Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Applied Forecasting

(LECTURENOTES)

Prof. Rozenn Dahyot

SCHOOL OF COMPUTER SCIENCE AND STATISTICS
TRINITY COLLEGE DUBLIN
IRELAND
https://www.scss.tcd.ie/Rozenn.Dahyot

Michaelmas Term 2017

Contents

1 Introduction **4**
1.1 Who Forecasts? . . . 4
1.2 Why Forecast? . . . 4
1.3 How to Forecast? . . . 4
1.4 What are the Steps in the Forecasting Procedure? . . . 5

2 Quantitative Forecasting **6**
2.1 When Can We Use Quantitative Methods? . . . 6
2.2 What are the Types of Quantitative Methods? . . . 6
2.3 Explanatory model and Time Series Forecasting . . . 6

I Data preparation and visualisation **8**

3 Preparation of Data **9**
3.1 Month Length Adjustment . . . 9
3.2 Trading Day Adjustment . . . 9

4 Visualisation tools for Time series **10**
4.1 Definitions . . . 10
4.2 Time Series Patterns . . . 11
4.3 Additional exercises . . . 12

II Ad-Hoc Algorithms: Holt-Winters Algorithms **16**

5 Single Exponential Smoothing Algorithm **18**
5.1 Notations . . . 18
5.1.1 Single Exponential Smoothing . . . 18
5.2 What does Exponential Smoothing Really Do? . . . 19
5.3 Exercises . . . 19

6 Double exponential Smoothing Algorithm **20**
6.1 Exercises . . . 20
6.2 Final comments . . . 21

7 Comparing Holt-Winters Forecasting Algorithms **22**
7.1 Definitions . . . 22
7.2 Exercises . . . 23

8 Holt-Winters' Exponential Smoothing with Seasonality **24**
8.1 Exercise . . . 24
8.2 Selecting the best Holt-Winters algorithms . . . 26

III Statistical models: ARIMA **27**

9 Linear Regression **29**
9.1 Regression with one explanatory variable . . . 29
9.2 Using Linear Regression to Make Forecasts . . . 31
9.2.1 Time as an explanatory variable . . . 31
9.2.2 Indicator variables: modelling seasonality . . . 31
9.3 Least Square algorithm in Matrix Form . . . 32
9.3.1 Least Squares for Linear regression . . . 32
9.3.2 Multiple Linear regression . . . 33

10 AR(p): Autoregressive Models **34**
10.1 Definition . . . 34
10.2 Prediction interval for AR(1) k steps ahead . . . 35

11 MA(q): Moving Average Processes **38**
11.1 Definitions . . . 38
11.2 Fitting an MA model . . . 38

12 ARMA(p,q): AutoRegressive Moving Average Models **39**
12.1 Definition . . . 39
12.2 Exercises . . . 39
12.3 Simulation of ARMA models . . . 39
12.4 Stationarity in mean and variance . . . 40
12.5 Conclusion . . . 42

13 Using ACF and PACF to select MA(q) or AR(p) models **43**
13.1 ACF and PACF . . . 43
13.2 Exercises: ACF and PACF for AR(1) and MA(1) . . . 44
13.3 Least Squares algorithm for MA models ? . . . 45

14 The backshift operator **46**
14.1 Definition . . . 46
14.2 Exercises. . . . 46

15 AIC and BIC **47**
15.1 Information Criterion . . . 47
15.2 R output . . . 48

16 ARIMA(p,d,q) **50**
16.1 Differencing a time series . . . 50
16.2 Integrating differencing into ARMA models . . . 50
16.3 Which $ARIMA(p,d,q)$ model do I use? . . . 52

17 Seasonal $ARIMA(p,d,q)(P,D,Q)_s$ **54**
17.1 Seasonal $ARIMA(p,d,q)(P,D,Q)_s$ . . . 54
17.2 Using ACF and PACF to identify seasonal ARIMAs . . . 55
17.3 How to select the best Seasonal ARIMA model? . . . 55
17.4 Conclusion . . . 55

18 Transforming a time series **60**

IV Conclusions **63**

19 Summary of the course **64**

20 Beyond Holt-Winters algorithms and ARIMA models **66**
20.1 ARCH GARCH . . . 66
20.2 Continuous time series modelling . . . 66
20.2.1 Brownian motion . . . 67
20.2.2 Gaussian processes . . . 67
20.3 Fourier analysis for time series . . . 68
20.4 Others techniques used for time series analysis and forecasting . . . 68

Chapter 1

Introduction

1.1 Who Forecasts?

- Aer Lingus — sales next year by class
- Superquinn — demand for oranges next month
- Government — household numbers in 2015
- Astronomer — predicting the effects of interplanetary travel
- Airbus — sales of the new A-380 super-jumbo over the next 20 years
- Trinity College — pension fund obligations in 2020
- ISP — internet routing schedule for the next 30 seconds
- You — what will be on the exam in June
- Meteorologist — climate in 2050

1.2 Why Forecast?

- Because there is often a time lag between knowing an event is going to happen and when it happens.
- If we can *forecast* the event accurately then we can *plan* appropriate action to deal with the event.
- The benefits of forecasting can be to: *save money, increase profits, improve quality of life, prevent death.....*

1.3 How to Forecast?

There are broadly 3 methods:

Quantitative: quantitative data are available about the events to be forecast. Some mathematical procedure is then used to make a forecast from the data. Such procedures vary from the very ad-hoc to formal statistical methods.

Example: predicting monthly inflation rates in 2001 based on historical rates and other economic data.

Qualitative: little or no quantitative data is available, however there is a lot of "knowledge" and expertise that are then used to make the forecast.

Example: an economist forecasting how a large increase in oil price will affect oil consumption.

Unpredictable: little or no information about the events to be forecast exists. Forecasts are made on the basis of speculation.

Example: predicting the effect of a new, very cheap, non-polluting form of energy on the world economy.

1.4 What are the Steps in the Forecasting Procedure?

(1) **Problem definition:** what do we want to forecast? Who wants the forecast? What will it be used for? How does the forecast fit into the organisation? What data do we need to make the forecast? Can we get all the data necessary within the time allowed? All these questions must be answered before we try to make a forecast. This is often the most difficult step in the forecasting task.

(2) **Gathering information:** there are two kinds, numerical data and the accumulated knowledge of experienced personnel connected with the quantity to be forecast.

(3) **Exploratory Analysis:** here we try to get a feel for the numerical data. We plot graphs, compute summary statistics, possibly do a "decomposition analysis" and look for correlations between variables. This helps us in the next step.

(4) **Selecting and fitting models to make the forecast:** we choose and fit several forecasting models. The models we pick may be based on information we revealed in the exploratory analysis.

(5) **Using and evaluating the forecast:** from the problem definition and other measures of definition, we choose a model that we consider best. Once the events to be forecast have occurred, we can then evaluate how well this model has done. On the basis of this, we may modify the model or even decide to use another for the next forecast.

In this course we will concentrate on Quantitative Forecasting.
Further, we will concentrate on the last 2 steps of the forecasting procedure: choosing and fitting models, making forecasts and evaluating them. We will look at several different forecasting models, how they can be used to make forecasts, and different ways to compare their performance.

Chapter 2

Quantitative Forecasting

2.1 When Can We Use Quantitative Methods?

Quantitative forecasting can be applied only if:

(1) Information about the past is available;

(2) The information is quantified as numerical data;

(3) The *continuity assumption* holds: this means that some aspects of the past will continue into the future.

2.2 What are the Types of Quantitative Methods?

Quantitative methods vary from:

- *intuitive* or *ad-hoc* methods. Example: an expert forecasts next month's inflation rate by looking at all available economic data and using her experience to make a reasoned prediction.
- *formal statistical procedures*, such as linear regression or more generally the fitting of a statistical model.

What are the advantages/disadvantages of each?

- Intuitive methods are easy to use, but vary from business to business and forecaster to forecaster, even with the same data. It is not easy to give estimates of the accuracy of the forecast.
- Formal methods are now inexpensive to implement and are now often more accurate than intuitive methods. They are easy to calculate and replicate with measures of uncertainty (prediction interval) associated with forecasts.

2.3 Explanatory model and Time Series Forecasting

We can also classify quantitative methods by the type of model used:

- **Explanatory models:** the quantity to be forecast has a relationship with variables in the data. Example: in forecasting next month's inflation rate, we assume:

$$\text{inflation rate} = f(\text{previous inflation rates}, \text{export level}, \text{GDP growth last year}, \\ \text{inflation in neighbouring countries}, \text{exchange rates}, \ldots, \text{error}).$$

  It assumes that any change in input variables will change the forecast in a predictable way (given by f). We have to discover the form of f. There are always random changes in inflation that cannot be forecast, so we always include "error" in the function f to account for these.

- **Time series:** unlike explanatory models, we make no attempt to discover what variables might affect the quantity to be forecast. We merely look at the values of the quantity over time (a *time series*), try to discover patterns in the series, and make a forecast by extrapolating them into the future. Thus, the inflation rate at month $t+1$ can be written:

$$\text{inflation rate}_{t+1} = g(\text{inflation rate}_t, \text{inflation rate}_{t-1}, \text{inflation rate}_{t-2}, \ldots, \text{error})$$

  This is often a good idea because the function f in the explanatory model can be very difficult to define, even approximately. Indeed the effect of the variables on the forecast may not be understood, and it may not be worthwhile or too expensive to try to understand it. It is therefore often better just to treat the time series as a "black box", and use a time series model.

In this course we will concentrate on Time series models.

Part I

Data preparation and visualisation

Chapter 3

Preparation of Data

Sometimes, time series need to be normalised or adjusted before trying to fit any model. Indeed artificial seasonal patterns may appear on monthly data just because months have a different duration.

3.1 Month Length Adjustment

This is a transformation, very useful sometimes with *monthly* data. Because different months are actually different lengths of time (28 – 31 days), a time plot of monthly data often shows seasonal behaviour that is due purely to this difference (particularly in February). This can mask more important effects that we are looking for.

The average month length is 365.25/12 days. The *Month Length Adjustment* transforms the y_i so that they represent the value over an average month length:

$$\begin{aligned} w_i &= y_i \times \text{ average month length /no. of days in month } i \\ &= y_i \times 365.25/(12 \times \text{no. of days in month } i) \end{aligned}$$

3.2 Trading Day Adjustment

Monthly data of quantities like sales can be affected by the number of trading days in the month. *Trading Day Adjustment* transforms a monthly series to represent sales over an average number of trading days per month:

$$w_i = y_i \times \frac{\text{no. of trading days in an average month}}{\text{no. of trading days in month } i}.$$

Chapter 4

Visualisation tools for Time series

4.1 Definitions

1.1 Definition (Time series) The sequence of values $\{y_t\}_{t=0,1,\cdots,n}$ over time is called a *time series*.

We typically want to forecast the next value in the series (*1-step ahead prediction*) or the value at k time periods in the future (*k-step ahead prediction*).

Example: monthly Australian beer production (in millions of litres). Table 4.1 presents the beer production in Australia from 1991 to 1995. Do you notice anything?

Month	1991	1992	1993	1994	1995
January	164	147	139	151	138
February	148	133	143	134	136
March	152	163	150	164	152
April	144	150	154	126	127
May	155	129	137	131	151
June	125	131	129	125	130
July	153	145	128	127	119
August	146	137	140	143	153
September	138	138	143	143	
October	190	168	151	160	
November	192	176	177	190	
December	192	188	184	182	

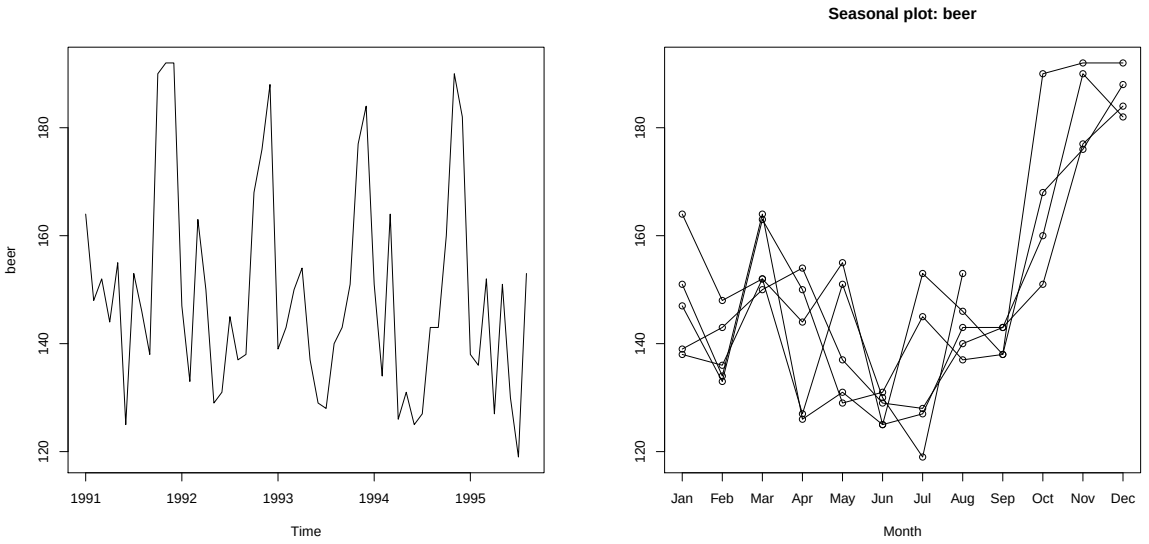
Table 4.1: Monthly Australian beer production (in millions of litres).

1.2 Definition The *time plot* is the plot of the series in order of time (i.e. the time is reported on the x-axis).

Figure 4.1(a) shows the time plot for the beer data.

1.3 Definition The *seasonal plot* shows the data from each season that are overlapping.

Fig. 4.1(b) shows the seasonal plot of the beer data.



(a) Time plot of the Beer data. R> `plot(beer)`

(b) Seasonal plot of the beer production over a year period. R> `seasonplot(beer)`

Figure 4.1: Monthly Australian beer production (in millions of litres).

4.2 Time Series Patterns

Time series can be decomposed onto several components:

(1) **Trend:** a long term increase or decrease occurs.

(2) **Seasonal:** series influenced by seasonal factors. Thus the series exhibits a behaviour that more or less repeats over a *fixed period of time*, such as a year. Such behaviour is easily demonstrated in a *seasonal plot*, where the data is plotted according to where in the seasonal cycle it was observed).

(3) **Cyclical (not addressed in this course):** series rises and falls regularly but these are *not of fixed period*. Example: economic data rises and falls according to the business cycle, but this cycle varies in length considerably.

(4) **Error:** this corresponds to random fluctuations that cannot be explained by a deterministic pattern.

Exercise. In the software R, split a time series (e.g. time series `beer`, `airpass`, and `dowjones`) into trend component, seasonality component and noise component, using the function `decompose`.

Exercise. What patterns do you recognise in the beer data ?

Exercise. Can you explain the increase of sales of beer at the end of the year ?

2.1 Definition The *AutoCorrelation Function (ACF)*

For a time series $y_1, y_2, \ldots, y_n$, the autocorrelation at lag k is:

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

where $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ is the mean of the series. The ACF can be plotted reporting the values r_k on the $y-axis$ with the lag k on the abscissa.

2.2 Definition The *Partial AutoCorrelation Function (PACF)* is another useful method to examine serial dependencies. This is an extension of the autocorrelation, where the dependence on the intermediate elements (those within the lag) is removed. The set of partial autocorrelations at different lags is called the *partial autocorrelation function* (PACF) and is plotted like the ACF.

Figure 4.2 shows time plot, ACF, and PACF for the beer data.

