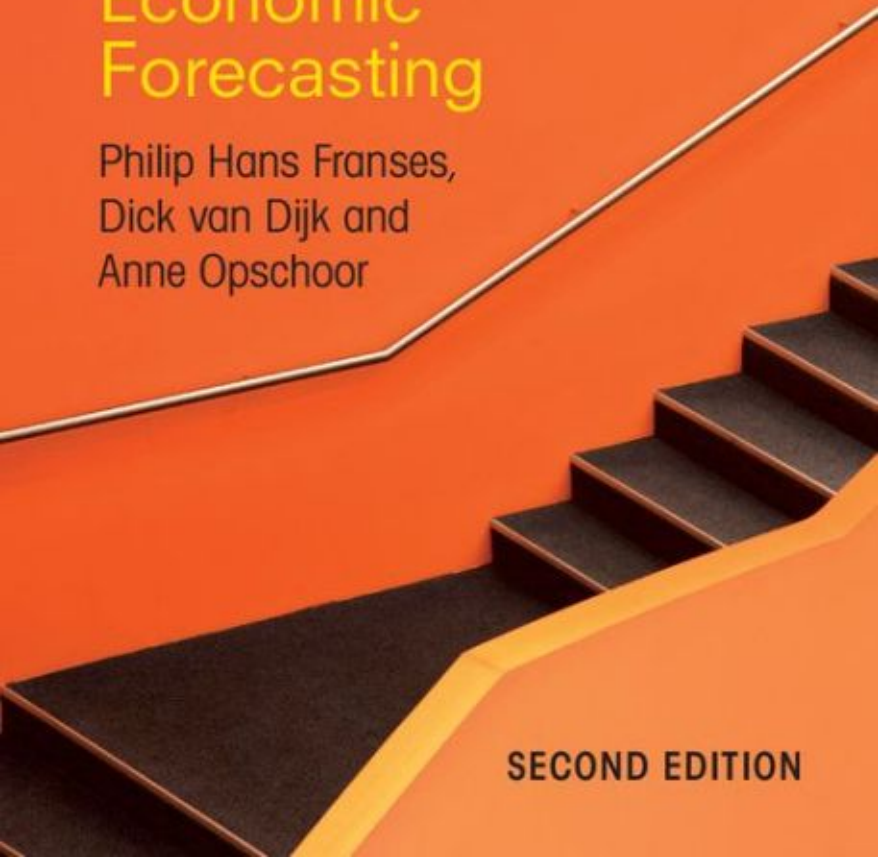
Time Series Models for Business and Economic Forecasting

Philip Hans Franses,
Dick van Dijk and
Anne Opschoor



SECOND EDITION

Time Series Models for Business and Economic Forecasting

With a new author team contributing decades of practical experience, this fully updated and thoroughly classroom-tested second edition textbook prepares students and practitioners to create effective forecasting models and master the techniques of time series analysis. Taking a practical and example-driven approach, this textbook summarises the most critical decisions, techniques and steps involved in creating forecasting models for business and economics. Students are led through the process with an entirely new set of carefully developed theoretical and practical exercises. Chapters examine the key features of economic time series, univariate time series analysis, trends, seasonality, aberrant observations, conditional heteroskedasticity and ARCH models, non-linearity and multivariate time series, making this a complete practical guide. A companion website with downloadable datasets, exercises and lecture slides rounds out the full learning package.

Philip Hans Franses is Professor of Applied Econometrics and Professor of Marketing Research at the Erasmus School of Economics.

Dick van Dijk is Professor of Financial Econometrics at the Erasmus School of Economics.

Anne Opschoor is completing a PhD at the Erasmus School of Economics and is an Assistant Professor at the Free University.

Time Series Models for Business and Economic Forecasting

SECOND EDITION

Philip Hans Franses, Dick van Dijk
and Anne Opschoor

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Published in the United States of America by Cambridge University Press, New York

Cambridge University Press is a part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521520911

First published 1998
Second edition published 2014

Printed in the United Kingdom by MPG Printgroup Ltd, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data

ISBN 978-0-521-81770-7 Hardback
ISBN 978-0-521-52091-1 Paperback

Contents

Figures

Tables

Preface

The econometric analysis of economic and business time series is a major field of research and application. The last few decades have witnessed an increasing interest in both theoretical and empirical developments in constructing time series models and in their important application in forecasting. This book aims at reviewing several important developments within the context of forecasting business and economic time series.

A full-blown textbook on all aspects of time series analysis will cover thousands of pages. For example, the field of unit root analysis has expanded in the last three decades with such a pace and variation that a book only on this topic would take more pages than the current book does. This book is therefore not intended to be a survey of all that is available and that can be done in time series analysis. Obviously, such a selection comes with a cost, that is, the discussion will sometimes not be as theoretically precise as some readers would have liked. Merely, it is our purpose that the readers should be able to generate their own forecasts from time series models that adequately describe the key features of the data, to evaluate these forecasts and to come up with suggestions for possible modifications if necessary. In some interesting cases, though, we also recommend further reading. To attain this, we make a selection between all the possible routes to constructing and evaluating time series models, between all the possible estimation methods, and between all the various tests that can be used. Basically, our choice is also motivated by the availability of methods in such statistical packages as Eviews, while sometimes a little bit of Gauss, R or Matlab programming is needed. In fact, all empirical results in this book are thus obtained. An additional motivation for our choice is given by our own practical experience in forecasting business and economic time series. This experience is also based on supervising projects of our econometrics undergraduate students during their internships at banks, investment companies, and consultancy agencies.

The second purpose with this book is that the reader will be able to get some understanding of novel approaches reported in recent and future issues of, say, the Journal of Time Series Analysis, Journal of Econometrics, Journal of Business and Economic Statistics, Journal of Forecasting, International Journal of Forecasting, Journal of Applied Econometrics and the Journal of the American Statistical Association.

It is hoped that the reader finds the material in this book helpful to understand why such new methods can be useful for forecasting.

Although this book amounts to an introduction to the field of time series analysis and forecasting, it is necessary that the reader has knowledge of introductory econometrics. Specifically, regression analysis, matrix algebra and various concepts in estimation should be included in that knowledge. This book should then be useful to advanced undergraduate students and graduate students in business and economics, but also to practitioners and applied economists who wish to obtain a first, but not too technical, impression of time series forecasting. In fact, most of the material has already been used in "Time Series Analysis" courses for third year undergraduate students at the Econometric Institute in Rotterdam ever since 1996.

The first edition of this book Franses (1998) contained material that has now been deleted. Periodic models for seasonal data are not included anymore and also an extensive discussion of common features has been deleted. On the other hand, more details on ARCH model and on non-linear models have been included. More importantly, this fully revised second edition contains sections with exercises and answers. These exercises match with those presented at past exams to our students.

This book was written during our affiliation with the Econometric Institute at the Erasmus University Rotterdam. This Institute is a very stimulating teaching and research environment. We wish to express our gratitude to our (then) colleagues Teun Kloek, Christiaan Heij, Herman van Dijk, Dennis Fok, Richard Paap, Andre Lucas, Marius Ooms and anonymous reviewers for their kind willingness to comment on some or all chapters.

Rotterdam, July 2013

Philip Hans Franses
Dick van Dijk
Anne Opschoor

1 Introduction and overview

This book concerns the construction of time series models for describing the dynamic properties of economic variables and for out-of-sample forecasting. The economic variables can originate from various subject areas in economics and business, including macro-economics, finance, and marketing. Specific examples of time series of interest are inflation rates, unemployment rates, stock market returns, and market shares. Out-of-sample forecasts for such variables are often needed to set policy targets. For example, the forecast for next year's inflation rate can lead to a change in the monetary policy of a central bank. A forecast of a company's market share in the next few months may lead to changes in the allocation of advertising budgets. The models in this book can be called econometric time series models because we use econometric methods for analysis.

Time series variables can display a wide variety of patterns. Typically, macroeconomic aggregates such as industrial production, consumption, and wages show an upward trending pattern. Industrial production, tourism expenditures, and retail sales, among many others, display a pronounced seasonal pattern, that is, tourism spending is usually largest during the summer and retail sales tend to peak around Christmas. Another feature is that certain observations on economic data look aberrant in the sense that they occur rarely and deviate strongly from the typical behavior of the variable. For example, if new car registrations are almost zero in a certain month because of a computer breakdown, this does not reflect the true sales of new cars. Similarly, stock markets can crash with daily returns as large as minus 20 percent. Another characteristic property of financial asset prices is that periods of large price movements alternate with relatively calm periods, suggesting that the volatility of these variables changes over time. Finally, many economic time series display asymmetric or non-linear behavior. Unemployment, for example, increases rapidly during recessions but declines only slowly during expansions.

It seems obvious that there is not a single time series model that, first, can describe all of the above features simultaneously and, second, is also reasonably accurate in out-of-sample forecasting. In fact, several models are available to describe each of these

features, and all these models can be used to generate forecasts. It is the key purpose of this book to survey the most relevant of these models. We discuss how they can be implemented in practice and how their possible merits for forecasting can be evaluated. It is our opinion that just like there is no uniformly best descriptive model there is no such forecasting model. Therefore, we restrict ourselves to presenting guidelines for the selection between available models for forecasting. As will become apparent from the empirical examples in Chapter 2, the specific features of economic time series often lead to an *a priori* selection of several possibly useful models. For example, certain time series models explicitly deal with seasonality, and such models are not useful for data without seasonal variation.

An important requirement for the model construction methods discussed in this book is that the practitioner has some time available to construct his or her forecasts. Indeed, if it is necessary to generate forecasts for several hundreds or even thousands of variables every day, it may be better to rely on one of the many automatic extrapolation schemes that are available, such as smoothing algorithms and exponentially weighted moving averages. We do not wish to claim that such methods are inferior or less useful, as in fact these methods often do very well, see Makridakis *et al.* (1982), Makridakis and Hibon (2000) and Koning *et al.* (2005), among others. We do claim that the decisions involved in constructing a descriptive time series model for a time series with specific features that is also useful for out-of-sample forecasting is difficult to formalize in automatic routines.

