MONASH University

ETC2420

Statistical methods in Insurance

Week 11.
Time series

14 October 2016

Outline

Week	Topic	Lecturer
1	Randomization & Hypothesis Testing I	Souhaib & Di
2	Hypothesis Testing II & Decision Theory	Souhaib
3	Statistical Distributions	Di
4	Model fitting & Linear regression	Di
5	Linear models	Di
6	Bootstrap, Permutation and Linear models	Di
	Multilevel models	Di
7	Generalized Linear models	Di
8	Compiling data for problem solving	Di
9	Bayesian Reasoning	Souhaib
10	Monte Carlo sampling methods	Souhaib
11	Time series	Souhaib
12	Project presentation	Souhaib

References

- Makridakis, Spyros G., Steven C. Wheelwright, and Rob J. Hyndman. 1998. **Forecasting: Methods and Applications**. Edited by John Wiley Sons. John Wiley & Sons.
- Hyndman, R.J. and Athanasopoulos, G. (2013) **Forecasting: principles and practice**. OTexts: Melbourne, Australia. `http://otexts.org/fpp/`.
- Chatfield, C. 2016. **The Analysis of Time Series: An Introduction**, Sixth Edition. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Time series

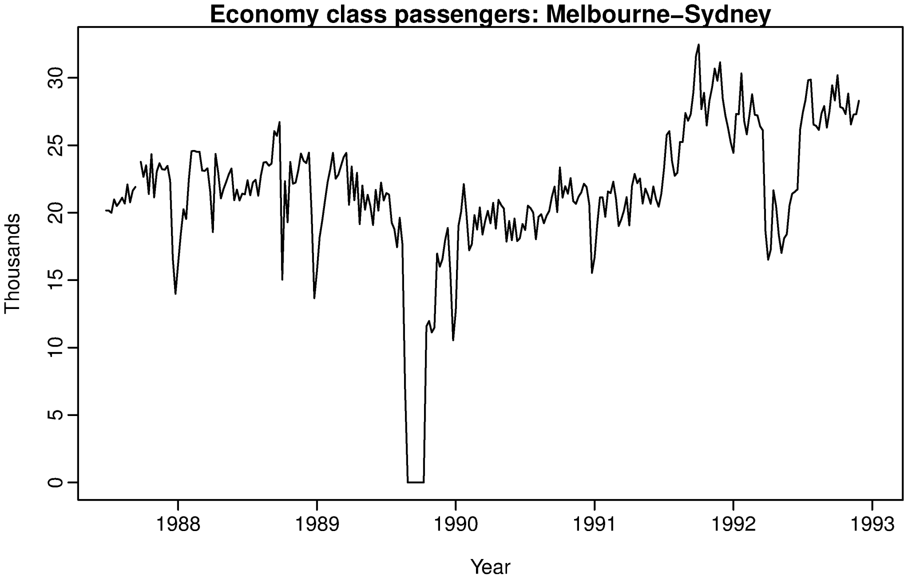
"A **time series** is a collection of observations made sequentially through time."

- Economic and financial time series
- Marketing time series
- Demographic time series
- Physical time series (meteorology, marine science, geophysics, etc)

