Introduction to Time Series

James Balamuta and Stephane Guerrier

2016-08-16

Contents

Chapter 1

Preface

Welcome to Introduction to Time Series with `R`!

1.1 A foreword

This book was designed for use in STAT 429, Time Series Analysis, at the University of Illinois at Urbana-Champaign. When possible, it would be best to always access the text online to be sure you are using the latest version. The online version affords additional features over the traditional PDF copy such as a scaling text, variety of font faces, and themed backgrounds. However, if you are in need of a local copy, a **pdf version** is also available.

Disclaimer: This book is under active development. As a result, errors may occur that range in severity from typos to broken code. If any of these issues arise, there are two options:

1. If you are familiar with GitHub and know RMarkdown, make a pull request and fix the issue yourself! (fastest resolution)

    - In the Online version, click the edit button in the top-left corner. 

2. Send an email to `balamut2 AT illinois DOT edu` and we will address issue.

1.2 Rendering Mathematical Formulae

Throughout the book, there will be mathematical symbols used to express the material. Depending on the version of the book, there are two different render engines.

- For the online version, the text uses MathJax to render mathematical notation for the web. In the event the formulae does not load for a specific chapter, first try to refresh the page. 9 times out of 10 the issue is related to the software library not loading quickly.
- For the pdf version, the text is built using the recommended AMS LaTeX symbolic packages. As a result, there should be no issue displaying equations.

An example of a mathematical rendering capabilities would be given as:

$$a^2 + b^2 = c^2$$

1.3 Mathematical Notation

The following notation will be adopted throughout the book.

- X denotes a (continuous) RV.
- X_t is X at time $t \in N$.
- $E(X_t)$ is the Mean of X at time t.
- $Var(X_t)$ is the Variance of X at time t.
- $X_1, X_2, \ldots, X_k$ are sequence of random variables.
- $f(x)$ denotes the density function of X and $f(x, y)$ denotes the joint density function of x and Y.
- $(X_t)_{t=1,\ldots,T} := (X_t) := (X_1, \ldots, X_T)$.

1.4 R Code Conventions

The code used throughout the book will predominately be `R` code.
To obtain a copy of `R`, go to the Comprehensive R Archive Network (CRAN) and download the appropriate installer for your operating system.

When `R` code is displayed it will be typeset using a `monospace` font with syntax highlighting enabled to ensure the differentiation of functions, variables, and so on. For example, the following adds 1 to 1

```
a = 1L + 1L
a
```

Each code segment may contain actual output from `R`. Such output will appear in grey font prefixed by `##`. For example, the output of the above code segment would look like so:

```
## [1] 2
```

Alongside the PDF download of the book, you should find the R code used within each chapter.

1.5 License

Figure 1.1: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Chapter 2

Introduction

Prévoir consiste à projeter dans l'avenir ce qu'on a perçu dans le passé. Henri Bergson

Generally speaking a *time series* (or stochastic process) corresponds to set of "repeated" observations of the same variable such as price of a financial asset or temperature in a given location. In terms of notation a time series is often writen as

$$(X_1, X_2, ..., X_T) \quad \text{or} \quad (X_t)_{t=1,...,T}\,.$$

The time index t generally the set $\mathbb{N}$ (or sometimes $\mathbb{Z}$). When $t \in \mathbb{R}$, a time series becomes a *continuous-time* stochastic process such a Brownian motion. In this class we limit our selves to *discrete-time* processes where a variable is measured sequentially at fixed and equally spaced intervals in time. This implies that we will assume t is not random (the time at which each observation is measure is know) and the time between two consequtive observation is constant.

Moreover, the term "time series" can, as we discussed, denote a sample or a set of observations but also a probability model for that sample. For example, on of the simplest probability model used in time series analysis is called a *white noise* process and is defined as

$$W_t \overset{iid}{\sim} N(0, \sigma^2).$$

This statement simply means that (X_t) is normally distribtued and independent over time. This model is quite unintersting but as we will very usefull to construct other (more interesting) models. Unlike white noise process, time series are typically *not* independent over time. Suppose that the temperature in Champaign is unusually low, then it is reasonable to assume that tomorrow's temperature will also be low. Such behaviour suggest a dependent over time. The time series methods we will discuss in this class consists parametric models used to characterize (or at least approximate) the joint distribution of (X_t). Often, time series models can be decomposed in what we called a *signal*, say (Y_t) and a *noise*, say (W_t), leading to the model

$$X_t = Y_t + W_t.$$

Typically, we have $E[Y_t] \neq 0$ while $E[W_t] = 0$ (although we may have $E[W_t|W_{t-1}, ..., W_1] \neq 0$). Such models impose some parametric structure which represent a convenient and flexible way of studying time series and evalute to which extent *future* value of the series can be forecasted. As we will see, predicting future values is one of the main aspects of time series analysis. However, making predictions is often a daunting taks or as famously stated by Nils Bohr:

"*Prediction is very difficult, especially about the future.*"

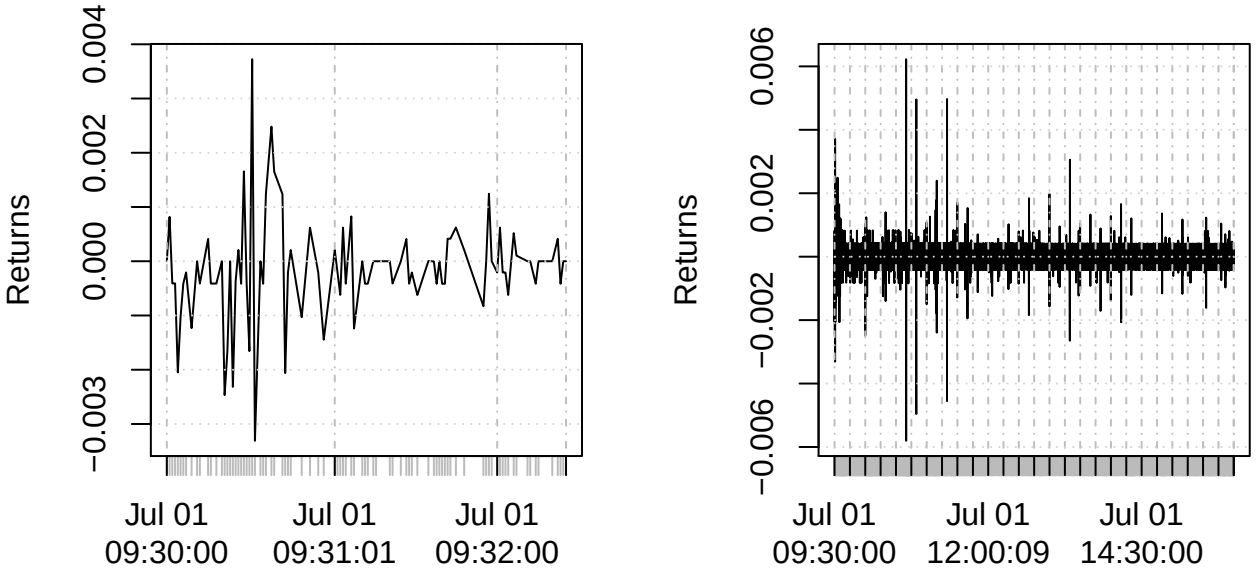
There are plenty of examples predictions which releved to be completely erroneous. For example, Irving Fisher, Professor of Economics at Yale University, famously predicted three days before the 1929 crash:

"*Stock prices have reached what looks like a permanently high plateau*".

Another example is Thomas Watson, president of IBM, who said in 1943:

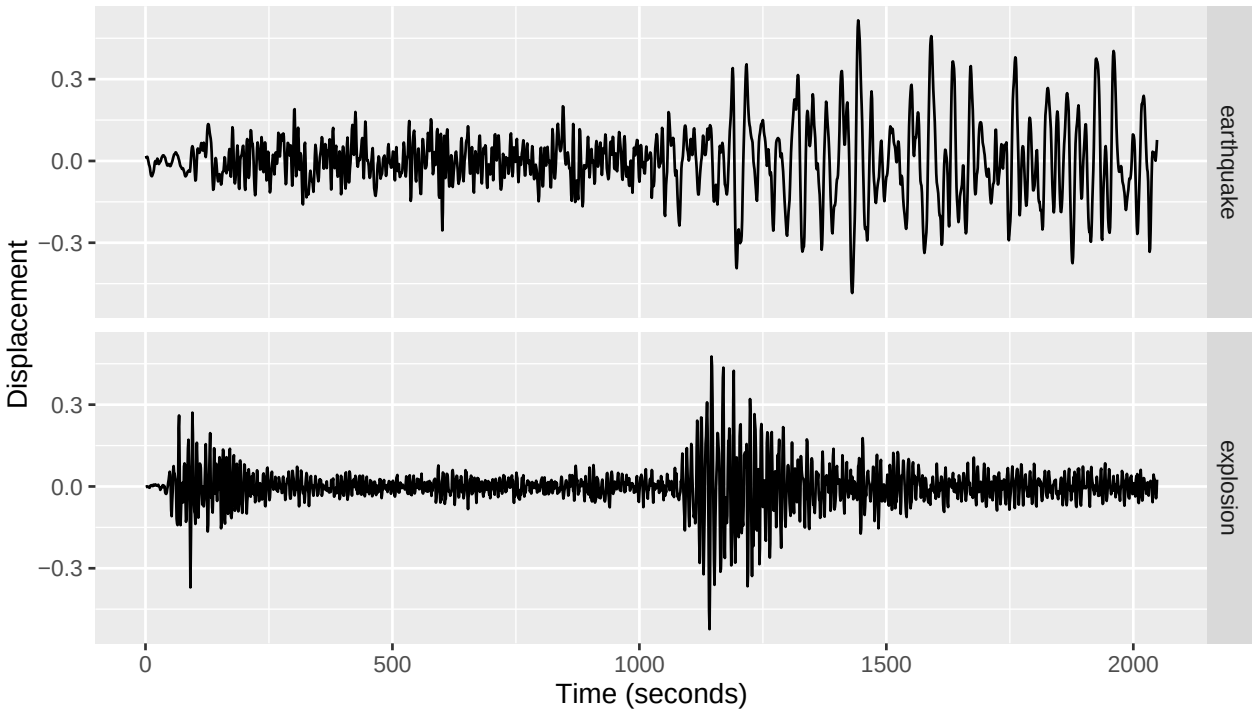
"*I think there is a world market for maybe five computers.*"

Examples of Time Series:



1. Stock Data from Johnson and Johson's Quarterly earnings…

2. Earthquake and explosion data



2.1 Exploratory Data Analysis (EDA) for Time Series

When dealing with relatively small time series (e.g. a few thousands), it is often useful to look at a graph of the original data. Such graphs can be informative to "detect" some features of a time series such as trends and the presence of outliers.

Indeed, a trend is typically deamed present in a time series when the data exhibit some form of long term increase or decrease or combination of increases or decreases. Such trends could be linear or non-linear and represent a important part of the "signal" of a model. Here are few examples of non-linear trends:

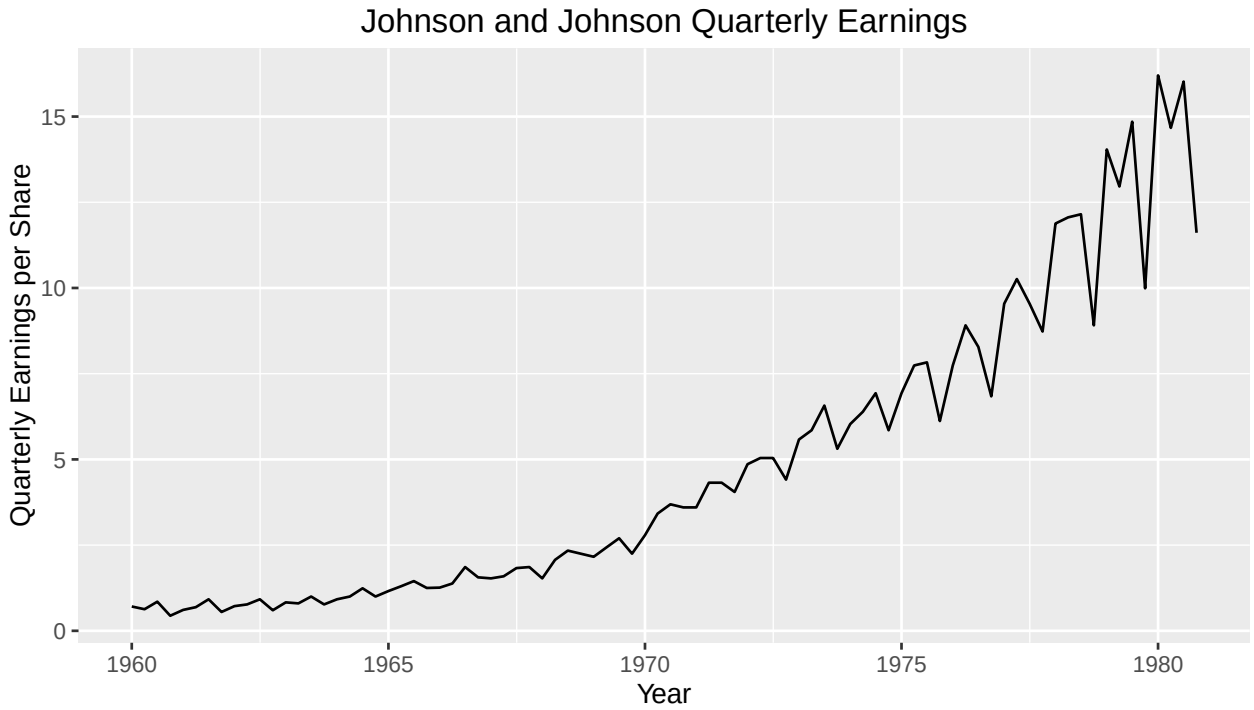
1. **Seasonal trends** (periodic): These are the cyclical patterns which repeat after a fixed/regular time period. This could be due to business cycles (e.g. bust/recession, recovery).

2. **Non-seasonal trends** (periodic): These patterns cannot be associated to seasonal variation and can for example to external variable. For example, impact of economic indicators on stock returns. Such trends are often hard to detect based on a graphical anaylsis of the data.

3. **"Other" trends**: These trends have typically no regular patterns and change statistical properties of a time series over a segment of time ("window"). A typicall example of such trends corresponds to vibrations observed before, during and after an earthquake.

