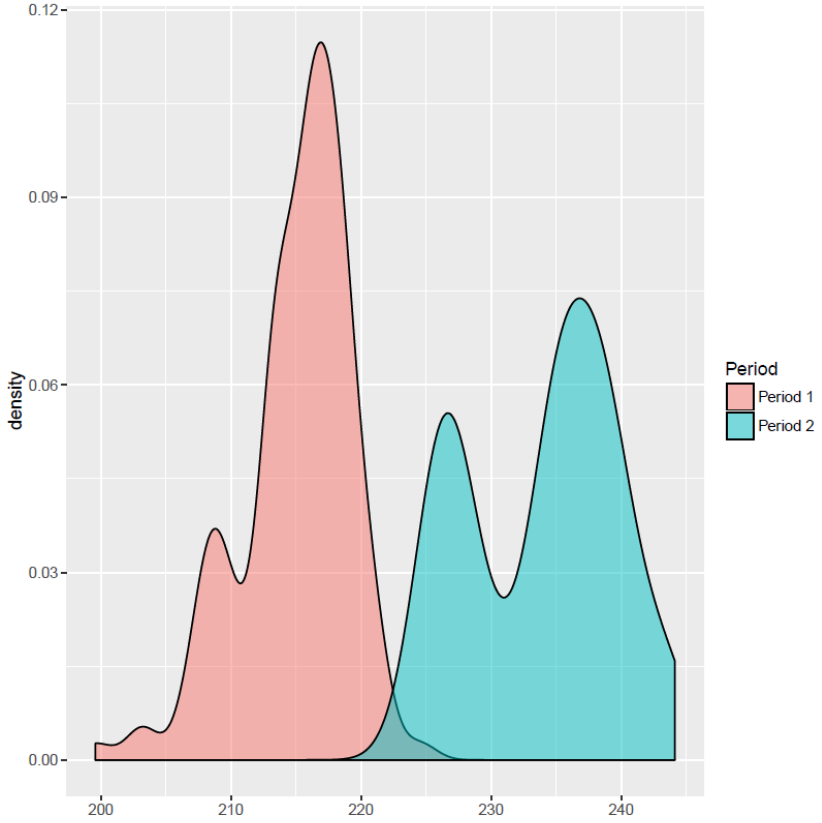


Figure: SPY ETF - Violation of Strict Stationarity



Example: Return of SPY is Stationary

Figure: SPY ETF Returns - Weak Stationarity

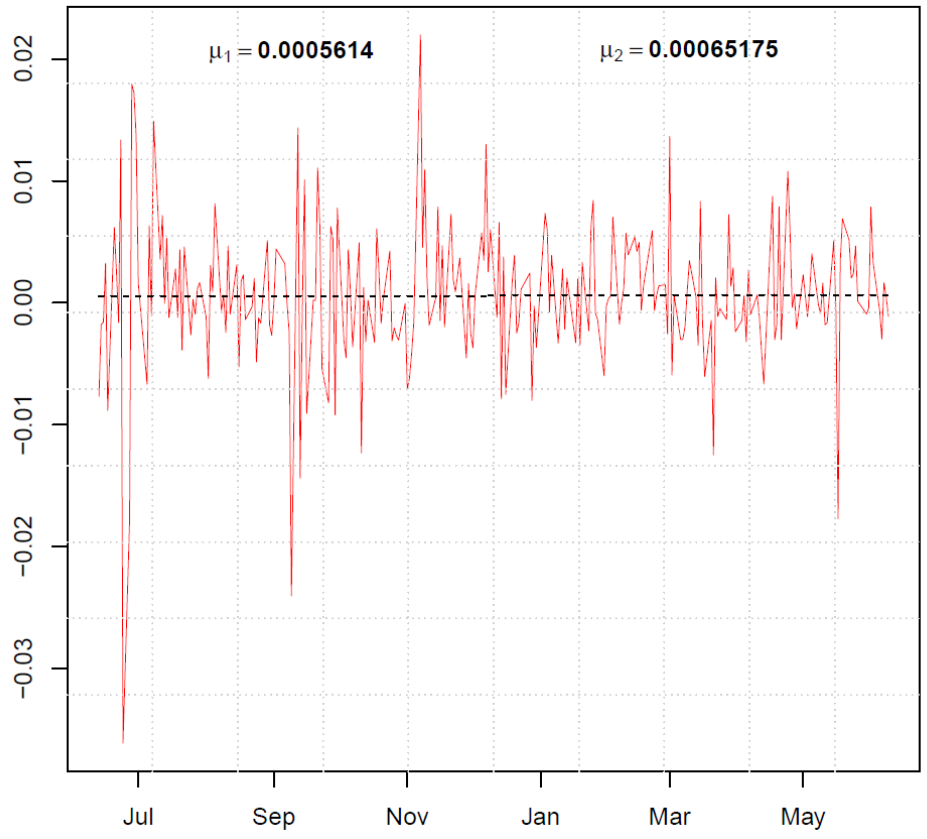
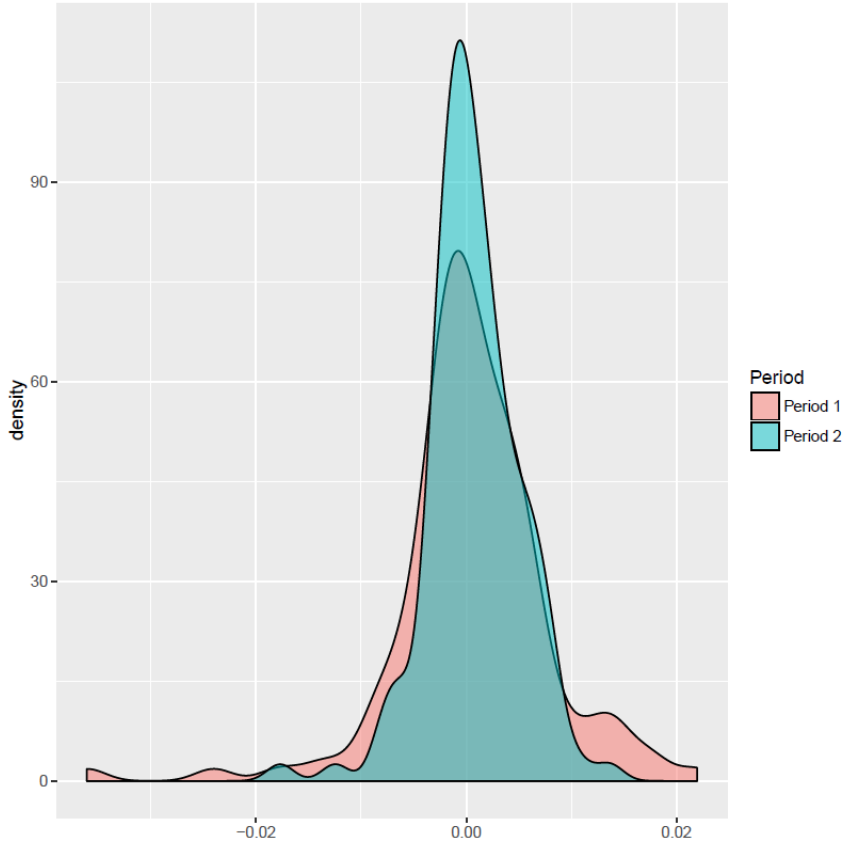