Model building

Time series variables in economics and business are observed at different frequencies. For example, estimates of gross domestic product (GDP) and other measures of economic output are available per quarter, inflation typically is measured at a monthly frequency, product sales may be given per week, while stock prices can be recorded daily. The empirical time series used in this book as running examples reflect these different possible frequencies. It turns out that to some extent the sampling frequency determines the importance of the features discussed above. For example, if quarterly observations on consumption are used, seasonal variation is a very important characteristic to be accounted for in a time series forecasting model, while this would not be the case if annual consumption were considered. Similarly, stock returns display clear signs of time-varying volatility at daily and weekly frequencies, but much less so at the monthly frequency.

The modeling strategy described in this book exploits the key property of economic time series that the sequence of the observations is determined by calendar time. For example, the observation on unemployment in 2009 always precedes observations in 2010 and later. This seems too obvious to mention, but we believe it is not. Indeed, the

value of unemployment in 2010 is likely to be influenced by that in 2009. Hence when analyzing time series data we should leave their "natural ordering" intact.

Throughout this book the value of a time series y at time t is denoted by y_t, where t takes integer values from 1 to T. Note that we assume that the time series observations are regularly spaced, that is, the time period between consecutive observations y_{t-1} and y_t is the same for all t. Models that allow for irregularly spaced data are of course available, but these are not dealt with here, see Parzen (1984) for useful reading. The key property of a time series is that observation y_t always comes after y_{t-1}. Therefore it makes sense to take y_{t-1} into account when analyzing y_t. This is in contrast to cross-sectional data, where the ordering of the data points does not matter.

Given that y_{t-1} is always measured prior to y_t, it is likely that part of the value of y_{t-1} is reflected in the value y_t. For example, it is unlikely that if this month's inflation rate is 10 percent, it will be –5 percent next month. In fact, it is more likely that it will be, say, between 8 and 12 percent. Another of putting this is that, for many time series variables, the observations y_{t-1} and y_t are correlated. Since these observations are measurements of the same variable, we say that y_t is correlated with itself. This is called autocorrelation. If there is such autocorrelation between y_t and y_{t-1}, we can exploit this correlation for forecasting. For example, if it holds that y_t on average equals $0.8y_{t-1}$ for all $t = 1, \ldots, T$, and y_t is again the inflation rate with a value of 2 percent in the Tth month, we may forecast next month's rate y_{T+1} as 1.6 per cent.

The above implies that a time series variable can be characterized by its autocorrelations. Given a sample of observations y_t for $t = 1, \ldots, n$, we can estimate these correlations simply by computing their sample counterparts. The key feature of time series analysis is that such empirical autocorrelations can be exploited to obtain a first impression of which model is possibly useful to describe and forecast the time series at hand. This follows from the fact that all time series models theoretically imply certain autocorrelation properties of the time series in case these models were the true data generating processes. For example, the so-called autoregressive model of order 1 for a non-trending time series y_t implies that the correlations between y_t and y_{t-k} decline exponentially towards zero as k increases, see Chapter 3 for details. When the estimated empirical autocorrelations suggest that this pattern holds, we may be inclined to consider such a first order autoregressive model in a first round of analysis. In brief, certain features of observed time series data suggest the possible adequacy of corresponding time series models. This resembles what medical doctors do. They know that the flu can come with fever, so if a patient is diagnosed to be feverish, it might be due to the flu.

In this book we focus on five key features of economic and business time series variables. These features are trends, seasonality, aberrant observations, conditional heteroskedasticity and non-linearity. Each of these features points towards the possible usefulness of certain classes of time series models. In order to keep matters simple, we

restrict attention to regression-based models like

$$y_t = \beta' x_t + \varepsilon_t, \qquad t = 1, 2, \ldots, T, \tag{1.1}$$

where y_t is the observed time series of interest, x_t is a $k \times 1$ vector of other observed time series, ε_t is an unobserved error, β is a $k \times 1$ vector of unknown parameters, and T is the sample size. Commonly, the x_t variables contain the past of y_t. Needless to say that (1.1) is an overly simplified version of the models considered later on, but here it serves as a useful illustration. When there are T observations available, and the current time point is $t = T$, we often wish to forecast h periods ahead or "out-of-sample", that is, we want to estimate y_{T+h}. Usually, the y_t variable, which is the variable of focal interest, is known. However, the x_t variables usually have to be selected from among a potentially large number of candidates. It will become clear in later Chapters that autocorrelations can be helpful to decide on the most appropriate components x_t. In this book, these x_t variables (or functions thereof) are usually assumed to be observable. There are also classes of time series models where x_t is unobserved or latent, and should be estimated as well. In these so-called unobserved components models such x_t variables can be labeled as "trend" or "seasonal fluctuations", see Harvey (1989) and Durbin and Koopman (2001) for excellent treatments of these models. Harvey (2006) provides a survey on forecasting with unobserved components models. Furthermore, also in order to limit the exposition, we confine ourselves to situations where the functional form of the relationship between y_t and x_t is known and can be characterized by a few parameters. In (1.1) this relationship is linear. In Chapters 7 and 8, we will discuss some non-linear time series models. For a detailed treatment of non-parametric methods, which allow for more flexible functional relationships between y_t and x_t, the interested reader is referred to Härdle *et al.* (1997), Fan and Gijbels (1970), and Pagan and Ullah (1999), among others.

Statistical method

There are two different approaches in analyzing time series. The first of these is called the frequency domain approach, which makes extensive use of spectral analysis. The key assumption within this approach is that any time series can be decomposed into a certain number of cyclical components with different frequencies. In fact, the lengths (and amplitudes) of these cycles can be exploited to characterize a time series. For example, a cycle of infinite length corresponds with a trend, see Granger (1966). Similarly, for a quarterly observed time series a cycle of four quarters corresponds with a seasonal pattern. Although spectral techniques can be useful to obtain an impression of the salient features of a time series and to describe, for example, business cycle

properties, these are not often explicitly considered for out-of-sample forecasting. For a thorough treatment of frequency domain techniques, see, for example, Priestley (1981).

The second approach to analyzing time series, which is much more common in economics and business is called the time domain approach. Within this time domain approach, the autocorrelation function plays a central role, see Box and Jenkins (1970). Although the area of time series analysis has expanded widely since 1970, the main ideas underlying Box and Jenkins' work remain valid. In the present book we confine the discussion to this time domain approach, also since it is easier to use for data with features such as outliers, non-linearity and seasonality. In fact, in those latter cases, the application and interpretation of spectral techniques is not at all trivial.

A final remark on the statistical method concerns Bayesian and classical statistical methods. Without taking a standpoint toward favoring either of these two approaches, in this book we will use only classical statistical methods. For treatments of Bayesian methods in statistics and econometrics, see Zellner (1970), Bernardo and Smith (1994), Poirier (1995), Geweke (2005), and Koop (2003), among others. Geweke and Whiteman (2006) provide a detailed discussion and overview of Bayesian forecasting.

The data

An obvious first step when forecasting economic and business time series is to collect the relevant data for model construction. Sometimes this is easy, but in many cases we have to make many decisions before a useful set of data is available. For example, how to define a market share when only the information of about 50 percent of the market is available? Or, how should we define the unemployment rate? Does unemployment also include persons who work less than five hours per week? After how many releases of GDP figures can we say that they are final?

In this book we use several empirical time series to illustrate various concepts and models. Examples of the series used in this book for illustration are annual GDP series for Latin American countries, daily Dow-Jones data, quarterly unemployment in the US, weekly observed market and distribution shares for a fast-moving consumer good, monthly new passenger car registrations, and four-weekly advertising expenditures on television and radio. The data can be obtained from http://people.few.eur.nl/djvandijk/tsmbef/data in EViews and Excel format. These data sets can be used by the reader to verify the empirical results reported here and also to try alternative modeling strategies and to examine the properties of other out-of-sample forecasts.

Forecasting

From a methodological point of view it is important to be precise about the goal of econometric time series modeling. In this book, we emphasize that time series models should both give an adequate description of the available in-sample data, and be useful for out-of-sample forecasting. The concept of adequate description will be discussed in detail in Chapter 3. When it comes to forecasting, a crucial assumption is the data in the sample that is used for model specification and estimation are similar to the out-of-sample data. If not, there is no point in spending much time on the construction of high-brow time series models, at least if the possible dissimilarities are not taken into account properly. It is crucial that we evaluate the stability of the forecasting model. For example, if all forecasts are too high or too low, we should obviously wish to re-specify the time series model.

An empirical strategy that can be helpful to assess the stability of the model and the modeling environment is the following. Suppose we have $T + P$ observations for a variable y_t to our disposition. We can then use the first T observations to construct a model and to estimate its parameters, and we can use the last P observations to evaluate its out-of-sample forecasting performance. Hence, we obtain forecasts of y_{T+h} for $h = 1, 2, \ldots, P$ from a model that is constructed using observations $y_1, y_2, \ldots, y_T$. The accuracy of these forecasts can be assessed by comparison with the true values observed for $y_{T+1}, y_{T+2}, \ldots, y_{T+P}$. When the forecasting performance is satisfactory, we may want to generate forecasts for future, unknown observations at times $T + P + 1, T + P + 2, \ldots$, where we typically re-estimate the model parameters using all $T + P$ available observations.

There are no strict guidelines for the choice of P, the number of forecasts used to evaluate the predictive accuracy of the constructed time series model. On the one hand, it is important that T is large enough to have reasonably precision for the estimates of the unknown parameters in the model. On the other hand, P should also be large enough to allow for a meaningful comparison of the predictive accuracy of various competing models. So, it is up to the user to decide on the appropriate values of T and P.

Outline of this book

The contents of this book are as follows. Chapter 2 surveys typical features of time series variables in economics and business. We limit this discussion to five such features, that is, trends, seasonality, aberrant data, time-varying variance, and non-linearity. These features correspond with a decreasing source of variation in economic time series. The most dominant source of variation is often the trend. The next dominant source is seasonal variation, while the smallest amount of variation often tends to be

due to non-linearity. There is no explicit treatment of a "business cycle" here since this sometimes corresponds with cyclical dynamics in the time series itself, sometimes with short-term deviations from a trend and sometimes as a specific type of non-linearity or the occurrence of outliers. Each of the five features suggests different model structures and model specification methods.

We cover each of these features in Chapters 4 to 8. As a prelude, Chapter 3 is concerned with a discussion of several important concepts in time series analysis. Intentionally, this discussion is far from being as technically rigorous as Anderson (1971), Fuller (1976) and Hamilton (1994). Instead the focus is on discussing those techniques which can be readily applied in practice. When necessary, we give references to studies that include proofs of formal asymptotics and other results.