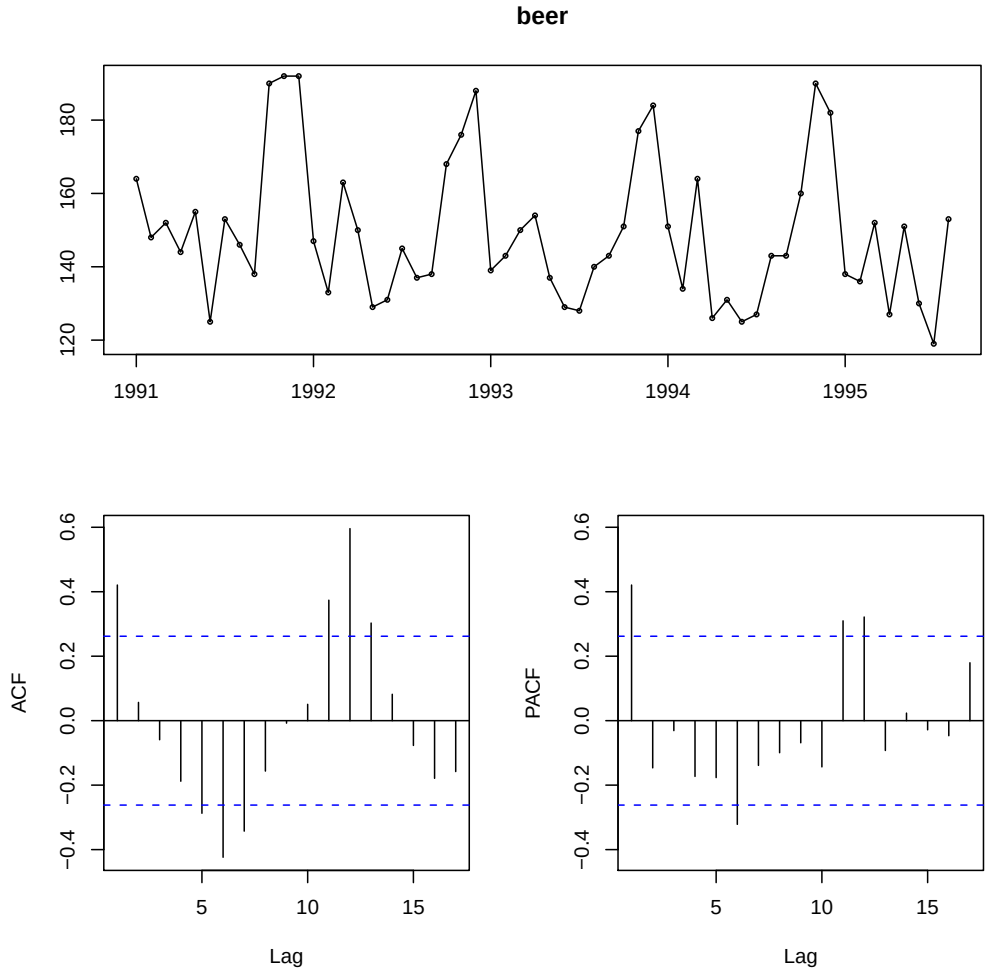


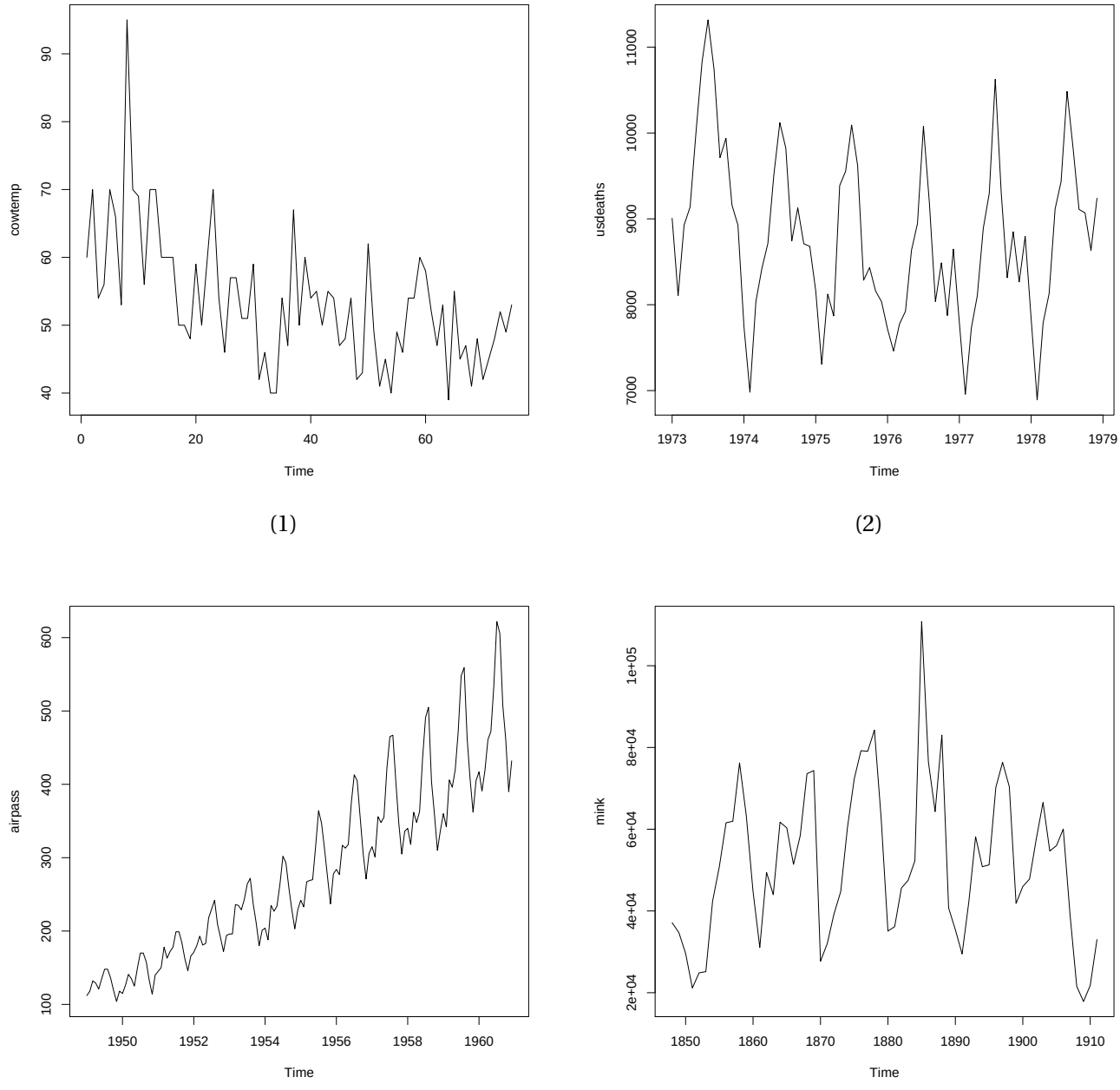
Figure 4.2: Displaying the time plot of the beer data with its ACF and PACF plots. R>`tsdisplay(beer)`

4.3 Additional exercises

(1) What are the patterns that you see in the time series 1 to 4 in figure 4.3?

(2) Identify the ACF functions ABCD given in figure 4.4 corresponding to the time plots 1 to 4 in figure 4.3.

(3) Identify the PACF functions abcd given in figure 4.5 corresponding to the time plots 1 to 4 in figure 4.3.



(1) (2)

(3) (4)

Figure 4.3: Time plots: (1) Daily morning temperature of a cow, (2) Accidental deaths in USA, (3) International airline passengers and (4) Annual mink trapping.

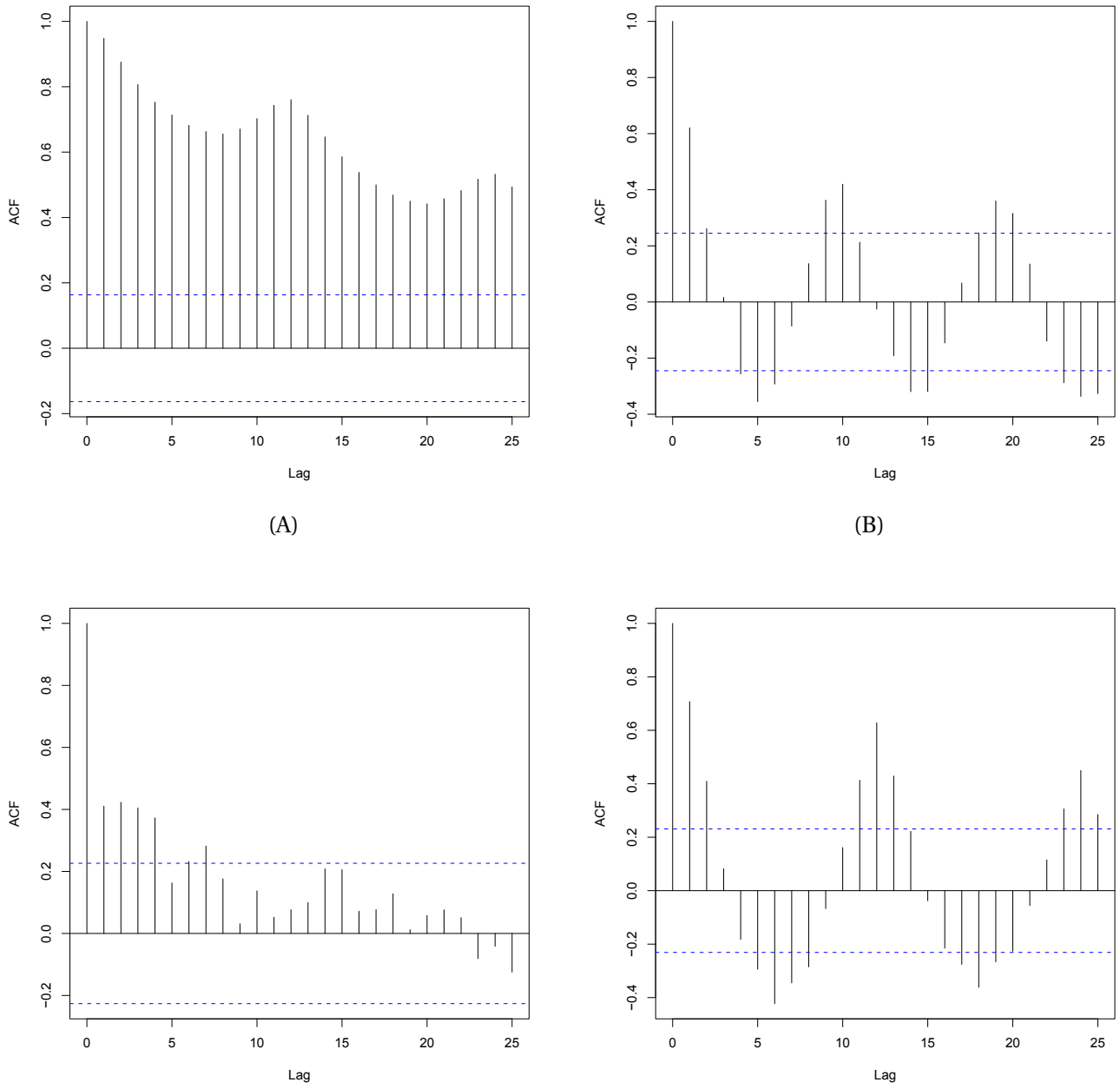
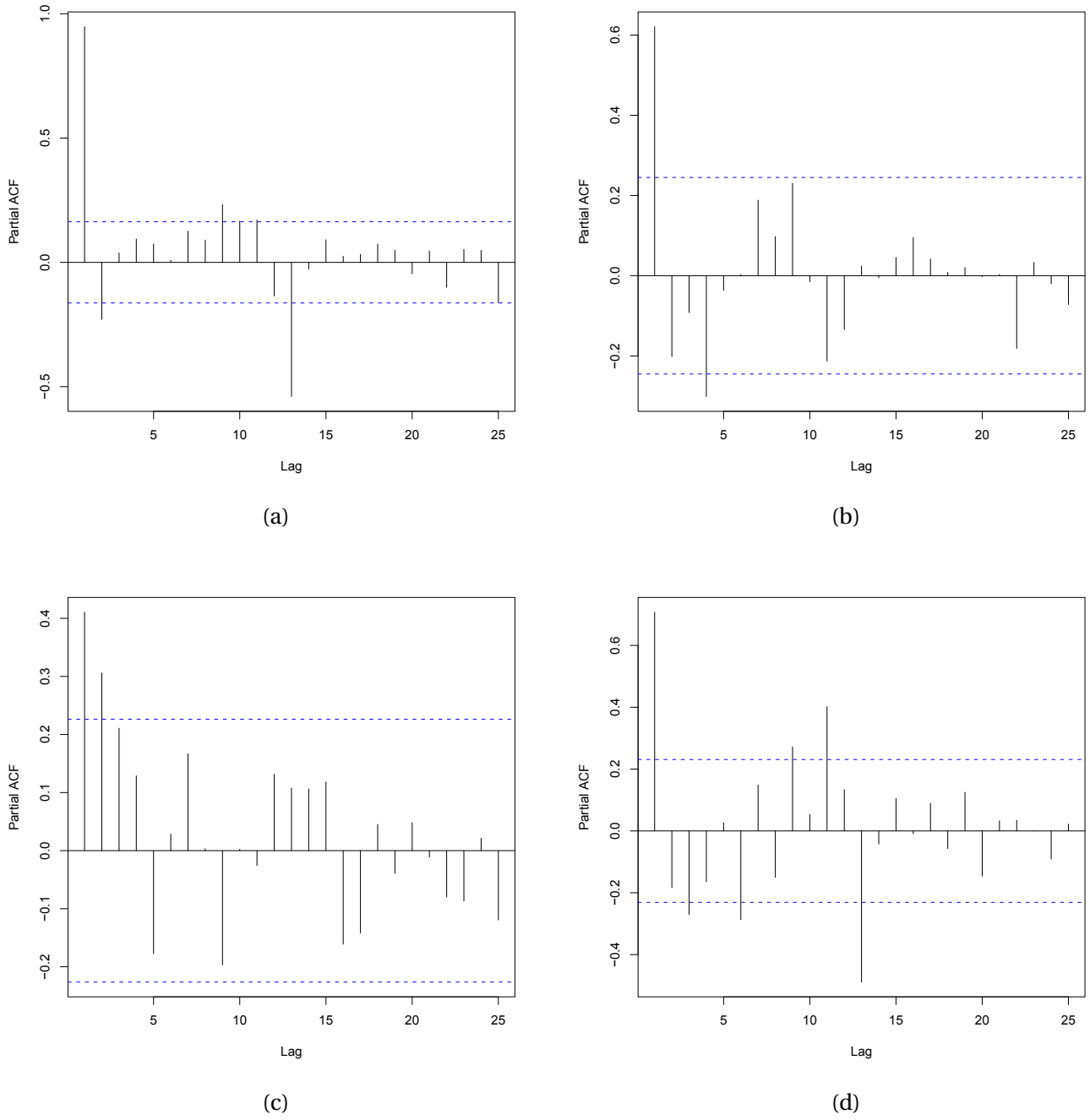


Figure 4.4: ACF. Example of R command for usdeaths time series `>acf(ts(usdeaths,freq=1),25)`.



(a) (b)

(c) (d)

Figure 4.5: PACF. Example of R command for usdeaths time series `>pacf(ts(usdeaths,freq=1),25)`.

Part II

Ad-Hoc Algorithms: Holt-Winters Algorithms

This part introduces a number of forecasting methods called Holt-Winters Algorithms which are not explicitly based on a probability models, and can be seen as being of an ad-hoc nature [1]. Chapter 5 introduces an algorithm suitable to be fitted to a time series with no or little trend and no seasonal patterns. A second algorithm is introduced in chapter 6 to deal with time series with trends but no seasonal patterns. Chapter 8 proposes two algorithms for dealing with time series presenting both a trend and a seasonal component. Chapter 7 proposes criteria that can be used to select the 'best algorithm' for a particular time series.

Chapter 5

Single Exponential Smoothing Algorithm

5.1 Notations

From now on in the course we use the following notation:

- $y_1, y_2, \ldots, y_n$ are the observed values of the time series
- y_n is the last value of the series to be observed i.e. we are currently at time n (months, quarters, years...)
- Forecasts for the value of the series at future times $n+1, n+2, \ldots$, using a model fitted to $y_1, \ldots, y_n$, are denoted F_{n+1}, $F_{n+2}, \ldots$. The k-step ahead forecast from time n would be F_{n+k}.
- Fitted values using the model are $F_1, \ldots, F_n$.
- The residuals or errors are $y_1 - F_1, \ldots, y_n - F_n$.

5.1.1 Single Exponential Smoothing

There is no obvious statistical model that we try to fit (by regression or another fitting technique). Exponential Smoothing is simply an algorithm for creating forecasts iteratively on the basis of how well one did with previous forecasts.

- Suppose we make a forecast F_t for the value of y_t (which is not yet observed).
- Now we observe y_t and wish to make a forecast F_{t+1}. We do this by taking our old forecast F_t and adjusting it using the error in forecasting y_t as follows:

$$F_{t+1} = F_t + \alpha(y_t - F_t),$$

  where α is between 0 and 1.
- The nearer α is to 1 then the larger the adjustment.
- We cannot forecast the first term in the series (since $F_1 = F_0 + \alpha(y_0 - F_0)$ and there is no F_0 or y_0). By convention, we fix $F_1 = y_1$ and only forecast y_2 onwards.

Init: $F_1 = y_1$ and choose $0 < \alpha < 1$
Forecast:
$F_{t+1} = F_t + \alpha\,(y_t - F_t)$
Until no more observation are available then
$F_{n+k} = F_{n+1},\ \forall k \geq 1$

Table 5.1: Simple Exponential Smoothing (SES) Algorithm.