Example: An example of a time series is, for example, the quarterly earnings of the company Johnson and Johson. In the figure below we present these earnings between 1960 and 1980:

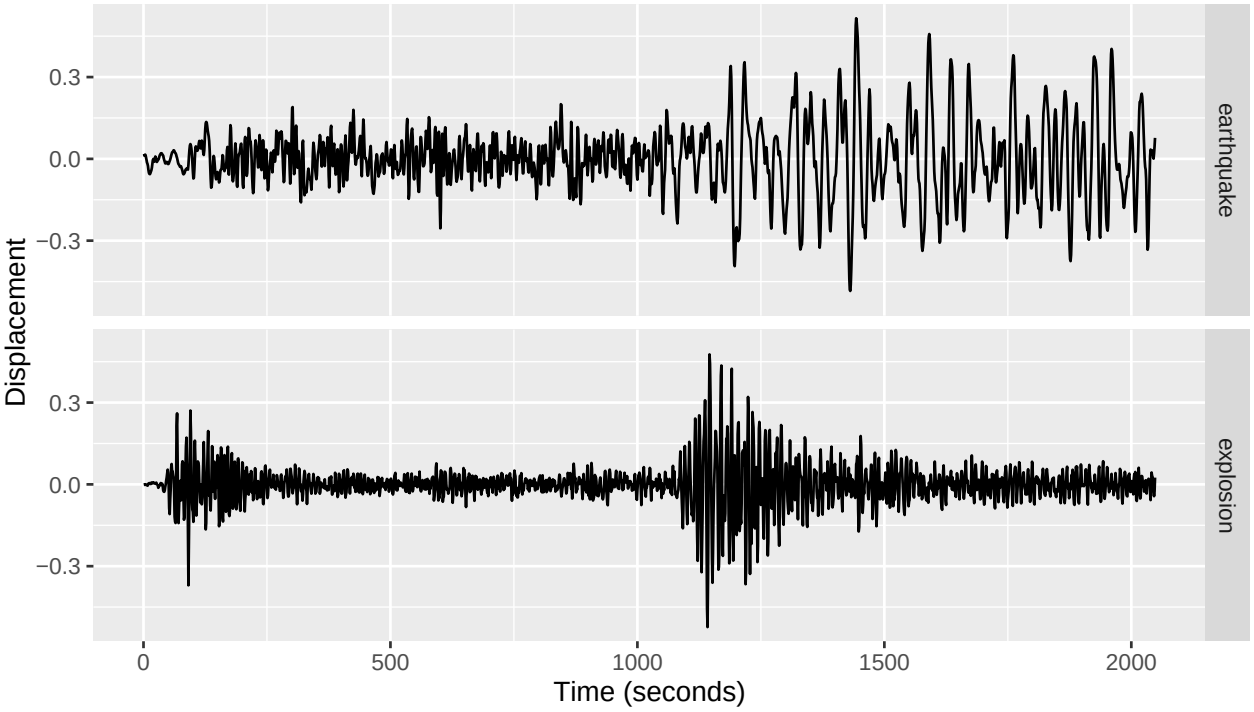
JAMES: can we "gmwm" this graph? Thanks!



It can clearly be observed that the data present a non-linear increasing trend as well as a yearly seasonal component. In addition, on can note that the *variability* of the data seems to increase with time. As we will such observations provide some valuable guilines to select a suitable models for such data.

Example: In the figure below we present TO COMPLETE

JAMES: can we "gmwm" this graph? Thanks!



ADD VISUAL DESCRIPTION

Change in time and outliers yields interesting results. These results can be seen as:

1. Change in Means
   - Change in means of a TS can be related to long-term, cyclical, and short-term trends.
2. Change in Variance
   - Change in variance can be related to change in the amplitude of the fluctuations of a TS.
3. Change in State
   - An event which causes change in statistical properties of TS for short term and long term! Some events cause abrupt changes in statistical properties of TS. They are often associated with "explosive" nature of TS.
4. Outliers
   - These are the "extreme" observations in the time series. May be related to data collection or change in state.

2.2 Basic Time Series Models

In this section, we introduce some simple time series models. Before doing so we define all the information avaiable up to time $t-1$ has Ω_t, i.e.

$$\Omega_t = (X_{t-1}, X_{t-2}, ..., X_0)\,.$$

As we will this compact notation is quite useful.

2.2.1 White noise processes

The building block for most time series models is the Gaussian white noise process, which can be defined as

$$W_t \overset{iid}{\sim} N\left(0, \sigma_w^2\right).$$

This definition implies that:

1. $E[W_t|\Omega_t] = 0$ for all t,
2. $\text{cov}\,(W_t, W_{t-h}) = \mathbf{1}_{h=0}\,\sigma^2$ for all t, h.

Therefore, this process present an absence of temporal (or serial) dependence and has a constant variance. This definition can be generalized in two sorts of processes, the *weak* and *strong* white noise. The process (W_t) is a weak white noise if

1. $E[W_t] = 0$ for all t,
2. $\text{var}\,(W_t) = \sigma^2$ for all t,
3. $\text{cov}\,(W_t, W_{t-h}) = 0$, for all t, and for all $h \neq 0$.

Note that this definition does not imply that W_t and W_{t-h} are independent (for $h \neq 0$) but simply uncorrelated. However, the notion of indepence is used to define a *strong* white noise as

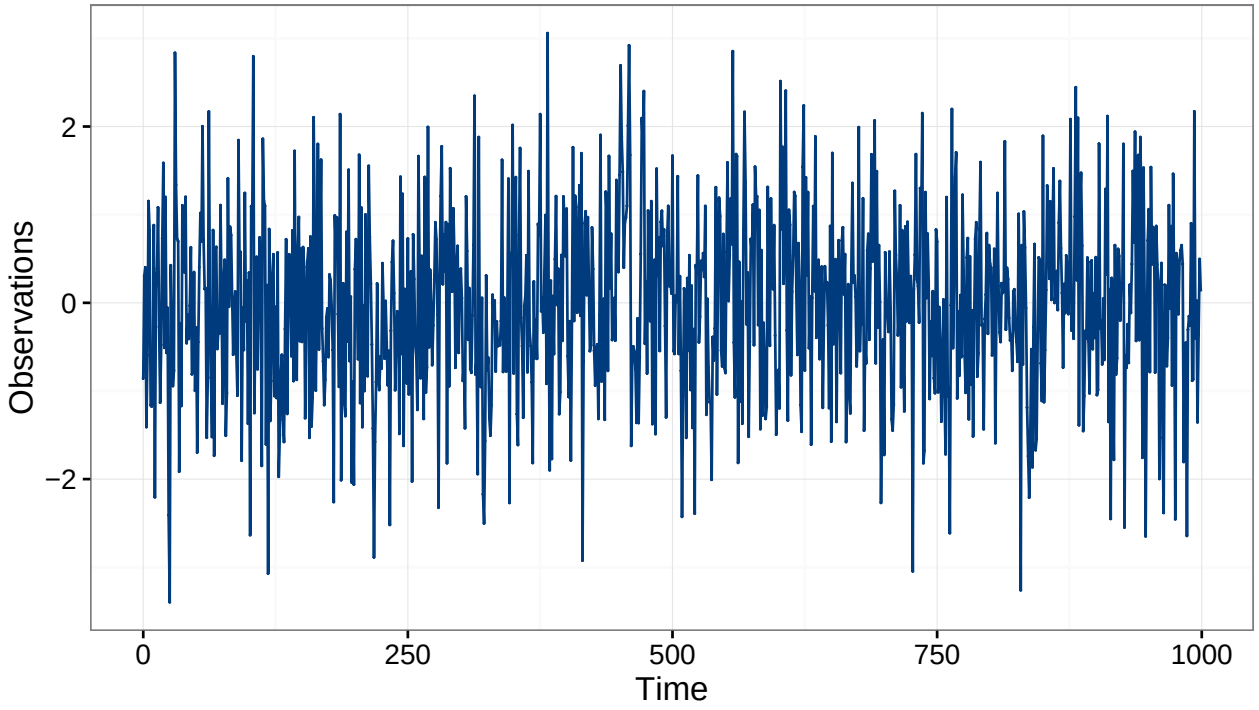
1. $E[W_t] = 0$ and $\text{var}(W_t) = \sigma^2 < \infty$, for all t,
2. $F(W_t) = F(W_{t-h})$, for all t, h (where $F(W_t)$ denotes the distribution of W_t),
3. W_t and W_{t-h} are independent for all t and for all $h \neq 0$.

It is clear from these definition that if a process is a strong white noise it is also a weak white noise. However, the converse is not true a shown in the following example:

Example: Let $X_t \overset{iid}{\sim} F_t$, where F_t denote a Student distribution with t degrees of freedom. Such process is a weak but not a strong white noise.

The code below presents an example of how to simulate a Gaussian white noise process

```
# This code simulate a gaussian white noise process
n = 100                                  # process length
sigma2 = 1                               # process variance
Xt = gen.gts(WN(sigma2 = 1), n = n)
plot(Xt)
```



We 1.

Chapter 3

Autocorrelation and Stationarity

After reading this chapter you will be able to:

- Describe independent and dependent data
- Interpret a processes ACF and CCF.

- Understand the notion of stationarity.
- Differentiate between Strong and Weak stationarity.
- Judge whether a process is stationary.

3.1 Dependency

Generally speaking, there is a dependence that within the sequence of random variables.

Recall the difference between independent and dependent data:

Definition: **Independence**

$X_1, X_2, \ldots, X_T$ are independent and identically distributed if and only if

$$P\left(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_T \leq x_T\right) = P\left(X_1 \leq x_1\right) P\left(X_2 \leq x_2\right) \cdots P\left(X_T \leq x_T\right) \tag{3.1}$$

for any $T \geq 2$ and $x_1, \ldots, x_T \in \mathbb{R}$.

Definition: **Dependence**

$X_1, X_2, \ldots, X_T$ are identically distributed but dependent, then

$$\left|P\left(X_1 < x_1, X_2 < x_2, \ldots, X_T < x_T\right) - P\left(X_1 < x_1\right) P\left(X_2 < x_2\right) \cdots P\left(X_T < x_T\right)\right| \neq 0 \tag{3.2}$$

for some $x_1, \ldots, x_T \in \mathbb{R}$.

3.1.1 Measuring (Linear) Dependence

There are many forms of dependency…

However, the methods, covariance and correlation, that we will be using are specific to measuring linear dependence. As a result, these tools are less helpful to measure monotonic dependence and they are much less helpful to measure nonlinearly dependence.

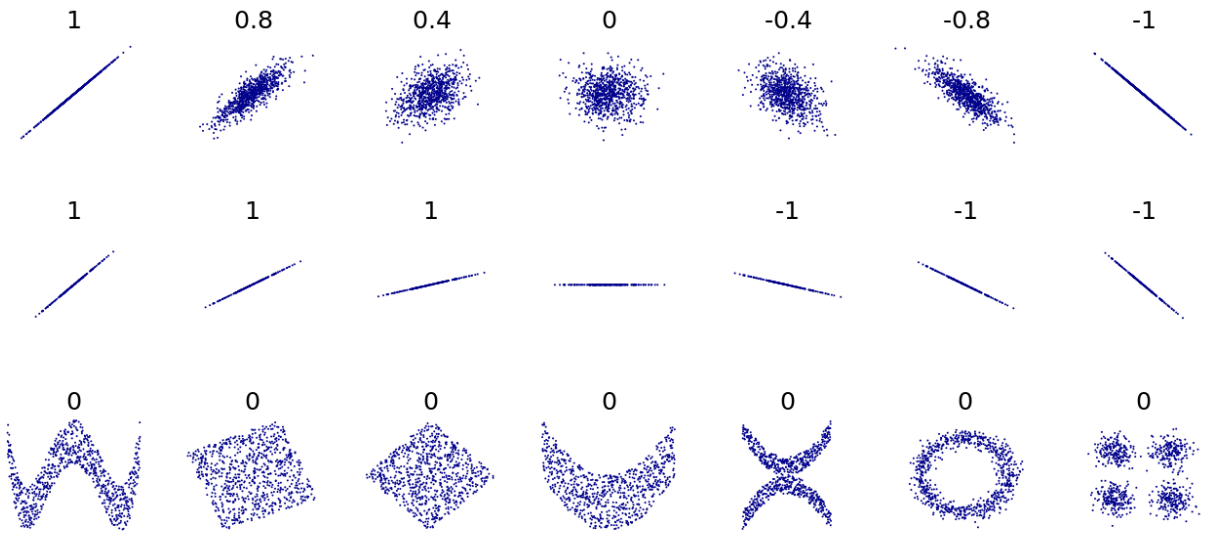


Figure 3.1: dependency

3.2 The Autocorrelation and Autocovariance Functions

Dependence between T different RV is difficult to measure in one shot! So we consider just two random variables, X_t and X_{t+h}. Then one (linear) measure of dependence is the covariance between (X_t, X_{t+h}). Since X is the same RV observed at two different time points, the covariance between X_t and X_{t+h} is defined as the Autocovariance.

3.2.1 Definitions

The **Autocovariance Function** of a series X_t is defined as

$$\gamma_x(t, t+h) = \operatorname{cov}(x_t, x_{t+h}).$$

Since we generally consider stochastic processes with constant zero mean we often have

$$\gamma_x(t, t+h) = E[X_t X_{t+h}].$$

We normally drop the subscript referring to the time series if it is clear to the time series the autocovariance function is referencing. For example, we generally use $\gamma(t, t+h)$ instead of $\gamma_x(t, t+h)$. Moreover, the notation is even further simplify when the covariance of X_t and X_{t+h} is the same as that of X_{t+j} and X_{t+h+j} (for $j \in \mathbb{Z}$), i.e. that the covariance depends only on the time between observations and not the absolute date t. This is an important property call *stationarity*, which will be discuss in the next section. In this case, we simply use to following notation:

$$\gamma(h) = \operatorname{cov}(X_t, X_{t+h}).$$

A few other remarks:

1. The covariance function is **symmetric**. That is, $\gamma(h) = \gamma(-h)$ since $\operatorname{cov}(X_t, X_{t+h}) = \operatorname{cov}(X_{t+h}, X_t)$.
2. Note that $\operatorname{var}(X_t) = \gamma(0)$.

3. We have that $|\gamma(h)| \leq \gamma(0)$ for all h. The proof of this inequality follows from Cauchy-Schwarz inequality, i.e.
$$\begin{aligned}(|\gamma(h)|)^2 = \gamma(h)^2 &= \left(E\left[\left(X_t - E[X_t]\right)\left(X_{t+h} - E[X_{t+h}]\right)\right]\right)^2 \\ &\leq E\left[\left(X_t - E[X_t]\right)^2\right] E\left[\left(X_{t+h} - E[X_{t+h}]\right)^2\right] = \gamma(0)^2.\end{aligned}$$
4. Just as any covariance, the $\gamma(h)$ is "scale dependent", $\gamma(h) \in \mathbb{R}$, or $-\infty \leq \gamma(h) \leq +\infty$
    1. If $|\gamma(h)|$ is "close" to 0, then they are "less dependent".
    2. If $|\gamma(h)|$ is "far" from 0, X_t and X_{t+h} are "more dependent".
5. $\gamma(h) = 0$ does not imply X_t and X_{t+h} are independent. This is only true in joint Gaussian case.

An important related statistic is the correlation of X_t with X_{t+h} or *autocorrelation* which is defined (for stationary processes) as

$$\rho(h) = \text{corr}(X_t, X_{t+h}) = \frac{\gamma(h)}{\gamma(0)}.$$

It is important to note that the above notation implies that the autocorrelation function is only a function of the lag h between observations. Thus, autocovariances and autocorrelations are one possible way to describe the joint distribution of a time series. Indeed, the correlation of X_t with X_{t+1} is an obvious measure of how *persistent* a time series is.