Most of the discussion in Chapter 3, as well as that in Chapters 4 to 8, deals with univariate time series. In Chapter 9 some of the concepts in Chapter 3 are extended to multivariate time series. In addition, the focus in this chapter is on common aspects across economic time series, in particular common trends. Finally, most chapters make mention of recent or even current research topics. The research area of time series analysis is very active, and we can expect many new developments in the near future.

2 Key features of economic time series

In this chapter the focus is on key features of business and economic time series. It also serves to introduce several of the empirical time series that will be used throughout this book as running examples.

The five key features of economic and business data that we consider are (i) trends, (ii) seasonality, (iii) influential data points, (iv) time-varying (conditional) variance (conditional heteroskedasticity) and (v) non-linearity. Typically, an economic time series displays at least one, but usually two or more of these features. To keep matters simple, however, in this chapter each series is analyzed for only one of these five features at a time. In later chapters, the various possible models for each of these features will sometimes be combined to illustrate the practice of time series modeling.

In this chapter, each of the five data features will be illuminated using simple regression-based calculations. This should not imply that these regression models are the best we can do. They are merely helpful tools to demonstrate the properties of the data. A second important tool in this chapter is graphical analysis. In many cases it is already quite helpful just to put the data in a graph with the values of the observations (or some transformation thereof) on the vertical axis and time on the horizontal axis. However, in case of many observations or of a time series with large variance, it may sometimes be more insightful to construct scatter or correlation plots. The latter shows the correlation between y_t and another variable x_t. Since time series analysis is our focus here, x_t can be replaced by, for example, y_{t-1}. In practical situations, we would advise to construct each of the graphs and to estimate each of the simple regression models in order to obtain an overall insight in the specific data properties.

The data that are considered in practice usually come in their original format, that is, they have not been transformed. In empirical time series analysis it is quite common to analyze data after applying the natural logarithmic transformation. Hence, if the original data are denoted by w_t, we usually model and forecast $y_t = \log(w_t)$, where log denotes the natural logarithm. One of the reasons for the log-transformation is that it makes an exponential trend to become a linear trend. Also, if the time series shows increasing variation over time, the log-transformation dampens this trend. When the

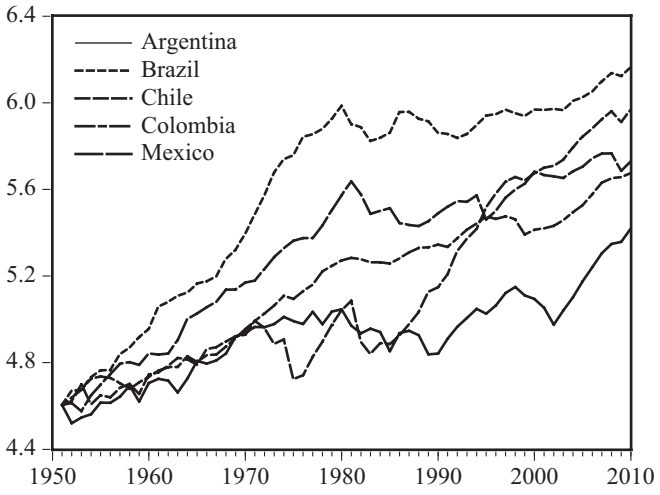


Figure 2.1: Annual indices of log real GDP per capita in Latin American countries, 1951–2010 (1951 = 100).

data are already in relative format, as in the case of the unemployment rate, interest rates or market shares, for example, the log transformation is usually not applied.

2.1 Trends

One of the key features of many economic and business time series is the trend, by which we mean, at least for the moment, that the data show a general tendency to increase or decrease over time. Such a trend can take different shapes. It can be upward or downward sloping, it can be steep or moderate, and it can be exponential or approximately linear. As will become clear below, for many time series the trend is the dominant source of variation, which makes it of crucial importance for out-of-sample forecasting. If a trend is wrongly incorporated in a time series model, forecasts will be poor, especially in the long run.

To illustrate the presence of trends in economic data, consider the five time series shown in Figure 2.1, which are annual indices of real gross domestic product (GDP) per capita (in logs) in the five largest Latin American economies for the sample period 1951–2010, that is, Argentina, Brazil, Chile, Colombia and Mexico. From the graphs it can be observed that over the complete 60-year sample period all five countries have grown considerably, although at different rates, with Brazil and Argentina showing the highest and lowest average growth rates, respectively. Furthermore, although the

Table 2.1: Trends in real GDP per capita in Latin American countries, 1951–2010

Country	$\hat{\beta}$ in regression: $y_t = \alpha + \beta t + u_t$		$\hat{\mu}$ in regression: $y_t - y_{t-1} = \mu + u_t$	
Argentina	1.06	(0.07)	1.42	(0.62)
Brazil	2.48	(0.14)	2.64	(0.51)
Chile	2.28	(0.12)	2.31	(0.75)
Colombia	1.78	(0.04)	1.76	(0.32)
Mexico	1.92	(0.09)	1.99	(0.53)

Note: The numbers in parentheses are estimated standard errors. All numbers are multiplied by 100.

(indexed) levels of real GDP per capita in Chile, Colombia and Mexico are approximately the same in the year 1995, such that their average growth rates are rather close, their developments during the preceding period were quite different. GDP growth in Colombia appears to have been rather stable at an almost constant pace. Mexico experienced rapid growth up to 1980 when a long recession started, which lasted until 1988 and was followed by a gradual recovery until another recession occurred in 1995. Finally, GDP growth in Chile was about the same as in Colombia until the early 1970s when the economy was hit by a severe recession, followed by a second one in the early 1980s. Starting in 1984, a steep recovery occurred that took GDP per capita back to the levels of Mexico and Colombia within a decade.

To quantify the trends in the five GDP per capita series, consider the following simple regression model

$$y_t = \alpha + \beta t + u_t, \qquad t = 1, 2, \ldots, T, \tag{2.1}$$

where α and β are unknown parameters and where u_t is an unknown residual error time series with mean zero. Note that the GDP series in Figure 2.1 show some cyclical behavior around their trends, indicating that u_t may be correlated with u_{t-1}, and hence suggesting misspecification of (2.1). The standard errors of the parameter estimates should therefore be interpreted with care.

The left-hand panel of Table 2.1 displays the estimates of β in (2.1), multiplied by 100 for convenience. It is clear that the upward trend is steepest for Brazil ($\hat{\beta} = 2.48$) and that average growth is lowest for Argentina ($\hat{\beta} = 1.06$). As expected, the estimates of β are fairly close for Chile, Colombia, and Mexico. Note, though, that the standard errors for Mexico and Chile are two and three times as large as for Colombia, reflecting the much steadier growth pattern of the latter country.

The simple regression model in (2.1) assumes that the trend in y_t can be represented by a linear deterministic trend $t = 1, 2, 3, \ldots$. An alternative method to obtain insight in the trend pattern is to directly consider the growth rate of the variable. In case $y_t = \log(w_t)$, it follows that

$$\begin{aligned} y_t - y_{t-1} &= \log(w_t/w_{t-1}) \\ &= \log(1 + (w_t - w_{t-1})/w_{t-1}) \\ &\approx (w_t - w_{t-1})/w_{t-1}, \end{aligned}$$

where the approximation is valid when the growth rate $(w_t - w_{t-1})/w_{t-1}$ is small. In industrialized countries, annual growth rates of macroeconomic aggregates typically range between -2% and 4%, which can be considered as small. Hence, the first differences of y_t approximately corresponds to the growth rate of w_t. Notice that the growth rate can also be useful to obtain more interpretable numbers. Usually it is less important to know that the value of the Dow Jones index is 15,000 or so, than that it is to know that the rate of change with respect to the week before is, say, 2 percent.

A trending pattern in an economic time series is reflected by an average growth rate that is different from zero. As an alternative to (2.1), we can therefore also consider the regression model

$$y_t - y_{t-1} = \mu + u_t \qquad t = 2, 3, \ldots, T. \tag{2.2}$$

Notice that, where the trend is captured by means of the term βt in (2.1), it appears through μ in (2.2). A result of looking at the growth rate $y_t - y_{t-1}$ instead of the (log) level y_t is that there are now $T - 1$ observations to be used to estimate μ. The series y_t is said to have a stochastic trend in case (2.2) is a better model for its trend behavior than (2.1). If regression (2.1) is more appropriate, the trend in y_t is called deterministic. This clearly demonstrates that the concept of a trend is defined only within the context of a specific model. In Chapter 4, we will return to this issue of modeling trends, and the practically very important choice between (2.1) and (2.2).

The second panel of Table 2.1 displays the estimates of μ in (2.2) for the five GDP per capita series shown in Figure 2.1. It appears that the estimates of μ are slightly larger than the estimates of β in (2.1) for Argentina, Brazil, Chile and Mexico, while it is somewhat smaller for Colombia. Again the growth rate of Brazil at more than 2.5 percent exceeds average growth in the other countries. Notice also that the estimated standard error of $\hat{\mu}$ in (2.2) is considerably larger than that of $\hat{\beta}$ in (2.1) for all countries.

The trends in Figure 2.1 are all of the familiar type, in the sense that many economic time series display only an upward sloping trend. It is also possible that a trend is less smooth and displays changes in its slope or even changes direction once in a while. For example, although it is clear from Figure 2.1 that GDP per capita for Brazil has an upward trend, it also appears that this trend was much steeper between 1950 and 1980 than during the last two decades of the sample period.

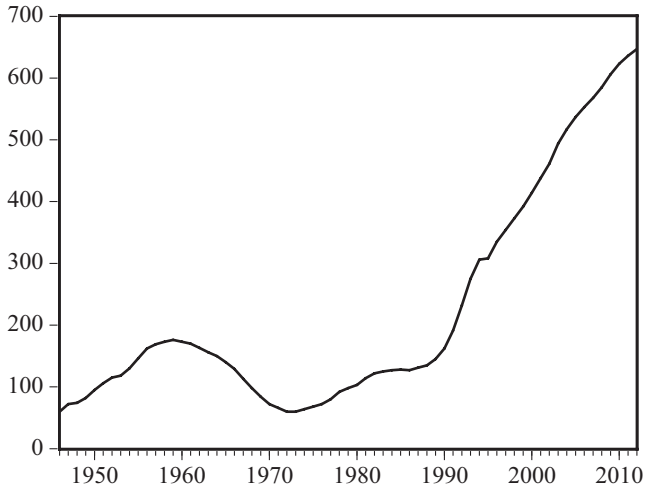


Figure 2.2: Annual stock of motorcycles (in thousands) in The Netherlands, 1946–2012.