Figure: SPY ETF Returns - Strict Stationarity



Test for Unit-Root

- It is important to plot the time series before running any tests of unit root.
- May start with checking the correlation between y(t) and y(t-1)

```
> cor(P_daily[-1],lag(P_daily)[-1])

           SPY.Close
SPY.Close    0.99238
```

- Formal test of Unit-Root: Augmented Dickey-Fuller Test
  - Null hypothesis is a unit root exist.
  - In R, use the function adfTest() from library(fUnitRoots)

Example: GNP Growth (Continued)

- See “GNP_AR.html”
- Step 8. Test the unit root in the original GNP data

ARIMA Model

- ARIMA is a combination of AR, MA and Integrated Model.
- ARIMA(p,d,q) is a model with:
  - p: order of the autoregressive part
  - d: degree of first differencing involved
  - q: order of the moving average part
- The following is a ARIMA(p,d,q) model

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

- Where y'_t is difference series (it has been differenced d times)

ARIMA Model Selection

- Selection the order of ARIMA model is difficult, happily, we have auto.arima() function from the library(forecast) can do this for us. The function forecast() gives us forecast of the model, with default confidence interval level = (80,95) shown in picture.
- auto.arima() function in R uses a variation of the Hyndman-Khandakar (Hyndman & Khandakar 2008) algorithm which combines unit root tests, minimisation of the AICs and MLE to obtain an ARIMA model.
- Instead, arima() function allows us to select our own model.

ARIMA Model Selection

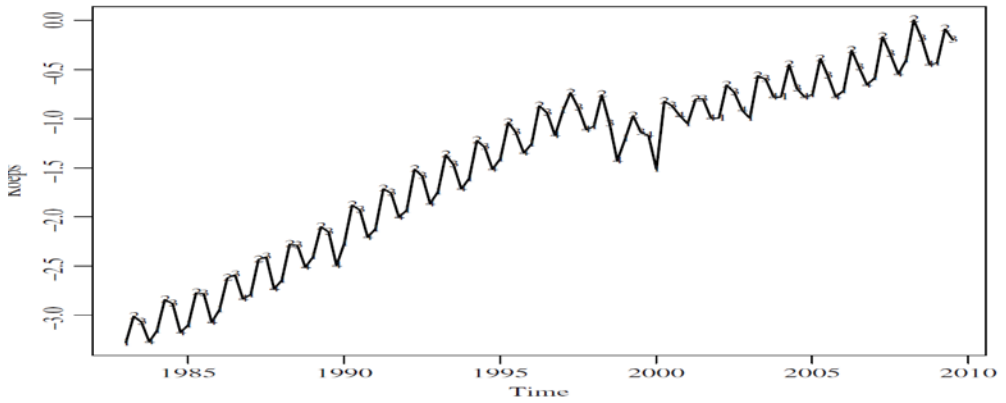
- 1. Plot the data and identify any unusual observations.
- 2. If the data are non-stationary, take first differences of the data until the data are stationary.
- 3. Examine the ACF/PACF: Is an ARIMA(p,d,0) or ARIMA(0,d,q) model appropriate?
  - Spikes in the PACF less than order p indicates AR(p) model
  - Spikes in ACF less than order q indicate MA(q) model.
- 4. Try your chosen model(s), and use the AIC to search for a better model.
- 5. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals. If they do not look like white noise, try a modified model.
- 6. Once the residuals look like white noise, calculate forecasts.

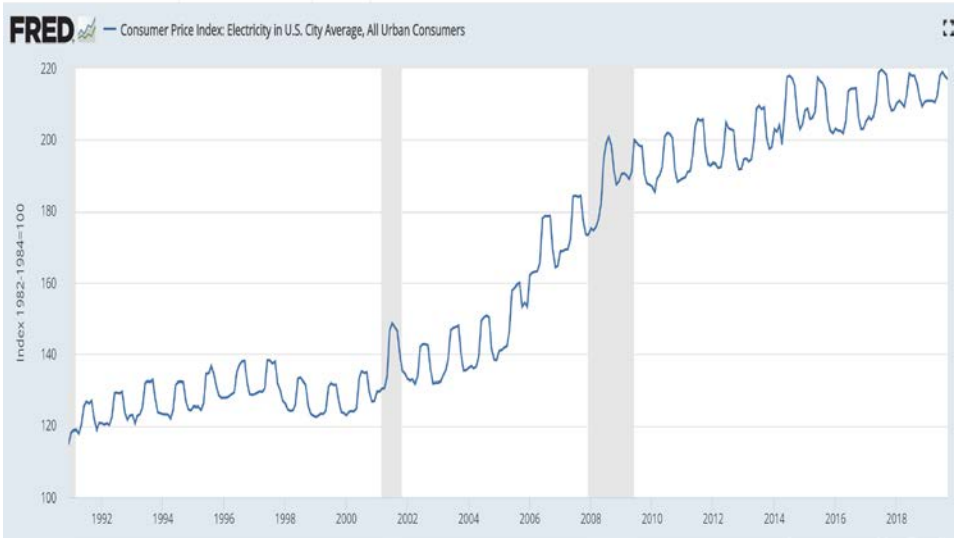
ARIMA Model Selection

- Example: “crude_oil_acf.html”
- Step 5. Use auto.arima() function to fit the crude oil return
- Step 6. Use arima to select the model by ourself.
- 1. Plot the price and identify unitroot.
- 2. Check the existence of unitroot of price. Check the existence of unitroot of return.
- 3. Examine the ACF/PACF of returns: Is an ARIMA(p,d,0) or ARIMA(0,d,q) model appropriate? Only focus on low orders.
- 4. Try different combinations of p and q and use the AICs to search for a better model.

Seasonal Models

- Some financial time series such as quarterly earnings per share of a company exhibits certain cyclical or periodic behavior. This type of series is called a seasonal time series.
- Quarterly earnings per share of the Coca-Cola Company (top): The seasonal pattern repeats itself every year so that the periodicity of the series is 4.
- Electricity Price (bottom): The periodicity is 12 months.





Seasonality

- Seasonality refers to a regular pattern of changes that repeat for S time periods, where S defines the number of timer periods until the pattern repeats again.
- Seasonality makes the time series nonstationary.
- Similar to first order difference to remove trend, taking a different of the time series at the periodicity usually can remove the seasonality and make the timeseries stationary.

Seasonal Differencing

- **Seasonal differencing** is defined as a difference between a value and a value with lag that is a multiple of *S*.
  - With $S = 12$, which may occur with monthly data, a seasonal difference is $x(t) - x(t-12)$
    - The differences (from the previous year) may be about the same for each month of the year giving us a stationary series.
  - With $S = 4$, which may occur with quarterly data, a seasonal difference is $x(t) - x(t-4)$
- Seasonal differencing removes non-stationarity.

After Seasonal Differencing

- Non-seasonal behavior will still matter…
- With seasonal data, it is likely that short run non-seasonal components will still contribute to the model.
- For example, in the monthly sales of cooling funs, sales in the previous month or two, along with the sales from the same month a year ago, may help predict this month's sales.
- The seasonal-ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model.
- The seasonal-ARIMA model is a linear combination of seasonal model, AR model, MA model and Unit Root Model.