5.2 What does Exponential Smoothing Really Do?

If we recursively apply the smoothing equation to F_{t+1}, we get:

$$\begin{aligned} F_{t+1} &= F_t + \alpha\,(y_t - F_t) \\ &= [F_{t-1} + \alpha\,(y_{t-1} - F_{t-1})] + \alpha\,(y_t - [F_{t-1} + \alpha\,(y_{t-1} - F_{t-1})]) \\ &= \alpha\,y_t + \alpha\,(1-\alpha)\,y_{t-1} + (1-\alpha)^2\,F_{t-1}, \end{aligned}$$

Now F_{t+1} is in terms of y_t, y_{t-1} and F_{t-1}. We can repeat this, replacing F_{t-1} by $F_{t-2} + \alpha(y_{t-2} - F_{t-2})$, to get F_{t+1} is in terms of y_t, y_{t-1}, y_{t-2} and F_{t-2}. Doing this replacement another $t-3$ times, we end up with F_{t+1} in terms of $y_1, \ldots, y_t$ and F_1, and the following equation for F_{t+1}:

$$F_{t+1} = \alpha\,y_t + \alpha\,(1-\alpha)\,y_{t-1} + \alpha\,(1-\alpha)^2\,y_{t-2} + \cdots + \alpha\,(1-\alpha)^{t-1}\,y_1 + (1-\alpha)^t\,F_1 \tag{5.1}$$

So *exponential smoothing forecasts are a weighted sum of* **all** *the previous observations.*

5.3 Exercises

(1) What is F_{t+1} when $\alpha = 0$? What happens as α increases to 1? What range of values must F_{t+1} lie in?

(2) Here is a short time series. Calculate the exponentially smoothed series and make a forecast for the next value in the sequence, using $\alpha = 0.5$ and $\alpha = 0.1$:

t	y_t	F_t (for $\alpha = 0.5$)	error	F_t (for $\alpha = 0.1$)	error
1	3	3	0	3	0
2	4				
3	2				
4					

(3) Can you make k-step ahead forecasts using exponential smoothing?

(4) Which observation is given the biggest weight in the formula (5.1) for F_{t+1}. Which is given the smallest? Is this sensible?

Chapter 6

Double exponential Smoothing Algorithm

Double exponential Smoothing (DES) Algorithm (also known as Holt's Linear Method) is an extension to the SES algorithm originally designed for time series with no trend nor seasonal patterns. It includes a term to model linear trends. Holt's method allows the estimates of level (L_t) and slope (b_t) to be adjusted with each new observation.

Init: $L_1 = y_1 \quad b_1 = y_2 - y_1 \quad F_1 = y_1$ and choose $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$

Compute and Forecast:

$$L_t = \alpha\ y_t + (1-\alpha)\ (L_{t-1} + b_{t-1})$$
$$b_t = \beta\ (L_t - L_{t-1}) + (1-\beta)\ b_{t-1}$$
$$F_{t+1} = L_t + b_t$$

Until no more observation are available then

$$F_{n+k} = L_n + k\ b_n,\ \forall k \geq 1$$

Table 6.1: Double Exponential Smoothing (Holt's Linear Model) Algorithm.

Note that no forecasts or fitted values can be computed until y_1 and y_2 have been observed. Also by convention, we let $F_1 = y_1$.

6.1 Exercises

(1) Calculate the level and slope of the series on the next page by Holt's linear method, using $\alpha = 0.5$ and $\beta = 0.1$. Compute, at each point, the 1-step ahead forecast $F_{t+1} = L_t + b_t$.

t	y_t	L_t	b_t	$F_t = L_{t-1} + b_{t-1}$	$y_t - F_t$
1	3	3	1	3	0
2	4			4	0
3	2				
4					

6.2 Final comments

In summary, exponential smoothing is good for forecasting data with no trend or seasonal patterns. If there is a linear trend, Holt's method (i.e. D.E.S) can be used. For data with a shift, exponential smoothing is able to adapt to the shift, but the speed at which it does so depends on α.

time series patterns:			
trend	no	yes	no/yes
seasonal	no	no	yes
noise	yes	yes	yes
Algorithms	SES	DES	
parameters	α	(α, β)	

Table 6.2: Holt-Winters Algorithms.

Chapter 7

Comparing Holt-Winters Forecasting Algorithms

- In Single exponential smoothing (SES), how do we pick a value of α?
- In Holt's Linear Method (DES), how do we pick values of α and β?
- When faced with many alternative forecast models, how do we decide which one to use?

7.1 Definitions

Given a time series $y_1, \ldots, y_n$, fit a model and compute the fitted values $F_1, \ldots, F_n$.

1.1 Definition (SSE) The Sum of Square Errors is defined by:

$$\text{SSE} = \sum_{t=1}^{n} (y_t - F_t)^2$$

In R, the computer selects the best forecasting model by finding the parameters, α for SES or (α, β) for DES, such that SSE is minimal. Other software may use other criteria such as RMSE and MAPE.

1.2 Definition (RMSE) The *root mean square error* of the model is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y_t - F_t)^2}. = \sqrt{\frac{\text{SSE}}{n}}$$

Note that the same parameters (α or (α, β)) would be found when minimising the SSE or the RMSE.

1.3 Definition (MAPE) The *mean absolute percent error* is:

$$\text{MAPE} = 100\frac{1}{n}\sum_{t=1}^{n} \left| \frac{y_t - F_t}{y_t} \right|.$$

It is written as a percentage. Again, pick the model with the smallest MAPE.

Other software may use MAPE for finding the best parameters. This would give slightly different estimates of the parameters than using the SSE/RMSE.

7.2 Exercises

(1) Here is a short time series and its exponentially smoothed forecast with $\alpha = 0.5$. Compute the error and then the RMSE and MAPE:

y_i	F_i	$y_i - F_i$	$(y_i - F_i)^2$	$\left\lvert\frac{y_i - F_i}{y_i}\right\rvert$
4	4			
3	4			
5	3.5			
7	4.25			
5	5.73			
6	5.26			
4	5.73			

Chapter 8

Holt-Winters' Exponential Smoothing with Seasonality

What about an exponential smoothing for data with a trend and seasonal behaviour? Winters generalised Holt's linear method to come up with such a technique, now called Holt Winters. A seasonal equation is added to Holt's linear method equations. It is done in two ways, additive (cf. table 8.1) and multiplicative (cf. table 8.2).

Init:

$L_s = \frac{1}{s}\sum_{i=1}^{s} y_i$

$b_s = \frac{1}{s}\left[\frac{y_{s+1}-y_1}{s} + \frac{y_{s+2}-y_2}{s} + \cdots + \frac{y_{2s}-y_s}{s}\right]$

$S_i = y_i - L_s,\ i = 1,\ldots,s$

and choose $0 \le \alpha \le 1$ and $0 \le \beta \le 1$ and $0 \le \gamma \le 1$

Compute for $t > s$:

level	$L_t = \alpha\ (y_t - S_{t-s}) + (1-\alpha)\ (L_{t-1} + b_{t-1})$
trend	$b_t = \beta\ (L_t - L_{t-1}) + (1-\beta)\ b_{t-1},$
seasonal	$S_t = \gamma\ (y_t - L_t) + (1-\gamma)\ S_{t-s}$
forecast	$F_{t+1} = L_t + b_t + S_{t+1-s}$

Until no more observationare available

and subsequent forecasts:

$F_{n+k} = L_n + k\ b_n + S_{n+k-s}$

Table 8.1: Seasonal Holt Winter's Additive Model Algorithm (noted SHW+).

s is the length of the seasonal cycle. We have to pick the values of α, β and γ. As with the other methods (i.e. SES and DES), we can use the SSE/RMSE or MAPE to choose the best values.

8.1 Exercise

In the table on the next page are the first 14 months beer production data. Since the data have a 12 month seasonal cycle, we initialise L_{12}, b_{12} and $S_1,\ldots,S_{12}$. Use the additive model formulae to calculate month 13 and 14's level, trend and seasonality, and make a 1-step ahead forecast for months 13, 14 and 15. Use $\alpha = 0.5, \beta = 0.3$ and $\gamma = 0.9$.

Init:

$L_s = \frac{1}{s}\sum_{i=1}^{s} y_i$

$b_s = \frac{1}{s}\left[\frac{y_{s+1}-y_1}{s} + \frac{y_{s+2}-y_2}{s} + \cdots + \frac{y_{2s}-y_s}{s}\right]$

$S_i = \frac{y_i}{L_s},\ i = 1,\ldots,s$

and choose $0 \le \alpha \le 1$ and $0 \le \beta \le 1$ and $0 \le \gamma \le 1$

Compute for $t > s$:

level	$L_t = \alpha\ \frac{y_t}{S_{t-s}} + (1-\alpha)\ (L_{t-1} + b_{t-1})$
trend	$b_t = \beta\ (L_t - L_{t-1}) + (1-\beta)\ b_{t-1},$
seasonal	$S_t = \gamma\ \frac{y_t}{L_t} + (1-\gamma)\ S_{t-s}$
forecast	$F_{t+1} = (L_t + b_t)\ S_{t+1-s}$

Until no more observation are available

and subsequent forecasts:

$F_{n+k} = (L_n + k \cdot b_n)\ S_{n+k-s}$

Table 8.2: Seasonal Holt Winter's Multiplicative Model Algorithm (noted SHW×).

Month No.	Production	Level L_t	Trend b_t	Seasonal S_t	Forecast F_t
1	164	–	–	5.75	–
2	148	–	–	−10.25	–
3	152	–	–	−6.25	–
4	144	–	–	−14.25	–
5	155	–	–	−3.25	–
6	125	–	–	−33.25	–
7	153	–	–	−5.25	–
8	148	–	–	−12.25	–
9	138	–	–	−20.25	–
10	190	–	–	31.75	–
11	192	–	–	33.75	–
12	192	158.25	−0.65	33.75	–
13	147				
14	133				
15	163				

8.2 Selecting the best Holt-Winters algorithms

For a time series, you select the Holt Winters algorithms with the smallest SSE or RMSE or MAPE as defined in chapter 7. To analyse a time series, identify the patterns that occur in the time series (is there a trend? is there seasonality? there is always noise!), and decide which algorithm(s) is the best suited i.e. when no seasonality select the best algorithm between SES and DES, and when there is seasonality, select the best algorithm between SHW+ and SHWx. To know more about Holt-Winters algorithms (in particular their limitations), read at Goodwin short paper [2].

Part III

Statistical models: ARIMA

This part is investigating one important class of linear statistical models called ARIMA. ARIMA models use the same hypotheses and the same approach as Linear regression, so chapter 9 (re-)introduces Linear regression and show how Linear regression could be used for Forecasting. Linear regression however requires the definition of explanatory variables, and the selection of informative explanatory variables can often only be done by domain experts. In addition to choosing these explanatory variables, one also needs to collect this data along with the time series of interest. On the contrary, ARIMA models only requires the times series to be recorded for a while, and no additional information is required for analysis.

In this part of the lecturenotes:

- chapter 10 introduces AutoRegressive models. The acronym used for these models is AR and corresponds to the first 2 letters of the acronym ARIMA.
- chapter 11 introduces Moving Average models. The acronym used for these models is MA and corresponds to the last 2 letters of the acronym ARIMA.
- chapter 13 presents the ACF and PACF in more details, and illustrates what sort of shapes these should have when dealing with time series following perfectly a AR or MA modelling.
- chapter 14 introduces the backshift operator. This operator is meant to simplifies the mathematical notations for ARIMA models.
- To analyse a time series, several ARIMA models can be suitable, and chapter 15 presents two criteria that can be used to select which model is the best to analyse the time series.
- AR, MA and ARMA models are not able to deal with time series with a slope, and chapter 16 presents how this limitation is overcome by the Integration of differencing. The I of integration corresponds to the I of the acronym ARIMA: AutoRegressive Integrated Moving Average models.
- Seasonal ARIMA models are then introduced in chapter 17 to model seasonal patterns.
- Chapter 18 presents techniques that can be used to transform a time series such that they become suitable for analysis with ARIMA models.

Chapter 9

Linear Regression

9.1 Regression with one explanatory variable

We have collected the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where the x_i are known *predictor* values set by the experimenter and y_i is the observed *response*. We wish to model y as a function of x. In simple linear regression we assume that y is related to x by:

$$y_i = a + b\, x_i + \epsilon_i,$$

with the assumptions that

- ϵ_i is the error and is normally distributed with mean 0 and unknown variance σ^2.
- ϵ_i and ϵ_j are independent when $i \neq j$.

Thus we say that the y_i follow a "pattern" $a + b\, x_i$ but with some random, unpredictable behaviour modelled as a normal distribution. Or, in other words, given a and b, y_i will be normally distributed with a mean $a + b\, x_i$ and variance σ^2.

(1) **Fitting the Model:** The best fitting values of α and β are the *least squares estimates* that minimise the Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

also known as the Sum of Square Errors (SSE). The estimates $(\hat{a}, \hat{b})$ are then computed such that:

$$\frac{\partial RSS}{\partial a} = 0 \quad \text{and} \quad \frac{\partial RSS}{\partial b} = 0$$

giving the **least squares** estimates:

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{a} = \bar{y} - \hat{\beta}\bar{x}, \tag{9.1}$$

and σ^2 is estimated from the sum of squares by the quantity:

$$s^2 = \frac{\widehat{RSS}}{(n-2)}$$

$\widehat{RSS}$ indicates that the RSS value is computed with the least squares estimate $(\hat{a}, \hat{b})$. The denominator $n-2$ corresponds to the degree of freedom in the errors $\{\epsilon_i\}_{i=1,\cdots,n}$: the estimation of 2 parameters (a, b) removes 2 degrees of freedom.

The means of x and y are simply estimated by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

(2) **Measuring the Strength of the Linear Relationship:**

1.1 Definition (correlation coefficient) Usually we calculate the *correlation coefficient*:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}.$$

If you note $\tilde{\mathbf{x}}$ a column vector gathering all the values $(x_1-\bar{x}, x_2-\bar{x}, \cdots, x_n-\bar{x})$, and $\tilde{\mathbf{y}}$ a column vector collecting all the values $(y_1-\bar{y}, y_2-\bar{y}, \cdots, y_n-\bar{y})$ then the correlation coefficient corresponds to:

$$r_{xy} = \frac{<\tilde{\mathbf{x}}, \tilde{\mathbf{y}}>}{\|\tilde{\mathbf{x}}\|\ \|\tilde{\mathbf{y}}\|}$$

where $<\tilde{\mathbf{x}}, \tilde{\mathbf{y}}>$ is the dot product between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. $\|\tilde{\mathbf{x}}\|$ (resp. $\|\tilde{\mathbf{y}}\|$) is the norm of the vector $\tilde{\mathbf{x}}$ (resp. $\tilde{\mathbf{y}}$). By definition of the dot product, we have:

$$<\tilde{\mathbf{x}}, \tilde{\mathbf{y}}> = \|\tilde{\mathbf{x}}\| \cdot \|\tilde{\mathbf{y}}\|\ \cos(\alpha)$$

where α is the angle between vectors $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. The correlation coefficient is then simplified to $r_{xy} = \cos\alpha$ and consequently has values between -1 (perfect negative correlation) and $+1$ (perfect positive correlation).

Another important measure when we do regression is the coefficient of determination.

1.2 Definition (coefficient of determination) The *coefficient of determination* is the correlation between the y_i and their predicted values from the fitted model $\hat{y}_i = \hat{a} + \hat{b}x_i$:

$$R^2 = r_{y\hat{y}}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

For simple linear regression, as explained here, $R^2 = r_{xy}^2$. This is not true for more general multivariate regression.

(3) **Evaluating Model Fit:** We look at the *residuals*. These are the difference between the observed and predicted values for y:

$$\epsilon_i = y_i - \hat{y}_i.$$

There are two problems with model fit that can arise:

- We have not fit the pattern of the data. Any unmodelled relationships between x and y appear in the residuals. A scatter plot of the residuals against the x_i should show up any unmodelled patterns.
- The normally distributed error assumption is not correct. The residuals should be independent and normally distributed with variance σ^2. A histogram of the residuals can usually verify this.

(4) **Outliers:** An observation that has an unusually large residual is an *outlier*. An outlier is an observation that has been predicted very badly by the model. *Standardised residuals* give a good indication as to whether an observation is an outlier. We should investigate if there is some special reason for the outlier occurring. Outliers can also cause problems because they can significantly alter the model fit, making it fit the rest of the data worse than it would otherwise.

(5) **Making Predictions:** Given a new value X, y should be normally distributed with mean $a + b\ X$ and variance σ^2. We replace a, b and σ^2 by their estimates and so forecast the value of y to be $\hat{y} = \hat{a} + \hat{b}X$. A 95% prediction interval turns out to be:

$$\hat{a} + \hat{b}X \pm 2\ s$$

(6) **Statistical Tests in Regression:** An *F-test* is used to determine if there is any significant linear relationship between x and y. In other word, it is checking if the hypothesis $b = 0$ is true or not. The test statistic is:

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)/(n-2)},$$

which is compared with the F-value having 1 and $n-2$ degrees of freedom.

9.2 Using Linear Regression to Make Forecasts

9.2.1 Time as an explanatory variable

We have a time series $y_1, y_2, \ldots, y_n$. We can fit a regression to the time plot i.e. where the x_i values are the time that the ith observation was taken. Usually our observations occur at equally spaced time intervals and we take that interval to be the unit of time, so that $x_i = i$.