Another example of changes in the trend pattern is given in Figure 2.2 where the annual stock of motorcycles in The Netherlands over the years 1946–2012 (measures on January 1) is displayed. From 1946 to about 1960, there is an upward trend due to an increasing popularity of motorcycles because of their successful use in World War II. From 1960 to 1973 the stock of motorcycles declines because of the increasing willingness to own a car. From 1974 onwards there is again a trend upwards, which in the last few years seems to explode. This can be attributed to the fact that car owners may want to own a motorcycle as an additional leisure vehicle. In sum, for this time series we can observe several (gradual) changes in the trend. In Chapter 4 it will appear that time series such as in Figure 2.2 may have a so-called double stochastic trend, that is, the direction of the trend is also a stochastic trend itself.

One possible approach to describe changing trends is to allow the parameters in either (2.1) or (2.2) to change over time. Usually we consider such an approach in case it is reasonable to assume that certain exogenous shocks may have changed the direction of the trend. For example, the oil price shock in the fourth quarter of 1979 may have changed the direction of the trend in macroeconomic variables, see Perron (1989), for example. This can be illustrated by the (quarterly, seasonally adjusted) index of US industrial production, for which some regression results for (2.1) and (2.2) for different sample periods are presented in Table 2.2. The left panel again considers regression (2.1), where the parameters are estimated for the complete sample 1959Q1–2012Q4 and for the sub-samples 1959Q1 to 1979Q4 and 1980Q1 to 2012Q4.

Table 2.2: Trends in US industrial production, 1959–2012

Sample	$\hat{\beta}$ in regression: $y_t = \alpha + \beta t + u_t$		$\hat{\mu}$ in regression: $y_t - y_{t-1} = \mu + u_t$	
1959.1–2012.4	0.676	(0.010)	0.687	(0.114)
1959.1–1979.4	1.027	(0.028)	1.009	(0.220)
1980.1–2012.4	0.605	(0.016)	0.484	(0.122)

Note: The numbers in parentheses are estimated standard errors. All numbers are multiplied by 100.

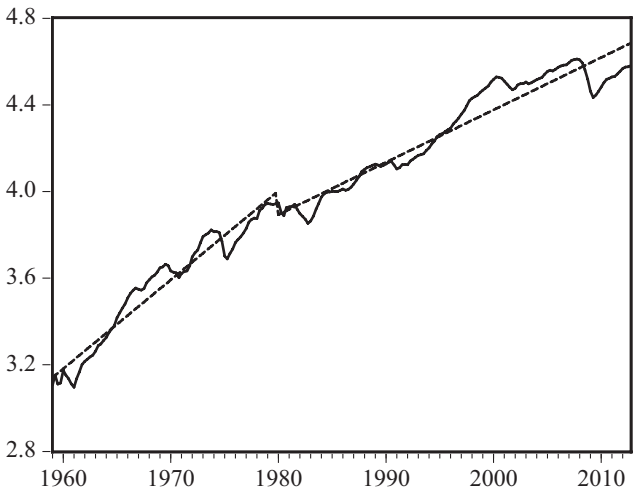


Figure 2.3: Quarterly (log) US industrial production index (seasonally adjusted), with changing deterministic trend after 1979Q4.

Clearly, the estimated trend parameter $\hat{\beta}$ in the first sample (1.027) exceeds that of the second sample (0.605). A similar conclusion holds for regression (2.2) in the right panel of Table 2.2 where growth in the period before the oil crisis is about 1 percent per quarter while it is only about 0.5 percent per quarter after that crisis. Again, the estimated standard error of $\hat{\mu}$ by far exceeds that of $\hat{\beta}$. The change in trend is visualized in Figure 2.3.

To conclude this introduction on trends in economic time series, it should be mentioned that there is no unique way to describe a trend. There are several different

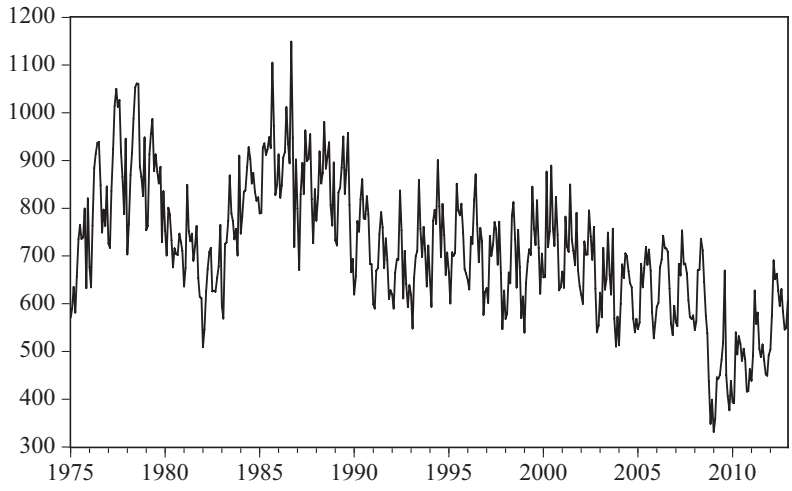


Figure 2.4: Monthly US new passenger car registrations (in thousands), January 1975–December 2012.

possibilities, with the deterministic trend in (2.1) and the stochastic trend in (2.2) being the two most popular approaches. As will be seen in subsequent chapters, different trend specifications have different impacts on forecasting. When comparing Figures 2.1 and 2.2 with Figure 2.3, it is also clear that it may not be easy to make a proper choice between the different versions of trend descriptions. A selection method to choose between different models will be discussed in Chapter 4. The issue of breaking or changing deterministic trends will be treated in Chapter 6.

2.2 Seasonality

When an economic time series is observed each month or quarter, it is often the case that such a series displays more or less regular variation across the different seasons of the year. Put differently, economic time series often have a pronounced seasonal pattern. Similar to the feature of a trend, where the precise meaning of "trend" depends on the model used to describe it, there is no precise definition of *seasonality*. Usually we refer to seasonality when observations in certain seasons display substantially differences compared to observations in other seasons. For example, the number of new passenger car registrations in the US are always relatively high in the summer months July and August and in December (the latter because of tax reasons), as can be observed from Figure 2.4. Hence we can say that this variable displays seasonality. It may also be that

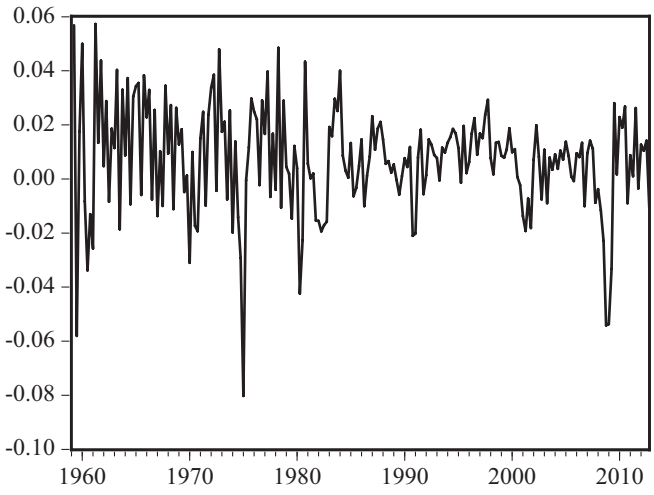


Figure 2.5: First differences of the (log of) US industrial production index (not seasonally adjusted).

seasonality is reflected in the variance of a time series. For example, for daily observed stock market returns the volatility seems often highest on Mondays, perhaps because investors have to digest three days of news instead of only one day.

The number of seasons is denoted by S. When a calendar year is considered the benchmark, which is the case throughout this book, S equals 4 and 12 for quarterly and monthly data, respectively. In case there are N complete years of data on y_t, the total number of observations T equals SN. Of course, other possibilities can also be considered. For example, in empirical finance day-of-the-week effects may be relevant, in which case N refers to the number of weeks and $S = 5$.

Seasonality is often noticeable rightaway from a simple graph of the time series, as in Figure 2.4. In other cases, more experience is needed to spot seasonal variation. For example, in Figure 2.5, showing the first difference $y_t - y_{t-1}$ of the (log) US industrial production index (seasonally unadjusted) over the period 1959Q1–2012Q4, it may not be evident how relatively important seasonality is. On the other hand, it seems that the pattern after the oil crisis in 1979Q4 is different from the pattern before. Notice also the dip in the first quarter of 1975, which seems to be an irregular observation.

More precise information about the importance and stability of the seasonal pattern in a given time series can often be obtained by plotting the observations for the different seasons as separate time series in a single graph. This so-called vector-of-seasons graph was introduced by Engle, Granger, Hylleberg and Lee (1993) and Franses (1994). In

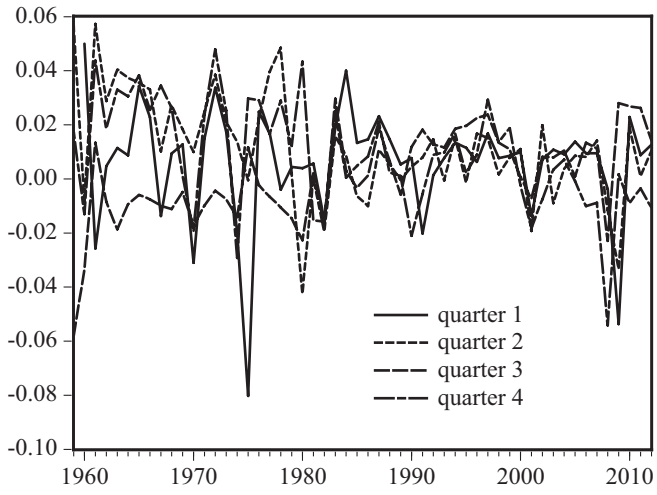


Figure 2.6: Vector-of-quarters representation of first differences of the (log of) US industrial production index (not seasonally adjusted).

case seasonal variation in a time series is large, we would expect the series for the different seasons to be distinct and relatively far apart. Changes in the seasonal pattern should result in time series for different seasons getting closer or drifting apart, or even crossing each other. Figure 2.6 displays the vector-of-quarters graph for the US industrial production growth rates. The fact that the lines for the different quarters are not clearly separated suggests that there is not much seasonal variation in this time series. On the other hand, closer inspection of the graph shows that before 1980, the time series for the second and fourth quarters were considerably above that for the third quarter, suggesting that indeed the seasonal pattern may have been more pronounced during the first half of the sample period, and has changed around 1980.