Remeber that just as with any correlation:

1. $\rho(h)$ is scale free.
2. $\rho(X_t, X_{t+h})$ is closer to $\pm 1 \Rightarrow (X_t, X_{t+h})$ "more dependent."
3. $|\rho(h)| \leq 1$ since $|\gamma(h)| \leq \gamma(0)$.
4. Causation and correlation are two very different things!

3.2.2 A Fundamental Representation

Autocovariances and autocorrelation also turn out to be a very useful tool because they are one of fundamental representations of time series. Indeed, if we consider a zero mean normally distrbuted process it is clear that its joint distribution is fully characterized by the autocariances $E[X_t X_{t+h}]$ (since the joint probability density only depends of these covariances). Once we know the autocovariances we know *everything* there is to know about the process and therefore:

> If two processes have the same autocovariance function, then they are the same process.

3.2.3 Admissible autocorrelation functions

Since the autocorrelation is related to a fundamental representation of time series it implies that one might be able to define a stochastic process by picking a set autocorrelation values. However, it turns out not every collection of numbers such as $\{\rho_1, \rho_2, ...\}$ is the autocorrelation of a process. Two conditions are required to ensure the validity of an autocorrelation sequence:

1. $\max_j |\rho_j| \leq 1$.
2. $\text{var}\left[\sum_{j=0}^{\infty} \alpha_j X_{t-j}\right] \geq 0$ for all $\{\alpha_0, \alpha_1, ...\}$.

The first condition is obvious and simply relects the fact that $|\rho(h)| \leq 1$ but the second is more difficult to verify. Let $\alpha_j = 0,\ j > 1$, then conditon 2 implies that

$$\text{var}\left[\alpha_0 X_t + \alpha_1 X_{t-1}\right] = \gamma_0 \begin{bmatrix}\alpha_0 & \alpha_1\end{bmatrix} \begin{bmatrix}1 & \rho_1 \\ \rho_1 & 1\end{bmatrix} \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix} \geq 0.$$

Thus, the matrix

$$\boldsymbol{A}_1 = \begin{bmatrix}1 & \rho_1 \\ \rho_1 & 1\end{bmatrix}$$

must be positive semi-definite. Therefore,

$$\det(\boldsymbol{A}_1) = 1 - \rho_1^2$$

implying that $|\rho_1| < 1$. Next, let $\alpha_j = 0,\ j > 2$, then we must verify that:

$$\text{var}\left[\alpha_0 X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2}\right] = \gamma_0 \begin{bmatrix}\alpha_0 & \alpha_1 & \alpha_2\end{bmatrix} \begin{bmatrix}1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1\end{bmatrix} \begin{bmatrix}\alpha_0 \\ \alpha_1 \\ \alpha_2\end{bmatrix} \geq 0.$$

Similarly, this implies that the matrix

$$\boldsymbol{A}_2 = \begin{bmatrix}1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1\end{bmatrix}$$

must be positive semi-definite. It is easy to verify that

$$\det(\boldsymbol{A}_2) = (1 - \rho_2)\left(-2\rho_1^2 + \rho_2 + 1\right).$$

It implies that $|\rho_2| < 1$ as well as

$$\begin{aligned}
&-2\rho_1^2 + \rho_2 + 1 \geq 0 \Rightarrow 1 > \rho_2 \geq 2\rho_1^2 - 1 \\
&\Rightarrow 1 - \rho_1^2 > \rho_2 - \rho_1^2 \geq -(1 - \rho_1^2) \\
&\Rightarrow 1 > \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \geq 1,
\end{aligned}$$

imlying that ρ_1 and ρ_2 must lie in a parabolic shaped region defined by the above inequalities. Therefore, the restrictions on the autocorrelation are very complicated providing a motivation for other form of fundamental representation.

3.3 Stationarity

3.3.1 Definitions

There are two kinds of stationarity which are commonly used. They are defined below:

- A process $\{X_t\}$ is **strongly stationary** or **strictly stationary** if the joint probability distribution of $\{X_{t-h}, ..., X_t, ..., X_{t+h}\}$ is independent of t for all h.

- A process $\{X_t\}$ is **weakly stationary**, **covariance stationary** or **second order stationary** if $E[X_t]$, $E[X_t^2]$ are finite and $E[X_t X_{t-h}]$ depends only on h and not on t.

These types of stationarity are **not equivalent** and the presence of **one kind of stationarity does not imply the other**. That is, a time series can be strongly stationary but not weakly stationary and vice versa. In some cases, a time series can be both strong and weakly stationary, this is happends for example in the (joint) Gaussian case. Stationarity of X_t matters, because **it provides the framework in which averaging dependent data makes sense**.

A few remarks:

- Strong stationarity $\not\Longrightarrow$ weak stationarity. *Example*: an iid Cauchy process is strongly but not weakly stationary.
- Weak stationarity $\not\Longrightarrow$ strong stationarity. *Example*: $X_{2t} = U_{2t}, X_{2t+1} = V_{2t+1} \forall t$ where $U_t \overset{iid}{\sim} N(1,1)$ and $V_t \overset{iid}{\sim} Exponential(1)$ is weakly stationary but **NOT** strongly stationary.
- Strong stationarity + $E[X_t]$, $E[X_t^2] < \infty \implies$ weak stationarity
- Weak stationarity + normality $\implies$ strong stationarity.

3.3.2 Assessing Weak Stationarity of Time Series Models

In order to verify if a process is weakly stationary, we must make sure the process satisfies:

1. $E[X_t] = \mu_t = \mu < \infty$,
2. $\text{var}[X_t] = \sigma_t^2 = \sigma^2 < \infty$,
3. $\text{cov}(X_t, X_{t+h}) = \gamma(h)$.

3.3.2.1 Example: Gaussian White Noise

It is easy to verify that a Gaussian white noise is stationary. Indeed, we have:

1. $E[X_t] = 0$,
2. $\gamma(0) = \sigma^2 < \infty$,
3. $\gamma(h) = 0$ for $|h| > 0$.

3.3.2.2 Example: Random Walk

To evaluate the stationarity of a random walk we first derive its properties:

1. $$E[X_t] = E[X_{t-1} + W_t] = E\left[\sum_{i=1}^{t} W_t + X_0\right]$$
$$= E\left[\sum_{i=1}^{t} W_t\right] + X_0 = X_0$$
Note, the mean here is constant since it depends only on the value of the first term in the sequence.
2. $$\text{var}(X_t) = \text{var}\left(\sum_{i=1}^{t} W_t + X_0\right) = \text{var}\left(\sum_{i=1}^{t} w_t\right) + \underbrace{\text{var}(X_0)}_{=0}$$
$$= \sum_{i=1}^{t} Var(w_t) = t\sigma_w^2.$$

where $\sigma_w^2 = \text{var}(W_t)$. Therefore, the variance has a dependence on time and we have:

$$\lim_{t\to\infty} \text{var}\,(X_t) = \infty.$$

As a result, the process is not weakly stationary.

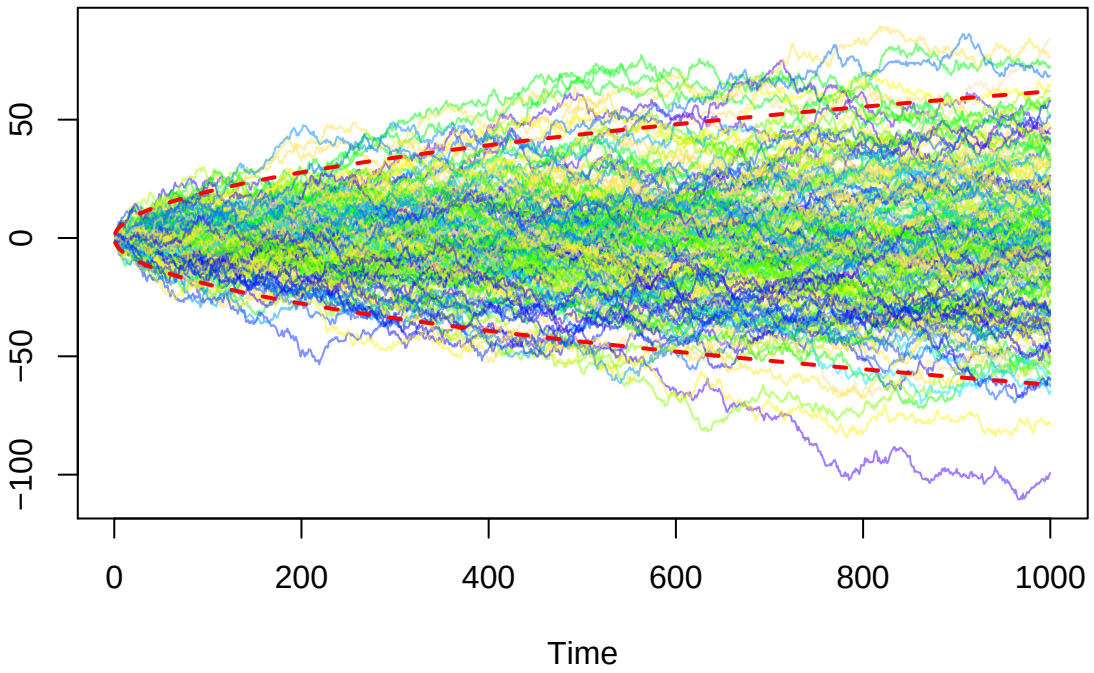
Continuing on just to obtain the covariance, we have:

$$\begin{aligned}\gamma(h) = Cov\,(y_t, y_{t+h}) &= Cov\left(\sum_{i=1}^{t} w_i, \sum_{j=1}^{t+h} w_j\right) \\ &= Cov\left(\sum_{i=1}^{t} w_i, \sum_{j=1}^{t} w_j\right) = \min(t, t+h)\,\sigma_w^2 \\ &= (t + \min(0, h))\,\sigma_w^2,\end{aligned}$$

which also illustrates that non-stationarity of a random walk.

In the following simulated example, we illustrate the non-stationary feature of such process:

```
# In this example, we simulate a large number of random walks
# Number of simulated processes
B = 200

# Length of random walks
n = 1000

# Output matrix
out = matrix(NA,B,n)

for (i in 1:B){
  # Simulate random walk
  Xt = cumsum(rnorm(n))

  # Store process
  out[i,] = Xt
}

# Plot random walks
plot(NA, xlim = c(1,n), ylim = range(out), xlab = "Time", ylab = " ")
color = sample(topo.colors(B, alpha = 0.5))
for (i in 1:B){
  lines(out[i,], col = color[i])
}

# Add 95% confidence region
lines(1:n, 1.96*sqrt(1:n), col = 2, lwd = 2, lty = 2)
lines(1:n, -1.96*sqrt(1:n), col = 2, lwd = 2, lty = 2)
```



The relationship between time and variance can clearly be observed in the above graph.

3.3.2.3 Example: MA(1)

To evaluate the stationarity of an MA(1) process we first derive its properties:

1.
$$\begin{aligned} E[y_t] &= E[\theta_1 w_{t-1} + w_t] \\ &= \theta_1 E[w_{t-1}] + E[w_t] = 0 \end{aligned}$$

2.
$$\begin{aligned} Cov(y_t, y_{t+h}) &= E[(y_t - E[y_t])(y_{t+h} - E[y_{t+h}])] \\ &= E[y_t y_{t+h}] - \underbrace{E[y_t]}_{=0}\underbrace{E[y_{t+h}]}_{=0} \\ &= E[(\theta_1 w_{t-1} + w_t)(\theta_1 w_{t+h-1} + w_{t+h})] \\ &= E\left[\theta_1^2 w_{t-1} w_{t+h-1} + \theta w_t w_{t+h} + \theta_1 w_{t-1} w_{t+h} + w_t w_{t+h}\right] \end{aligned}$$

$$E[w_t w_{t+h}] = \operatorname{cov}(w_t, w_{t+h}) + E[w_t]E[w_{t+h}] = 1_{\{h=0\}}\sigma_w^2$$

$$\Rightarrow Cov(y_t, y_{t+h}) = \left(\theta_1^2 1_{\{h=0\}} + \theta_1 1_{\{h=1\}} + \theta_1 1_{\{h=-1\}} + 1_{\{h=0\}}\right)\sigma_w^2$$

$$\gamma(h) = \begin{cases} \left(\theta_1^2 + 1\right)\sigma_w^2 & h = 0 \\ \theta_1 \sigma_w^2 & |h| = 1 \\ 0 & |h| > 1 \end{cases}$$

Therefore, an MA(1) process is weakly stationary since both the mean and variance are constant over time. In addition, we can easily obtain the autocorrelation function which is given by

$$\Rightarrow \rho(h) = \begin{cases} 1 & h = 0 \\ \frac{\theta_1 \sigma_w^2}{(\theta_1^2+1)\sigma_w^2} = \frac{\theta_1}{\theta_1^2+1} & |h| = 1 \\ 0 & |h| > 1 \end{cases}$$

Interestingly, we can note that $|\rho(1)| \leq 0.5$.

3.3.2.4 Example: MA(1)

Consider the AR(1) process given as:

$$y_t = \phi_1 y_{t-1} + w_t, \text{ where } w_t \overset{iid}{\sim} WN\left(0, \sigma_w^2\right)$$

This process was shown to simplify to:

$$y_t = \phi^t y_0 + \sum_{i=0}^{t-1} \phi_1^i w_{t-i}$$

In addition, we add the requirement that $|\phi_1| < 1$. This requirement allows for the process to be stationary. If $\phi_1 \geq 1$, the process would not converge. This way the process will be able to be written as a geometric series that converges:

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}, \;\; |r| < 1$$

Next, we demonstrate how crucial this property is:

$$\begin{aligned}
\lim_{t\to\infty} E\left[y_t\right] &= \lim_{t\to\infty} E\left[\phi^t y_0 + \sum_{i=0}^{t-1} \phi_1^i w_{t-i}\right] \\
&= \lim_{t\to\infty} \underbrace{\phi^t y_0}_{|\phi|<1 \Rightarrow t\to\infty \;=\; 0} + \sum_{i=0}^{t-1} \phi_1^i \underbrace{E\left[w_{t-i}\right]}_{=0} \\
&= 0 \\
\lim_{t\to\infty} Var\left(y_t\right) &= \lim_{t\to\infty} Var\left(\phi^t y_0 + \sum_{i=0}^{t-1} \phi_1^i w_{t-i}\right) \\
&= \lim_{t\to\infty} \underbrace{Var\left(\phi^t y_0\right)}_{=0 \text{ since constant}} + Var\left(\sum_{i=0}^{t-1} \phi_1^i w_{t-i}\right) \\
&= \lim_{t\to\infty} \sum_{i=0}^{t-1} \phi_1^{2i} Var\left(w_{t-i}\right) \\
&= \lim_{t\to\infty} \sigma_w^2 \sum_{i=0}^{t-1} \phi_1^{2i} \\
&= \sigma_w^2 \cdot \underbrace{\frac{1}{1-\phi^2}}_{\text{Geometric Series}}
\end{aligned}$$

This leads us to being able to conclude the autocovariance function is:

$$\begin{aligned} Cov\left(y_t, y_{t+h}\right) &= Cov\left(y_t, \phi y_{t+h-1} + w_{t+h}\right) \\ &= Cov\left(y_t, \phi y_{t+h-1}\right) \\ &= Cov\left(y_t, \phi^{|h|} y_t\right) \\ &= \phi^{|h|} Cov\left(y_t, y_t\right) \\ &= \phi^{|h|} Var\left(y_t\right) \\ &= \phi^{|h|} \frac{\sigma_w^2}{1 - \phi_1^2} \end{aligned}$$

Both the mean and autocovariance function do not depend on time and, thus, the AR(1) process is stationary if $|\phi_1| < 1$.