9.2.2 Indicator variables: modelling seasonality

2.1 Definition (indicator variable) A indicator variable is a binary variable that takes only 0 or 1 as possible values.

There is another way to model seasonality that does not require the autoregressive idea presented in chapter 10. Think of monthly data with a yearly seasonal cycle. For each month in the year, we can define an *indicator* variable for instance variable 'Jan' for January, 'Feb' for February, etc. e.g.:

$$\text{Jan}_i = \begin{cases} 1 & \text{if } i \text{ is corresponding to the month of January} \\ 0 & \text{otherwise} \end{cases}$$

We then fit by linear regression the model:

$$y_i = a + b\ i + \gamma_1\ \text{Jan}_i + \gamma_2\ \text{Feb}_i + \cdots + \gamma_{12}\ \text{Dec}_i + \epsilon_i.$$

EXERCISE: If month i is January, what does the above equation reduce to? What if the month is February?

The parameters $\gamma_1, \ldots, \gamma_{12}$ represent a *monthly* effect, the same in that month for all the series, that is the departure from the trend-cycle in that month. There's one technical matter with the above model. One of the monthly terms is not a *free* parameter. We can in fact only fit 11 of the 12 monthly effects and the rest is absorb by the term $a + b\ i$. In other word we need only 11 binary varables to encode the 12 months:

I choose to eliminate the January effect — you are free to choose any other if you wish — so the model we use is:

$$y_i = a + b\ i + \gamma_2\ \text{Feb}_i + \cdots + \gamma_{12}\ \text{Dec}_i + \epsilon_i.$$

EXERCISE: What is the trend-cycle component of the model? What is the seasonal component?
EXERCISE: what is the earliest value in the series that we can compute a predicted value for?

EXERCISE: we have quarterly data with a yearly seasonal component. What model would you fit using this method?

9.3 Least Square algorithm in Matrix Form

9.3.1 Least Squares for Linear regression

This is the case when only one explanatory variable x is used to explain y:

$$y = a + bx + \epsilon \tag{9.2}$$

Having collected n observations $\{(x_i, y_i)\}_{i=1,\cdots,n}$, we can write the following linear system:

$$\begin{cases} y_1 = a + b\ x_1 + \epsilon_1 \\ y_2 = a + b\ x_2 + \epsilon_2 \\ \vdots \\ y_n = a + b\ x_n + \epsilon_n \end{cases}$$

this system can be rewritten as:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{\mathrm{X}} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\Theta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

Minimising the RSS corresponds to finding Θ such that

$$\begin{aligned} \widehat{\Theta} &= \arg\min_\Theta \left\{RSS = \textstyle\sum_{i=1}^n \epsilon_i^2 = \|\boldsymbol{\epsilon}\|^2\right\} \\ &= \arg\min_\Theta \left\{RSS = \|\mathbf{y} - \mathrm{X}\Theta\|^2\right\} \\ &= \arg\min_\Theta \left\{RSS = (\mathbf{y} - \mathrm{X}\Theta)^T(\mathbf{y} - \mathrm{X}\Theta)\right\} \\ &= \arg\min_\Theta \left\{RSS = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathrm{X}\Theta - \Theta^T\mathrm{X}^T\mathbf{y} - \Theta^T\mathrm{X}^T\mathrm{X}\Theta\right\} \end{aligned}$$

To find the minimum, we differentiate and find the solution such that the derivative is zero. We use here differentiation w.r.t. a vector Θ:

$$\begin{aligned} \frac{d\ RSS}{d\Theta} &= 0 - (\mathbf{y}^T\mathrm{X})^T - \mathrm{X}^T\mathbf{y} + \mathrm{X}^T\mathrm{X}\Theta + (\mathrm{X}^T\mathrm{X})^T\Theta \\ &= -2\mathrm{X}^T\mathbf{y} + 2\mathrm{X}^T\mathrm{X}\Theta \end{aligned}$$

using table 9.1. So the estimate of Θ such that the derivative of the RSS is zero is:

$$\widehat{\Theta} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathbf{y} \quad \text{(Least Square estimate)} \tag{9.3}$$

you can check that equation (9.3) gives the same result as equation (9.1).

y	$\frac{\partial y}{\partial x}$
Ax	A^T
$x^T A$	A
$x^T x$	$2x$
$x^T Ax$	$Ax + A^T x$

Table 9.1: Useful vector derivative formulas

9.3.2 Multiple Linear regression

Solution in equation (9.3) remains the same when considering multiple linear regression: X and Θ just need to be expanded. For instance considering the case of 2 explanatory variables:

$$y = a + b\,x + c\,z + \epsilon$$

having collected observations $\{(y_i, x_i, z_i)\}_{i=1,\cdots,n}$, matrix X becomes

$$\mathrm{X} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \\ 1 & x_n & z_n \end{bmatrix}$$

and Θ is

$$\Theta = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Chapter 10

AR(p): Autoregressive Models

10.1 Definition

1.1 Definition An *autoregressive* model is a very common model for time series. Consider a series $y_1, y_2, \ldots, y_n$. An autoregressive model of order p (denoted AR(p)) states that y_i is the *linear function of the previous p values of the series* plus an error term:

$$y_i = \phi_0 + \phi_1\, y_{i-1} + \phi_2\, y_{i-2} + \cdots + \phi_p\, y_{i-p} + \epsilon_i,$$

where $\phi_1, \ldots, \phi_p$ are weights that we have to define or determine, and ϵ_i are normally distributed with zero mean and variance σ^2.

Note: the formula only holds for $i > p$. We have to define $y_1, y_2, \ldots, y_p$ before we can use the formula. We'll concentrate on the simplest model, the AR(1), where:

$$y_i = \phi_0 + \phi_1\, y_{i-1} + \epsilon_i.$$

For fitting an AR(1) Model, we have the observations $y_1, \cdots, y_n$ that defines the linear system of $n-1$ equations:

$$\begin{cases} y_2 &= \phi_0 + \phi_1\, y_1 + \epsilon_2 \\ y_3 &= \phi_0 + \phi_1\, y_2 + \epsilon_3 \\ \vdots & \vdots \\ y_n &= \phi_0 + \phi_1\, y_{n-1} + \epsilon_n \end{cases}$$

(1) Define $x_i = y_{i-1}$; this is called the *lagged* series. Note that x_i is only defined for $i = 2, \ldots, n$. It is NOT defined for $i = 1$, since there is no y_0.

(2) The AR(1) model is then:

$$y_i = \phi_0 + \phi_1\, x_i + \epsilon_i.$$

This is just the linear regression model! So, we can fit this model by doing a linear regression of the series against the lagged series. That will give us the best values for the parameters $\hat{\phi}_0$ and $\hat{\phi}_1$, and an estimate s^2 for σ^2. We could also do an F-test to verify if there is a significant relationship.

(3) NOTE: because x_1 does not exist, the regression is fitted on $n-1$ points $(x_2, y_2), \ldots, (x_n, y_n)$.

(4) Our fitted values for the series are then

$$\hat{y}_i = \hat{\phi}_0 + \hat{\phi}_1 x_i = \hat{\phi}_0 + \hat{\phi}_1 y_{i-1},$$

for $i = 2, \ldots, n$. We cannot fit a value to y_1 because there is no y_0!

(5) We estimate σ^2 by s^2:

$$s^2 = \frac{1}{n-1-2}\sum_{i=2}^{n}(y_i - \hat{\phi}_0 - \hat{\phi}_1 y_{i-1})^2 = \frac{1}{n-3}\sum_{i=2}^{n}(y_i - \hat{\phi}_0 - \hat{\phi}_1 y_{i-1})^2$$

Note that we had only $n-1$ equations in the linear system used to estimate $(\hat{\phi}_0, \hat{\phi}_1)$, and there are 2 parameters $(\hat{\phi}_0, \hat{\phi}_1)$ in our model. Thus a 95% prediction interval for y_i when x_i is known, is

$$\hat{\phi}_0 + \hat{\phi}_1 x_i \pm 2s.$$

10.2 Prediction interval for AR(1) k steps ahead

EXERCISE: We observe $y_1, \ldots, y_n$. We fit an AR(1) model

$$y_i = \hat{\phi}_0 + \hat{\phi}_1 \; y_{i-1} + \epsilon_i$$

(1) What is our forecast for y_{n+1}? What is the 95% prediction interval?

Ans. *According to the AR model:*

$$y_{n+1} = \hat{\phi}_0 + \hat{\phi}_1 \; y_n + \epsilon_{n+1}$$

We dont know the value of $\epsilon_{n+1} \sim \mathcal{N}(0, s^2)^*$*, but we know* $\hat{\phi}_0$, $\hat{\phi}_1$ *and* y_n*. So*

$$y_{n+1} = \underbrace{\hat{\phi}_0 + \hat{\phi}_1 \; y_n}_{\text{forecast } \hat{y}_{n+1}} \pm 2 \; s$$

with $s^2 = \frac{\sum_{i=2}^{n}\epsilon_i^2}{n-3}$.

(2) Forecast for y_{n+2}?

Ans. *According to the AR model:*

$$y_{n+2} = \hat{\phi}_0 + \hat{\phi}_1 \; y_{n+1} + \epsilon_{n+2}$$

We dont know y_{n+1} *(we just know a prediction* $\hat{y}_{n+1}$*) so we replacing* y_{n+1} *by its expression w.r.t* y_n*:*

$$\begin{aligned} y_{n+2} &= \hat{\phi}_0 + \hat{\phi}_1 \; y_{n+1} + \epsilon_{n+2} \\ &= \hat{\phi}_0 + \hat{\phi}_1 \; (\hat{\phi}_0 + \hat{\phi}_1 y_n + \epsilon_{n+1}) + \epsilon_{n+2} \\ &= \underbrace{\hat{\phi}_0 + \hat{\phi}_1 \; \hat{\phi}_0 + \hat{\phi}_1^2 y_n}_{\text{Forecast } \hat{y}_{n+2}} + \underbrace{\hat{\phi}_1 \epsilon_{n+1} + \epsilon_{n+2}}_{\text{error term}} \end{aligned}$$

Note that the Forecast is the part that we can compute (i.e. we know the values of $\hat{\phi}_0, \hat{\phi}_1, y_n$ *) whereas we dont know the values of the errors, we only know how these behave statistically.*

(3) What is the prediction interval for y_{n+2}?

Ans. *From the previous question, we know the forecast* $\hat{y}_{n+2}$ *and the error on this forecast. We need to estimate the variance of the error. First lets compute its mean*†*:*

$$\mathbb{E}[\hat{\phi}_1 \epsilon_{n+1} + \epsilon_{n+2}] = \hat{\phi}_1 \mathbb{E}[\epsilon_{n+1}] + \mathbb{E}[\epsilon_{n+2}]$$

$^*\mathcal{N}(0, s^2)$ indicates a normal distribution with mean 0 and estimated variance s^2.

$^\dagger\mathbb{E}[\cdot]$ is the expectation and it is a linear operator.

We know $\epsilon_{n+1} \sim \mathcal{N}(0, s^2)$ *and* $\epsilon_{n+2} \sim \mathcal{N}(0, s^2)$ *so* $\mathbb{E}[\epsilon_{n+1}] = 0$ *and* $\mathbb{E}[\epsilon_{n+2}] = 0$. *Now lets compute the variance of the error term:*

$$\mathbb{E}[(\hat{\phi}_1\epsilon_{n+1} + \epsilon_{n+2})^2] = \hat{\phi}_1^2 \underbrace{\mathbb{E}[\epsilon_{n+1}^2]}_{=s^2} + 2\hat{\phi}_1 \underbrace{\mathbb{E}[\epsilon_{n+1}\,\epsilon_{n+2}]}_{=0} + \underbrace{\mathbb{E}[\epsilon_{n+2}^2]}_{=s^2}$$

The expectation of $\epsilon_{n+1} \times \epsilon_{n+2}$ *is 0 because we assume independence of the residuals. So the 95% confidence interval is*

$$y_{n+2} = \hat{y}_{n+2} \pm 2\, s\sqrt{(\hat{\phi}_1^2 + 1)}$$

We see that the confidence interval is getting larger as we move further in the future from the last observation available y_n.

(4) How would we go about forecasting k steps ahead, that is y_{n+k}? What is the prediction interval?

we know that

$$y_{n+1} = \underbrace{\hat{\phi}_0 + \hat{\phi}_1\, y_n}_{\text{Forecast } \hat{y}_{n+1}} \quad \underbrace{\pm 2s}_{\text{confidence interval}}$$

and

$$y_{n+2} = \underbrace{\hat{\phi}_0 + \hat{\phi}_0 \times \hat{\phi}_1 + \hat{\phi}_1^2\, y_n}_{\text{Forecast } \hat{y}_{n+2}} \quad \underbrace{\pm 2s\sqrt{1 + \hat{\phi}_1^2}}_{\text{confidence interval}}$$

and

$$\begin{aligned} y_{n+3} &= \hat{\phi}_0 + \hat{\phi}_1 y_{n+2} + \epsilon_{n+3} \\ &= \hat{\phi}_0 + \hat{\phi}_1(\hat{\phi}_0 + \hat{\phi}_1 y_{n+1} + \epsilon_{n+2}) + \epsilon_{n+3} \\ &= \hat{\phi}_0 + \hat{\phi}_1(\hat{\phi}_0 + \hat{\phi}_1(\hat{\phi}_0 + \hat{\phi}_1 y_n + \epsilon_{n+1}) + \epsilon_{n+2}) + \epsilon_{n+3} \\ &= \underbrace{\hat{\phi}_0 + \hat{\phi}_1\hat{\phi}_0 + \hat{\phi}_1^2\hat{\phi}_0 + \hat{\phi}_1^3 y_n}_{\text{Forecast } \hat{y}_{n+3}} + \underbrace{\hat{\phi}_1^2\epsilon_{n+1} + \hat{\phi}_1\epsilon_{n+2} + \epsilon_{n+3}}_{\text{error term}} \end{aligned}$$

so

$$y_{n+3} = \hat{\phi}_0 + \hat{\phi}_1\hat{\phi}_0 + \hat{\phi}_1^2\hat{\phi}_0 + \hat{\phi}_1^3 y_n \pm 2s\sqrt{1 + \hat{\phi}_1{}^2 + \hat{\phi}_1{}^4}$$

So we propose the following formula:

$$y_{n+k} = \underbrace{\hat{\phi}_0\left(\sum_{i=1}^{k}\phi_1^{i-1}\right) + \phi_1^k\, y_n}_{\text{forecast}} + \underbrace{\sum_{i=1}^{k}\hat{\phi}_1^{i-1}\epsilon_{n+k-i-1}}_{\text{error term}} \tag{10.1}$$

implying:

$$y_{n+k} = \underbrace{\hat{\phi}_0\left(\sum_{i=1}^{k}\phi_1^{i-1}\right) + \phi_1^k\, y_n}_{\text{forecast}} \underbrace{\pm 2s\sqrt{\sum_{i=1}^{k}\phi_1^{2(i-1)}}}_{\text{Confidence interval}} \tag{10.2}$$

By induction, we can show that equation (10.1) is valid at step $k+1$:

$$\begin{aligned} y_{n+k+1} &= \hat{\phi}_0 + \hat{\phi}_1\, y_{n+k} + \epsilon_{n+k+1} \\ &= \hat{\phi}_0 + \hat{\phi}_1\, \left(\hat{\phi}_0\left(\textstyle\sum_{i=1}^{k}\phi_1^{i-1}\right) + \phi_1^k\, y_n + \textstyle\sum_{i=1}^{k}\hat{\phi}_1^{i-1}\epsilon_{n+k-i-1}\right) + \epsilon_{n+k+1} \\ &= \hat{\phi}_0 + \hat{\phi}_1\hat{\phi}_0\left(\textstyle\sum_{i=1}^{k}\phi_1^{i-1}\right) + \phi_1^{k+1}\, y_n + \hat{\phi}_1\textstyle\sum_{i=1}^{k}\hat{\phi}_1^{i-1}\epsilon_{n+k-i-1} + \epsilon_{n+k+1} \\ &= \hat{\phi}_0\left(\textstyle\sum_{i=1}^{k+1}\phi_1^{i-1}\right) + \phi_1^{k+1}\, y_n + \textstyle\sum_{i=1}^{k+1}\hat{\phi}_1^{i-1}\epsilon_{n+k+1-i-1} \end{aligned}$$

Note that the width of the confidence interval depends on the term:

$$\sum_{i=1}^{k} \phi_1^{2(i-1)}$$

We recognise here a geometric series and its limit is:

$$\lim_{k\to\infty} \sum_{i=1}^{k} \phi_1^{2(i-1)} = \frac{1}{1-\phi_1^2}$$

So for an AR(1) model, the confidence interval is growing up to a finite limit (it is bounded). Check this result for AR, MA, ARMA models using simulations R.

Chapter 11

MA(q): Moving Average Processes

11.1 Definitions

1.1 Definition A *moving average model of order 1* is a time series model defined as follows:

$$y_t = \psi_0 - \psi_1 \epsilon_{t-1} + \epsilon_t$$

where ϵ_t are *independent* errors, normally distributed with mean 0 and variance σ^2: $\mathcal{N}(0,\sigma^2)$.