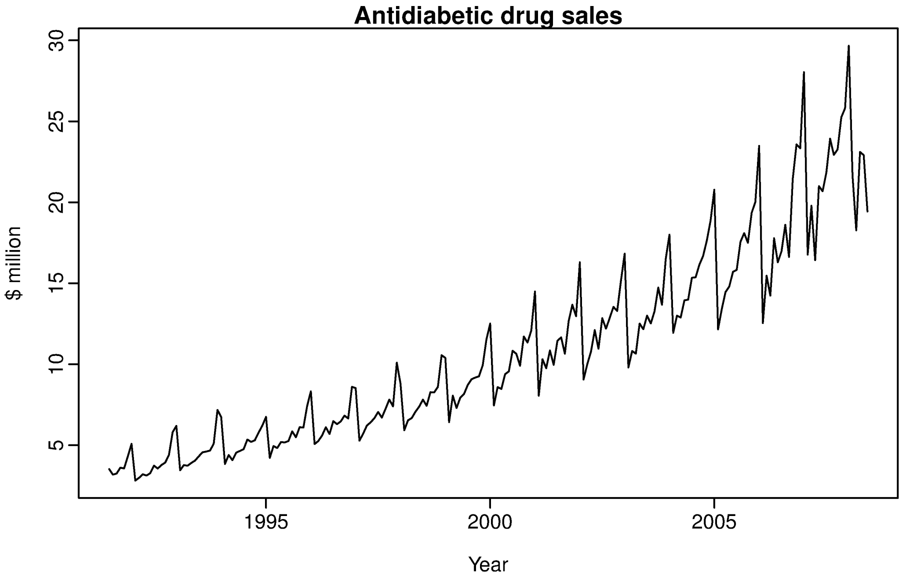
Time series example

ANSETT AUSTRALIA

Time series example



Time series example



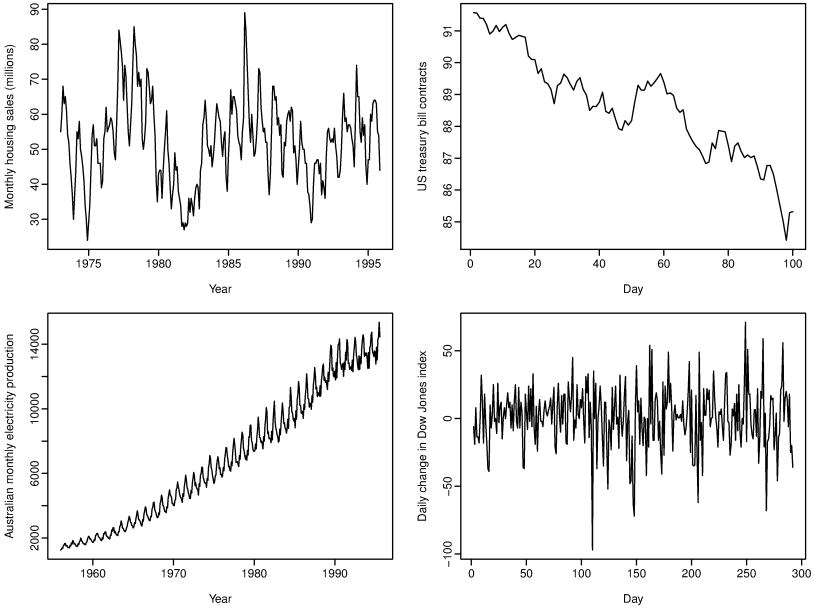
Time series patterns

- A **trend** exists when there is a **long-term increase or decrease** in the data. There is a trend in the antidiabetic drug sales data shown above.
- A **seasonal pattern** occurs when a time series is affected by **seasonal factors** such as the time of the year or the day of the week.
- A **cycle** occurs when the data exhibit **rises and falls that are not of a fixed period**. These fluctuations are usually due to economic conditions and are often related to the business cycle.

Time series patterns

- If the fluctuations **are not of fixed period** then they are **cyclic**; if the **period is unchanging** and associated with some aspect of the calendar, then the pattern is **seasonal**.
- In general, the **average length of cycles** is longer than the length of a seasonal pattern, and the **magnitude of cycles** tends to be more variable than the magnitude of seasonal patterns.

Time series patterns



Stochastic processes

Let $\mathcal{T}$ be a subset of $[0, \infty)$. A family of random variables $\{Y_t\}_{t \in \mathcal{T}}$, indexed by $\mathcal{T}$, is called a **stochastic (or random) process**.

When $T = \mathbb{N}$, $\{Y_t\}_{t \in \mathcal{T}}$ is said to be a **discrete-time process**, and when $T = [0, \infty)$, it is called a **continuous-time process**.