If we assume that the AR(1) process is stationary, we can derive the mean and variance in another way. Without a loss of generality, we'll assume $y_0 = 0$.

Therefore:

$$\begin{aligned} y_t &= \phi_t y_{t-1} + w_t \\ &= \phi_1 \left(\phi_1 y_{t-2} + w_{t-1}\right) + w_t \\ &= \phi_1^2 y_{t-2} + \phi_1 w_{t-1} + w_t \\ &\vdots \\ &= \sum_{i=0}^{t-1} \phi_1^i w_{t-i} \end{aligned}$$

$$\begin{aligned} E\left[y_t\right] &= E\left[\sum_{i=0}^{t-1} \phi_1^i w_{t-i}\right] \\ &= \sum_{i=0}^{t-1} \phi_1^i \underbrace{E\left[w_{t-i}\right]}_{=0} \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var\left(y_t\right) &= E\left[\left(y_t - E\left[y_t\right]\right)^2\right] \\ &= E\left[y_t^2\right] - \left(E\left[y_t\right]\right)^2 \\ &= E\left[y_t^2\right] \\ &= E\left[\left(\phi_1 y_{t-1} + w_t\right)^2\right] \\ &= E\left[\phi_1^2 y_{t-1}^2 + w_t^2 + 2\phi_1 y_t w_t\right] \\ &= \phi_1^2 E\left[y_{t-1}^2\right] + \underbrace{E\left[w_t^2\right]}_{=\sigma_w^2} + 2\phi_1 \underbrace{E\left[y_t\right]}_{=0} \underbrace{E\left[w_t\right]}_{=0} \\ &= \underbrace{\phi_1^2 Var\left(y_{t-1}\right) + \sigma_w^2 = \phi_1^2 Var\left(y_t\right) + \sigma_w^2}_{\text{Assume stationarity}} \end{aligned}$$

$$\begin{aligned} Var\left(y_t\right) &= \phi_1^2 Var\left(y_t\right) + \sigma_w^2 \\ Var\left(y_t\right) - \phi_1^2 Var\left(y_t\right) &= \sigma_w^2 \\ Var\left(y_t\right)\left(1 - \phi_1^2\right) &= \sigma_w^2 \\ Var\left(y_t\right) &= \frac{\sigma_w^2}{1 - \phi_1^2} \end{aligned}$$

3.3.3 Esimtation of the Mean Function

If a time series is stationary, the mean function is constant and a possible estimator of this quantity is given by

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t.$$

This estimator is clearly unbiased and has the following variance:

$$
\begin{aligned}
\operatorname{var}\left(\bar{X}\right) &= \operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) \\
&= \frac{1}{n^2}\operatorname{var}\left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}_{1\times n}\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}_{n\times 1}\right) \\
&= \frac{1}{n^2}\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}_{1\times n}\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & & \vdots \\ \vdots & & \ddots & \vdots \\ \gamma(n-1) & \cdots & \cdots & \gamma(0) \end{bmatrix}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}. \\
&= \frac{1}{n^2}\left(n\gamma(0) + 2(n-1)\gamma(1) + 2(n-2)\gamma(2) + \cdots + 2\gamma(n-1)\right) \\
&= \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma(h)
\end{aligned}
$$

In the white noise case, the above formula reduces to the usual $\operatorname{var}\left(\bar{X}\right) = \operatorname{var}(X_t)/n$.

3.3.4 Sample Autocovariance and Autocorrelation Functions

A natural estimator of the **autocovariance function** is given as:

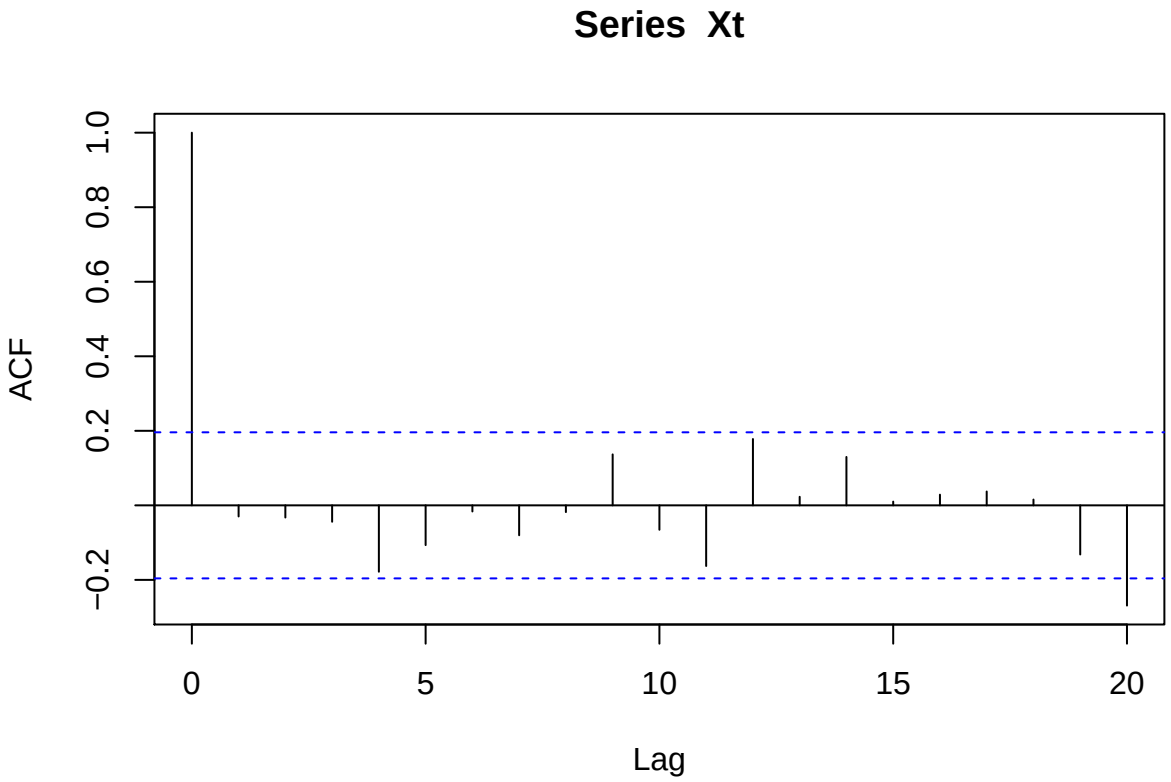
$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}\left(X_t - \bar{X}\right)\left(X_{t+h} - \bar{X}\right)$$

leading the following "plug-in" estimator of the **autocorrelation function**

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

A graphical representation of the autocorrelation function is often the first step of any time series analysis (assuming the process to be stationary). Consider the following simulated example:

```
# Simulate iid gaussian RV (i.e. white noise)
Xt = rnorm(100)

# Compute autocorrelation
acf_Xt = acf(Xt)

# Plot autocorrelation
plot(acf_Xt)
```

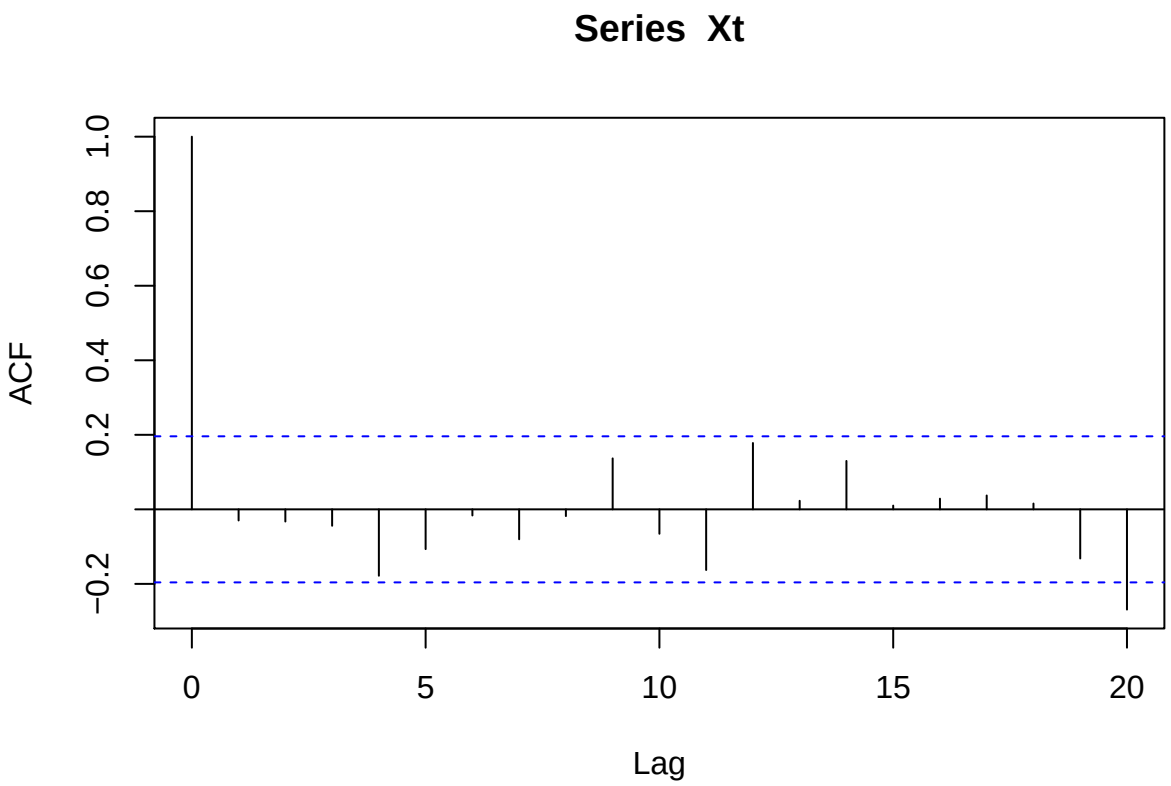


In this example, the true autocorrelation at lag h ($|h| > 0$) is equal 0 but obviously the estimated autocorrelations are random variables and are not equal to their true value. It would therefore be usefull to have have some knowledge about the variability of the sample autocorrelations (under some conditions) to assess whether the data comes from a completely random series or presents some significant correlation at some lags. The following result provide an asymptotic solution to this problem:

> If X_t is white noise with finite fourth moment, then $\hat{\rho}(h)$ is approximately normally distributed with mean 0 and variance T^{-1} for all fixed h.

Using on this result, we now have an approximate method to assess whether peaks in sample autocorrelation are significant by determining whether the observed peak lies outside the interval $+/- 2/\sqrt{T}$ (i.e. an approximate 95% confidence interval). Returning to our previous example:

```
# Plot autocorrelation with confidence bands
plot(acf_Xt)
```

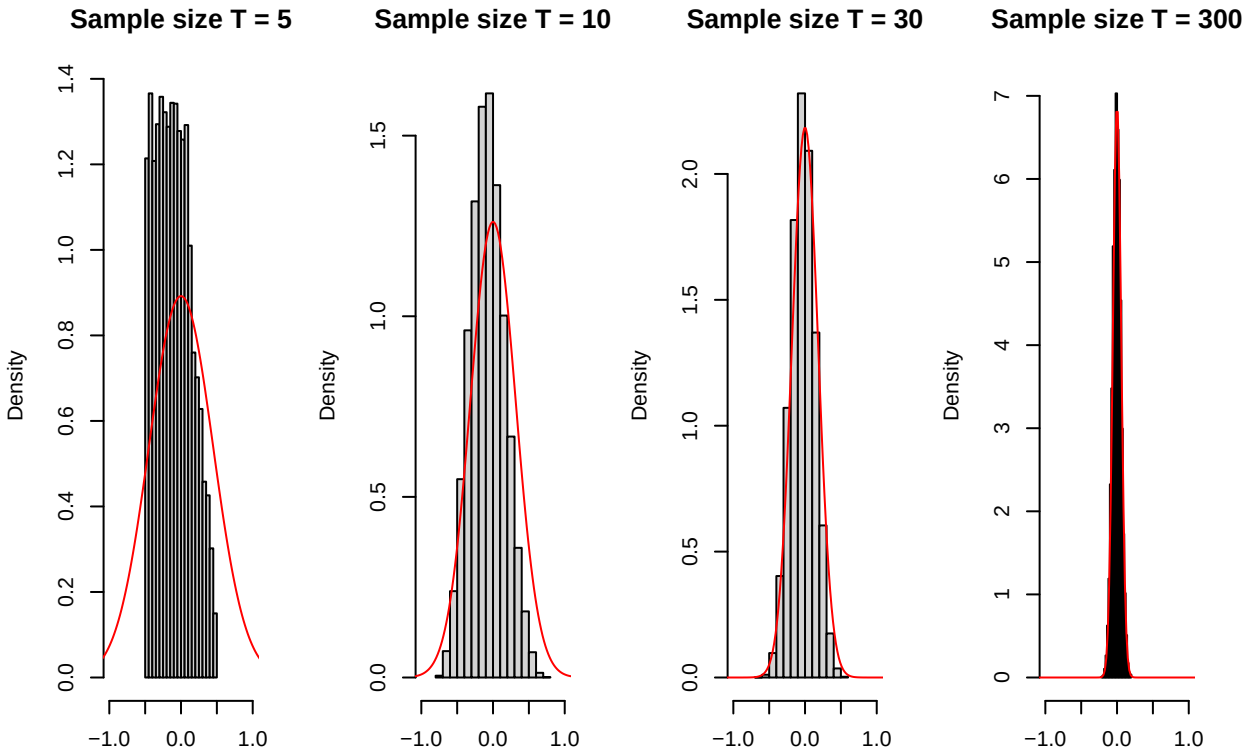


It can now be observed that most peaks lies within the interval $+/-2/\sqrt{T}$ suggesting that the true data generating process is completely random (in the linear sense).

Unfortunately, this method is asymptotic (it relies on the central limit theorem) and there no "exact" tools that can be used in this case. In the simulation study below consider the "quality" of this result for $h = 3$ considering different sample sizes:

```
# Number of Monte Carlo replications
B = 10000

# Define considered lag
h = 3

# Sample size considered
T = c(5,10,30,300)

# Initialisation
result = matrix(NA,B,length(T))

# Set seed
set.seed(1)

# Start Monte Carlo
for (i in 1:B){
  for (j in 1:length(T)){
    # Simluate process
    Xt = rnorm(T[j])

    # Save autocorrelation at lag h
    result[i,j] = acf(Xt, plot = FALSE)$acf[h+1]
```

```
  }
}

# Plot results
par(mfrow = c(1,length(T)))
for (i in 1:length(T)){
  # Estimated empirical distribution
  hist(result[,i], col = "lightgrey", main = paste("Sample size T =",T[i]), probability = TRUE, xlim = c

  # Asymptotic distribution
  xx = seq(from = -10, to = 10, length.out = 10^3)
  yy = dnorm(xx,0,1/sqrt(T[i]))
  lines(xx,yy, col = "red")
}
```



It can clearly be observed that asymptotic approximation is quite poor when $T = 5$ but as the sample size increases the approximation becomes more appropriate and is nearly perfect with $T = 300$.