1.2 Definition A *moving average model of order q*, noted MA(q), is a time series model defined as follows:

$$y_t = \psi_0 - \psi_1 \epsilon_{t-1} - \psi_2 \epsilon_{t-2} - \cdots - \psi_q \epsilon_{t-q} + \epsilon_t$$

11.2 Fitting an MA model

The errors are now used as explanatory variables in MA models! Lets assume this simplified MA(1) model:

$$y_t = \psi_1 \epsilon_{t-1} + \epsilon_t$$

Assuming we have observed the first n values of a time series the $y_1, \cdots, y_n$, then we can write the following system of n equations with the convention $\epsilon_0 = 0$:

$$\begin{cases} y_1 &= \psi_1 \epsilon_0 + \epsilon_1 \\ y_2 &= \psi_1 \epsilon_1 + \epsilon_2 \\ y_3 &= \psi_1 \epsilon_2 + \epsilon_3 \\ \vdots & \vdots \\ y_n &= \psi_1 \epsilon_{n-1} + \epsilon_n \end{cases} \equiv \begin{cases} y_1 &= \epsilon_1 \\ y_2 &= \psi_1 y_1 + \epsilon_2 \\ y_3 &= \psi_1 (y_2 - \psi_1 y_1) + \epsilon_3 \\ \vdots & \vdots \\ y_n &= \psi_1 y_{n-1} - \psi_1^2 y_{n-2} + \cdots + (-\psi_1)^{n-1} y_1 + \epsilon_n \end{cases}$$

We estimate the parameter ψ_1 by minimising the sum of squares errors again, however the systems of equations is non-linear w.r.t. the parameter ψ_1 (powers of ψ_1 appears in the expression). More complex numerical methods can perform this estimation (out of scope in this class). The simple Least Squares algorithm used for Linear Regression (cf. chapter 9) and AR models cannot be used when there is an MA component in the model.

Chapter 12

ARMA(p,q): AutoRegressive Moving Average Models

12.1 Definition

1.1 Definition Combining AR and MA models, we can define ARMA(p,q) models as:

$$y_t = \phi_0 + \phi_1\, y_{t-1} + \cdots + \phi_p\, y_{t-p} + \psi_0 - \psi_1\, \epsilon_{t-1} - \cdots - -\psi_q\, \epsilon_{t-q} + \epsilon_t$$

with p the order of the AR part, and q the order of the MA part. ψ_0 and ϕ_0 can be put together to define a unique constant c:

$$y_t = c + \phi_1\, y_{t-1} + \cdots + \phi_p\, y_{t-p} - \psi_1\, \epsilon_{t-1} - \cdots - -\psi_q\, \epsilon_{t-q} + \epsilon_t$$

Note with ARMA(p,q) models, it is difficult to identify orders of the AR(p) and MA(q) parts using the ACF/PACF functions.

The parameters $\{\phi_i\}$ are computed by minimising the sum of square errors (the algorithm is out of scope of the course).

12.2 Exercises

(1) Identify the following equations as MA(q), AR(p) or ARMA(p,q) identifying the orders p and q:

(a) $y_t = \phi_0 + \phi_{12}\, y_{t-12} + \epsilon_t$

(b) $y_t = \psi_0 + \psi_{12}\, \epsilon_{t-12} + \epsilon_t$

(c) $y_t = c + \phi_{12} y_{t-12} + \psi_{12}\, \epsilon_{t-12} + \epsilon_t$

(d) $y_t = \phi_0 + \phi_1\, \epsilon_{t-1} + \phi_{12}\, \epsilon_{t-12} + \epsilon_t$

(2) Assume an MA(1) model, what is the expectation $\mathbf{E}[y_t]$? Is it stationary in mean (i.e. is $\mathbf{E}[y_t]$ changing with the time t)?

(3) Assume an MA(2) model, what is the expectation $\mathbf{E}[y_t]$? Is it stationary in mean?

(4) Assume an AR(1) model, what is the expectation $\mathbf{E}[y_t]$ (consider when $t \to \infty$)? Is it stationary in mean?

12.3 Simulation of ARMA models

Figures 12.1, 12.2 and 12.3 shows simulations of AR and ARMA models. We can notice that there is no trend appearing on these simulated ARMA models.

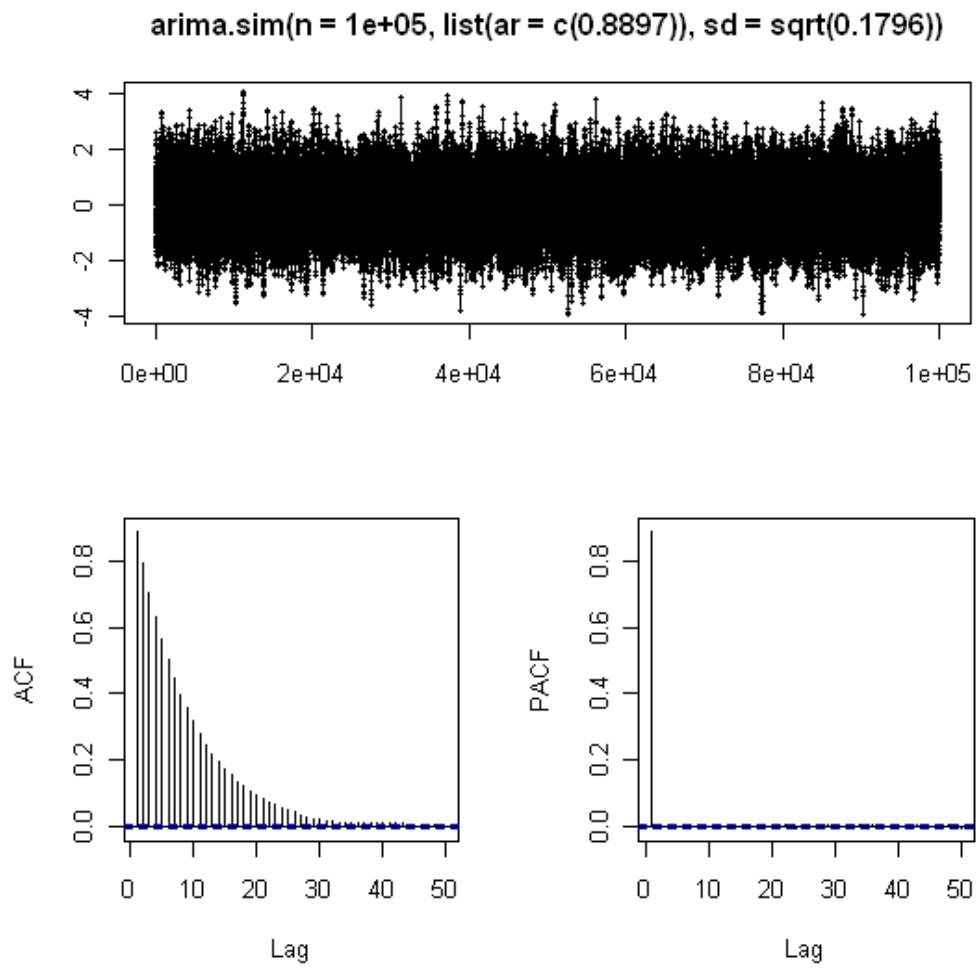


Figure 12.1: Simulation of AR(1) : `>tsdisplay(arima.sim(n=100000,list(ar = c(0.8897)),sd = sqrt(0.1796)))`

12.4 Stationarity in mean and variance

4.1 Definition (Stationary in mean) A time series is called *stationary in mean* if it randomly fluctuates about a *constant* mean level.

EXERCISE:

(1) Are the simulated times series in figures 12.1, 12.2 and 12.3 stationary in mean?

(2) what types of patterns of a time series would imply that it is not stationary in mean?

Stationary series are 'nice' because they are not complicated. They are easily modelled and fitted (an ARMA model will usually do) and we do not have to worry about seasonal or trend/cycle.

In fact, in time series modelling, the term *stationarity* has a more general meaning. There are three key parts that make a series stationary:

- The mean is constant (stationary in mean)
- The variance is finite (stationary in variance)

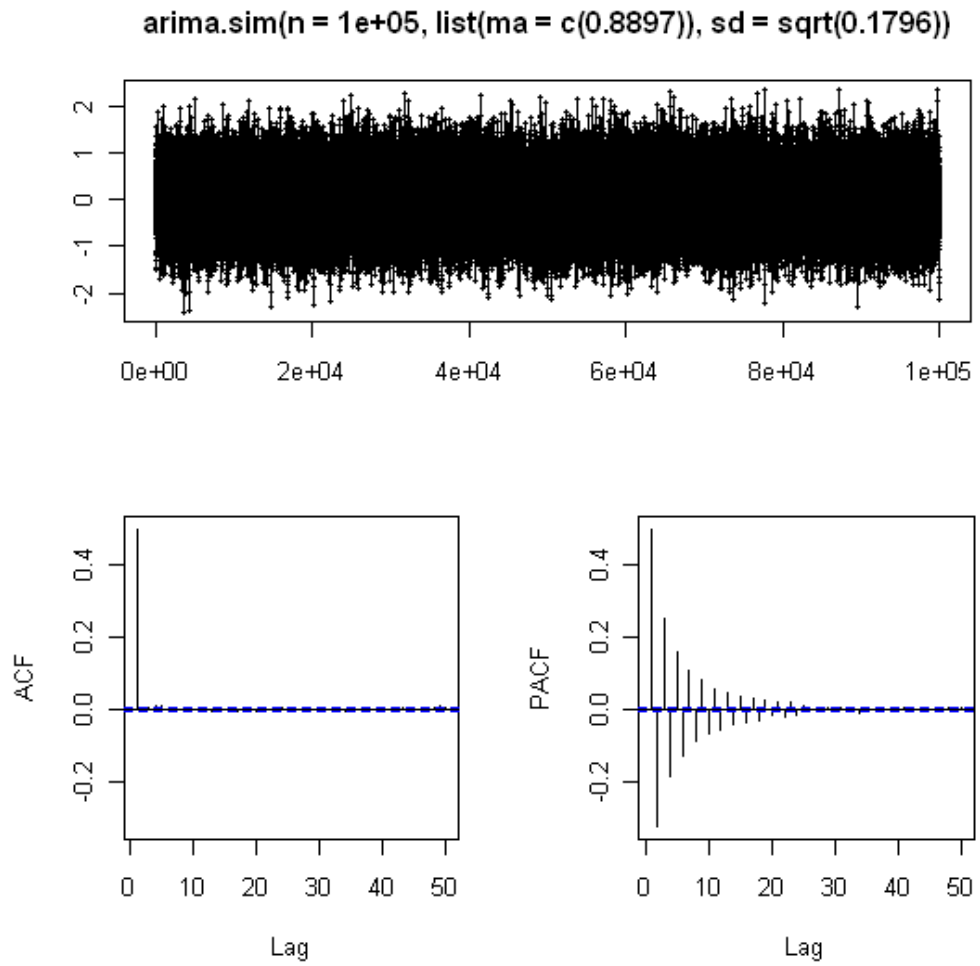


Figure 12.2: Simulation of MA(1) : `>tsdisplay(arima.sim(n=100000,list(ma = c(0.8897)),sd = sqrt(0.1796)))`

- The correlation between values in the time series depends *only* on the time distance between these values. (stationary in autocorrelation)

We spend most of our time discussing the first two.

4.2 Definition (Stationarity in variance) In addition to stationarity in mean, a time series is said to be *stationary in variance* if the variance in the time series does not change with time.

EXERCISE:

(1) Are the simulated times series in figures 12.1, 12.2 and 12.3 stationary in variance?

(2) As well as non-stationarity in both mean and variance, series can also be: non-stationary in mean and stationary in variance; or stationary in mean and non-stationary in variance; or non-stationary in both. Sketch time series with each of these three properties.

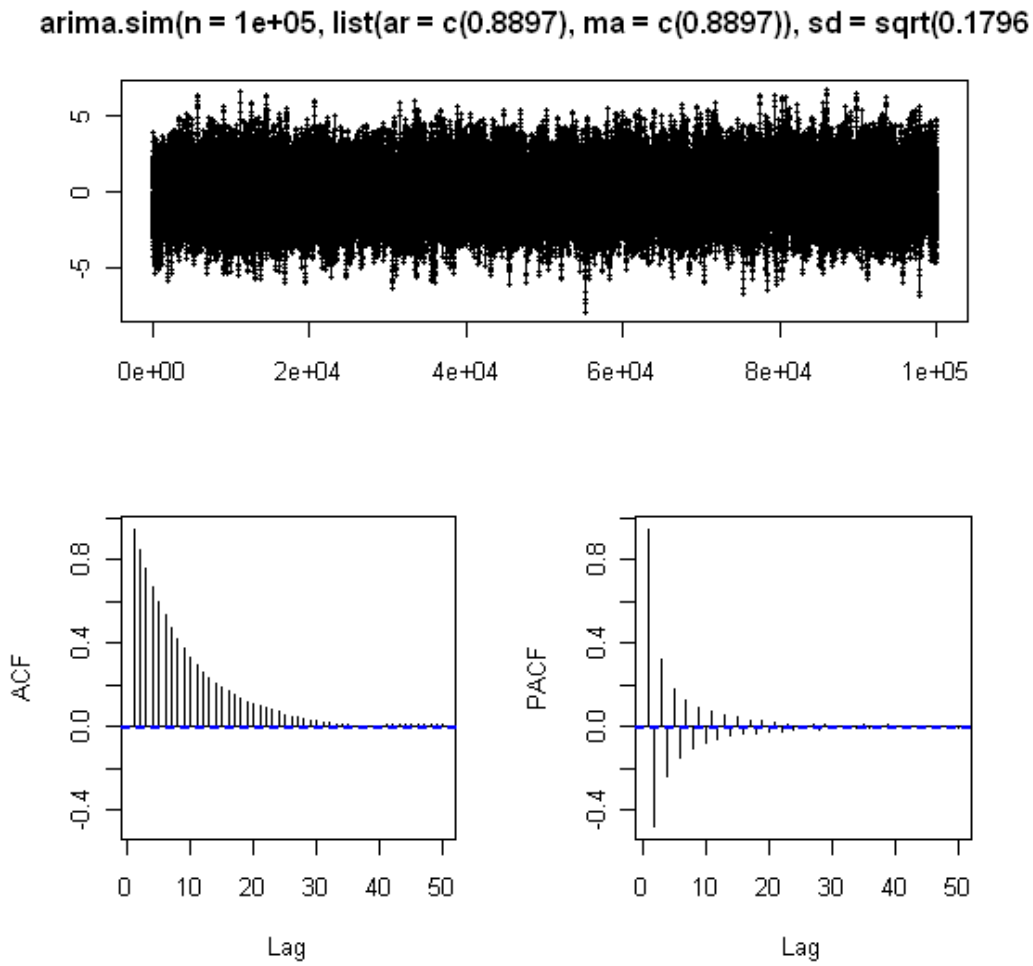


Figure 12.3: Simulation of ARMA(1,1) : `>tsdisplay(arima.sim(n=100000,list(ar=c(.8897),ma = c(0.8897)),sd = sqrt(0.1796)))`

12.5 Conclusion

ARMA models are not able to handle time series that are not stationary in mean and variance. In other word, ARMA models should only be fitted to time series that are stationary in mean (i.e. no trend or no seasonal pattern) and stationary in variance.

Chapter 13

Using ACF and PACF to select MA(q) or AR(p) models

The principle way to determine which AR or MA model is appropriate is to look at the ACF and PACF of the time series. The table 13.1 gives the theoretical behaviour of these functions for different MA and AR models. An informal way to pick a model is to compute the ACF and PACF for a time series and match it to the behaviour in the table 13.1. This rule is however difficult to use when the time series is explained by an ARMA model (combined effect of AR and MA).

MODEL	**ACF**	**PACF**
AR(1)	Exponential decay: on +ve side if $\phi_1 > 0$ and alternating in sign, starting on −ve side, if $\phi_1 < 0$.	Spike at lag 1, then 0; +ve spike if $\phi_1 > 0$ and −ve spike if $\phi_1 < 0$.
AR(p)	Exponential decay or damped sine wave. The exact pattern depends on the signs and sizes of $\phi_1, \ldots, \phi_p$.	Spikes at lags 1 to p, then zero.
MA(1)	Spike at lag 1, then 0; +ve spike if $\psi_1 < 0$ and −ve spike if $\psi_1 > 0$.	Exponential decay: on +ve side if $\psi_1 < 0$ and alternating in sign, starting on +ve side, if $\phi_1 < 0$.
MA(q)	Spikes at lags 1 to q, then zero.	Exponential decay or damped sine wave. The exact pattern depends on the signs and sizes of $\psi_1, \ldots, \psi_q$.

Table 13.1: Shapes of ACF and PACF to identify AR or MA models suitable to fit time series.

13.1 ACF and PACF

1.1 Definition (ACF) At lag k, the ACF is computed by:

$$ACF(k) = \frac{\mathbb{E}\left[\left(y_t - \mathbb{E}[y_t]\right)\left(y_{t-k} - \mathbb{E}[y_{t-k}]\right)\right]}{\sqrt{\mathbb{V}ar[y_t]\mathbb{V}ar[y_{t-k}]}}$$

In time series, we may want to measure the relationship between Y_t and Y_{t-k} when the effects of other time lags $1, 2, \ldots, k-1$ have been removed. The autocorrelation does not measure this. However, *Partial autocorrelation* is a way to measure this effect.