When T is a singleton (say $T = \{1\}$), the process $\{Y_t\}_{t \in \mathcal{T}}$ is really just a single **random variable**. When T is finite (e.g., $T = 1, 2, ..., n$), we obtain a **random vector**.

Stochastic processes

We may regard an **observed time series** as one **realization** of the stochastic process, and is denoted y_t for $t = 1, \ldots, T$ (if time is discrete).

Time series analysis is mainly concerned with evaluating the **properties of the underlying stochastic process** from this observed time series, even though we only observe a single realization.

Stochastic processes

$$\{Y_t\}_{t \in \mathcal{T}}$$

- Mean function

$$\mu_t = \mathbb{E}[Y_t]$$

- Variance function

$$\sigma_t^2 = \mathrm{Var}(Y_t)$$

- Autocovariance function (ACVF)

$$\begin{aligned}\gamma(t_1, t_2) &= \mathrm{Cov}(Y_{t_1}, Y_{t_2}) \\ &= E[(Y_{t_1} - \mu_{t_1})(Y_{t_2} - \mu_{t_2})]\end{aligned}$$

Strictly stationary processes

A stochastic process is **strictly stationary** if the joint distribution of $Y_{t_1}, \ldots, Y_{t_k}$ is the same as the joint distribution of $Y_{t_1+\tau}, \ldots, Y_{t_k+\tau}$ for all $t_1, \ldots, t_k, \tau$.

The above definition holds for any value of k. For $k = 1$, we have

$$\mu_t = \mu \text{ and } \sigma_t^2 = \sigma^2$$

For $k = 2$, the joint distribution of Y_{t_1} and Y_{t_2} depends only on the time difference $t_2 - t_1 = \tau$, also called the **lag**. Thus the ACVF also depends only on $t_2 - t_1$ and may be written as

$$\begin{aligned} \gamma(\tau) &= \mathrm{Cov}(Y_t, Y_{t+\tau}) \\ &= E[(Y_t - \mu)(Y_{t+\tau} - \mu)] \end{aligned}$$

which is called the **ACVF at lag** τ.

Strictly stationary processes

The size of an autocovariance coefficient depends on the units in which X_t is measured. Thus, for interpretative purposes, it is helpful to standardize the ACVF to produce a function called the **autocorrelation function** (ACF) defined by

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$$

which measures the correlation between X_t and $X_{t+\tau}$.

Weakly stationary processes

A process is called second-order stationary (or weakly stationary) if its mean is constant and its ACVF depends only on the lag, i.e.

$$\mu_t = \mu$$

$$\mathrm{Cov}(Y_t, Y_{t+\tau}) = \gamma(\tau)$$

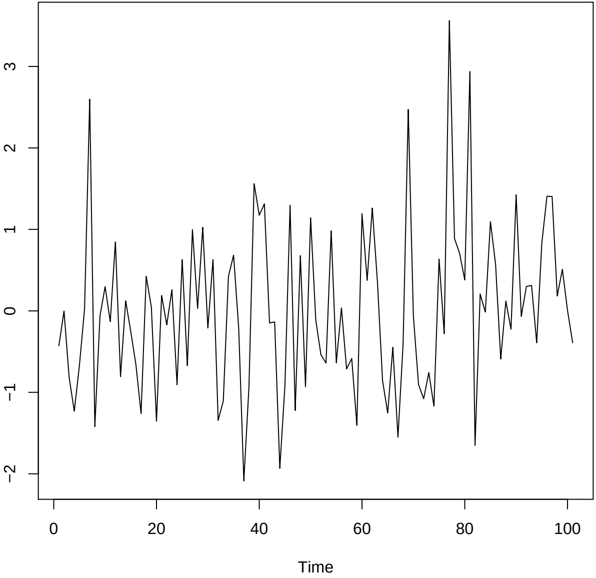
No requirements are placed on moments higher than second order.