Seasonal ARIMA Model (Advanced)

- The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model.
- One shorthand notation for the model is

$$\text{ARIMA}\ (p, d, q) \times (P, D, Q)S$$

- p = non-seasonal AR order
- d = non-seasonal differencing
- q = non-seasonal MA order
- P = seasonal AR order
- D = seasonal differencing
- Q = seasonal MA order
- S = time span of repeating seasonal pattern

Seasonal ARIMA

- In a seasonal ARIMA model, AR and MA terms predict x_t using data values and errors at times with lags that are multiples of S.
  - With monthly data and an annual trend ($S = 12$), a seasonal first order autoregressive model would use x_{t-12} to predict x_t. For example, if we were selling ice cream, we might predict August sales using last years August sales. Similarly, you could use the past two Augusts and include x_{t-24}. Then this is seasonal AR(2) model.
  - A seasonal first order MA(1) model (with $S = 12$) would use e(t−12) as a predictor. A seasonal second order MA(2) model would use e(t−12) and e(t−24), where e are iid normal random variables.

How to identify a Seasonal ARIMA model

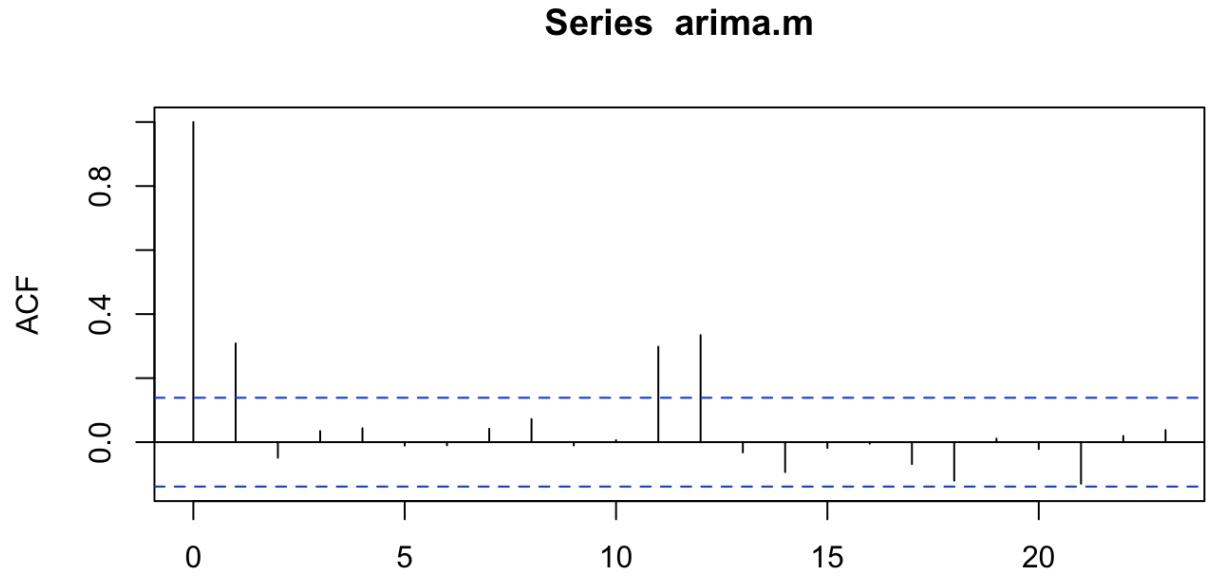
- 1. Do a time series plot of the data. Examine it for global trends and seasonality.
- 2. Do any necessary differencing
  - If there is seasonality and no trend take a difference of lag *S*. For example, take a 12th difference for monthly data with seasonality.
  - If there is a linear trend and no obvious seasonality, take a first difference. If there is a curved trend, consider a transformation of the data before differencing.
  - If there is both trend and seasonality, apply both a non-seasonal and seasonal difference to the data, as two successive operations. For example:
  - `diff1 = diff(x, 1)`
  - `diff1_and_12 = diff(diff1, 12)`

How to identify a seasonal ARIMA model

- 3. Examine the ACF and PACF of the differenced data (if necessary). We can begin to make some basic guesses about the most appropriate model at this time.
  - *non-seasonal terms*: Examine the early lags(1, 2, 3, ⋯) to judge non-seasonal terms. Spikes in the ACF (at low lags) indicate non-seasonal MA terms. Spikes in the PACF (at low lags) indicated possible non-seasonal AR terms.
  - *Seasonal terms*: Examine the patterns across lags that are multiples of *S*. For example, for monthly data, look at lags 12, 24, 36 (probably won't need to look much past the first two or three seasonal multiples). Judge the ACF and PACF at the seasonal lags in the same way you do for the earlier lags.

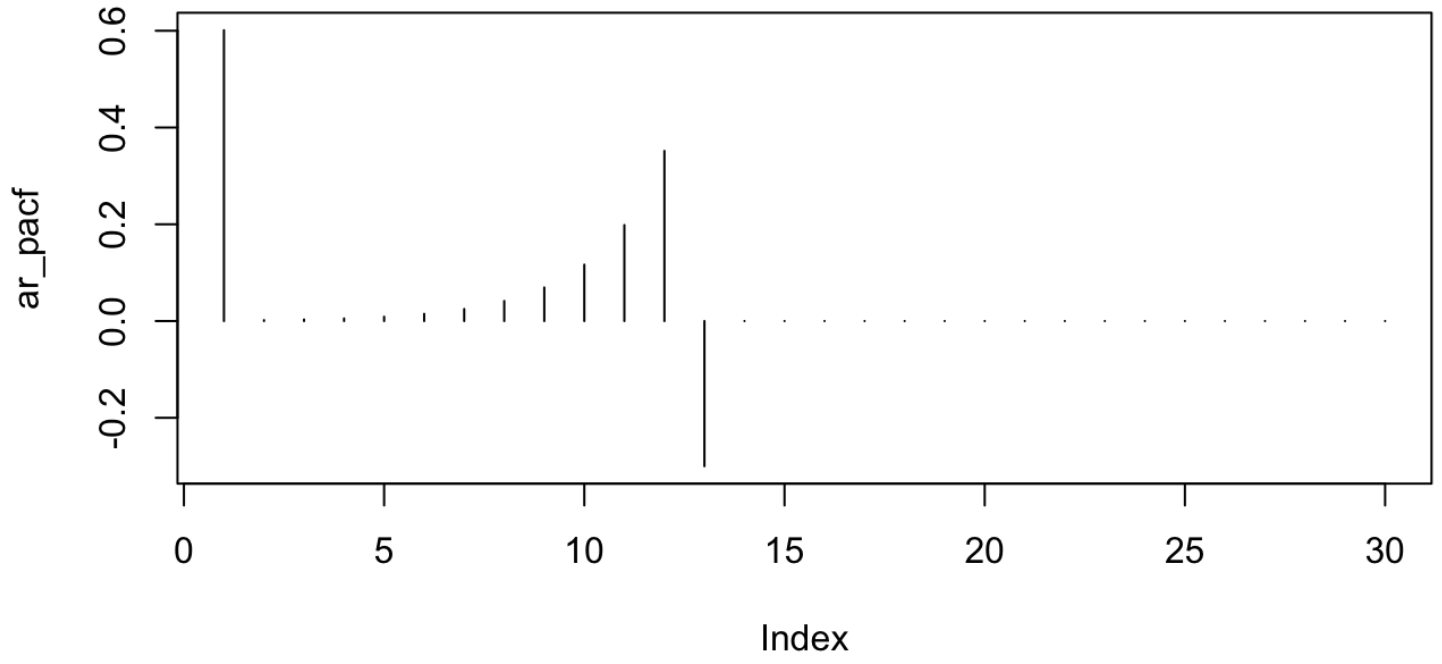
How to identify a Seasonal ARIMA model

- The AR model and MA model will have different ACF and PACF shape.
- ARIMA(0, 0, 1) x $(0, 0, 1)_{12}$
- The model includes seasonal and non-seasonal MA(1) terms, no differencing and no AR terms, and the seasonal period is $S = 12$.
- The spike at lag 11 is because y(t-11) is a function of e(t-12) due to the non-seasonal MA(1), while y(t) also contains e(t-12) due to the seasonal MA(1)



How to identify a Seasonal ARIMA model

- Example 2: ARIMA(1, 0, 0) x $(1, 0, 0)_{12}$
- The model include seasonal and non-seasonal AR(1) model.