1.2 Definition (PACF) The partial autocorrelation of a time series at lag k is denoted α_k and is found as follows:

(1) Fit a linear regression of y_t to the first k lags (i.e. fit an AR(k) model to the time series):

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_k y_{t-k} + \epsilon_t$$

(2) Then $\alpha_k = \hat{\phi}_k$, the fitted value of ϕ_k from the regression (Least Squares).

The set of partial autocorrelations at different lags is called the *partial autocorrelation function* (PACF) and is plotted like the ACF.

13.2 Exercises: ACF and PACF for AR(1) and MA(1)

(1) Assuming AR(1) model with $\phi_0 = 0$, show that the PACF coefficients are zeros when $k > 1$.

Ans. *By definition, the model is (ignoring the constant term ϕ_0) :*

$$y_t = \phi_1 \, y_{t-1} + \epsilon_t$$

Computing the PACF at order 2 for instance, implies to fit a AR(2) model to our AR(1). This is easily done:

$$y_t = \phi_1 \, y_{t-1} + 0 \, y_{t-2} + \epsilon_t$$

therefore the PACF coefficient at lag 2, is 0. The same reasoning can be used for any $k > 1$. At lag $k = 1$, the PACF coefficient is ϕ_1. This explains the shape of the PACF you have for a simulated AR(1) model using R.

(2) Lets assume a MA(1) model with $\psi_0 = 0$

- what is $\mathbb{E}[y_t]$?
  Ans.
  $$\begin{aligned}\mathbb{E}[y_t] &= \mathbb{E}[\psi_1 \, \epsilon_{t-1} + \epsilon_t] \quad \textit{By def. of our MA(1)} \\ &= \psi_1 \, \mathbb{E}[\epsilon_{t-1}] + \mathbb{E}[\epsilon_t] \quad \textit{Expectation is a linear operator} \\ &= \psi_1 \, 0 + 0 \quad \textit{Since } \epsilon_t \sim \mathcal{N}(0, \sigma^2) \; \forall \, t \textit{ (i.e. expectation of the errors is 0)} \\ &= 0\end{aligned}$$
- What is the variance of y_t?
  Ans.
  $$\begin{aligned}\mathbb{V}ar[y_t] &= \mathbb{E}[(y_t - \mathbb{E}[y_t])^2] \quad \textit{By def. of variance} \\ &= \mathbb{E}[(y_t)^2] \quad \textit{since } \mathbb{E}[y_t] = 0 \\ &= \mathbb{E}[(\psi_1 \, \epsilon_{t-1} + \epsilon_t)^2] \quad \textit{By def. of our MA(1)} \\ &= \mathbb{E}[\psi_1^2 \, \epsilon_{t-1}^2 + \epsilon_t^2 + 2\psi_1 \, \epsilon_{t-1}\epsilon_t] \\ &= \psi_1^2 \, \mathbb{E}[\epsilon_{t-1}^2] + \mathbb{E}[\epsilon_t^2] + 2\psi_1 \, \mathbb{E}[\epsilon_{t-1}\epsilon_t] \\ &= \psi_1^2 \, \sigma^2 + \sigma^2 + 2\psi_1 \, 0 \quad \textit{Using the hypothesis on the errors}\end{aligned}$$

  Remember that all errors ϵ follow a Normal distribution with mean 0 ($\mathbb{E}[\epsilon_t] = 0, \; \forall \, t$), and variance σ^2. In addition, the errors are independent from each other i.e.:

  $$\mathbb{E}[\epsilon_{t_1}\epsilon_{t_2}] = 0 \quad \forall \, t_1 \neq t_2$$

- What is the covariance of y_t and y_{t-k}?
  Ans.

$$\begin{aligned}
\mathbb{C}ov[y_t, y_{t-k}] &= \mathbb{E}[(y_t - \mathbb{E}[y_t])(y_{t-k} - \mathbb{E}[y_{t-k}])] \quad \textit{By def. of covariance} \\
&= \mathbb{E}[(y_t)(y_{t-k})] \quad \textit{Because } \mathbb{E}[y_t] = 0 \quad \forall\, t \\
&= \mathbb{E}[(\psi_1\epsilon_{t-1} + \epsilon_t)(\psi_1\epsilon_{t-1-k} + \epsilon_{t-k})] \quad \textit{Because of our MA(1) model} \\
&= \mathbb{E}[\psi_1^2\epsilon_{t-1-k}\epsilon_{t-1} + \psi_1\epsilon_{t-1-k}\epsilon_t + \psi_1\epsilon_t\epsilon_{t-1-k}\epsilon_t + \epsilon_{t-1}\epsilon_{t-k})] \\
&= \psi_1^2 \underbrace{\mathbb{E}[\epsilon_{t-1-k}\epsilon_{t-1}]}_{=0,\forall k\geq 1} + \psi_1 \underbrace{\mathbb{E}[\epsilon_{t-k}\epsilon_{t-1}]}_{0,\forall k>1;\ \sigma^2 for k=1} + \psi_1 \underbrace{\mathbb{E}[\epsilon_t\epsilon_{t-1-k}]}_{0,\forall k} + \underbrace{\mathbb{E}[\epsilon_t\epsilon_{t-k})]}_{0,\forall k>0}
\end{aligned}$$

  so $\mathbb{C}ov[y_t, y_t] = (\psi_1^2 + 1)\sigma^2$, $\mathbb{C}ov[y_t, y_{t-1}] = \psi_1\sigma^2$ *and* $\mathbb{C}ov[y_t, y_{t-k}] = 0,\ \forall k > 1$.

- What is the correlation of y_t and y_{t-k}?

  Ans. *The correlation is the covariance divided by the variances:*

$$\mathbb{C}orr[y_t, y_{t-k}] = \frac{\mathbb{C}ov[y_t, y_{t-k}]}{\sqrt{\mathbb{V}ar[y_t]\mathbb{V}ar[y_{t-k}]}} = \begin{cases} 1 & \textit{if } k = 0 \\ \frac{\psi_1}{\psi_1^2+1} & \textit{if } k = 1 \\ 0 & \textit{otherwise } t > 1 \end{cases}$$

(3) Conclude about the form of the ACF function for a MA(1) models?

Ans. *The ACF plots the lags k on the x-axis, and the y-axis reports the correlation* $\mathbb{C}orr[y_t, y_{t-k}]$.

13.3 Least Squares algorithm for MA models ?

Consider an MA(1) (with $\psi_0 = 0$ for simplication)

$$y_t = \psi_1\, \epsilon_{t-1} + \epsilon_t$$

we need to write this with lagged series of y_t (for which we have the observations $y_1, \cdots, y_n$). The model can be rewritten:

$$\begin{aligned}
y_t &= \psi_1\, y_{t-1} - \psi_1^2\, \epsilon_{t-2} + \epsilon_t \\
&= \psi_1\, y_{t-1} - \psi_1^2\, y_{t-2} + \psi_1^3\, \epsilon_{t-3} + \epsilon_t \\
&\vdots \\
y_t &= \psi_1\, y_{t-1} - \psi_1^2\, y_{t-2} + \cdots + (-1)^t \psi_1^{t-1}\, y_1 + \psi_1^t\, \epsilon_0 + \epsilon_t
\end{aligned}$$

Assuming $\epsilon_0 = 0$, y_t is a weighted average of all the past observations, and the expression is not linear w.r.t. the parameter to estimate ψ_1 (powers of ψ_1 appear in the equation). Hence the Least square algorithm used for estimation with AR models cannot be used.

Chapter 14

The backshift operator

14.1 Definition

1.1 Definition (Backshift operator) In what follows, it will be very useful to denote a lagged series by using the *backshift operator B*:

$$By_t = y_{t-1},$$

For lags of length k, we apply B k times:

$$y_{t-2} = By_{t-1} = B(By_t) = B^2 y_t; \text{ in general } B^k y_t = y_{t-k}.$$

We can use B to express differencing:

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1-B)y_t.$$

The great power of the backshift operator is that it is *multiplicative*:

$$(1-B)(1-B^s)y_t = (1-B-B^s+B^{s+1})y_t = y_t - y_{t-1} - y_{t-2} + y_{t-s-1},$$

14.2 Exercises.

(1) Write an MA(1) model with the backshif operator

(2) Write an AR(1) model with the backshif operator

(3) Write an MA(q) model with the backshif operator

(4) Write an AR(p) model with the backshif operator

(5) Write an ARMA(p,q) model with the backshif operator

Chapter 15

AIC and BIC

In the lab, we have tried to find the best ARMA models by using the ACF and PACF graphs to identify the AR(p) and MA(q) components. Several ARMA(p,q) were then tested until all ACF and PACF coefficients becomes negligeable, and also when the time plot of the residuals looks like noise.

One way to allow you to choose any ARMA model is simply to consider a lot of different ARMA models, fit them, and choose the one that has the smallest mean square error (as we have done before when picking the best parameter value in exponential smoothing, etc.) There is a problem with this though; we can always make the MSE smaller by adding another MA or AR term! So if we did this, then we would just keep finding more and more complicated models that fit better and better!

Clearly, what we want is a compromise between a model that fits well but does not have too many parameters. There is no one way to do this, but one technique is to use *information criterion*.

15.1 Information Criterion

1.1 Definition (AIC and BIC criteria) We define two types of information criterion: the *Bayesian Information Criterion* (BIC) and the *Akaike Information Criterion* (AIC). In AIC and BIC, we choose the model that has the minimum value of:

$$AIC = -2\log(L) + 2m,$$

$$BIC = -2\log(L) + m\log n,$$

where

- L is the likelihood of the data with a certain model,
- n is the number of observations and
- m is the number of parameters in the model. The number of parameters is $m = p + q$ for an ARMA(p,q) model.

We see that as the model gets more complicated, the model will fit better, and $-2\log(L)$ gets smaller, but m gets bigger. The best model will be the one that achieves a compromise between the two. Finally, often the likelihood is difficult to calculate, but there is a useful approximation:

$$-2\log(L) \approx n(1+\log(2\pi)) + n\log(s^2),$$

where n is the number of observations in the series and s^2 is the estimated variance of the residuals after fitting the model. Therefore we find the model where

$$\text{AIC} \approx n(1+\log(2\pi)) + n\log(s^2) + 2m$$

is the smallest. Again, a forecasting computer package will allow you to hunt for the ARMA model with the smallest AIC or BIC.

15.2 R output

Tables 15.1 and 15.1 show the R outputs when fitting a AR(1) and AR(2) to the dowjones data.

```
Series: dowjones
ARIMA(1,0,0) with non-zero mean

Call: arima(x = dowjones, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.9964   116.0329
s.e.  0.0045     4.8878

sigma^2 estimated as 0.1974:  log likelihood = -49.86
AIC = 105.72   AICc = 106.04   BIC = 112.79
```

Table 15.1: Output in R of `arima(dowjones,order=c(1,0,0))`.

```
Series: dowjones
ARIMA(2,0,0) with non-zero mean

Call: arima(x = dowjones, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      1.4990  -0.5049   115.7854
s.e.  0.0993   0.1000     4.1654

sigma^2 estimated as 0.1483:  log likelihood = -38.96
AIC = 85.91   AICc = 86.46   BIC = 95.34
```

Table 15.2: Output in R of `arima(dowjones,order=c(2,0,0))`.

Understanding the R outputs:

(1) what are the coefficients ar1 (and ar2) ? what is the intercept?

(2) Write down the mathematical equation of the models fitted in both cases.

(3) What is sigma?

(4) What is log likelihood?

Note that the AIC and BIC are given.

Chapter 16

ARIMA(p,d,q)

An ARMA model is not suitable for fitting a times with a trend: Remember a time series showing a trend is not stationary in mean and ARMA models are only suited for time series stationary in mean and variance. Differencing is an operation that can be applied to a time series to remove a trend. If after differencing the time series looks stationary in mean and variance, then an ARMA(p,q) model can be used. Section 16.1 presents differencing (of order d) and section 16.2 extends the ARMA(p,q) models to the ARIMA(p,d,q) models.

16.1 Differencing a time series

Consider a time series y_t, the first order differencing is defined as

$$y'_t = y_t - y_{t-1}$$

We can use B to express differencing:

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1-B)y_t.$$

Exercise. Express the second order difference $y''_t = y'_t - y'_{t-1}$ in terms of the backshift operator B. Conclude on the differencing of order d.

Visualisation. Figures 16.1, 16.2 and 16.3 shows the dowjones time series before and after differencing with $d = 1$ and $d = 2$.

16.2 Integrating differencing into ARMA models

2.1 Definition (Autoregressive integrated moving average (ARIMA(p,d,q))) Trends in time series can be removed by differencing the time series. This differencing is integrated into the ARMA models creating the ARIMA models. ARIMA(p,d,q) define models with an AutoRegressive part of order p, a Moving average part of order q and having applied d order differencing:

$$\underbrace{(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)}_{AR(p)}\ \underbrace{(1-B)^d}_{I(d)}\ y_t = c + \underbrace{(1-\psi_1 B-\psi_2 B^2-\cdots-\psi_q B^q)}_{MA(q)}\ \epsilon_t$$

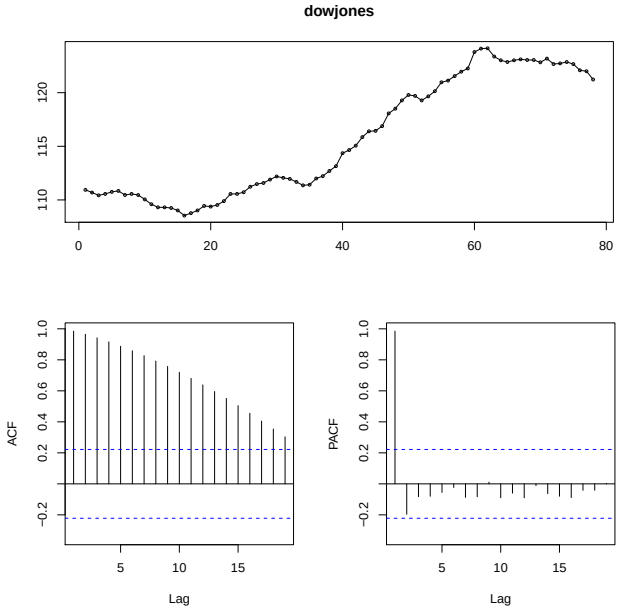


Figure 16.1: Dowjones time series. `>tsdisplay(dowjones)`

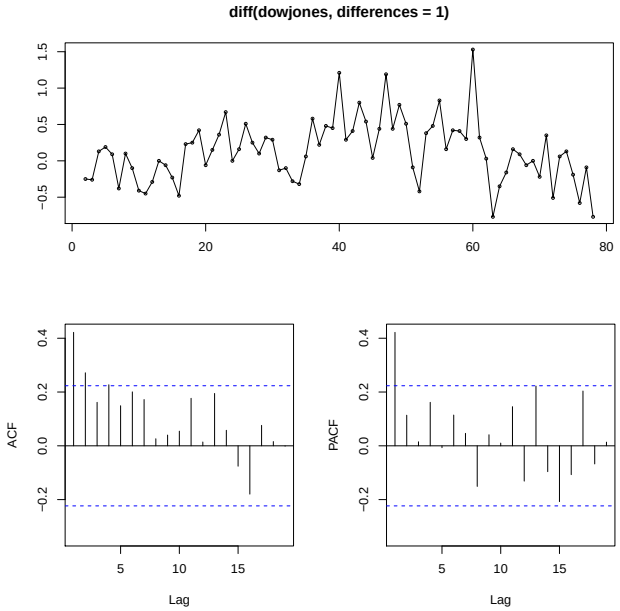


Figure 16.2: Differencing of the Dowjones time series $d = 1$. `>tsdisplay(diff(dowjones,differences=1))`

Example: Random Walk The Random walk is an ARIMA(0,1,0):

$$(1-B)^1 y_t = \epsilon_t$$

or

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

Exercises

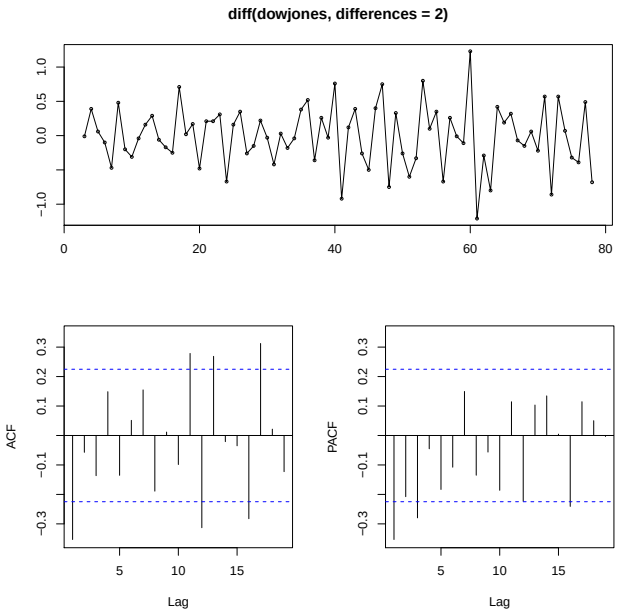


Figure 16.3: Differencing of the Dowjones time series $d = 2$. `>tsdisplay(diff(dowjones,differences=2))`

(1) In the table overpage are some data to which we wish to fit the ARIMA(1,1,1) model

$$(1 - 0.4\ B)(1 - B)\ y_t = 0.1 + (1 - 0.9\ B)\ \epsilon_t.$$

If we let $x_t = (1 - B) y_t$ be the differenced series, we can fit the simpler ARMA(1,1) model:

$$(1 - 0.4\ B)\ x_t = 0.1 + (1 - 0.9\ B)\ \epsilon_t.$$

We can re-write this as $x_t = 0.1 + 0.4\ x_{t-1} - 0.9\ \epsilon_{t-1} + \epsilon_t$, and so create fitted values $\hat{x}_t$. With these fitted values we can back transform to get fitted values $\hat{y}_t = \hat{x}_t + y_{t-1}$. Use these facts to fill in the table.