3.4 Joint Stationarity

Two time series, say (X_t) and (Y_t), are said to be jointly stationary if they are each stationary, and the cross-covariance function

$$\gamma_{XY}(t, t+h) = Cov(X_t, Y_{t+h}) = \gamma_{XY}(h)$$

is a function only of lag h.

The cross-correlation function for jointly stationary times can be expressed as:

$$\rho_{XY}(t, t+h) = \frac{\gamma_{XY}(t, t+h)}{\sigma_{X_t}\sigma_{Y_{t+h}}} = \frac{\gamma_{XY}(h)}{\sigma_{X_t}\sigma_{Y_{t+h}}} = \rho_{XY}(h)$$

Chapter 4

Basic Models

4.1 The Backshift Operator

Definition: **Backshift Operator**

The **Backshift Operator** is helpful when manipulating time series. When we backshift, we are changing the indices of the time series. e.g. $t \to t-1$. The operator is defined as:

$$Bx_t = x_{t-1}$$

If we were to repeatedly apply the backshift operator, we would receive:

$$\begin{aligned} B^2 x_t &= B\left(Bx_t\right) \\ &= B\left(x_{t-1}\right) \\ &= x_{t-2} \end{aligned}$$

We can generalize this behavior as:

$$B^k x_t = x_{t-k}$$

The backshift operator is helpful for later decompositions in addition to making differencing operations more straightforward.

4.2 White Noise

The process name of white noise has meaning in the notion of colors of noise. Specifically, the white noise is a process that mirrors white light's flat frequency spectrum. So, the process has equal frequencies in any interval of time.

Definition: **White Noise**

w_t or ε_t is a **white noise process** if w_t are uncorrelated identically distributed random variables with $E\left[w_t\right] = 0$ and $Var\left[w_t\right] = \sigma^2$, for all t. We can represent this algebraically as:

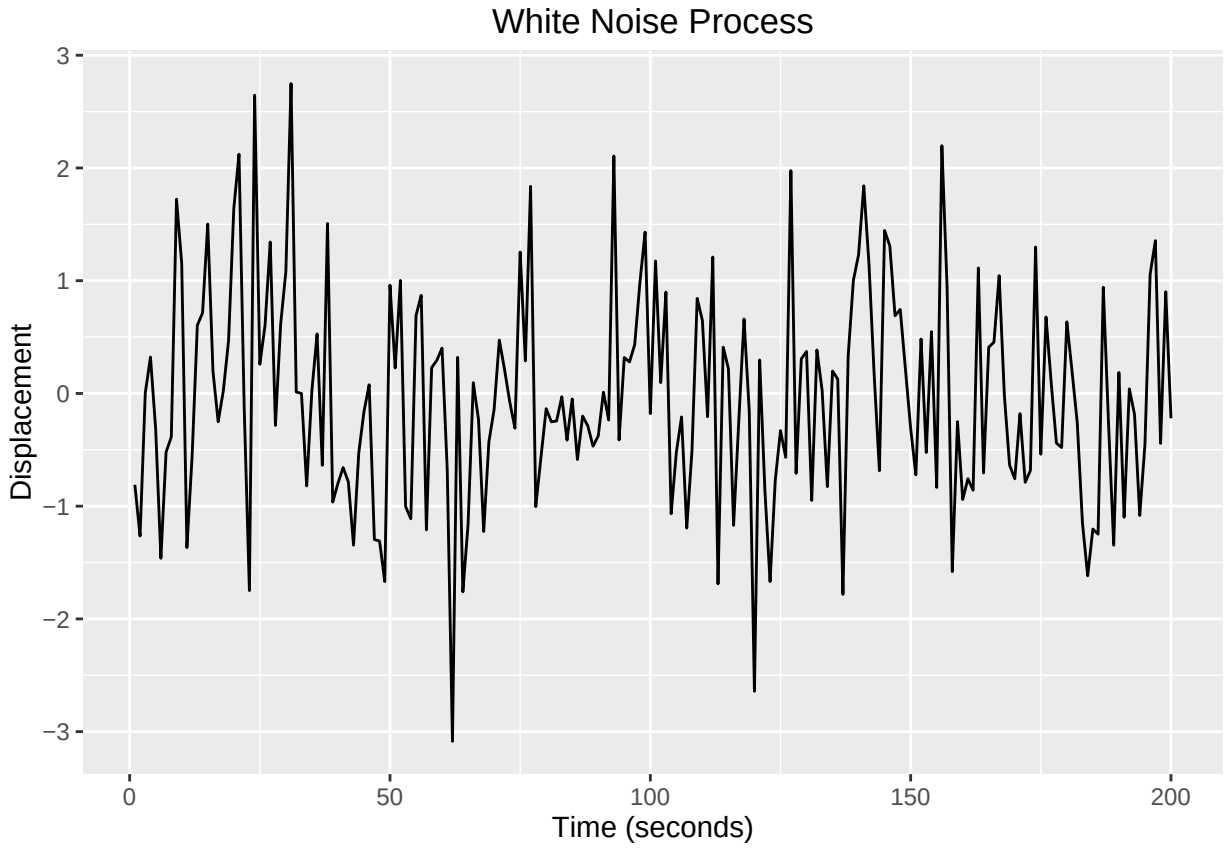
$$y_t = w_t,$$

where $w_t \overset{id}{\sim} WN\left(0, \sigma_w^2\right)$

Now, if the w_t are **Normally (Gaussian) distributed**, then the process is known as a **Gaussian White Noise** e.g. $w_t \overset{iid}{\sim} N\left(0, \sigma^2\right)$

To generate gaussian white noise use:

```
set.seed(1336)           # Set seed to reproduce the results
n  = 200                 # Number of observations to generate
wn = ts(rnorm(n,0,1))    # Generate Guassian white noise.

autoplot(wn) +
  ggtitle("White Noise Process") +
  ylab("Displacement") + xlab("Time (seconds)")
```



4.3 Moving Average Process of Order q = 1 a.k.a MA(1)

Definition: **Moving Average Process of Order (q = 1)**

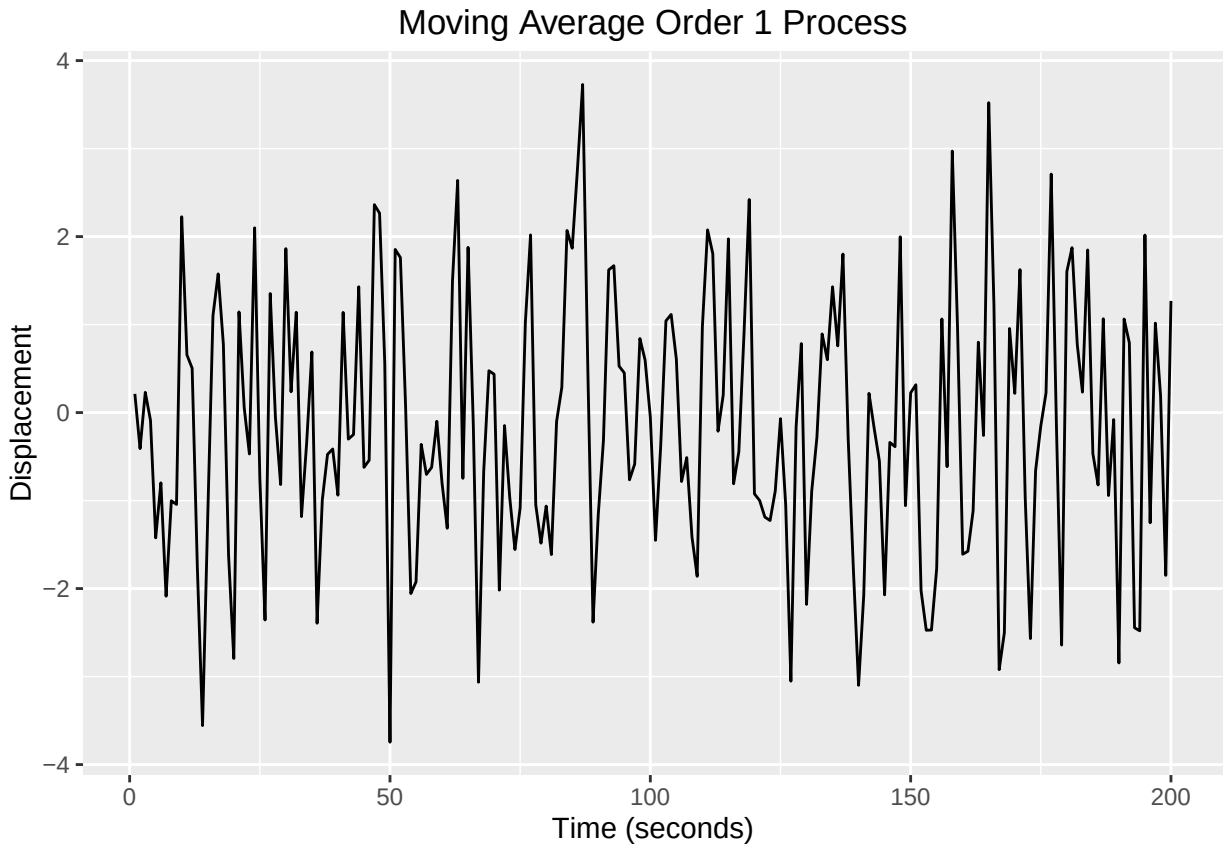
The concept of a **Moving Average Process of Order q** is a way to remove "noise" and emphasize the signal. The moving average achieves this by taking the local averages of the data to produce a new smoother time series series. The newly created time series is more descriptive, but it does influence the dependence within the time series.

This process is generally denoted as **MA(1)** and is defined as:

$$y_t = \theta_1 w_{t-1} + w_t,$$

where $w_t \overset{iid}{\sim} WN\left(0, \sigma_w^2\right)$

```r
set.seed(1345)  # Set seed to reproduce the results
n      = 200    # Number of observations to generate
sigma2 = 2      # Controls variance of Guassian white noise.
theta  = 0.3    # Handles the theta component of MA(1)


# Generate a white noise
wn = rnorm(n+1, sd = sqrt(sigma2))

# Simulate the MA(1) process
ma = rep(0, n+1)
for(i in 2:(n+1)) {
  ma[i] = theta*wn[i-1] + wn[i]
}

ma = ts(ma[2:(n+1)])      # Remove first item

autoplot(ma) +
  ggtitle("Moving Average Order 1 Process") +
  ylab("Displacement") + xlab("Time (seconds)")
```



4.4 Drift

Definition: **Drift**

A **drift process** has two components: time and a slope. As more points are accumlated over time, the drift will match the common slope form.

Specifically, the drift process has the following form:

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

with the initial condition $y_0 = c$.

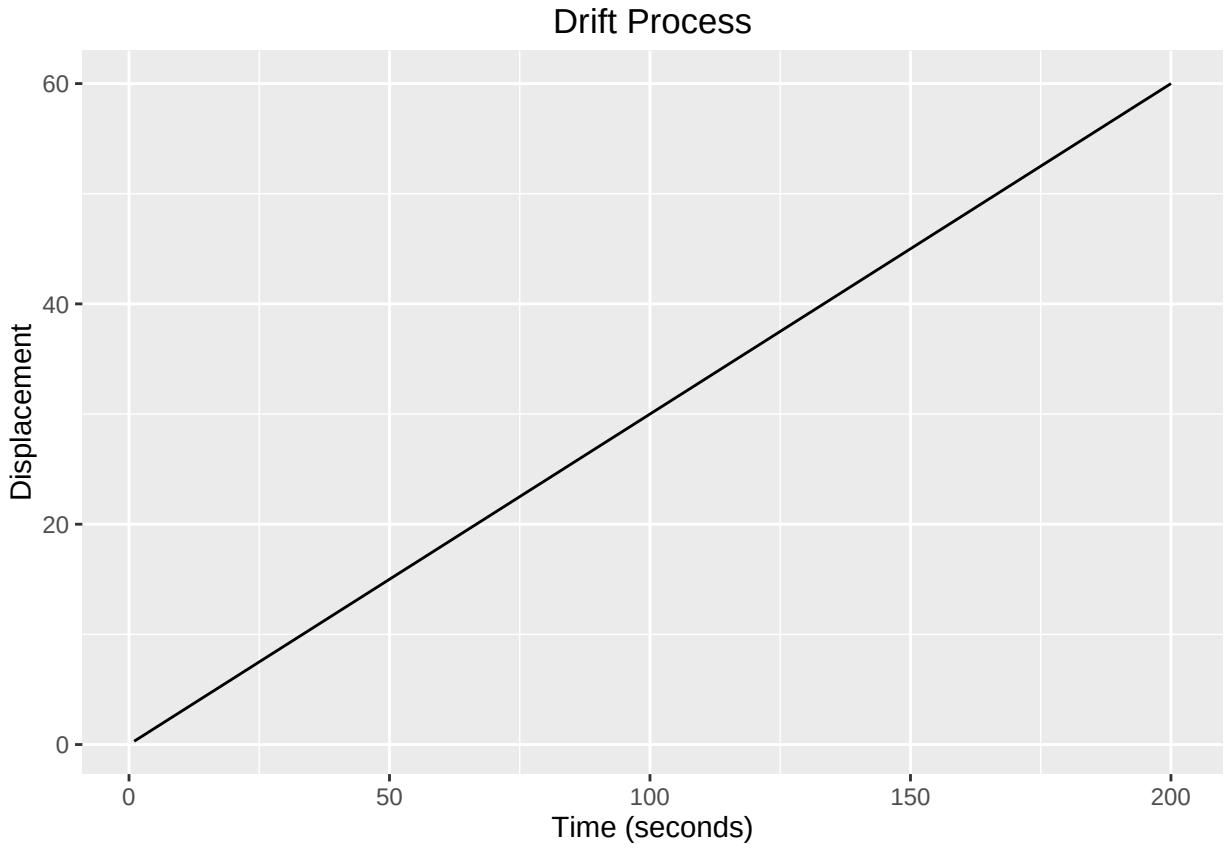
The process can be simplified using **backsubstitution** to being:

$$\begin{aligned} y_t &= y_{t-1} + \delta \\ &= (y_{t-2} + \delta) + \delta \\ &\vdots \\ &= \sum_{i=1}^{t} \delta + y_0 \\ y_t &= t\delta + c \end{aligned}$$

Again, note that a drift is similar to the slope-intercept form a linear line. e.g. $y = mx + b$.