Example I: purely random process

A **purely random** process (white noise) is a sequence of random variables, $\{Y_t\}$, which are mutually independent and identically distributed with $\mathbb{E}[Y_t] = 0$ and $\text{Var}(Y_t) = \sigma_Y^2$.

$$\gamma(k) = \text{Cov}(Y_t, Y_{t+k}) = \begin{cases} \sigma_Y^2 & k = 0 \\ 0 & k = \pm 1, \pm 2, \ldots \end{cases}$$

Example I: purely random process



Example II: random walk process

Suppose that $\{Z_t\}$ is a white noise process with $\mathbb{E}[Z_t] = \mu_Z$ and $\text{Var}[Z_t] = \sigma_Z^2$, a **random walk** (white noise) process $\{Y_t\}$ is given by

$$Y_t = Y_{t-1} + Z_t$$

where $Y_1 = Z_1$.

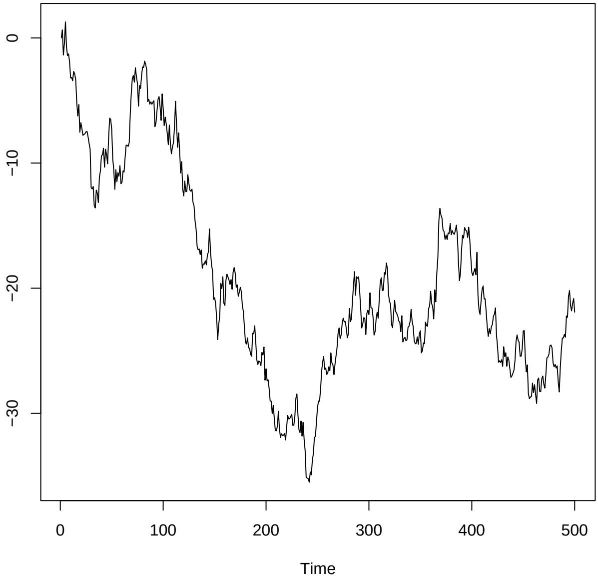
$$\mathbb{E}[Y_t] = t\mu_Z$$

and

$$\text{Var}(Y_t) = t\sigma_Z^2$$

The random walk process is **non-stationary**.

Example II: random walk process



Example III: autoregressive process

Suppose that $\{Z_t\}$ is a white noise process with $\mathbb{E}[Z_t] = 0$ and $\text{Var}[Z_t] = \sigma_Z^2$.

A process $\{Y_t\}$ is said to be an **autoregressive process** of order p if

$$Y_t = \phi_1 Y_{t-1} + \cdots + Y_{t-p} + \phi_p Z_t.$$

Example III: autoregressive process

If $p = 1$:

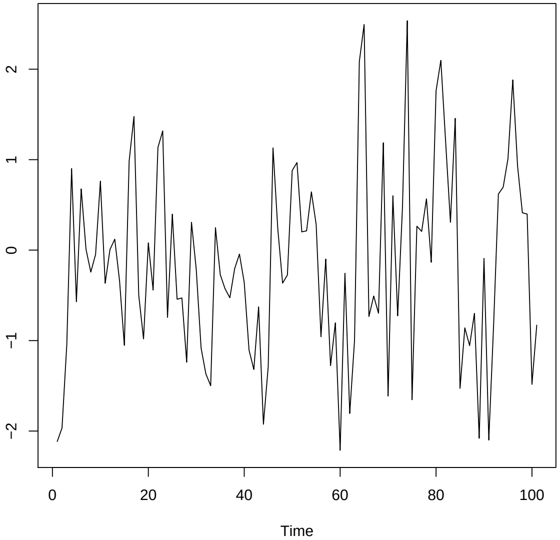
$$Y_t = \phi Y_{t-1} + Z_t.$$

$$\begin{aligned}
\mathbb{E}[Y_t] &= 0, \\
\mathrm{Var}[Y_t] &= \sigma_Z^2(1 + \phi^2 + \phi^4 + \dots) \overset{|\phi|<1}{=} \sigma_Z^2 \frac{1}{1-\phi^2}, \\
\gamma(k) &\overset{|\phi|<1}{=} \phi^k \sigma_Y^2 \\
\rho(k) &\overset{|\phi|<1}{=} \phi^k \quad k = 0, 1, 2, \dots, ...
\end{aligned}$$

Exercice: plot the ACF for different values of k.