(2) Here is the ARIMA(1,1,1) model:

$$(1 - \phi_1\ B)\ (1 - B)\ y_t = c + (1 - \psi_1\ B)\epsilon_t.$$

Expand this equation and apply the backshift operators to get an equation for y_t in terms of y_{t-1}, y_{t-2}, ϵ_t and ϵ_{t-1}.

16.3 Which $ARIMA(p,d,q)$ model do I use?

There are a very large number of ARIMA models. Which one is appropriate for a given data set? There are some things to bear in mind. First, values of p, q or d of more than 3 are very rarely needed. Second, it is often the case that many different ARIMA models give more or less the same predictions, so there is some flexibility in the choice of p, d and q. The following approach can be followed:

(1) Plot the data.

(2) Look to see if the data is stationary, that is they are scattered randomly about a constant mean level. Also look at the ACF and PACF (stationarity is implied by the ACF or PACF dropping quickly to zero).

Time t	Data y_t	Differenced data $x_t = y_t - y_{t-1}$	Fitted values $\hat{x}_t = 0.1 + 0.4x_{t-1} - 0.9\,\epsilon_{t-1}$	Error $\epsilon_t = x_t - \hat{x}_t$	Fitted values $\hat{y}_t = \hat{x}_t + y_{t-1}$
1	9.5	–	–	–	–
2	13.7	4.2	0	4.2	
3	8.7	-5.0			
4	16.1	7.4			
5	15.3	-0.8			
6	12.2	-3.1			

Table 16.1: ARIMA(1,1,1) $(1-0.4B)(1-B)y_t = 0.1 + (1-0.9B)\epsilon_t$.

(3) If there is non-stationarity, such as a trend (we're ignoring seasonal behaviour for the moment!), difference the data. Practically, at most two differences need to be taken to reduce a series to stationarity. Verify stationarity by plotting the differenced series and looking at the ACF and PACF.

(4) Once stationarity is obtained, look at the ACF and PACF to see if there is any remaining pattern. Check against the theoretical behaviour of the MA and AR models to see if they fit. This will give you an ARIMA model with either no MA or no AR component i.e. ARIMA(0,d,q) or ARIMA(p,d,0).

(5) If there is no clear MA or AR model, an ARMA model will have to be considered. These can in general not be guessed from the ACF and PACF, other methods are needed, based on the ideas of minimising Information Criterion (AIC or BIC).

Chapter 17

Seasonal $ARIMA(p,d,q)(P,D,Q)_s$

Time series having a trend and/or a seasonal pattern are not stationary in mean. We extend ARMA(p,q) models in section 16.2 to allow removing a trend before fitting an ARMA model. Section 17.1 extends further these new models to allow seasonal pattern to be modelled.

17.1 Seasonal $ARIMA(p,d,q)(P,D,Q)_s$

As things stand, ARIMA models cannot really cope with seasonal behaviour; we see that, compared to ARMA models, ARIMA(p,d,q) only models time series with trends. We will incorporate now seasonal behaviour and present a general definition of the Seasonal ARIMA models.

1.1 Definition (Seasonal Autoregressive integrated moving average : $ARIMA(p,d,q)(P,D,Q)_s$**)** Seasonal ARIMA models are defined by 7 parameters $ARIMA(p,d,q)(P,D,Q)_s$

$$\underbrace{(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)}_{AR(p)}\underbrace{(1-\beta_1 B^s-\beta_2 B^{2s}-\cdots-\beta_P B^{Ps})}_{AR_s(P)}\underbrace{(1-B)^d}_{I(d)}\underbrace{(1-B^s)^D}_{I_s(D)} y_t = c+\underbrace{(1-\psi_1 B-\psi_2 B^2-\cdots-\psi_q B^q)}_{MA(q)}\underbrace{(1-\theta_1 B^s-\theta_2 B^{2s}-\cdots-\theta_Q B^{Qs})}_{MA_s(Q)}\epsilon_t \quad (17.1)$$

where

- $AR(p)$ Autoregressive part of order p
- $MA(q)$ Moving average part of order q
- $I(d)$ differencing of order d
- $AR_s(P)$ Seasonal Autoregressive part of order P
- $MA_s(Q)$ Seasonal Moving average part of order Q
- $I_s(D)$ seasonal differencing of order D
- s is the period of the seasonal pattern appearing i.e. $s = 12$ months in the Australian beer production data.

The idea behind the seasonal ARIMA is to look at what are the best explanatory variables to model a seasonal pattern. For instance lets consider the Australian beer production that shows a seasonal pattern of

period 12 months. Then to predict the production at time t, y_t, the explanatory variables to consider are:

$$y_{t-12}, y_{t-24}, \cdots, \text{and / or } \epsilon_{t-12}, \epsilon_{t-24}, \cdots$$

For seasonal data, it might also make sense to take differences between observations at the same point in the seasonal cycle i.e. for monthly data with annual cycle, define differences (D=1)

$$y_t - y_{t-12}.$$

or (D=2)

$$y_t - 2y_{t-12} + y_{t-24}.$$

17.2 Using ACF and PACF to identify seasonal ARIMAs

You can use ACF and PACF to identify P or Q:

- For $ARIMA(0,0,0)(P,0,0)_s$, you should see major peaks on the PACF at s, $2s$, ...Ps. On the ACF, the coefficients at lags s, $2s$, ...Ps, $(P+1)s$, ... should form an exponential decrease, or a damped sine wave. See examples figures 17.1 and 17.2.
- $ARIMA(0,0,0)(0,0,Q)_s$, you should see major peaks on the ACF at s, $2s$, ...Qs. On the PACF, the coefficients at lags s, $2s$, ...Qs, $(Q+1)s$... should form an exponential decrease, or a damped sine wave. See examples figures 17.3 and 17.4.

When trying to identify P or Q, you should ignore the ACP and PACF coefficients other than s, $2s$, ...Ps,.. or s, $2s$, ...Qs,.... In other word, look only at the coefficients computed for multiples of s.

17.3 How to select the best Seasonal ARIMA model?

It is sometimes not possible to identify the parameters p,d,q and P,D,Q using visualisation tools such as ACF and PACF. Using the BIC as the selection criterion, we select the ARIMA model with the lowest value of the BIC. Using the AIC as the selection criterion, we select the ARIMA model with the lowest value of the AIC.

17.4 Conclusion

We have now defined the full class of statistical models $ARIMA(p,d,q)(P,D,Q)_s$ studied in this course. ARMA(p,q) can only be applied to time series stationnary in mean, hence the extension to $ARIMA(p,d,q)(P,D,Q)_s$ (introducing d,D,P,Q,s) allowed us to make the time series stationary in mean. Unfortunatly, we still are not able to deal with time series that are not stationary in variance. We propose some possible solutions in the next chapter.

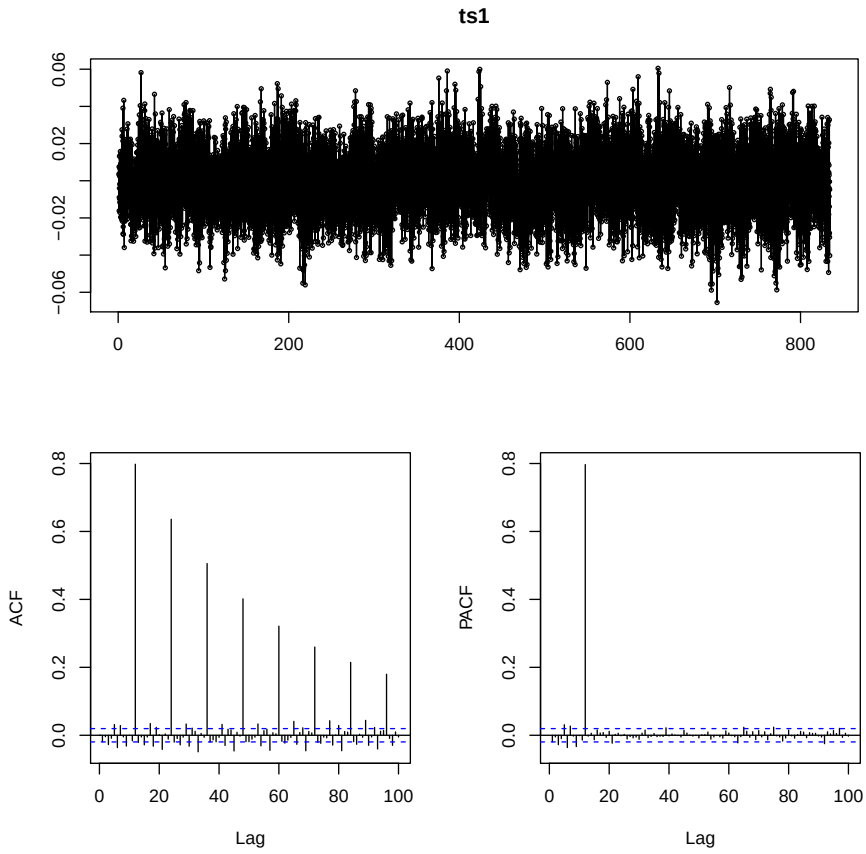


Figure 17.1: Simulation $ARIMA(0,0,0)(1,0,0)_{12}$

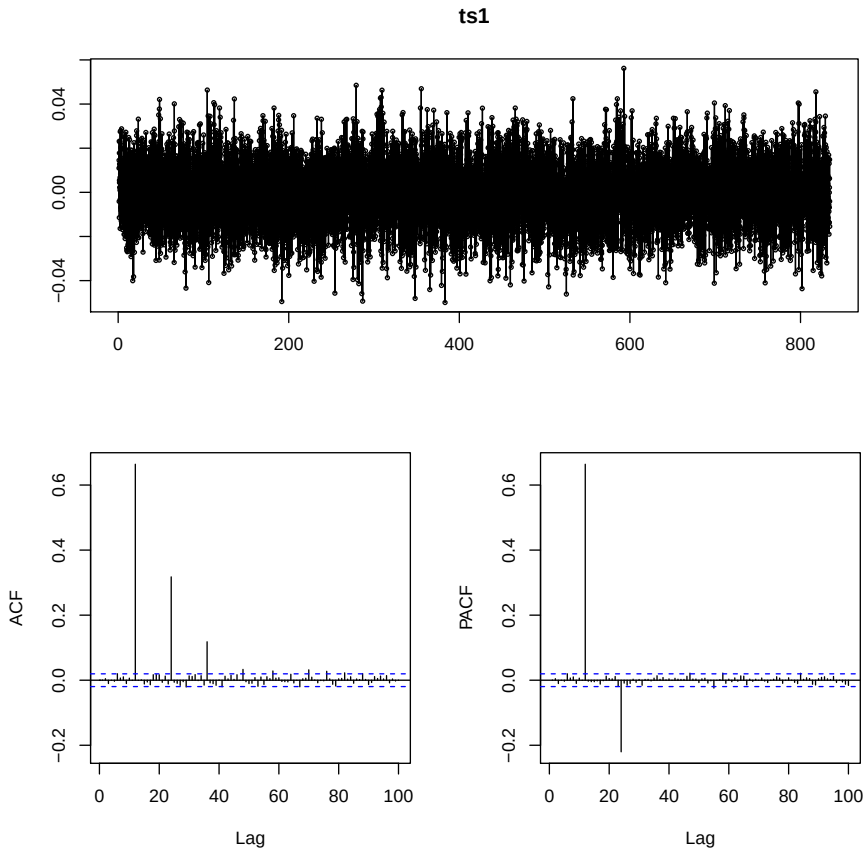


Figure 17.2: Simulation $ARIMA(0,0,0)(2,0,0)_{12}$

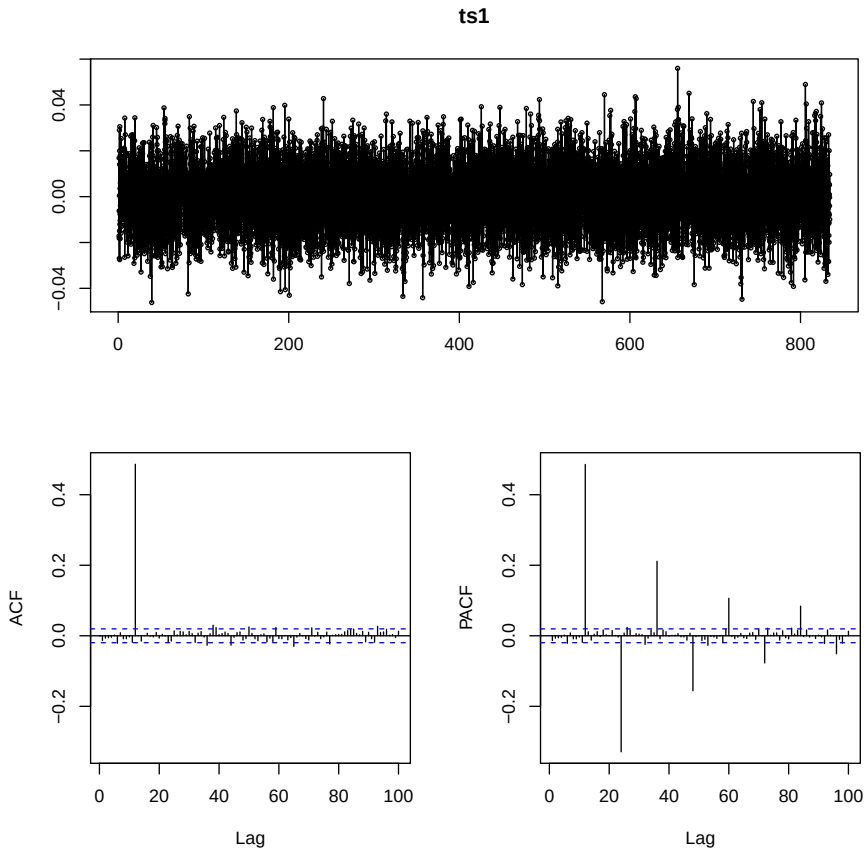


Figure 17.3: Simulation $ARIMA(0,0,0)(0,0,1)_{12}$

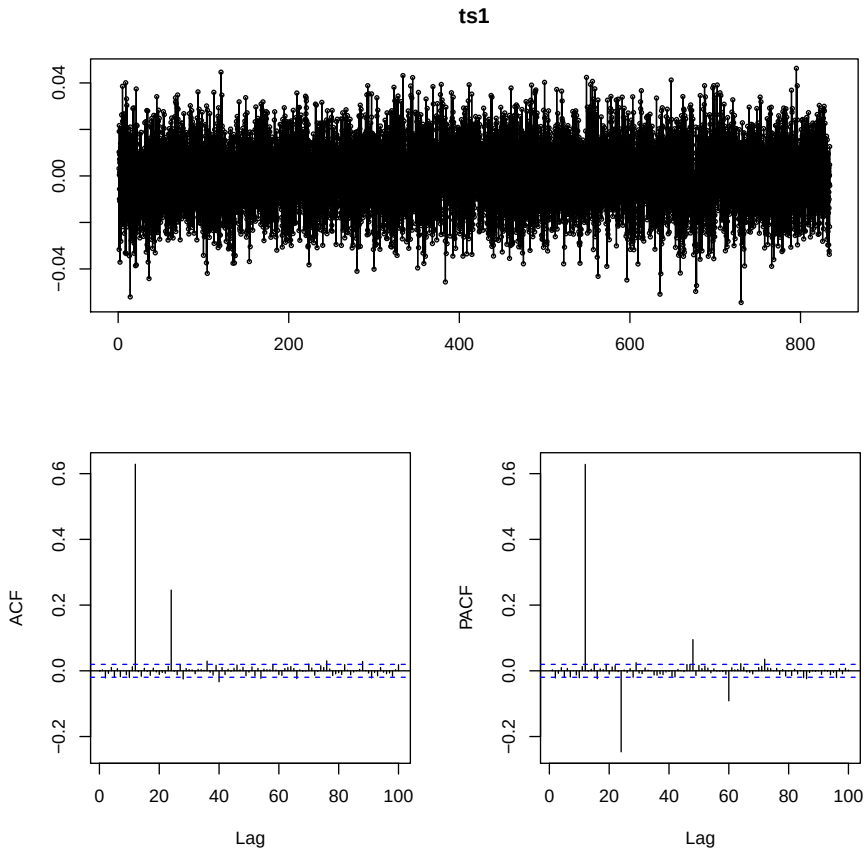


Figure 17.4: Simulation $ARIMA(0,0,0)(0,0,2)_{12}$

Chapter 18

Transforming a time series

The seasonal arima$(p,d,q)(P,D,Q)$ can only deal with time series that are stationary in variance. Before using these models, several transformation can make a time series stationary in variance.