To generate a drift use:

```
n     = 200             # Number of observations to generate
drift = .3              # Drift Control
dr    = ts(drift*(1:n)) # Generate drift sequence (e.g. y = drift*x + 0)

autoplot(dr) +
  ggtitle("Drift Process") +
  ylab("Displacement") + xlab("Time (seconds)")
```



4.5 Random Walk

In 1906, Karl Pearson coined the term 'random walk' and demonstrated that "the most likely place to find a drunken walker is somewhere near his starting point." Empirical evidence of this phenomenon is not too hard to find on a Friday night in Champaign.

Definition: **Random Walk**

A **random walk** is defined as a process where the current value of a variable is composed of the past value plus an error term that is a white noise. In algebraic form,

$$y_t = y_{t-1} + w_t$$

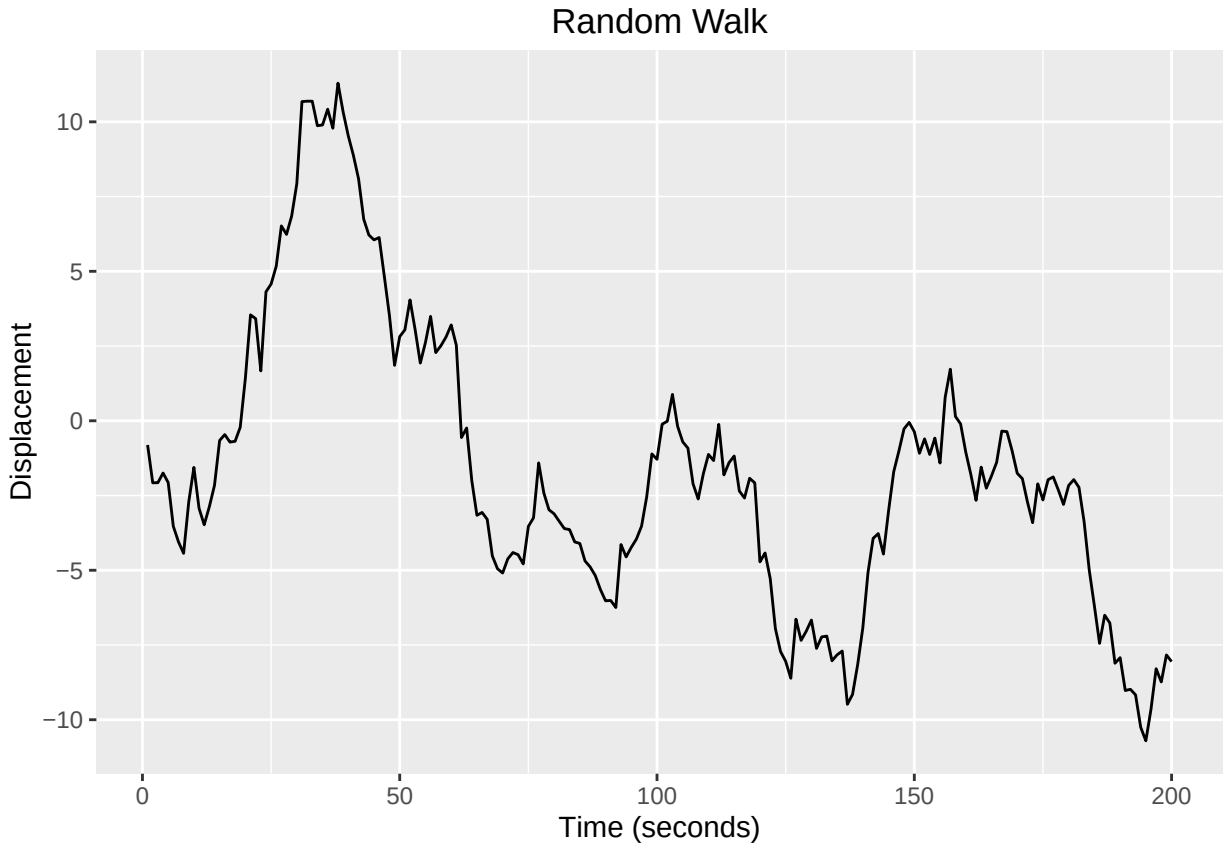
with the initial condition $y_0 = c$.

The process can be simplified using **backsubstitution** to being:

$$\begin{aligned} y_t &= y_{t-1} + w_t \\ &= (y_{t-2} + w_{t-1}) + w_t \\ &\vdots \\ y_t &= \sum_{i=1}^{t} w_i + y_0 = \sum_{i=1}^{t} w_i + c \end{aligned}$$

To generate a random walk, we use:

```r
set.seed(1336)        # Set seed to reproduce the results
n  = 200              # Number of observations to generate
w  = rnorm(n,0,1)     # Generate Guassian white noise.
rw = ts(cumsum(w))    # Cumulative sum

# Create a data.frame to graph in ggplot2
autoplot(rw) +
  ggtitle("Random Walk") +
  ylab("Displacement") + xlab("Time (seconds)")
```



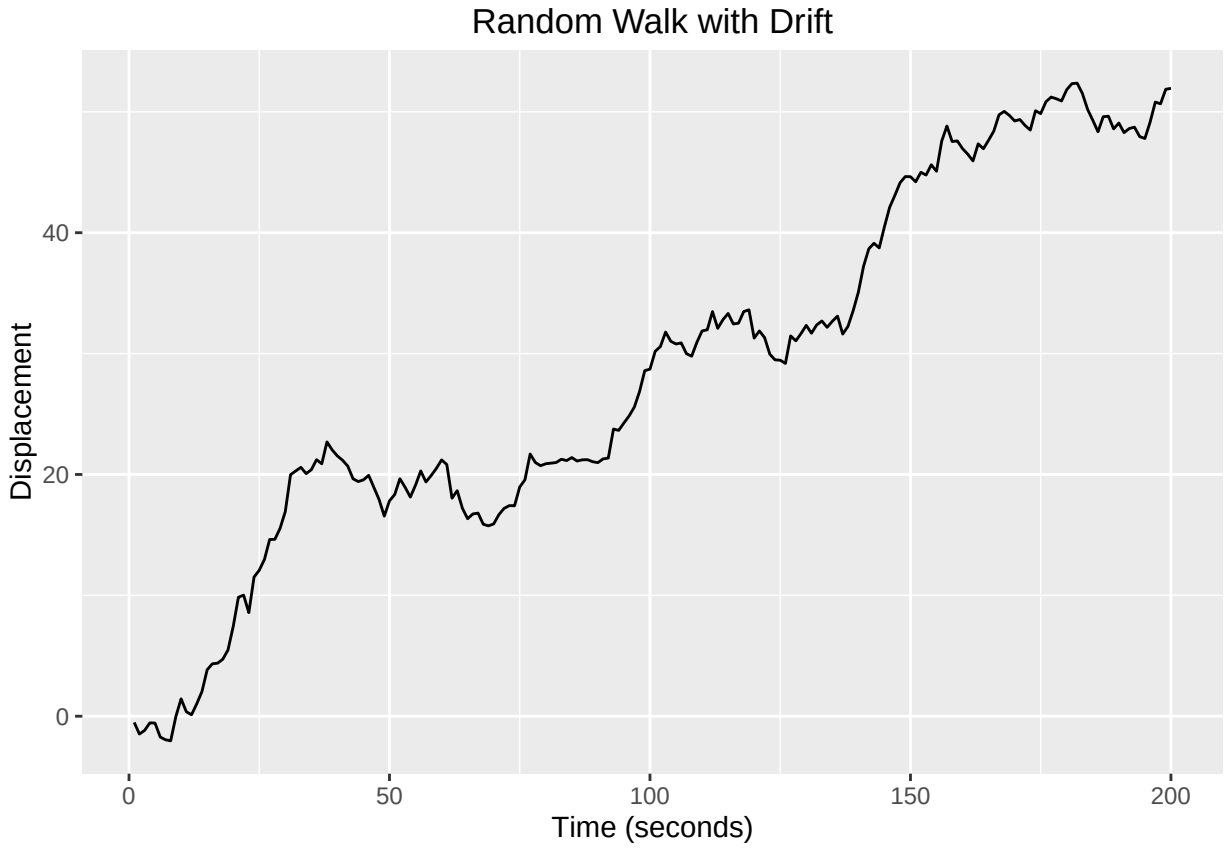
4.6 Random Walk with Drift

In the previous case of a random walk, we assumed that drift, δ, was equal to 0. What happens to the random walk if the drift is not equal to zero? That is, what happens with the initial condition $y_0 = c$?

$$\begin{aligned} y_t &= y_{t-1} + w_t + \delta \\ &= (y_{t-2} + w_{t-1} + \delta) + w_t + \delta \\ &\vdots \\ y_t &= \sum_{i=1}^{t} (w_i + \delta) + y_0 = \sum_{i=1}^{t} w_i + t\delta + c \end{aligned}$$

To generate a random walk with drift we use:

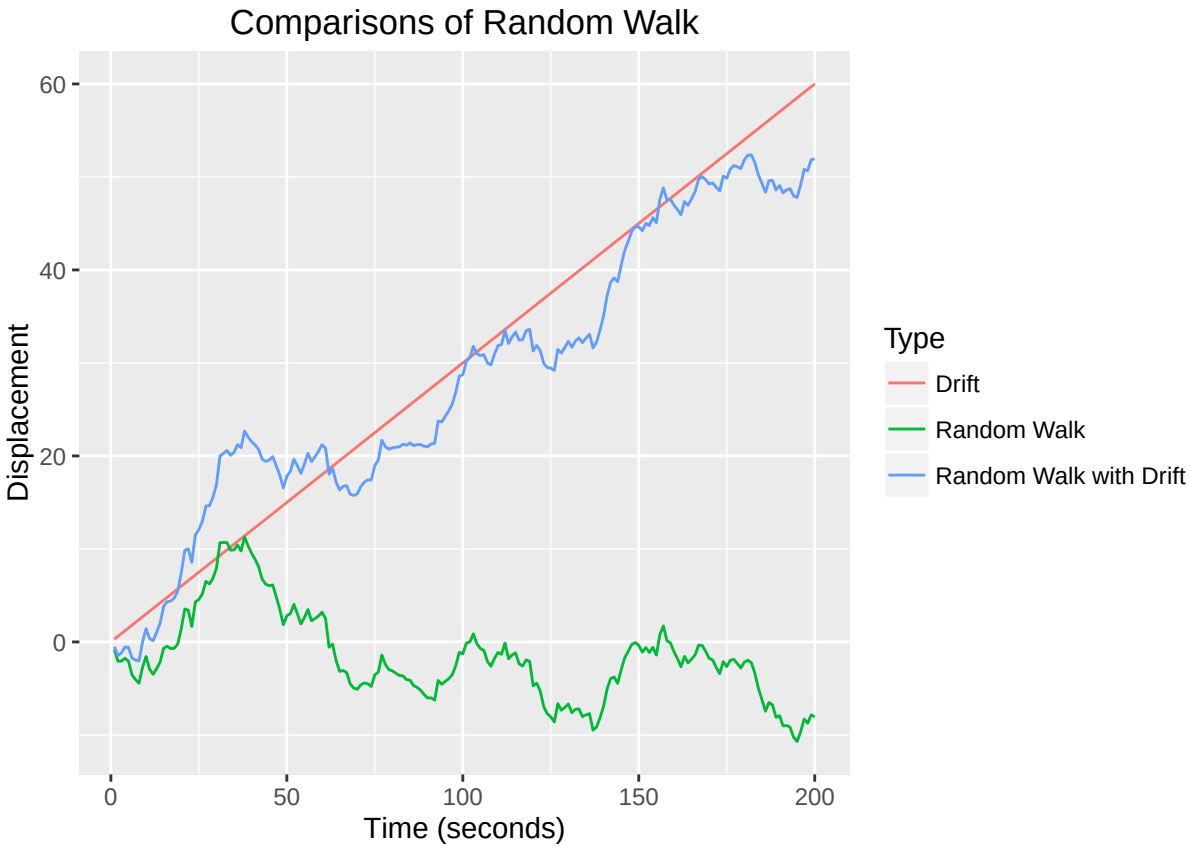
```
set.seed(1336)        # Set seed to reproduce the results
n     = 200           # Number of observations to generate
drift = .3            # Drift Control

w = rnorm(n,0,1)      # Generate Guassian white noise.
wd = w + drift        # Add a drift
rwd = ts(cumsum(wd)) # Cumulative sum

# Create a data.frame to graph in ggplot2
autoplot(rwd) +
  ggtitle("Random Walk with Drift") +
  ylab("Displacement") + xlab("Time (seconds)")
```



Notice the difference the drift makes upon the random walk:

```
# Add identifiers
drift.df = data.frame(Index = 1:n, Data = drift*(1:n), Type = "Drift")

rw.df = data.frame(Index = 1:n, Data = rw, Type = "Random Walk")

rwd.df = data.frame(Index = 1:n, Data = rwd, Type = "Random Walk with Drift")

combined.df = rbind(drift.df, rw.df, rwd.df)

ggplot(data = combined.df, aes(x = Index, y = Data, colour = Type)) +
  geom_line() +
```

```
  ggtitle("Comparisons of Random Walk") +
  ylab("Displacement") + xlab("Time (seconds)")
```



4.7 Autoregressive Process of Order p = 1 a.k.a AR(1)

Definition: **Autoregressive Process of Order p = 1**

This process is generally denoted as **AR(1)** and is defined as: $y_t = \phi_1 y_{t-1} + w_t$,

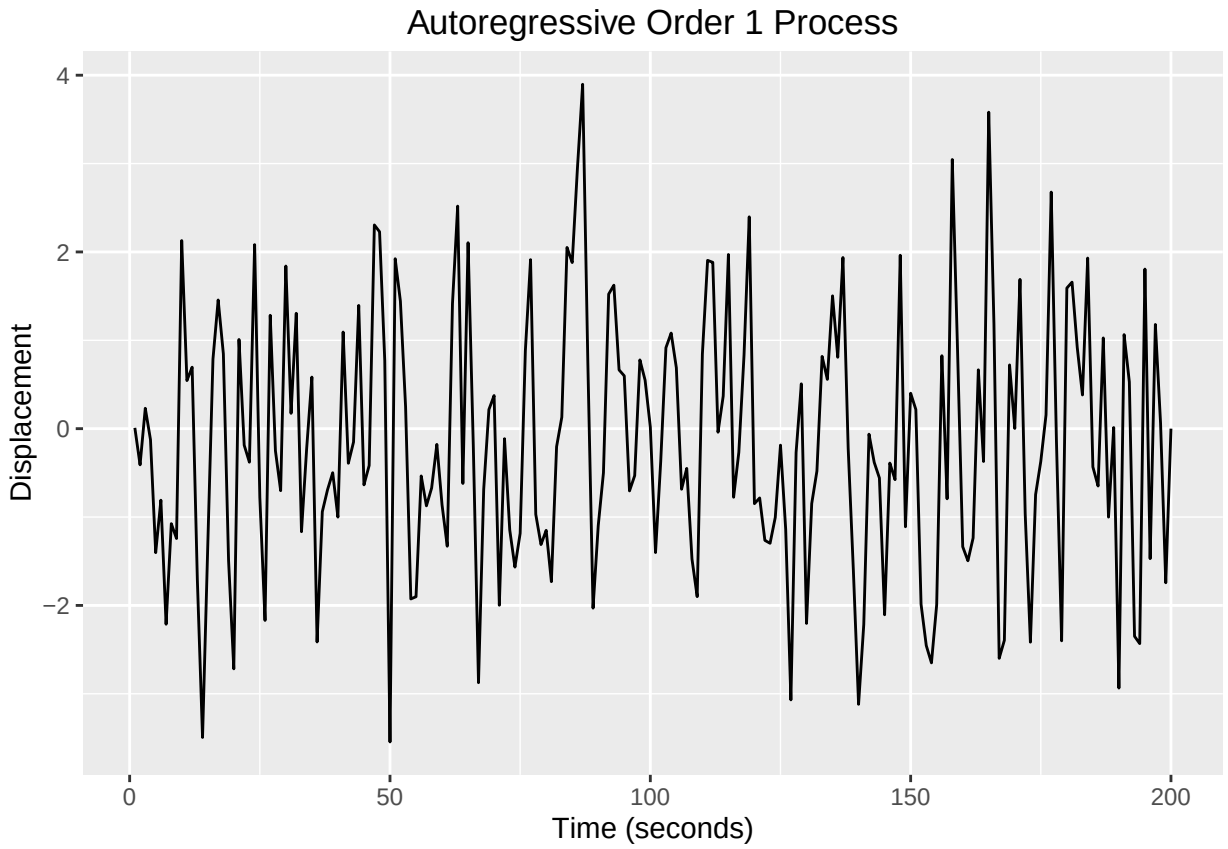
where $w_t \overset{iid}{\sim} WN\left(0, \sigma_w^2\right)$

If $\phi_1 = 1$, then the process is equivalent to a random walk.