Example: Coca-Cola Earnings

- See “Earnings.html”
- Step 1. Load the earnings eps.cola.Rdata and visualize the data
- Step 2. Take the first difference to detrend and take the seasonal difference. Check the ACF and PACF to detect model order.
- Step 3. Estimate the ARIMA model with seasonality
- Step 4. Model checking
- Step 5. Out-of-sample forecasting

Cointegration Time Series

- 2 Time Series
- Weekly US 3-month and 6-month treasury bill rates from January 2, 1959 to April 16, 2010.
- The two series move closely, and also exhibit certain differences.
- Cointegration try to model the relationship between two time series that moves together.

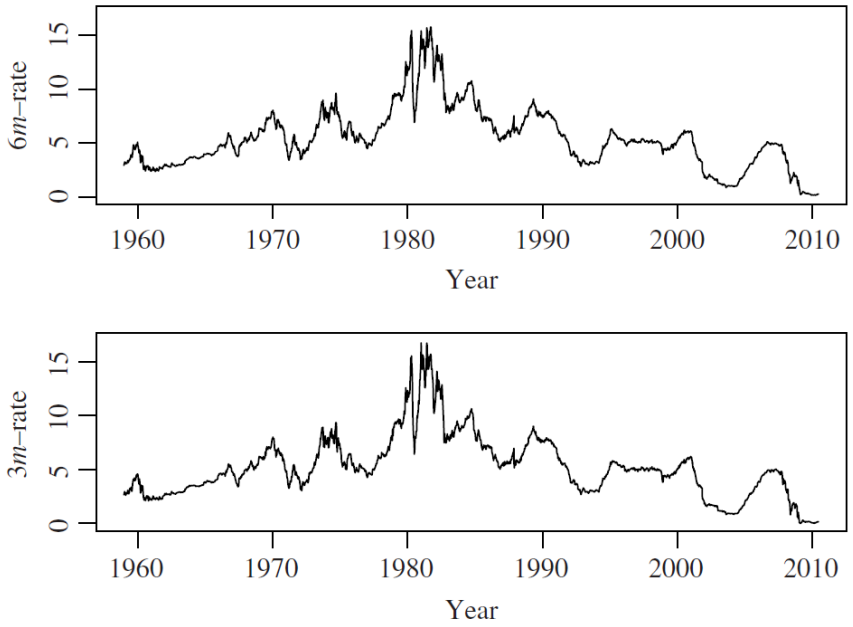


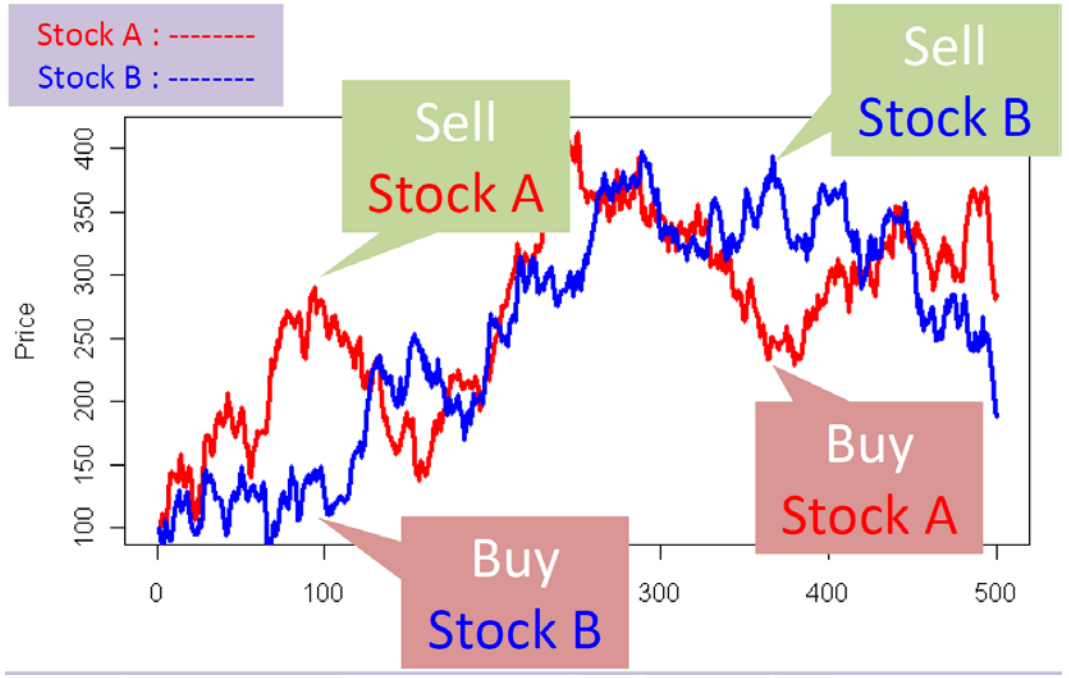
Figure 2.4. Weekly rates, from the secondary market, of the US 3-month and 6-month treasury bills from January 2, 1959 to April 16, 2010.

Cointegration and Pair Trading

- Cointegration is the key of the important class of investing strategy: Pair Trading.
- What is pair trading?
  - Quantitative group at Morgan Stanley Around 1980s
- Pair Trading is market neutral trading strategy: a market neutral trading strategy enabling traders to profit from virtually any market conditions: uptrend, downtrend, or sideways movement.
- Market Neutral: The profit and loss of a portfolio is independent of the market portfolio value changes.