The previous chapters have proposed to make a time series stationary in mean by first removing a trend by differentiation, and second by removing a seasonal pattern by considering AR and MA models combined with a seasonal differencing. In this section we focus on making the time series stationary in variance (when needed). Observe figure 18.1. This time series shows both a trend and a seasonal component therefore it is not stationary in mean. Note how the amplitude of the seasonal variation increase overtime from year to year: this is an indicator of a time series that may not be stationary in variance.

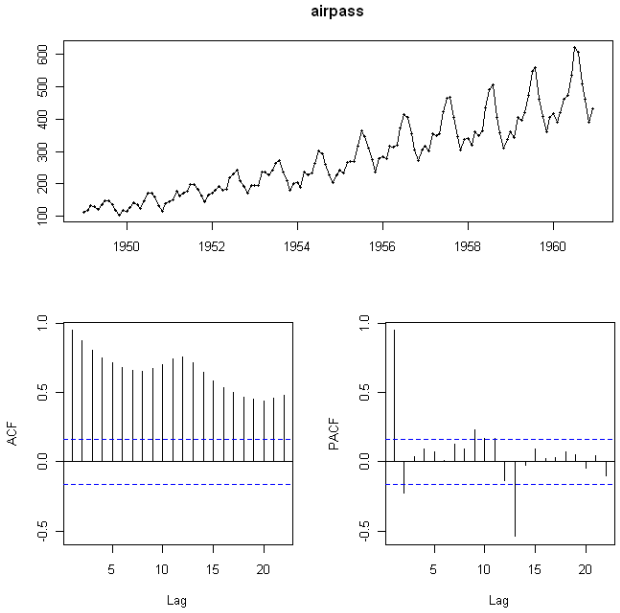


Figure 18.1: Monthly totals of international airline passengers (1949 to 1960) (time series `airpass`).

Mathematical functions can be applied to the time series to make them stationary in variance. Four such transformations are commonly used, and reduce variance by differing amounts. Which one to use depends on how much the variance is increasing with time.

Square root	$\sqrt{y_i}$	$\downarrow$
Cube root	$\sqrt[3]{y_i}$	Increasing
Logarithm	$\log(y_i)$	Strength
Negative Reciprocal	$-1/y_i$	$\downarrow$

Square root and logarithm are the most common.

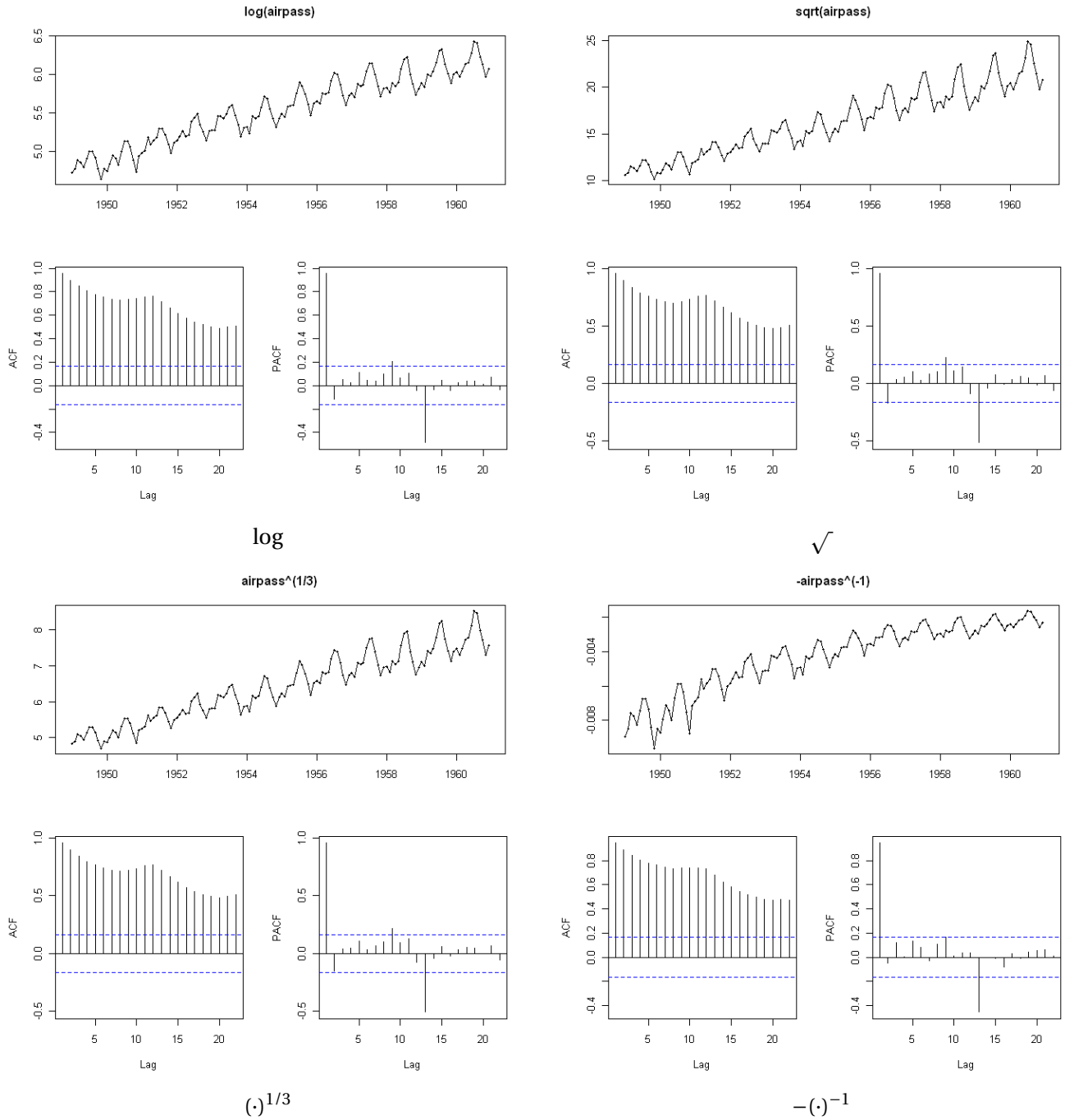


Figure 18.2: Transformations of `airpass`.

EXERCISE: look at the four transformations as applied to the `airpass` time series (figure 18.2). Which transformation is best at stabilising the variance?

Final remarks: It may be difficult to assess non stationarity in variance from the time series itself so an efficient alternative is to fit an ARIMA model (e.g. using `auto.arima` R function) to the time series and its various transformations, and visualise in each case the residuals: these should be having a constant mean 0 and a fixed variance over time when ARIMA is a suited model.

Part IV

Conclusions

Chapter 19

Summary of the course

We have introduced the Holt-Winters algorithms and the ARIMA models as two classes of techniques to analyse time series. Some Holt-Winters algorithms have an equivalent ARIMA models (cf. table 19.1). Figure 19.1 provides a summary to the content of this course. Remember that all these methods rely on the hypothesis that somewhat what has happened in the past repeats itself in the future (continuity hypothesis).

Simple Exponential Smoothing	$\equiv$	ARIMA(0,1,1)
Holt's linear method	$\equiv$	ARIMA(0,2,2)
Holt-Winters' additive method	$\subset$	ARIMA(0,1,s+1)$(0,1,0)_s$
Holt-Winters' multiplicative method		no ARIMA equivalent

Table 19.1: Final remarks about Holt-Winters Algorithms and ARIMA models ([3] p. 373).

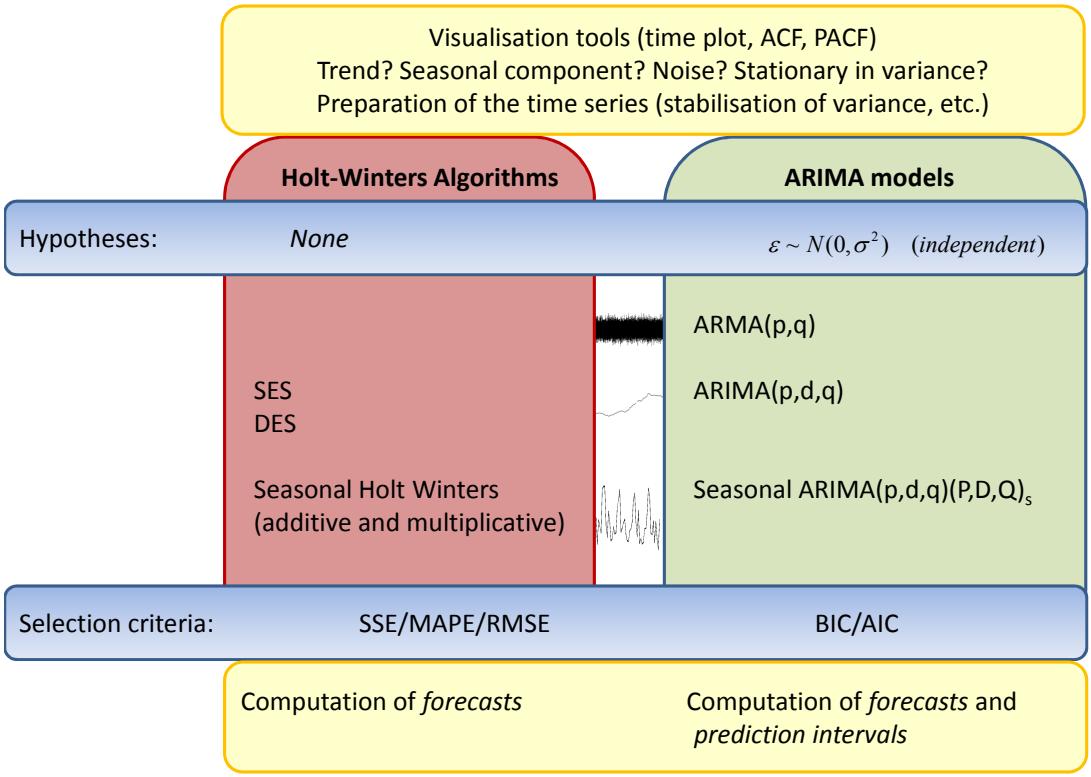


Figure 19.1: Course overview.

Chapter 20

Beyond Holt-Winters algorithms and ARIMA models

20.1 ARCH GARCH

In some parts of the course, we have come across series which are non-stationary in variance, and we have transformed the data in some way. These transformations rely on the non-stationarity in variance being highly systematic. For example, in order for the log transformation to work the variability in the data must be increasing as time goes on. It is easy to envisage situations where the variability gets bigger and smaller over time in a non-systematic way.

1.1 Definition (ARCH) *Autoregressive Conditional Heteroscedastic (ARCH) models* allow the variance term to have its own random walk, for instance extending AR(1):

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0, \sigma_t^2)$$

so that

$$\sigma_t^2 = \gamma_0 + \gamma_1\, \sigma_{t-1}^2$$

The net effect is to allow the variability in y_t to vary with time; each y_t has its own σ_t. The parameters γ_0 and γ_1 control the amount of variation in y_t at time t by controlling the size of σ_t.

1.2 Definition (GARCH) A further generalisation exists, known as *Generalised Autoregressive Conditional Heteroscedastic (GARCH) models*. In this version, the variance terms are given their own ARMA process (rather than the AR process in ARCH models), based on the squares of the error terms. A simple version would now have:

$$\epsilon_t \sim N(0, \sigma_t^2),\ \sigma_t^2 = \gamma_0 + \gamma_1 \sigma_{t-1}^2 + \beta_1 \epsilon_{t-1}^2.$$

where ϵ_t has the same meaning as in a standard MA process. Further generalisations exist which allow non-linear, non-stationary processes be applied to the variance terms.

20.2 Continuous time series modelling

Here we relax the assumption where y_t can only exist at discrete time points $t = 1, 2, \ldots, n$. There are many scenarios where we could observe y at irregular time points; we will now write $y(t)$ to describe the measurement of y at a continuous time t. Many simple (and, conversely, incredibly complicated) time series models can be written in this format.

20.2.1 Brownian motion

Perhaps the simplest of the continuous time models is that of Brownian motion, also known as a Weiner process. Here we define the differences between data values at different time points to be normally distributed with a very specific mean and autocovariance structure. We can write it as:

$$y(t) - y(t-s) \sim \mathcal{N}(0, s).$$

Thus the change in value of the series is dependent on the time distance between the values. An obvious extension is:

$$y(t) - y(t-s) \sim \mathcal{N}(\mu, s\sigma^2),$$

to allow for more complicated variation. In this context the μ parameter is known as a *drift* parameter as it controls the general direction of the data values as time proceeds.

20.2.2 Gaussian processes

We have already come across one very simple continuous time model in the shape of linear regression. If we let $y(t) = \alpha + \beta t + \epsilon_t$ (so $y_t \sim N(\alpha + \beta t, \sigma^2)$), we already have a technique for determining the value at all times t. This is an example of a simple Gaussian process (GP). In general we say that $y(t)$ follows a GP if it has the following properties:

(1) $y(t) \sim \mathcal{N}(\mu(t), \sigma^2)$,

(2) $Cov(y(t), y(t-k)) = f(k) \times \sigma^2$.

Note that this process is not stationary unless $\mu(t)$ is constant. The linear regression example above has $\mu(t) = \alpha + \beta t$ and $f(k) = 0$ if $k > 0$ and $f(0) = 1$. We can build more complicated GPs by allowing a more general structure for $f(k)$; the autocorrelation function. One choice for $f(k)$ is known as the Gaussian autocorrelation function:

$$f(k) = e^{-\frac{k^2}{h}},$$

where h is a bandwidth term which determines how far apart $y(t)$ and $y(t-k)$ have to be for them still to be correlated. In practice we often write down the the distribution for a GP using a multivariate normal distribution:

$$\begin{pmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_n) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(t_1) \\ \mu(t_2) \\ \vdots \\ \mu(t_n) \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & f(t_1 - t_2) & \dots & f(t_n - t_1) \\ f(t_2 - t_1) & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ f(t_1 - t_n) & \dots & \dots & 1 \end{pmatrix} \right).$$

We still only have a few parameters to estimate. We have those concerning $\mu(t)$ (2 if we use $\mu(t) = \alpha + \beta t$), we have h for the bandwidth, and we have σ^2. We could again use maximum likelihood to determine these parameters. Furthermore, if we write the model in matrix notation, we have:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{F}).$$

This allows us to use a very handy theoretical result from the normal distribution to predict data at time points we have not observed, $y(t^*)$:

$$y(t^*)|\boldsymbol{y} \sim \mathcal{N}\left(\mu(t^*) + \boldsymbol{f}(\boldsymbol{t} - t^*)^T \boldsymbol{F}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \sigma^2\left(1 - \boldsymbol{f}(\boldsymbol{t} - t^*)^T \boldsymbol{F}^{-1} \boldsymbol{f}(\boldsymbol{t} - t^*)\right)\right).$$

More generally, we can now create predictive distributions for time points other than those observed. Gaussian processes are widely used in practice by Bayesian statisticians as they can be used as prior distributions over random functions.

20.3 Fourier analysis for time series

A useful way of thinking about time series that have a seasonal component is to use the *frequency* domain rather than the time domain as we have been using. Consider a time series written as:

$$y_t = a\cos(\omega t) + b\sin(\omega t) + \epsilon_t.$$

We would now have a nice oscillating time series model with the same normally-distributed error, ϵ_t, as before. Indeed, it has been shown that you can write *any* function using this frequency domain approach by adding together lots of different sine and cosine parts (known as harmonics):

$$y_t = \sum_{j=1}^{k} a_j \cos(\omega_j t) + b_j \sin(\omega_j t) + \epsilon_t.$$

Now we have k harmonics and the series is written as a sum of terms (determining the number of harmonics we require is not always easy). We can adjust the seasonality of the different harmonics by changing the ω_j terms. For example, if we had daily data with a yearly seasonal trend we could set the first harmonic $\omega_1 = 2\pi/365$. We can make the model more complicated by letting the data determine the ωs via, for example, maximum likelihood.

20.4 Others techniques used for time series analysis and forecasting

- Functional Data Analysis
- Neural Networks for time series
- Kalman Filtering

Bibliography

[1] C. Chatfield. *Time series Forecasting*. Chapman& Hall/CRC, 2001.

[2] Paul Goodwin. The holt-winters approach to exponential smoothing: 50 years old and going strong. *Foresight: The International Journal of Applied Forecasting*, (19):30–33, 2010.

[3] S. Makidakis, S.C. Wheelwright, and R.J. Hyndman. *Forecasting; Methods and Applications*. Wiley, 1998.