The process can be simplified using **backsubstitution** to being:

$$\begin{aligned} y_t &= \phi_t y_{t-1} + w_t \\ &= \phi_1 \left(\phi_1 y_{t-2} + w_{t-1}\right) + w_t \\ &= \phi_1^2 y_{t-2} + \phi_1 w_{t-1} + w_t \\ &\vdots \\ &= \phi^t y_0 + \sum_{i=0}^{t-1} \phi_1^i w_{t-i} \end{aligned}$$

```r
set.seed(1345)  # Set seed to reproduce the results
n      = 200    # Number of observations to generate
sigma2 = 2      # Controls variance of Guassian white noise.
phi    = 0.3    # Handles the phi component of AR(1)

wn = rnorm(n+1, sd = sqrt(sigma2))

# Simulate the MA(1) process
ar = rep(0,n+1)
for(i in 2:n) {
  ar[i] = phi*ar[i-1] + wn[i]
}

ar = ts(ar[2:(n+1)])

autoplot(ar) +
  ggtitle("Autoregressive Order 1 Process") +
  ylab("Displacement") + xlab("Time (seconds)")
```



Chapter 5

ARMA

5.1 Definition

5.2 MA / AR Operators

5.3 Redundancy

5.4 Causal + Invertible

5.5 Estimation of Parameters

There are two primary methods for estimating parameters: Maximum Likelihood Estimation and Method of Moments.

Definition Consider $X_n = (X_1, X_2, \ldots, X_n)$ with the joint density $f(X_1, X_2, \ldots, X_n; \theta)$ where $\theta \in \Theta$. Given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is observed, we have the likelihood function of θ as

$$L(\theta) = L(\theta | x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n | \theta)$$

If the X_i are iid, then the likelihood simplifies to:

$$L(\theta) = \prod_{i=1}^{n} f\left(x_i | \theta\right)$$

However, that's a bit painful to maximize with calculus. So, we opt to use the log of the function since derivatives are easier and the logarithmic function is always increasing. Thus, we traditionally use:

$$l\left(\theta\right) = \log\left(L\left(\theta\right)\right) = \sum_{i=1}^{n} \log\left(f\left(x_i | \theta\right)\right)$$

From maximizes the likelihood function $L(\theta)$, we get the **maximum likelihood estimate (MLE)** of θ. So, we end up with a value that makes the observed data the "most probable."

Note: The likelihood function is **not** a probability density function.

Chapter 6

$AR(1)$ with mean μ

Consider an $AR(1)$ process given as $y_t = \phi y_{t-1} + w_t$, $w_t \overset{iid}{\sim} N\left(0, \sigma^2\right)$, with $E\left[y_t\right] = 0$, $|\phi| < 1$.

Let $x_t = y_t + \mu$, so that $E\left[x_t\right] = \mu$.

Then, $x_t - \mu = y_t$. Substituting in for y_t, we get:

$$\begin{aligned} y_t &= \phi y_{t-1} + w_t \\ \underbrace{(x_t - \mu)}_{=y_t} &= \phi \underbrace{(x_{t-1} - \mu)}_{=y_t} + w_t \\ x_t &= \mu + \phi\left(x_{t-1} - \mu\right) + w_t \end{aligned}$$

In this case, x_t is an $AR(1)$ process with mean μ.

This means that we have:

1. $E\left[x_t\right] = \mu$
2. $$\begin{aligned} Var\left(x_t\right) &= Var\left(x_t - \mu\right) \\ &= Var\left(y_t\right) \\ &= Var\left(\sum_{j=0}^{\infty} \phi^j w_{t-j}\right) \\ &= \sum_{j=0}^{\infty} \phi^{2j} Var\left(w_{t-j}\right) \\ &= \sigma^2 \sum_{j=0}^{\infty} \phi^{2j} \\ &= \frac{\sigma^2}{1-\phi^2}, \text{ since } |\phi| < 1 \text{ and } \sum_{k=0}^{n} ar^k = \frac{a}{1-r} \end{aligned}$$

So, $x_t \sim N\left(\mu, \frac{\sigma^2}{1-\phi^2}\right)$.

Note that the distribution of x_t is normal and, thus, the density function of x_t is given by:

$$\begin{aligned} f(x_t) &= \sqrt{\frac{1-\phi^2}{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \frac{1-\phi^2}{\sigma^2} \cdot (x_t - \mu)^2\right) \\ &= (2\pi)^{-\frac{1}{2}} \left(\sigma^2\right)^{-\frac{1}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2} \cdot \frac{1-\phi^2}{\sigma^2} \cdot (x_t - \mu)^2\right) \quad [1] \end{aligned}$$

We'll call the last equation [1].

Chapter 7

Conditioning time $x_t|x_{t-1}$

Now, consider $x_t|x_{t-1}$ for $t > 1$.

The mean is given by:

$$\begin{aligned} E\left[x_t|x_{t-1}\right] &= E\left[\mu + \phi\left(x_{t-1} - \mu\right) + w_t|x_{t-1}\right] \\ &= \mu + \phi\left(x_{t-1} - \mu\right) \end{aligned}$$

This is the case since $E\left[x_{t-1}|x_{t-1}\right] = x_{t-1}$ and $E\left[w_t|x_{t-1}\right] = 0$

Now, the variance is:

$$\begin{aligned} Var\left(x_t|x_{t-1}\right) &= Var\left(\mu + \phi\left(x_{t-1} - \mu\right) + w_t|x_{t-1}\right) \\ &= \underbrace{Var\left(\mu + \phi\left(x_{t-1} - \mu\right)|x_{t-1}\right)}_{=0} + Var\left(w_t|x_{t-1}\right) \\ &= Var\left(w_t\right) \\ &= \sigma^2 \end{aligned}$$

Thus, we have: $x_t \sim N\left(\mu + \phi\left(x_{t-1} - \mu\right), \sigma^2\right)$.

Again, note that the distribution of x_t is normal and, thus, the density function of x_t is given by:

$$\begin{aligned} f\left(x_t\right) &= \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \cdot \left[\left(x_t - \mu\right) - \phi\left(x_{t-1} - \mu\right)\right]^2\right) \\ &= \left(2\pi\right)^{-\frac{1}{2}}\left(\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} \cdot \left[\left(x_t - \mu\right) - \phi\left(x_{t-1} - \mu\right)\right]^2\right) \quad [2] \end{aligned}$$

And for this equation we'll call it [2].

Chapter 8

MLE for σ^2 on $AR(1)$ with mean μ

Whew, with all of the above said, we're now ready to obtain an MLE estimate on an $AR(1)$.

Let $\vec{\theta} = \begin{bmatrix} \mu \\ \phi \\ \sigma^2 \end{bmatrix}$, then the likelihood of $\vec{\theta}$ is given by $x_1, \ldots, x_T$ is:

$$\begin{aligned} L\left(\vec{\theta}|x_1, \ldots, x_T\right) &= f\left(x_1, \ldots, x_T|\vec{\theta}\right) \\ &= f\left(x_1\right) \cdot \prod_{t=2}^{T} f\left(x_t|x_{t-1}\right) \end{aligned}$$

The last equality is the result of us using a lag 1 of "memory." Also, note that $x_t|x_{t-1}$ must have $t > 1 \in \mathbb{N}$. Furthermore, we have dropped the parameters in the densities, e.g. $\vec{\theta}$ in $f(\cdot)$, to ease notation.

Using equations [1] and [2], we have:

$$L\left(\vec{\theta}|x_1, \ldots, x_T\right) = (2\pi)^{-\frac{T}{2}} \left(\sigma^2\right)^{-\frac{T}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\left(1-\phi^2\right)\left(x_t-\mu\right)^2 + \sum_{t=2}^{T}\left[\left(x_t-\mu\right) - \phi\left(x_{t-1}-\mu\right)\right]^2\right]\right)$$

For convenience, we'll define:

$$S\left(\mu, \phi\right) = \left(1-\phi^2\right)\left(x_t-\mu\right)^2 + \sum_{t=2}^{T}\left[\left(x_t-\mu\right) - \phi\left(x_{t-1}-\mu\right)\right]^2$$

Fun fact, this is called the "**unconditional** sum of squares."

Thus, we will operate on:

$$L\left(\vec{\theta}|x_1, \ldots, x_T\right) = (2\pi)^{-\frac{T}{2}} \left(\sigma^2\right)^{-\frac{T}{2}} \left(1-\phi^2\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} S\left(\mu, \phi\right)\right)$$

Taking the log of this yields:

$$\begin{aligned} l\left(\vec{\theta}|x_1, \ldots, x_T\right) &= \log\left(L\left(\vec{\theta}|x_1, \ldots, x_T\right)\right) \\ &= -\frac{T}{2}\log\left(2\pi\right) - \frac{T}{2}\log\left(\sigma^2\right) + \frac{1}{2}\left(1-\phi^2\right) - \frac{1}{2\sigma^2} S\left(\mu, \phi\right) \end{aligned}$$

Now, taking the derivative and solving for the maximized point gives:

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2} l\left(\vec{\theta}|x_1,\ldots,x_T\right) &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} S\left(\mu,\phi\right) \\
0 &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} S\left(\mu,\phi\right) \\
\frac{T}{2\sigma^2} &= \frac{1}{2\sigma^4} S\left(\mu,\phi\right) \\
\sigma^2 &= \frac{1}{T} S\left(\mu,\phi\right)
\end{aligned}$$

Thus, the MLE for $\hat{\sigma}^2 = \frac{1}{T} S\left(\hat{\mu},\hat{\phi}\right)$, where $\hat{\mu}$ and $\hat{\phi}$ are the MLEs for μ,ϕ that are obtained numerically via either *Newton Raphson* or a *Scoring Algorithm.* (More details in a numerical recipe book.)

Chapter 9

Conditional MLE on $AR(1)$ with mean μ

A common strategy to reduce the dependency on numerical recipes is to simplify $l\left(\vec{\theta}|x_1,\ldots,x_T\right)$ by using $l^*\left(\vec{\theta}|x_1,\ldots,x_T\right)$:

$$\begin{aligned} l^*\left(\vec{\theta}|x_1,\ldots,x_T\right) &= \prod_{t=2}^{T}\log\left(f\left(x_t|x_{t-1}\right)\right) \\ &= \prod_{t=2}^{T}\log\left(\left(2\pi\right)^{-\frac{1}{2}}\left(\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\cdot\left[\left(x_t-\mu\right)-\phi\left(x_{t-1}-\mu\right)\right]^2\right)\right) \\ &= -\frac{\left(T-1\right)}{2}\log\left(2\pi\right)-\frac{\left(T-1\right)}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{t=2}^{T}\left[\left(x_t-\mu\right)-\phi\left(x_{t-1}-\mu\right)\right]^2 \end{aligned}$$

Again, for convenience, we'll define:

$$S_c\left(\mu,\phi\right)=\sum_{t=2}^{T}\left[\left(x_t-\mu\right)-\phi\left(x_{t-1}-\mu\right)\right]^2$$

Fun fact, this is called the "**conditional** sum of squares."

So, we will use:

$$l^*\left(\vec{\theta}|x_1,\ldots,x_T\right)=-\frac{\left(T-1\right)}{2}\log\left(2\pi\right)-\frac{\left(T-1\right)}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}S_c\left(\mu,\phi\right)$$

Taking the derivative with respect to μ gives:

$$\begin{aligned}
\frac{\partial}{\partial\mu}l^*\left(\vec{\theta}|x_1,\ldots,x_T\right) &= -\frac{1}{2\sigma^2}\sum_{t=2}^{T}2\left[\left(x_t-\mu\right)-\phi\left(x_{t-1}-\mu\right)\right]\left(\phi-1\right)\\
&= \frac{1-\phi}{\sigma^2}\sum_{t=2}^{T}\left[\left(x_t-\mu\right)-\phi\left(x_{t-1}-\mu\right)\right]\\
&= \frac{1-\phi}{\sigma^2}\sum_{t=2}^{T}\left(x_t-\phi x_{t-1}-\mu\left(1-\phi\right)\right)\\
&= -\frac{\left(1-\phi\right)^2}{\sigma^2}\mu\left(T-1\right)+\frac{\left(1-\phi\right)}{\sigma^2}\sum_{t=2}^{T}\left(x_t-\phi x_{t-1}\right)
\end{aligned}$$

Solving for μ^* gives:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\mu}l^*\left(\vec{\theta}|x_1,\ldots,x_t\right)\\
0 &= -\frac{\left(1-\phi\right)^2}{\sigma^2}\mu^*\left(T-1\right)+\frac{\left(1-\phi^*\right)}{\sigma_*^2}\sum_{t=2}^{T}\left(x_t-\phi^* x_{t-1}\right)\\
\frac{\left(1-\phi^*\right)^2}{\sigma_*^2}\mu^*\left(T-1\right) &= \frac{\left(1-\phi^*\right)}{\sigma_*^2}\sum_{t=2}^{T}\left(x_t-\phi^* x_{t-1}\right)\\
\mu^*\left(1-\phi^*\right)\left(T-1\right) &= \sum_{t=2}^{T}\left(x_t-\phi^* x_{t-1}\right)\\
\mu^* &= \frac{1}{\left(1-\phi^*\right)\left(T-1\right)}\sum_{t=2}^{T}\left(x_t-\phi^* x_{t-1}\right)\\
\mu^* &= \frac{1}{1-\phi^*}\left[\underbrace{\frac{1}{T-1}\sum_{t=2}^{T}x_t}_{=\bar{x}_{(2)}}-\underbrace{\frac{\phi^*}{T-1}\sum_{t=2}^{T}x_{t-1}}_{=\bar{x}_{(1)}}\right]\\
\hat{\mu}^* &= \frac{1}{1-\phi^*}\left(\bar{x}_{(2)}-\phi\bar{x}_{(1)}\right)
\end{aligned}$$