In case simple graphs are not informative enough to highlight possible seasonal variation, we can rely on a version of (2.2), like

$$y_t - y_{t-1} = \mu_1 D_{1,t} + \mu_2 D_{2,t} + \cdots + \mu_S D_{S,t} + u_t \qquad t = 2, 3, \ldots, T, \tag{2.3}$$

where $D_{s,t}$ is a seasonal dummy variable with

$$D_{s,t} = \begin{cases} 1 & \text{when } t = (n-1)S + s \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

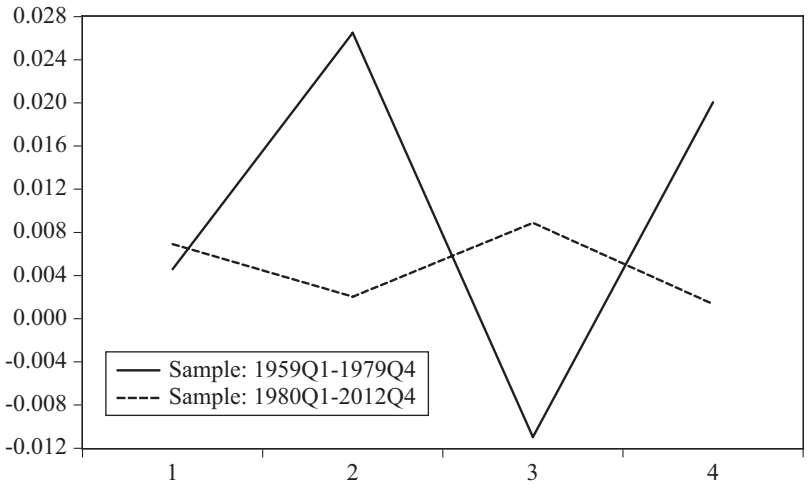


Figure 2.7: Changing seasonality in US industrial production.

with $s = 1, 2 \ldots, S$ and $n = 1, 2, \ldots, N$. In words, $D_{s,t}$ takes the value 1 when calendar period t corresponds with season s, and $D_{s,t}$ is equal to 0 otherwise. The unknown parameters μ_s, $s = 1, \ldots, S$, in (2.3) then represent the average value of the first difference $y_t - y_{t-1}$ for season s. In case the u_t process does not contain information on seasonality, we may also consider the R^2 of (2.3) as giving an indication of the "amount of deterministic seasonality", see Miron (1996).

We estimate the regression (2.3) for the growth rates of US industrial production over the sample period 1959Q1–2012Q4. As this concerns a quarterly time series, S equals 4. The estimates of the average growth rates in the different quarters (multiplied by 100) are $\hat{\mu}_1 = 0.60$, $\hat{\mu}_2 = 1.15$, $\hat{\mu}_3 = 0.12$, and $\hat{\mu}_4 = 0.86$, where the standard error equals 0.27. These confirm that average growth is highest during the second and fourth quarters, and lowest during the third. The R^2 of the regression is only 0.037, suggesting that seasonal variation does not affect US industrial production to a very large extent. Given the tentative evidence in Figures 2.5 and 2.6 of possibly changing seasonal patterns in the series, it is also useful to fit the model (2.3) for different subsamples. We split the sample period into 1959Q1–1979Q4 and 1980Q1 to 2012Q4, with the resulting estimates of the parameters μ_1 to μ_4 for these two sub-samples shown in Figure 2.7.

From Figure 2.7 we can observe indeed the seasonal pattern appears to be different before and after 1980. The growth rates in quarters 2 and 4 are largest in the first subsample (0.027 and 0.020). Average growth in the first quarter is almost equal to zero, while it even is negative in quarter 3 (-0.011). The variation in growth rates across

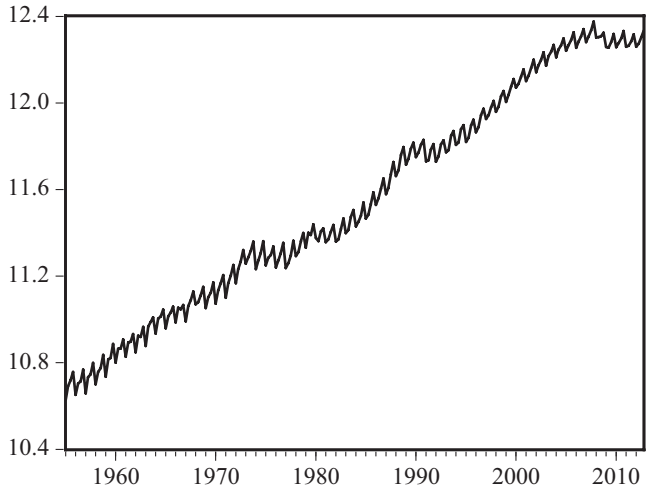


Figure 2.8: Quarterly UK household final consumption expenditures in the UK (in logs, not seasonally adjusted), 1955Q1–2012Q4.

quarters is much smaller in the second sub-sample, with estimates ranging between 0.001 for quarter 4 and 0.009 for quarter 3. The R^2 values for the two sub-samples are 0.341 and 0.043, respectively, confirming that seasonality in US industrial production has become much less pronounced in more recent years.

Another example of a time series that displays marked seasonality is the log of quarterly household final consumption expenditures in the UK over the period 1955Q1–2012Q4, which is depicted in Figure 2.8. Next to an upward sloping trend, which seems to lose strength during the recessions around 1976, 1982 and 1991, but which regains its path during the intermittent expansions, there seems clear visual evidence of seasonality. This is borne out even clearer by the first differences $y_t - y_{t-1}$ in Figure 2.9, with the corresponding vector-of-quarter graphs shown in Figure 2.10. The latter graph also suggests that the seasonal pattern may have changed after 1980, especially due to a decline of the average growth rate in the second quarter and an increase during the third.

The relative importance of seasonality in UK consumption is confirmed by the results of regression (2.3), where for the sub-samples 1955Q1–1979Q4 and 1980Q1–2012Q4 we obtain R^2 values of 0.922 and 0.914, respectively. The two sets of estimates of the parameters μ_1 to μ_4 for this quarterly series are given in Figure 2.11. Relative to the first sample, the growth rate in the second quarter becomes somewhat smaller in the second sub-sample, with an off-setting increase in growth during the third quarter. Put

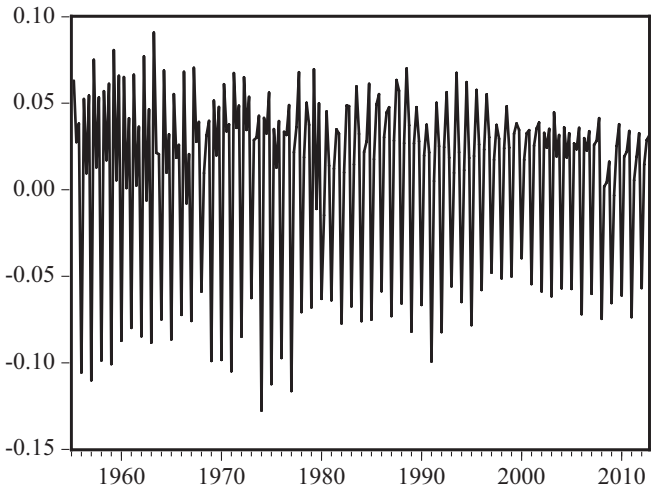


Figure 2.9: First differences of (the log of) quarterly UK household consumption (not seasonally adjusted), 1955Q1–2012Q4.

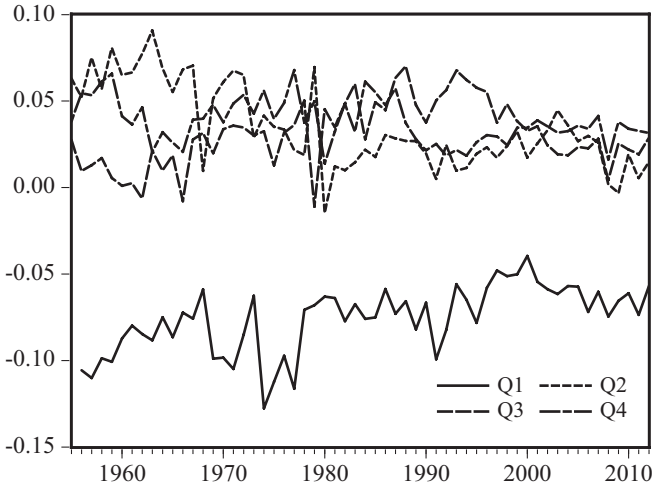


Figure 2.10: Vector-of-quarters graph of first differences of UK household consumption.

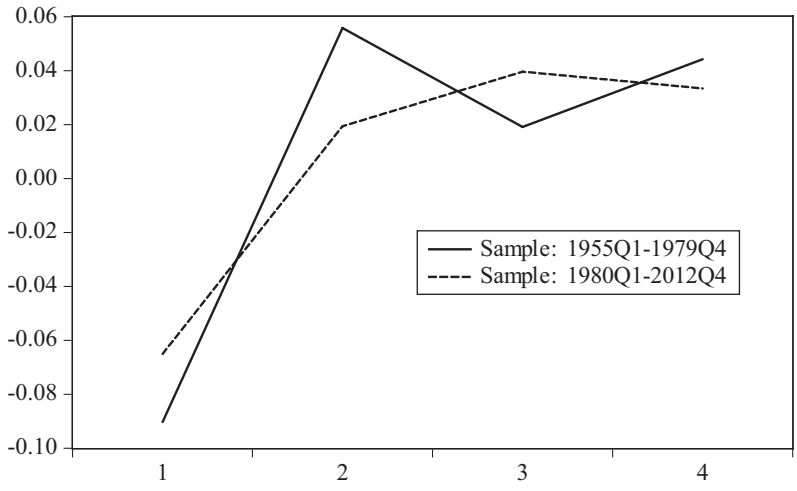


Figure 2.11: Changing seasonality in UK household consumption.

differently, consumption appears to shift from the second to the third quarter. Looking again at Figure 2.7, the changes in the seasonal pattern for UK consumption are less substantial than those for US industrial production.