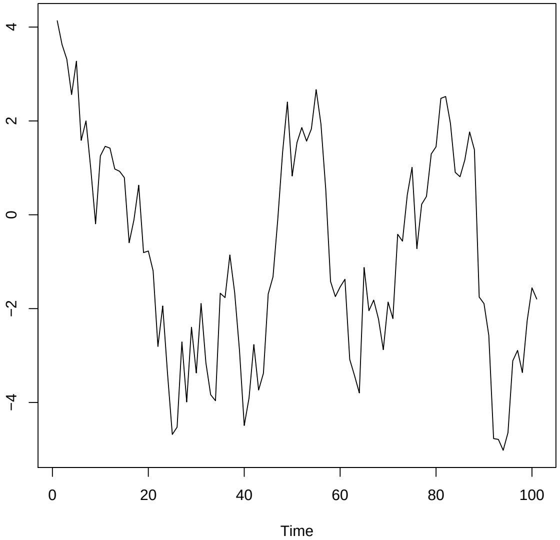
Example III: autoregressive process

$\phi = 0.2$



Example III: autoregressive process

$\phi = 0.9$



Sample ACF and correlogram

$$\gamma(k) = \text{Cov}(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)]$$

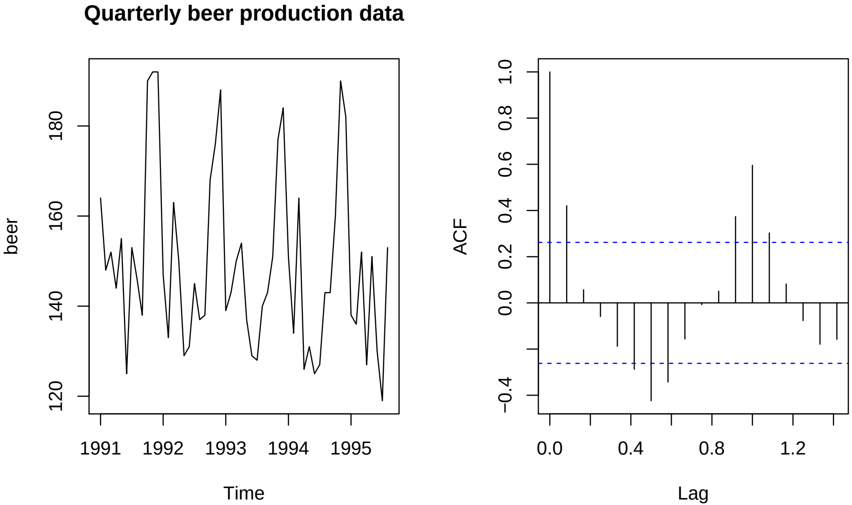
$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{E[(Y_t - \mu)(Y_{t+k} - \mu)]}{E[(Y_t - \mu)^2]}$$

The sample ACF is given by

$$r_k = \frac{\sum_{t=1}^{T-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \text{ where } \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

A **correlogram**, also known as an **autocorrelation plot**, is a plot of the sample autocorrelations r_k versus the lags k for $k = 0, 1, \ldots$.