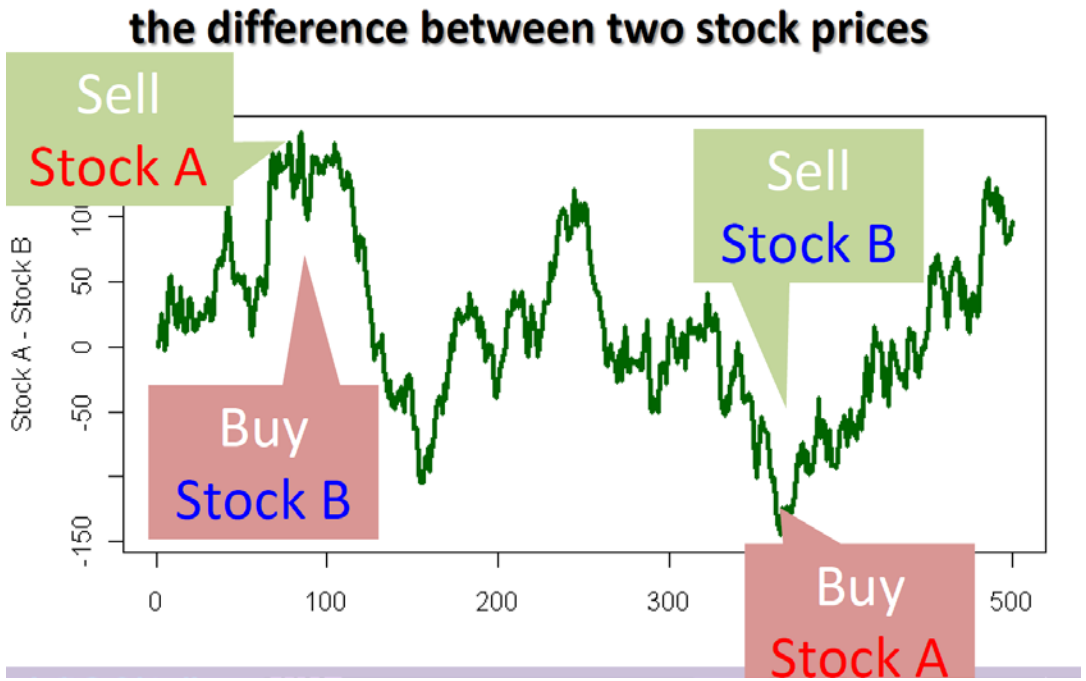
Cointegration and Pair Trading

- Review: Buy APPL and sell MSFT at the same time. Assume AAPL and MSFT prices moves closely together.
- Pair Trading: select two stocks which move similarly, sell high priced stock buy low priced stock.



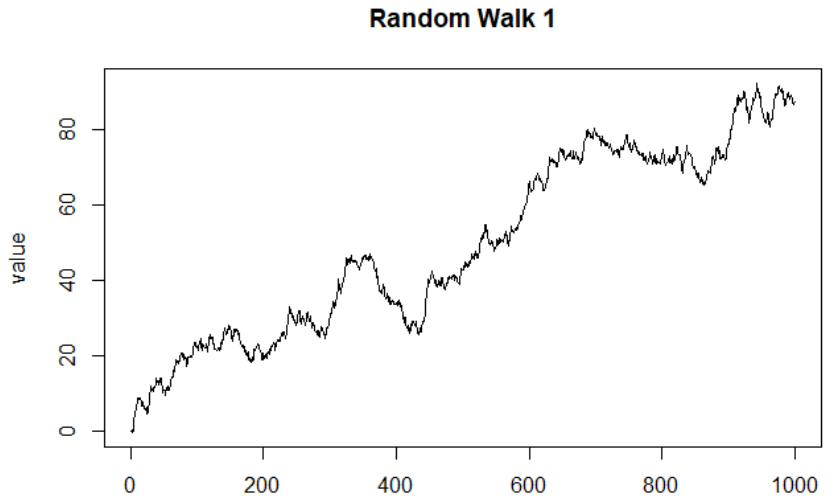
Cointegration and Pair Trading

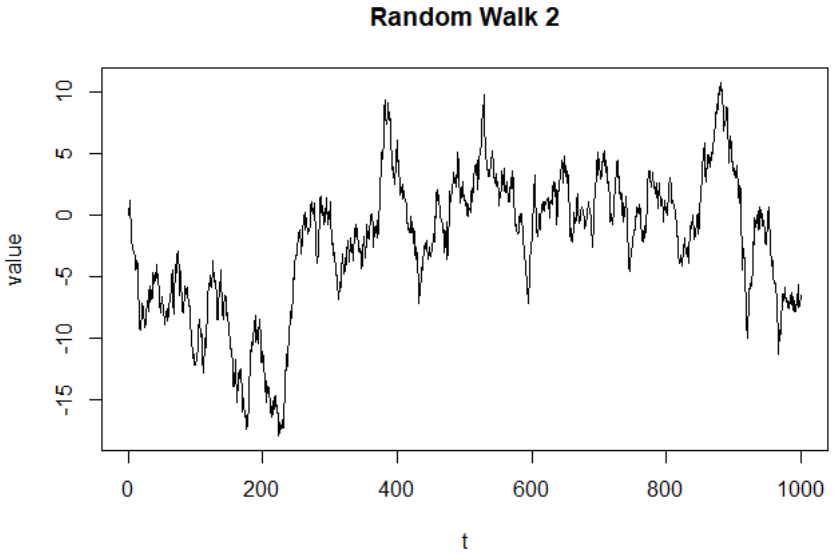
- Monitor the difference between two stocks' prices.



Cointegration and Pair Trading

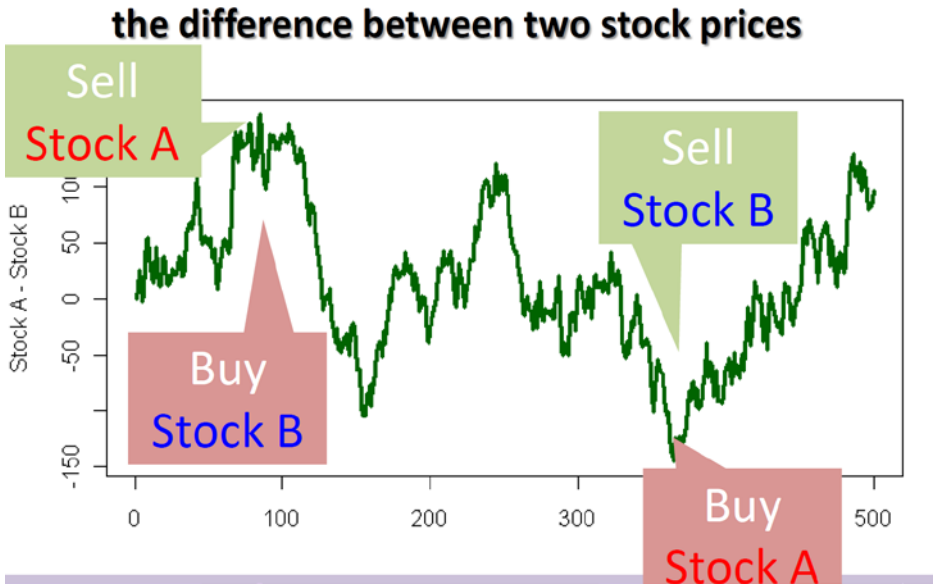
- However, the relationship between two price process might not be reliable.
  - Example: Random walk process
- Q1: How could we select out the right pairs?
- A1: Cointegration
  - Cointegration statistically tell us that two processes are moving together and the relationship is reliable.





Cointegration and Pair Trading

- Q2: How many shares should we buy or sell each stock if two stocks are a pair?
- A2: Depends on the cointegration model—the exact linear relationship between two stocks prices.