A final example of time series with pronounced seasonality is given in Figure 2.12, showing (the logs of) four-weekly advertising expenditures on radio and television in The Netherlands for 1978.01–1994.13. For these two time series it is clear that television advertising displays quite some seasonal fluctuation throughout the entire sample, where possibly there are some changes towards the end of the sample, and that radio advertising has seasonality only in about the last five years. This last period marks the introduction of an additional commercial network (RTL4) in The Netherlands. Furthermore, there seems to be a structural break in the radio series around observation 53. This break is related to an increase in radio broadcasting minutes in January 1982. Additionally, there is some visual evidence that the trend changes over time. In Chapter 6 we return to analyzing the consequences of such mean shifts on time series modeling.

To investigate the effect of the introduction of RTL4 on seasonality in television advertising, consider again the results of the regression in (2.3) with $S = 13$ for two sub-samples, which are depicted in Figure 2.13.

It is clear from Figure 2.13 that seasonality seems to increase in the second sub-sample since then the estimates of μ_s, $s = 1, 2, \ldots, 13$ show more variation.

Generally, it appears that many seasonally observed business and economic time series display seasonality in the sense that the observations in certain seasons have

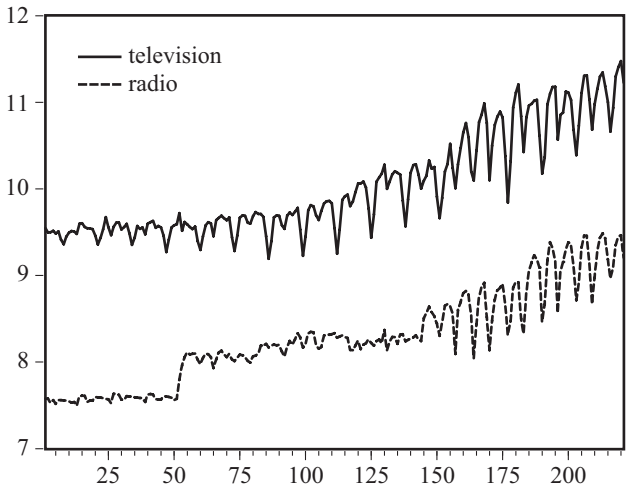


Figure 2.12: Four-weekly advertising expenditures on radio and television in the Netherlands (1978.1–1994.13).

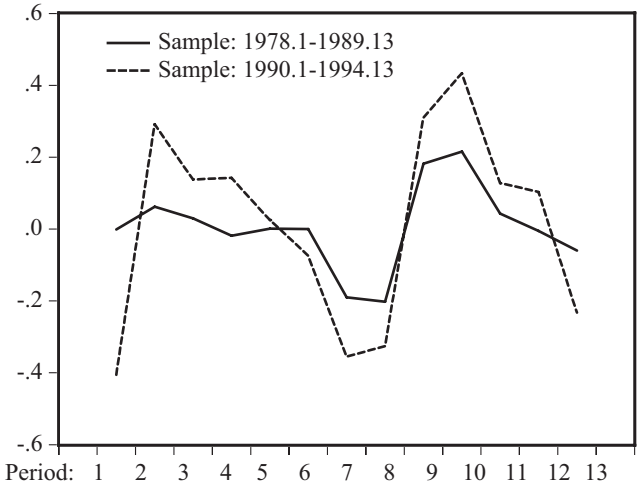


Figure 2.13: Changing seasonality in television advertising expenditures.

properties that differ from those data points in other seasons. A second feature of many economic time series is that seasonality changes over time. Sometimes these changes appear abrupt, as is the case for advertising on the radio in Figure 2.12, and sometimes such changes occur only slowly, as is the case for UK consumption. In Chapter 5, we will review methods to describe and forecast economic time series with changing seasonality.

2.3 Aberrant observations

The radio advertising series in Figure 2.12 indicates that time series features may change over time. For this series, the mean of the observations in the first part of the sample is smaller than that in the second part of the sample, and in the third part there even emerges a seasonal pattern that was previously absent. Obviously, such changing patterns should be taken into account when forecasting out-of-sample. For example, if we were to analyze the entire sample of the radio advertising series with a model that does not allow for seasonality, it is likely that we obtain rather inaccurate forecasts for certain seasons. Furthermore, when the mean shift around observation 53 is neglected, we would also expect a systematic bias in forecasts. This implies that we should take account of the possibility that there are periods or sub-samples that can make time series modeling difficult.

Changes in the features that are present in a time series are not necessarily permanent. It may also be that individual observations deviate from the regular patterns. Such data points are called aberrant observations or outliers. As will become clear below, such outliers can have a major impact on time series modeling and forecasting. As an illustration, consider the monthly time series of revenue passenger-kilometers flown (RPK; in millions) for all European airlines, for the sample period 1993.1–2012.12, shown in Figure 2.14. RPK, defined as the product of the total number of passengers flown on a certain route and of the corresponding distance, is an important measure of revenue in the airline industry. The time series shown concerns the aggregate RPK across all international and domestic flights operated by European airlines.

The time series displays a clear upward trend, reflecting the substantial increase in air traffic during these two decades. Also, a pronounced seasonal pattern is present with higher values during the summer months, presumably due to holiday flights. Closer inspection of the graph suggests that both the trend and the seasonal fluctuations are not perfectly regular though. Specifically, the declines during the Fall of 2001 and 2008 seem considerably larger than in other years, such that the time series declines to a lower level. In addition, the seasonal pattern seems somewhat erratic in the Spring of the years 2003 and 2010.

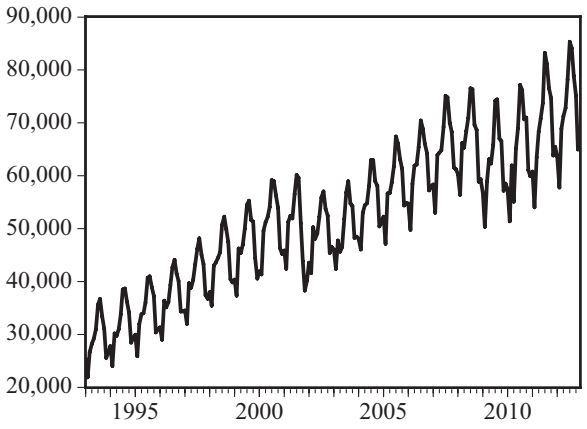


Figure 2.14: Monthly revenue passenger-kilometers flown (in millions) for all European airlines, 1993.1-2012.12.

These deviations from the regular pattern may in fact not be so obvious from the graph of the original time series as in Figure 2.14, given that the trend and seasonality are quite dominant. It may be helpful to consider transformations to bring out these observations more clearly. For that purpose, consider annual growth rates, that is, the difference between the natural logarithm of RPK in a given month and the log RPK in the same month of the previous year. In other words, if y_t denotes the natural log of the RPK measure, we consider $y_t - y_{t-12}$. The reason for considering this particular transformation is that it removes both the trend and the seasonal pattern. Figure 2.15 shows the resulting time series.

We observe that, on average, the annual growth rate is equal to about 10 percent per year until 2001. Then in September 2001 a sudden drop occurs, with the growth rate falling to -21 percent in October of that year. Obviously this is a consequence of 9/11, following which many flights were cancelled or at least flown with much less passengers. Another substantial decline in RPK occurred in the period September 2008–March 2009, with growth rates dropping below -10 percent. Obviously this may be attributed to the financial crisis and subsequent 'Great Recession' that unraveled during this period. A notable difference with the sudden drop in the Fall of 2001 is that the decline in 2008–2009 occurred much more gradually. Finally, an isolated spike in April 2010, with a growth rate of -14 percent, is observed. This is the result of the eruption of the Eyjafjallajøkull volcano in Iceland, which caused an enormous disruption to air travel across western and northern Europe. From 14–20 April, ash

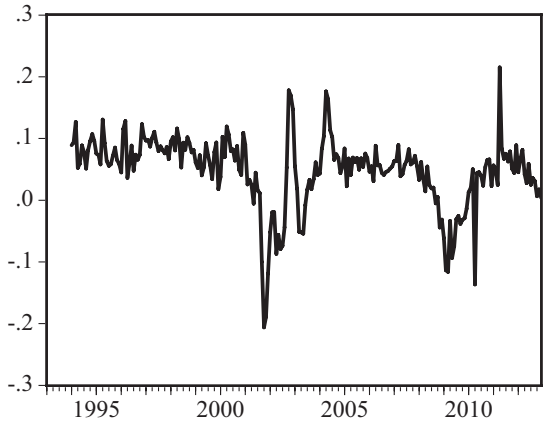


Figure 2.15: Annual growth rates of revenue passenger-kilometers flown (in millions) for all European airlines, 1994.1–2012.12.

covered large areas of northern Europe when the volcano erupted. About 20 countries closed their airspace to commercial jet traffic, affecting more than 100,000 travellers.

Another transformation that may be useful is the monthly growth rate $y_t - y_{t-1}$, as this also removes the trend from the time series. In this case, the resulting time series is still dominated by the strong seasonal fluctuations, see Figure 2.16, although the deviating pattern in the Fall of 2001 can already be seen from this graph.

Further insight may be obtained when the seasonal pattern is removed, by inspecting the residuals from the regression of the monthly growth rates on monthly dummy variables, as in (2.3). This leads to Figure 2.17. We clearly observe the effect of the volcano eruption in April 2010, leading to a large negative residual in the month and a large positive residual in May 2010, when air traffic immediately returned to its normal level. Also the drop following 9/11 and the subsequent recovery are visible in the form of a sequence of large residuals (both negative and positive). Interestingly, the substantial decline between September 2008–March 2009 is much less obvious. As we will see later, this is related to the more gradual nature of this decline. A final interesting observation, which was not seen in the earlier graphs, is that the residual for December 1999 is approximately equal to -10 percent. Possibly due to fears that the Y2K problem would affect the computerized airplane operating systems many people avoided flying during that month, resulting in a growth rate substantial below its usual level.