Sample ACF and correlogram



Sample ACF and correlogram

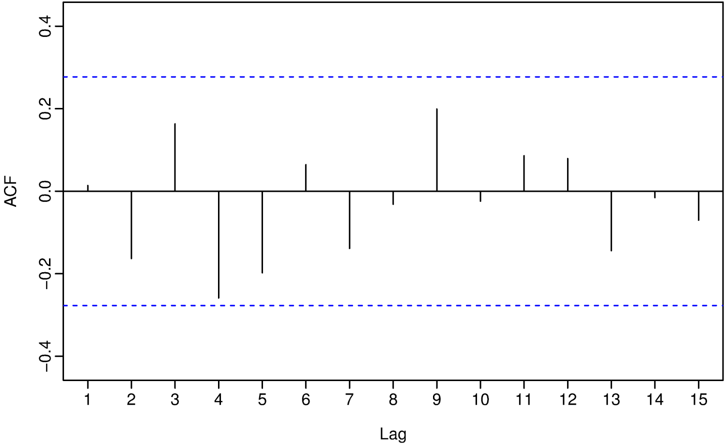
If $y_1, \ldots, y_T$ are i.i.d., it can be shown that

- $\mathbb{E}[r_k] \simeq -\frac{1}{N}$
- $\text{Var}(r_k) = \frac{1}{N}$
- r_k is asymptotically normally distributed under weak conditions

We can check for randomness by plotting approximate 95% confidence limits at $-\frac{1}{N} \pm \frac{2}{\sqrt{N}}$, which is often approximated to $\pm\frac{2}{\sqrt{N}}$.

Sample ACF and correlogram

Correlogram for $y_1, \ldots, y_{50}$ i.i.d.:



Stationarity and differencing

One way of removing non-stationarity is through the method of **differencing**. We define the differenced series as the *change* between each observation in the original series:

$$y_t^{'} = y_t - y_{t-1},$$

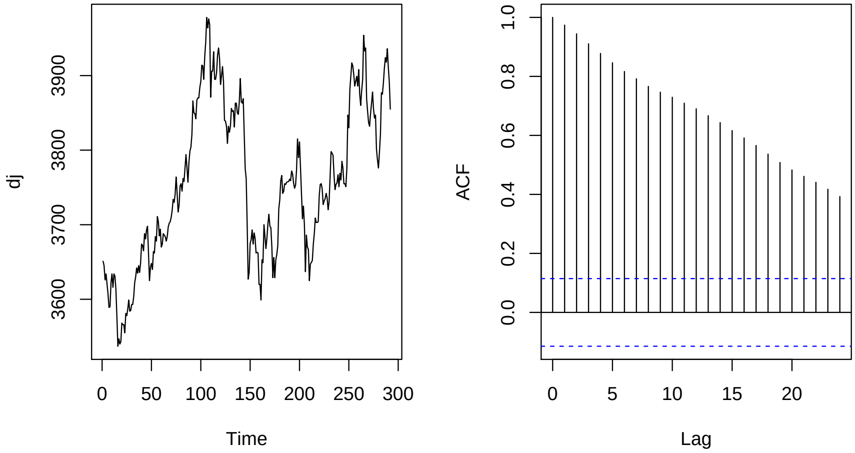
where $t = 2, \dots, T$.
Occasionally, it may be necessary to difference the data a second time:

$$y_t^{''} = y_t^{'} - y_{t-1}^{'} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2},$$

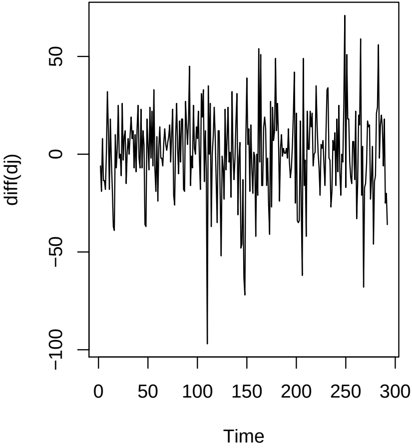
where $t = 3, \dots, T$.
Seasonal differencing can be useful with seasonal data. For example, with monthly data, we can compute