When T is large, we have the following:

$$\bar{x}_{(1)}\approx\bar{x},\bar{x}_{(2)}\approx\bar{x}$$

$$\begin{aligned}
\hat{\mu}^* &= \frac{1}{1-\phi^*}\left(\bar{x}-\phi^*\bar{x}\right)\\
&= \frac{\bar{x}}{1-\phi^*}\left(1-\phi^*\right)\\
&= \bar{x}
\end{aligned}$$

Taking the derivative with respect to σ^2 and solving for σ^2 gives:

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} l^* \left(\vec{\theta} | x_1, \ldots, x_T\right) &= -\frac{(T-1)}{2\sigma_*^2} + \frac{1}{2\sigma_*^4} S_c\left(\mu, \phi\right) \\
0 &= -\frac{(T-1)}{2\sigma_*^2} + \frac{1}{2\sigma_*^4} S_c\left(\mu, \phi\right) \\
\frac{(T-1)}{2\sigma_*^2} &= \frac{1}{2\sigma_*^4} S_c\left(\mu, \phi\right) \\
\hat{\sigma}_*^2 &= \frac{1}{T-1} S_c\left(\hat{\mu}^*, \hat{\phi}^*\right)
\end{aligned}$$

Taking the derivative with respect to ϕ gives:

$$\begin{aligned}
\frac{\partial}{\partial \phi} l^* \left(\vec{\theta} | x_1, \ldots, x_T\right) &= -\frac{1}{2\sigma^2} \sum_{t=2}^{T} -2\left[\left(x_t - \mu\right) - \phi\left(x_{t-1} - \mu\right)\right]\left(x_{t-1} - \mu\right) \\
&= \frac{1}{\sigma^2} \sum_{t=2}^{T} \left[x_t - \phi x_{t-1} - \mu\left(1-\phi\right)\right]\left(x_{t-1} - \mu\right) \\
&= \frac{1}{\sigma^2} \sum_{t=2}^{T} \left[x_t x_{t-1} - \phi x_{t-1}^2 - \mu\left(1-\phi\right) x_{t-1} - \mu x_t + \mu\phi x_{t-1} + \mu^2\left(1-\phi\right)\right] \\
&= \frac{1}{\sigma^2} \left[\begin{array}{l} \displaystyle\sum_{t=2}^{T} x_t x_{t-1} - \phi \sum_{t=2}^{T} x_{t-1}^2 - \mu\left(1-\phi\right)\left(T-1\right)\bar{x}_{(1)} \\ -\mu\left(T-1\right)\bar{x}_{(2)} + \phi\mu\left(T-1\right)\bar{x}_{(1)} + \mu^2\left(1-\phi\right)\left(T-1\right) \end{array}\right]
\end{aligned}$$

Solving for ϕ gives:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \phi} l^* \left(\vec{\theta} | x_1, \ldots, x_T\right) \\
0 &= \sum_{t=2}^{T} x_t x_{t-1} - \hat{\phi}^* \sum_{t=2}^{T} x_{t-1}^2 - \left(\bar{x}_{(2)} - \hat{\phi}^* \bar{x}_{(1)}\right)\left(T-1\right)\bar{x}_{(1)} - \frac{\bar{x}_{(2)} - \hat{\phi}^* \bar{x}_{(1)}}{1-\hat{\phi}^*}\left(T-1\right)\bar{x}_{(2)} \\
&\quad + \hat{\phi}^* \frac{\bar{x}_{(2)} - \hat{\phi}^* \bar{x}_{(1)}}{1-\hat{\phi}^*}\left(T-1\right)\bar{x}_{(1)} + \left(\frac{\bar{x}_{(2)} - \hat{\phi}^* \bar{x}_{(1)}}{1-\hat{\phi}^*}\right)^2 \left(1-\hat{\phi}^*\right)\left(T-1\right) \\
&\vdots \\
&\text{Magic} \\
&\vdots \\
\hat{\phi}^* &= \frac{\displaystyle\sum_{t=2}^{T} \left(x_t - \bar{x}_{(2)}\right)\left(x_{t-1} - \bar{x}_{(1)}\right)}{\displaystyle\sum_{t=2}^{T} \left(x_{t-1} - \bar{x}_{(1)}\right)^2}
\end{aligned}$$

When T is large, we have:

$$\sum_{t=2}^{T}\left(x_t-\bar{x}_{(2)}\right)\left(x_t-\bar{x}_{(1)}\right)\approx\sum_{t=2}^{T}\left(x_t-\bar{x}\right)\left(x_{t-1}-\bar{x}\right)$$
$$\sum_{t=2}^{T}\left(x_{t-1}-\bar{x}_{(1)}\right)^2\approx\sum_{t=1}^{T}\left(x_t-\bar{x}\right)^2$$

$$\hat{\phi}^*=\frac{\sum_{t=2}^{T}\left(x_t-\bar{x}_{(2)}\right)\left(x_t-\bar{x}_{(1)}\right)}{\sum_{t=2}^{T}\left(x_{t-1}-\bar{x}_{(1)}\right)^2}\approx\frac{\sum_{t=2}^{T}\left(x_t-\bar{x}\right)\left(x_{t-1}-\bar{x}\right)}{\sum_{t=1}^{T}\left(x_t-\bar{x}\right)^2}=\hat{\rho}\left(1\right)$$

Consider a time series given by $x_t\sim ARMA(p,q)$. This gives us with a paramter space Ω that looks like so:

$$\vec{\varphi}=\begin{bmatrix}\phi_1\\ \vdots\\ \phi_p\\ \theta_1\\ \vdots\\ \theta_q\\ \sigma^2\end{bmatrix}$$

In order to estimate this parameter space, we must assume the following three conditions:

1. The process is casual
2. The process is invertible
3. The process has Gaussian innovations.

Innovations are a time series equivalent to residuals. That is, an innovation is given by $x_t-\hat{x}_t^{t-1}$, where $\hat{x}_t^{t-1}$ is the prediction at time t given $t-1$ observations and x_t is the true value observed at time t.

There are two main ways of performing such an estimation of the parameter space.

1. Method of Moments (MoM)
2. Maximum Likelihood / Least Squares Estimation [MLE / LSE]

9.1 Method of Moments

The goal behind the estimation with Method of Moments is to match the theoretical moment (e.g. $E\left[x_t^k\right]$) with the sample moment (e.g $\frac{1}{n}\sum_{i=1}^{n}x_i^k$), where k denotes the moment.

This method often leads to suboptimal estimates for general ARMA models. However, it is quite optimal for $AR(p)$.

9.1.1 Method of Moments - AR(p)

Consider an $AR(p)$ process represented by:

$$x_t=\phi_1x_{t-1}+\cdots+\phi_px_{t-p}+w_t$$

where $w_t \sim N(0, \sigma^2)$

To begin, we find the Covariance of the process when $h > 0$:

$$\begin{aligned} Cov\left(x_{t+h}, x_t\right) &\overset{(h>0)}{=} Cov\left(\phi_1 x_{t+h-1} + \cdots + \phi_p x_{t+h-p} + w_{t+h}, x_t\right) \\ &= \phi_1 Cov\left(x_{t+h-1}, x_t\right) + \cdots + \phi_p Cov\left(x_{t+h-p}, x_t\right) + Cov\left(w_{t+h}, x_t\right) \\ &= \phi_1 \gamma\left(h-1\right) + \cdots + \phi_p \gamma\left(h-p\right) \end{aligned}$$

Now, we turn our attention to the variance of the process:

$$\begin{aligned} Var\left(w_t\right) &= Cov\left(w_t, w_t\right) \\ &= Cov\left(w_t, w_t\right) + \underbrace{Cov\left(\phi_1 x_{t-1}, w_t\right)}_{=0} + \cdots + \underbrace{Cov\left(\phi_p x_{t-p}, w_t\right)}_{=0} \\ &= Cov\left(\underbrace{\phi_1 x_{t-1} + \cdots + \phi_p x_p + w_t}_{=x_t}, w_t\right) \\ &= Cov\left(x_t, w_t\right) \\ &= Cov\left(x_t, x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_p\right) \\ &= Cov\left(x_t, x_t\right) - \phi_1 Cov\left(x_t, x_{t-1}\right) - \cdots - \phi_p Cov\left(x_t, x_{t-p}\right) \\ &= \gamma\left(0\right) - \phi_1 \gamma\left(1\right) - \cdots - \phi_p \gamma\left(p\right) \end{aligned}$$

Together, these equations are known as the **Yule-Walker** equations.

9.1.2 Yule-Walker

Definition

Equation form:

$$\begin{aligned} \gamma\left(h\right) &= \phi_1 \gamma\left(h-1\right) - \cdots - \phi_p \gamma\left(h-p\right) \\ \sigma^2 &= \gamma\left(0\right) - \phi_1 \gamma\left(1\right) - \cdots - \phi_p \gamma\left(p\right) \\ h &= 1, \ldots, p \end{aligned}$$

Matrix form:

$$\begin{aligned} &\Gamma \vec{\phi} = \vec{\gamma} \\ &\sigma^2 = \gamma\left(0\right) - \vec{\phi}^T \vec{\gamma} \end{aligned}$$

$$\vec{\phi} = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}_{p \times 1}, \vec{\gamma} = \begin{bmatrix} \gamma\left(1\right) \\ \vdots \\ \gamma\left(p\right) \end{bmatrix}_{p \times 1}, \Gamma = \left\{\gamma\left(k-j\right)\right\}_{j,k=1}^{p}$$

More aptly, the structure of Γ looks like the following:

$$\Gamma = \begin{bmatrix} \gamma\left(0\right) & \gamma\left(-1\right) & \gamma\left(-2\right) & \cdots & \gamma\left(1-p\right) \\ \gamma\left(1\right) & \gamma\left(0\right) & \gamma\left(-1\right) & \cdots & \gamma\left(2-p\right) \\ \gamma\left(2\right) & \gamma\left(1\right) & \gamma\left(0\right) & \cdots & \gamma\left(3-p\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma\left(p-1\right) & \gamma\left(p-2\right) & \gamma\left(p-3\right) & \cdots & \gamma\left(0\right) \end{bmatrix}_{p \times p}$$

Note, that we are able to use the above equations to effectively estimate $\vec{\phi}$ and σ^2.

$$\begin{bmatrix} \hat{\vec{\phi}} = \hat{\Gamma}^{-1}\hat{\vec{\gamma}} \\ \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\vec{\gamma}}^T\hat{\Gamma}^{-1}\hat{\vec{\gamma}} \end{bmatrix} \rightarrow \text{Yule - Walker Estimates}$$

For the second equation, we are effectively substituting in the first equation for $\hat{\vec{\phi}}$, hence the quadratic form $\hat{\vec{\gamma}}^T\hat{\Gamma}^{-1}\hat{\vec{\gamma}}$.

With this being said, there are a few nice asymptotic properties that we obtain for an $AR(p)$.

1. $\sqrt{T}\left(\hat{\vec{\phi}} - \vec{\phi}\right) \underset{t\to\infty}{\xrightarrow{L}} N\left(\vec{0}, \sigma^2\Gamma^{-1}\right)$
2. $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$

Yule-Walker estimates are optimal in the sense that they have the smallest asymptotic variance i.e.

$$Var\left(\sqrt{T}\hat{\vec{\phi}}\right) = \sigma^2\Gamma^{-1}$$

However, they are not necessarily optimal with small sample sizes.

Conceptually, the reason for this optimality result is a consequence from the linear dependence between moments and variables.

This is not true for MA or ARMA, which are both nonlinear and suboptimal.

9.1.3 Estimates

Consider x_t as an $MA(1)$ process: $x_t = \theta w_{t-1} + w_t, w_t \overset{i.i.d}{\sim} N\left(0, \sigma^2\right)$

Finding the covariance when $h = 1$ gives:

$$\begin{aligned} Cov(x_t, x_{t-1}) &= Cov(\theta w_{t-1} + w_t, \theta w_{t-2} + w_{t-1}) \\ &= Cov(\theta w_{t-1}, w_{t-1}) \\ &= \theta\sigma^2 \end{aligned}$$

Finding the variance (e.g. $h = 0$) gives:

$$\begin{aligned} Cov(x_t, x_t) &= Cov(\theta w_{t-1} + w_t, \theta w_{t-1} + w_t) \\ &= \theta^2 Cov(w_{t-1}, w_{t-1}) + \underbrace{2\theta Cov(w_{t-1}, w_t)}_{=0} + Cov(w_t, w_t) \\ &= \theta^2\sigma^2 + \sigma^2 \\ &= \sigma^2\left(1 + \theta^2\right) \end{aligned}$$

This gives us the MA(1) ACF of:

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \frac{\theta}{\theta^2+1} & h = \pm 1 \end{cases}$$

With this in mind, let's solve for possible θ values:

$$\begin{aligned} \rho(1) &= \frac{\theta}{\theta^2 + 1} \\ \Rightarrow \theta &= \left(\theta^2 + 1\right)\rho(1) \\ \theta &= \rho(1)\theta^2 + \rho(1) \\ 0 &= \rho(1)\theta^2 - \theta + \rho(1) \end{aligned}$$

Yuck, that looks nasty. Let's dig out an ol' friend from middle school known as the quadratic formula:

$$\theta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Applying the quadratic formula leads to:

$$\begin{aligned} a = \rho(h), b = -1, c = \rho(h) \\ \theta = \frac{1 \pm \sqrt{1^2 - 4\rho(h)\rho(h)}}{2\rho(h)} \\ \theta = \frac{1 \pm \sqrt{1 - 4[\rho(h)]^2}}{2\rho(h)} \end{aligned}$$

Thus, we have two possibilities:

$$\begin{aligned} \theta_1 = \frac{1 + \sqrt{1 - 4[\rho(h)]^2}}{2\rho(h)} \\ \theta_2 = \frac{1 - \sqrt{1 - 4[\rho(h)]^2}}{2\rho(h)} \end{aligned}$$

To ensure invertibility, we mandate that $|\rho(1)| < \frac{1}{2}$. Thus, we opt for θ_2.

So, our estimator is:

$$\hat{\theta} = \frac{1 - \sqrt{1 - 4[\hat{\rho}(1)]^2}}{2\hat{\rho}(1)}$$

Furthermore, it can be shown that:

$$\sqrt{T}\left(\hat{\theta} - \theta\right) \underset{T \to \infty}{\overset{L}{\to}} N\left(0, \frac{1 + \theta^2 + 4\theta^4 + \theta^6 + \theta^8}{(1 - \theta^2)^2}\right)$$

So, this is not a really optimal estimator…

9.2 Prediction (Forecast)

Bibliography