The main message from the above example is that a given time series may contain different types of aberrant observations, which give rise to specific deviating patterns

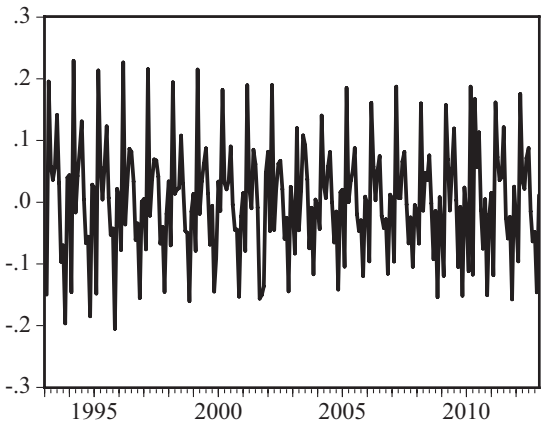


Figure 2.16: Monthly growth rates of revenue passenger-kilometers flown (in millions) for all European airlines, 1993.2–2012.12.

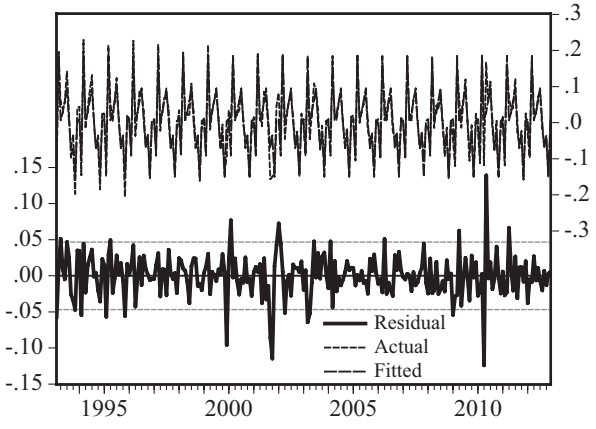


Figure 2.17: Monthly growth rates of revenue passenger-kilometers flown (in millions) for all European airlines, 1993.2–2012.12.

in the time series. In Chapter 6, we will return to a discussion of the effect of such aberrant data points on modeling and forecasting. Chapter 6 also discusses methods to detect several types of aberrant observations, and methods to take account of such data points for forecasting.

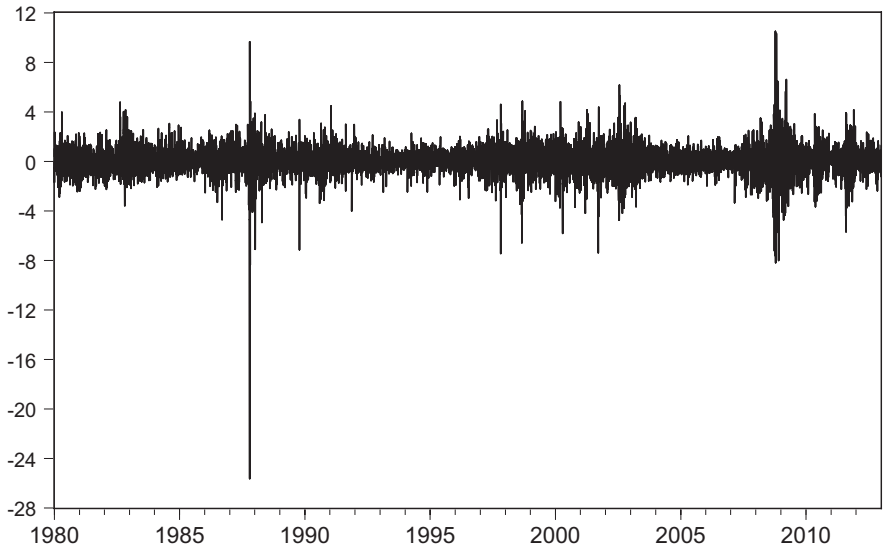


Figure 2.18: Daily returns on the Dow Jones index, (January 1, 1980–December 31, 2012).

2.4 Conditional heteroskedasticity

A fourth feature of economic time series, and in particular of returns on financial assets, is that their volatility changes over time. Relatively volatile periods, characterized by large price changes and, hence, large returns, alternate with more tranquil periods in which prices remain more or less stable and returns are, consequently, small. This feature commonly is referred to as *volatility clustering* or *conditional heteroskedasticity*. A possible explanation for the occurrence of volatility clustering relates to the arrival of news. Intuitively, in an efficient financial market asset prices reflect all currently available information, and prices change only due to the arrival of new information. A characteristic feature of such news is that it does not arrive at a constant rate, but that turmoil periods alternate with periods with very little news. In addition, investors may require some time to properly digest arriving news, leading to extended periods of substantial price changes.

Consider for example the time series of daily returns on the Dow-Jones index, from January 1, 1980 to December 31, 2012, as given in Figure 2.18. The observation that stands out most clearly with a return of -25.6% corresponds with the stock market crash on Monday October 19, 1987. Immediately following this observation, we can find several data points that are large in absolute value. Also in other parts of the sample

clusters of observations with large variance can be observed. These occur, for example, in 1998 due to the crisis in Russia, in 2000 due to the collapse of the internet bubble, and in the Fall of 2008 following the collapse of Lehman Brothers. The first half of the 1990s as well the period 2003-2008 stand out as long periods with low variance. The presence of such volatility clustering suggests that a model for the (conditional) variance of the time series can be useful.

Allowing for the possibility that high and low volatility of returns tend to persist for prolonged periods of time, we may consider the presence of so-called conditional heteroskedasticity. This time series feature can be examined by the simple regression

$$(y_t - y_{t-1})^2 = \alpha + \rho(y_{t-1} - y_{t-2})^2 + u_t, \qquad t = 3, 4, \ldots, T, \qquad (2.5)$$

where y_t denotes the log of the Dow Jones index such that $(y_t - y_{t-1})^2$ represents the variance of the returns. For the sample of 8610 daily returns shown in Figure 2.18, we obtain $\hat{\rho} = 0.103(0.011)$, where the estimated standard error is given in parentheses. This suggests that there is positive (albeit small) autocorrelation in the variance of the returns. Additionally, when $(y_t - y_{t-1})^2$ is replaced by the absolute returns $|y_t - y_{t-1}|$, the regression in (2.5) yields $\hat{\rho} = 0.197(0.011)$, also pointing towards the presence of volatility clustering.

In case of volatility clustering, we may wish to exploit this in order to forecast future volatility. Since this variable is a measure of risk, such forecasts can be useful to evaluate investment strategies. Furthermore, it can be useful for decisions on exercising options or other derivatives. Time series models for conditional volatility are often applied in practice, and some of these models will be discussed in Chapter 7.

2.5 Non-linearity

The fifth and final feature of economic time series that is treated in this book is non-linearity. Although the best definition of non-linearity may perhaps be "everything different from linearity", for economic time series we usually rely on concepts such as regime-switching or so-called state dependence. The latter indicates that the behavior of a time series is different depending on the current "state of the world". These states can be defined in several different ways, and the relevant state classification depends on the time series variable of interest. For example, periods with high and low volatility in a financial market may be considered as the relevant states for a time series of stock returns, as discussed in the previous section. For macroeconomic time series such as GDP and unemployment, economic recessions and expansions (that is, different phases of the business cycle) may be more appropriate.

When the behavior of a time series is different across such (discrete) states, we call this regime-switching behavior. It should be mentioned that certain states (such as a

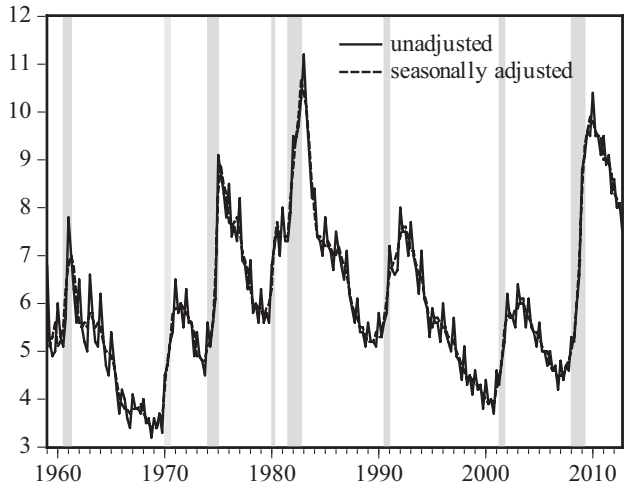


Figure 2.19: Quarterly US unemployment rate among men of 16 years and over, 1959Q1–2012Q4 (seasonally adjusted and not seasonally adjusted).

recession) may concern relatively few observations. Regime-switching can also show similarity with structural breaks. For example, the oil crisis in the fourth quarter of 1979 appears to have changed the trend and seasonality in US industrial production, see Figures 2.3, 2.5, and 2.7, which may be interpreted as a (permanent) change of regime.

Non-linear behavior is often quite obvious for certain macroeconomic time series, and that is when the first differences $y_t - y_{t-1}$ take different average values across states. Consider for example the quarterly US unemployment rate among men of 16 years and over for the period 1959Q1–2012Q4, as shown in Figure 2.19. Clearly, unemployment sometimes increases quite rapidly, especially in periods corresponding with economic recessions (which are the shaded areas in Figure 2.19: 1970, 1974, 1980, 1981–2, 1990–1, 2001 and 2008–9), while it decreases much more slowly in times of expansions. This asymmetry can be formalized by estimating the parameters in the following simple regression

$$y_t - y_{t-1} = \mu_1 I_t[\text{expansion}] + \mu_2 I_t[\text{recession}] + u_t, \qquad t = 2, 3, \ldots, T, \tag{2.6}$$

where y_t denotes the unemployment rate in quarter t (without taking logs), and I_t[expansion] and I_t[recession] are indicator variables taking the value 1 when quarter t is in an expansion and recession, respectively, and 0 otherwise. This implies that μ_1 and

μ_2 in (2.6) represent the average change in unemployment during expansions and recessions, respectively. For the seasonally adjusted US unemployment rate over the period 1959Q1–2012Q4, we find estimates of these parameters equal to $\hat{\mu}_1 = -0.093(0.018)$ and $\hat{\mu}_2 = 0.581(0.044)$, indicating that when the economy is in recession, the unemployment rate increases much faster than when it goes down during expansions. This regression result confirms the visual impression from Figure 2.19.