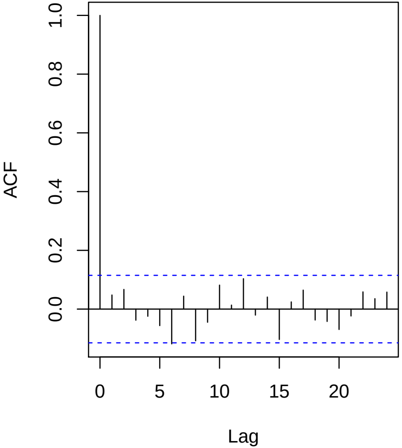
$$y_t^{'} = y_t - y_{t-12}$$

Stationarity and differencing

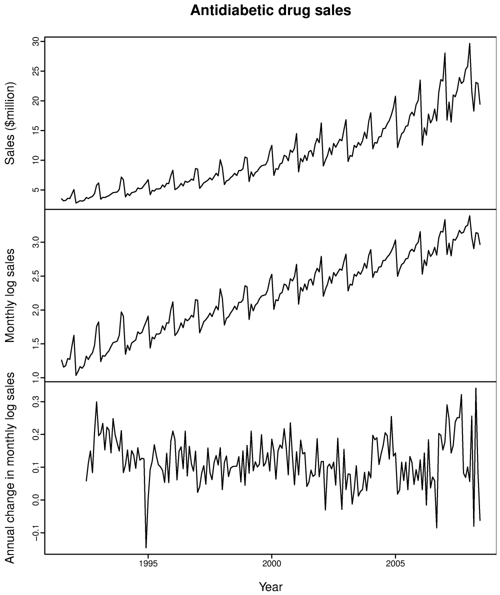


Stationarity and differencing





Stationarity and differencing



The bootstrap procedure

- $X_1, \dots, X_n$ where $X_i \overset{i.i.d.}{\sim} P$
- Draw B independent bootstrap samples

$$\boldsymbol{X}^{*(b)} = \{X_1^{*(b)}, \dots, X_n^{*(b)}\} \quad b = 1, \dots, B.$$

  where $X_i^{*(b)} \overset{i.i.d.}{\sim} \hat{P}$.
- Evaluate the bootstrap replications:

$$\hat{\theta}^{*(b)} = s(\boldsymbol{X}^{*(b)}) \quad b = 1, \dots, B.$$

- Estimate the quantity of interest from the distribution of the $\hat{\theta}^{*(b)}$

The block bootstrap

1. $X_1, \ldots, X_T \sim P_T$
2. Suppose $m \equiv T/l$ is an integer. Partition the time series in m blocks of size l:

$$\underbrace{\{X_1, \ldots, X_l\}}_{\mathbf{Y}_1}, \underbrace{\{X_{l+1}, \ldots, X_{2l}\}}_{\mathbf{Y}_2}, \ldots, \underbrace{\{X_{(m-1)\times l+1}, \ldots, X_T\}}_{\mathbf{Y}_m}$$

- If the time series is **stationarity**, each block has the same (l-dimensional joint) distribution P_l.
- If the time series is **weakly dependent** (i.e. $\rho(k) \to 0$ when $k \to \infty$) of the original sequence, the blocks are approximately independent for large values of l.

The block bootstrap

3. Draw B independent bootstrap samples:

$$\boldsymbol{Y}^{*(b)} = \{\boldsymbol{Y}_1^{*(b)}, \ldots, \boldsymbol{Y}_m^{*(b)}\} \text{ where } \boldsymbol{Y}_i^{*(b)} \overset{i.i.d.}{\sim} \hat{P}_l, b = 1, \ldots, B.$$

4. Evaluate the bootstrap replications:

$$\hat{\theta}^{*(b)} = s(\boldsymbol{Y}^{*(b)}) \quad b = 1, \ldots, B.$$

5. Estimate the quantity of interest from the distribution of the $\hat{\theta}^{*(b)}$