Cointegration

- What is cointegration?
- Statistical property of multiple time series: Two time series is cointegrated if their linear difference is stationary.
- X_t, Y_t are cointegrated if

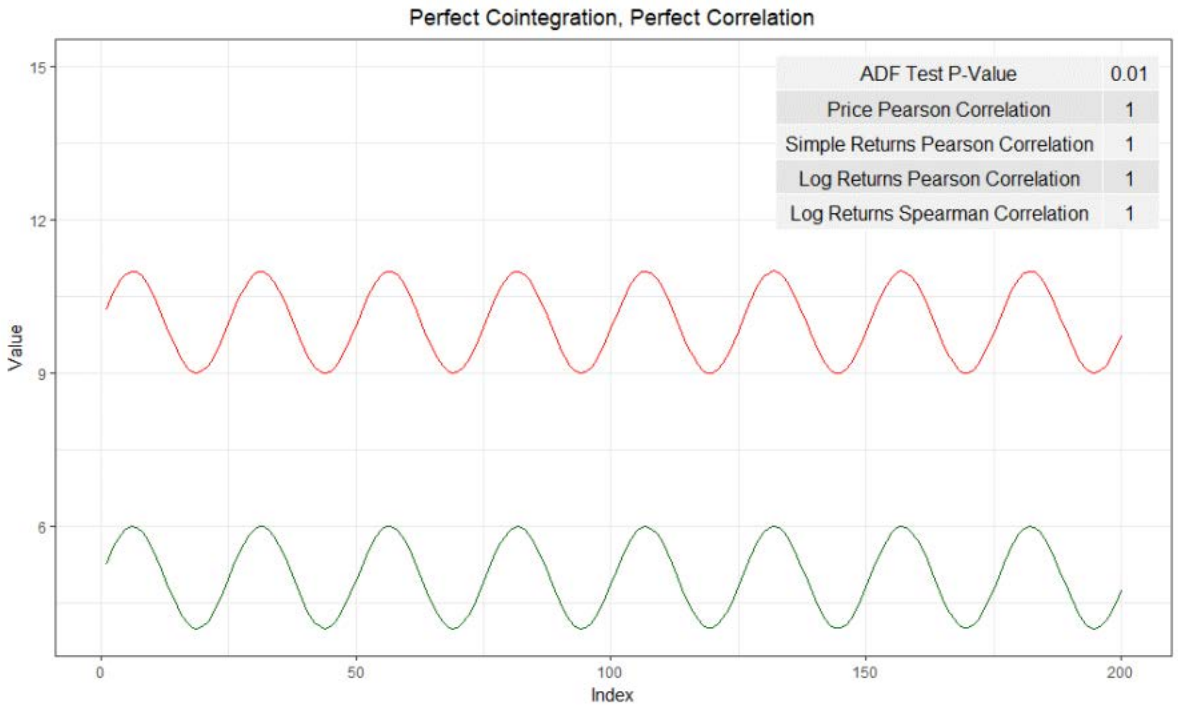
$$u_t = Y_t - (\alpha + \beta X_t)$$
$$u_t : \sim I(0), \text{ stationary process}$$
$$X_t, Y_t : \sim I(1)$$

Cointegration is Not Correlation

- Convention:
  - Use correlation to describe the co-movement of return
  - Use cointegration to describe co-movement of price
- Time series that are correlated in returns are usually also cointegrated in prices, but it’s not always the case and vice versus.

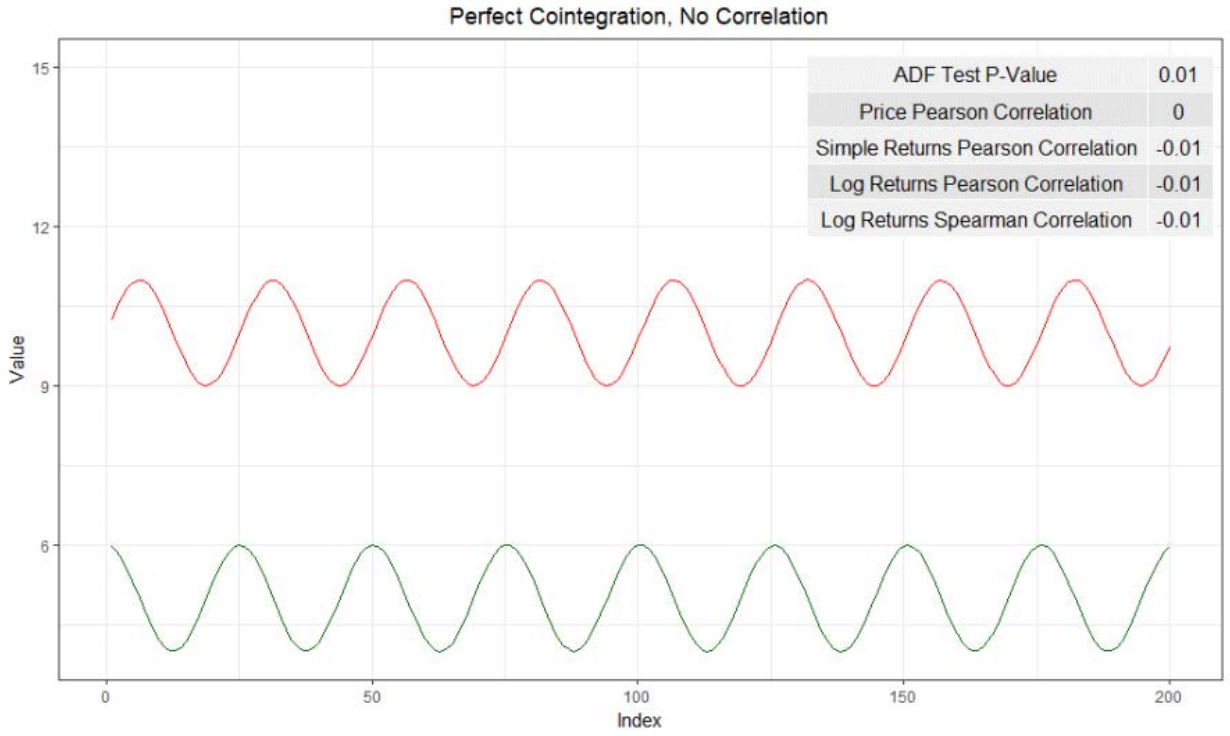
Cointegration is Not Correlation

- Red and green lines can be views as two stock prices.
- Example 1: Perfect Cointegration in price and Perfect Correlation in Returns



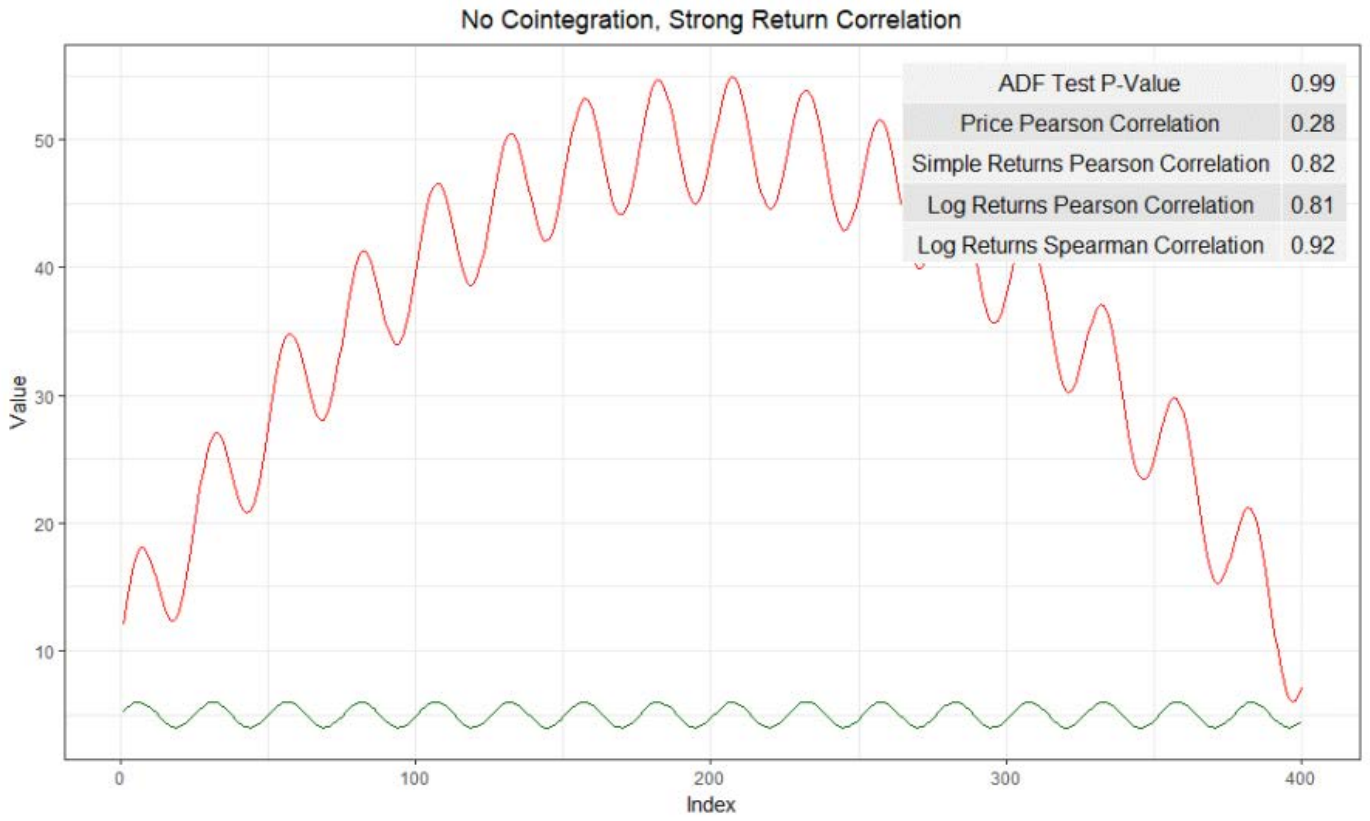
Cointegration is Not Correlation

- Example 2: Perfect Cointegration in Prices but No Correlation in Returns
  - By shifting the green price series



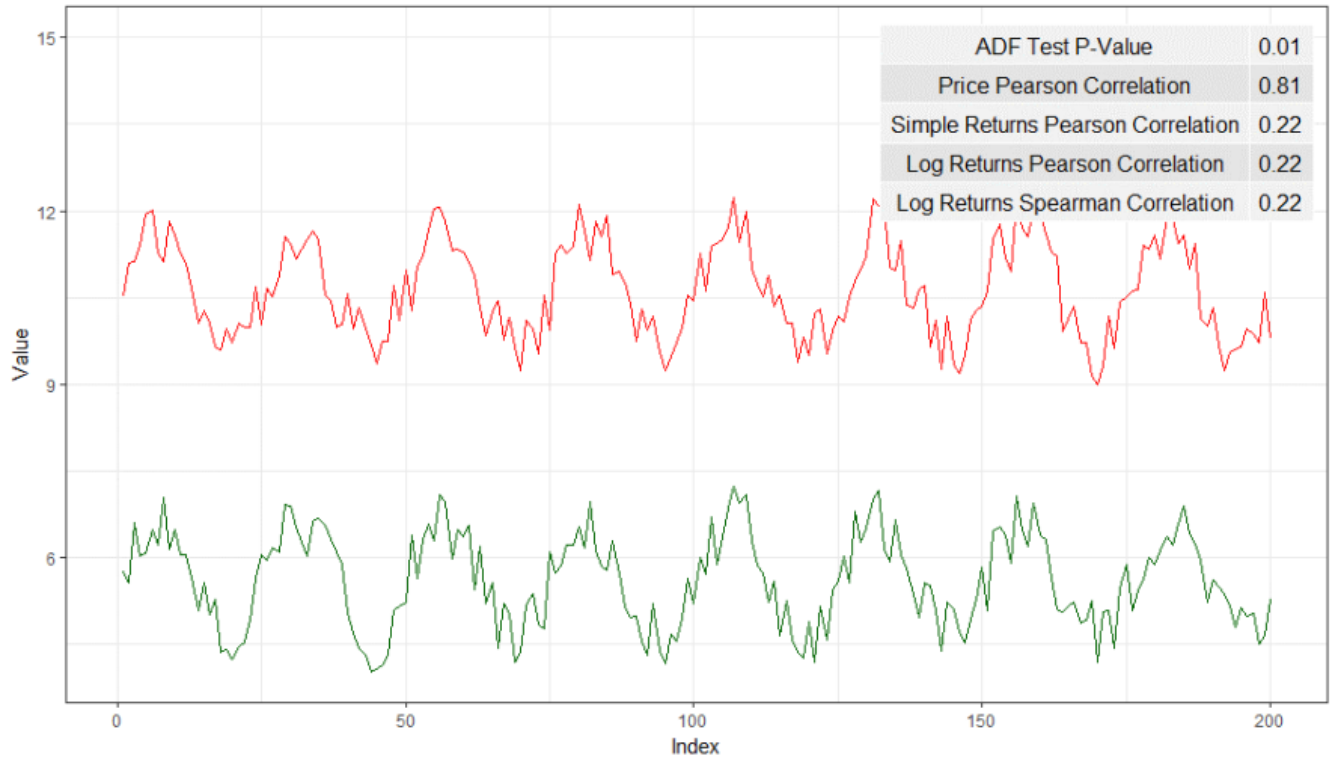
Cointegration is Not Correlation

- Example 3: No Cointegration in Prices but high Correlation in Returns.



Cointegration is Not Correlation

- Real time series: add noises/trend/seasonality, etc.



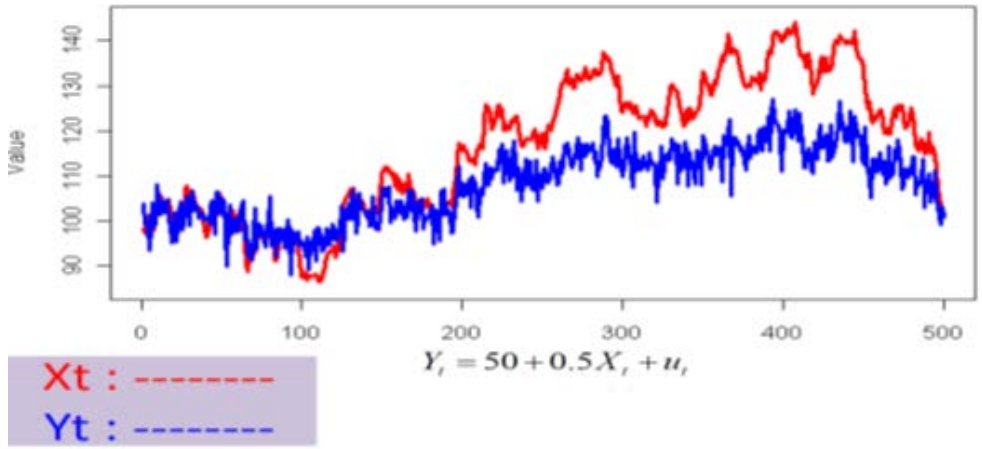
Example of Cointegrated Time Series

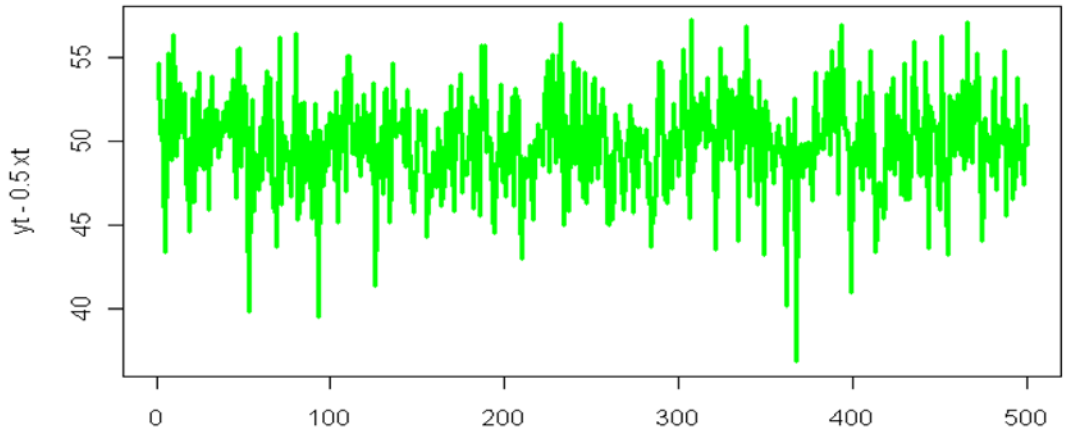
$$u_t = Y_t - (\alpha + \beta X_t)$$

$$u_t : \sim I(0), \text{ stationary process}$$

$$X_t, Y_t : \sim I(1)$$

- The residual u(t) seems to be stationary and mean reverting.





Plot : ut = Yt – 0.5 Xt

Q2: How Many Shares to Buy/Sell?

- For two stocks with prices: X_t and Y_t
- Define

$$Spread_t = \log(Y_t) - (\alpha + \beta \log(X_t))$$

- X_t and Y_t are cointegrated stocks: the spread is of I(0) and usually assumed to be a stationary process with long-term average 0.
- If the spread(t-1) < very low: Buy 1 share of Y and sell beta shares X:
  - P&L:

$$log(Y_t) - log(Y_{t-1}) - \beta(logX_t - logX_{t-1}) = Spread_t - Spread_{t-1}$$

- If the spread(t-1) > very high: Buy beta shares X and sell 1 share of Y:
  - P&L:

$$-log(Y_t) + log(Y_{t-1}) + \beta(logX_t - logX_{t-1}) = -Spread_t + Spread_{t-1}$$

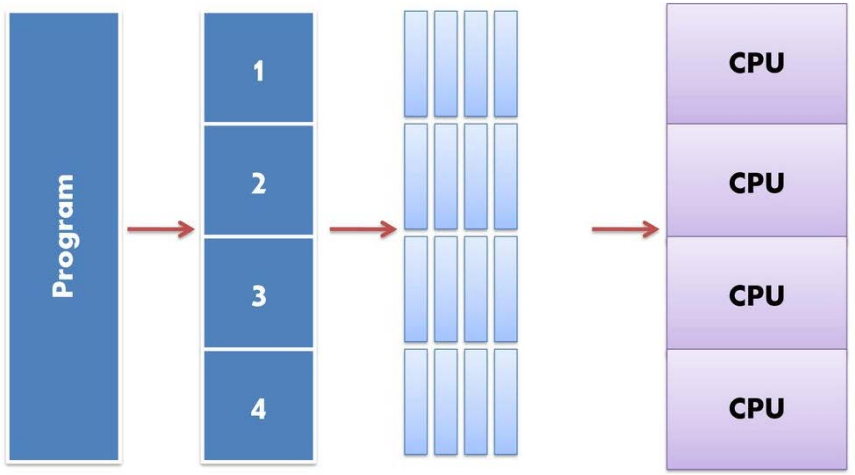
- Two Steps:
  - 1. Alpha and Beta are determined by linear regression: log(Y)~a+b*log(X)+e
  - 2. The spread process: Researchers usually assume Gaussian, AR, MA, or ARMA or other stationary models with long run average 0.

Q1: How could we select out the right pairs?

- Dickey–Fuller test:
  - If alpha and beta are know, we could use adfTest from unitroot.
  - If we don't know alpha and beta, we need to first do a regression of log(Y) to log(X) and check for the unitroot in the regression residuals.
- Johansen Test: Likelihood ratio based tests for linear relationship between time series
- H0: No Cointegration, reject if test statistics is large than critical values
- In R: library(urca)
  - ca.jo(x), where x is the data matrix to be investigated for cointegration.
  - Johansen allows testing on multiple time series cointegrations structure:
  - Eg. aX+bY+cZ = I(0)
  - Only need to test for pairs (two) time series in pair trading.

Q1: How could we select out the right pairs?

- Target: Find the cointegrated stocks from many stocks.
- Difficulty: When the universe is large, there are many possible pairs and the calculation speed is slow.
- Solution: Use parallel computation
- In R: use library(doParallel) and library(foreach)
- The foreach library is mainly use for parallel computation but could also be used as for loops with results automatically combined/organised.



Example: Pair Trade

- Target: Find the cointegrated stocks from many stocks.
- See “Cointegration.html”
- Step 1. Use the crsp.sample.Rdata which contains 10 stocks prices. Take log on prices
- Step 2. Construct the 45 (10*9/2) potential pairs and examine the cointegration for one pair as an example.
  - combn(permno,2) to generate all possible pairs of stock code
  - Wrap up the task as user defined function to be sent to each CPU.
  - Each task is to do the ca.jo() cointegration test and report the test statistics, and the critical values at 90%, 95% and 99% confidence level.

Example: Pair Trade

- Step 3. Use foreach and doParallel to compute the cointegration for all pairs
  - Load the packages doParallel and foreach, tell computer the number of CPU to use simultaneously by registerDoParallel(cores = 4)
  - foreach(i, .combine, .packages) %dopar% sometask
- Step 4. Select cointegrated pairs
- Step 5. Estimate the hedge ratio using linear regression, and get the spread between two stocks
  - Check for unit root in residuals using adfTest()

Example: Pair Trade

- Step 6. Create buy sell signals and calculate returns
  - First assume the spread is gaussian process.
  - Strategy: Sell 1 Y and buy b X when spread is large/vice versus.
- Step 7. Create buy sell signals and calculate returns
  - Assume the spread is following AR(1) process.
- Strategy: Sell 1 Y and buy b X when the realized spread is larger than the 1-step forecasted spread using AR(1) model
- Step 7. Use another window to estimate hedge ratio to avoid looking forward bias