A second example is given by the daily returns on the Dow Jones index, as depicted in Figure 2.18. As discussed in the previous section, a prominent feature of this time series is volatility clustering, in the sense that prolonged periods with high and low volatility alternate. Investors require higher returns for investments with higher risk, suggesting that the mean return should also differ across high- and low-volatility states. This can be investigated by means of a slight modification of the regression in (2.6), where the dependent variable is taken to be the daily return $y_t - y_{t-1}$ measured in percent, which is regressed on indicator variables $I_t[(y_t - y_{t-1})^2 > 1]$ and $I_t[(y_t - y_{t-1})^2 \leq 1]$ that define the periods of high and low volatility. The threshold value is set to 1 as this is close to the average squared return over the sample period. For the time series of daily returns over the period January 1, 1980 - December 31, 2005, this renders estimates $\hat{\mu}_1 = 0.095(0.026)$ and $\hat{\mu}_2 = 0.020(0.015)$, with standard errors in parentheses. Hence, we find that indeed the average return during the high-volatility period is approximately five times as large as the average return during the low-volatility period. Notice that here we have fixed the so-called threshold value that is, the border between the low and high volatility regimes. In practice one may allow for more flexibility, for example by treating the threshold value as an unknown parameter that is to be estimated along with the other parameters in the model. Additionally, one may also wish to allow for more than two regimes.

When a non-linear time series model is used to describe and forecast an economic variable, much more effort is usually needed to specify the model and to estimate its parameters. A key reason for this is that allowing for non-linearity leads to a wide variety of possible models. In practice therefore we often start with computing several diagnostic measures (in a linear model) to obtain a first and tentative impression of what type of non-linear model could be useful. In Chapter 8, we will discuss several non-linear time series models that are often used in practice. Furthermore, we will review tools that can guide model selection.

2.6 Common features

Many time series in economics and business usually have at least one of the above five features. For example, the time series graph of the US unemployment rate in

Figure 2.19 shows that this variable displays seasonality, non-linearity and possibly also some aberrant observations. As another example, the advertising expenditures series in Figure 2.12 seem to display structural breaks, changing seasonality, and trends. Finally, the daily Dow Jones returns series may be conditionally heteroskedastic, but there may also be aberrant observations (such as Monday October 19, 1987).

In case of univariate time series modeling, that is, when we only analyze and forecast y_t given its own past, we should take account of all of the observed features, which can be illuminated by simple auxiliary regressions and specific insightful graphs (as shown before in this chapter). In case we want to incorporate information from the past patterns of variables as, say, x_t and z_t to describe and forecast y_t, and possibly to use the past of y_t to describe and forecast x_t and z_t, we should consider so-called multivariate time series models. Some of the important aspects of these models are discussed in Chapter 9.

Given all the possible time series features, many decisions obviously have to be made when implementing multivariate models. One way to reduce the number of decisions is to search for the presence of so-called common features across time series. Additionally, such common features may themselves be the focus of interest. For example, a relevant question for the GDP per capita series in the five Latin American countries, as displayed in Figure 2.1, is whether these have common growth rates or whether Brazil significantly outgrows the other four countries. For the same five series, one may wonder whether Chile, Colombia, and Mexico have in fact the same trend, of whichever nature, be it deterministic or stochastic.

Figure 2.20 displays the monthly log prices of gold and silver for the period January 1978 – December 2012. Given that these two commodities can be close substitutes (for investors, for example), one may expect their prices to move together over time. Because of international pricing agreements, temporary overproduction or changes in investor sentiment, prices can display trending behavior, although the direction of these trends may not be constant over time. This expectation is reflected by the graphs in Figure 2.20 where both gold and silver experience some periods with increasing prices, while also prices sometimes tend downwards for longer periods (as for example between 1980 and 1985 and between 1996 and 2000). Finally, there are sequences of months with very large increases or decreases in prices, especially during the first years of the sample period. Hence, next to trends we may expect the gold and silver prices to display volatility clustering, which can be common or not. Figure 2.21 shows the time series of daily gold and silver returns over the same period. Roughly speaking, it appears that often periods with high and low volatility for the two series coincide. If volatility clustering indeed is common, this can be interpreted as saying that the risk involved in buying or selling these two commodities is about the same.

Chapter 9 of this book is mainly dedicated to common trends (which is also called cointegration). The simple framework that is often used to investigate common

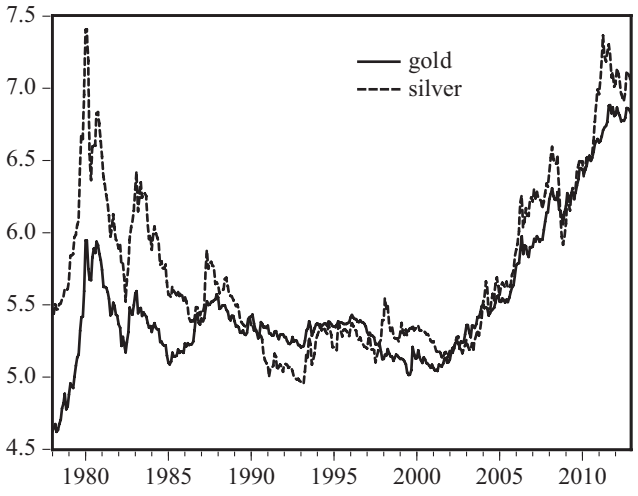


Figure 2.20: Monthly log prices of gold and silver, 1978.1–2012.12.

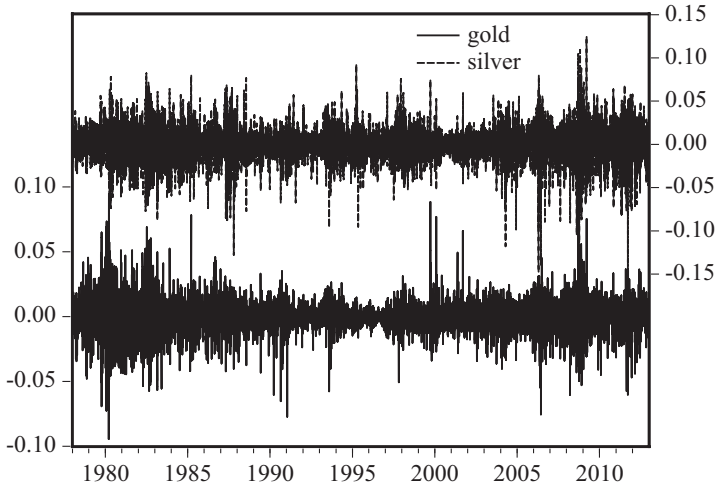


Figure 2.21: Daily returns of gold and silver, January 1, 1978–December 31, 2012.

features is

$$y_t = \beta x_t + u_t, \tag{2.7}$$

where y_t and x_t have a certain feature and the residual time series u_t perhaps not. For example, y_t and x_t may have a common trend, or they both individually have a trend, but u_t does not. Key topics of Chapter 9 are how to find a proper value of β and how to investigate the features of the unobserved u_t series.

CONCLUSION

In this chapter we have illustrated through several examples that time series in economics and business tend to display such features as trends, seasonality, aberrant observations, time-varying volatility and/or non-linearity. Before we can construct sensible out-of-sample forecasts of such time series, the relevant features should be described by a time series model that is specifically designed for that purpose. In Chapters 4 to 8, we will consider several models that are useful for describing a single specific time series feature. Needless to say, in practice such models usually need to be combined into a single model that is able to deal with several features jointly. The discussion in these five chapters is limited to univariate time series models. Before we turn to these specific models, in Chapter 3 we first give an overview of some of the key concepts in univariate time series analysis in general. Chapter 9 extends this Chapter 3 to multivariate time series models. The empirical time series already used in this introductory chapter will also be used in illustrative examples throughout the remainder of the book.

3 Useful concepts in univariate time series analysis

In this chapter we discuss several concepts that are useful for the analysis of time series in business and economics. Examples of such concepts are autoregressive moving average models, autocorrelation functions, parameter estimation, diagnostic measures, model selection, and forecasting. In this chapter these concepts are treated within the context of non-trending, non-seasonal, and linear univariate time series with constant variance. Although none of the five features highlighted in Chapter 2 will be dealt with explicitly, the above concepts are generally useful and often can be adapted to accommodate the apparent empirical features of the time series at hand.

The technical detail in this chapter is kept at a moderate level. The main focus is on explaining why the concepts are useful, how the relevant methods can be implemented in practice, and how the outcomes can be interpreted. It is not our intention to downplay the importance of formal asymptotic results for the techniques reviewed here, but our primary goal is to keep the discussion accessible, having in mind the actual use of these techniques in empirical time series modeling. We recommend the interested reader to consult more advanced time series textbooks such as Anderson (1971), Fuller (1976), Abraham and Ledolter (1983), Granger and Newbold (1986), Box *et al.* (1994), and Hamilton (1994). Also, there are many other concepts in time series analysis that we do not treat in detail here. The content of the present chapter merely reflects what should be a useful basis for modeling real-world time series, such as those discussed in the previous chapter. It is our experience that with the tools outlined in this chapter, it is possible to construct a time series model that is useful for forecasting and to understand how such models can be modified so that features such as seasonality and non-linearity, for example, can be incorporated.

Preliminaries

As before, the univariate time series of interest is denoted by y_t, where y_t can be $\log(w_t)$, with w_t being the originally observed time series. Observations on y_t are available for $t = 1, 2, \ldots, T$, that is, T denotes the sample size. The key aspect of a

time series, which contrasts with cross-sectional data, is that the y_t data are observed sequentially, that is, the observation y_{t-1} becomes available before the observation y_t, which in turn predates y_{t+1}. In particular, at time $t-1$ the observation at time t is yet unknown, while all observations up to and including $t-1$ are known. We denote the available history of a time series up to time $t-1$ as the set $\mathcal{Y}_{t-1} \equiv \{y_1, y_2, \ldots, y_{t-1}\}$. The fact that time series data are observed in a sequence suggests that there can be information in the set $\mathcal{Y}_{t-1}$ that can be exploited to explain or forecast y_t. For example, if the stock of motorcycles is 500,000 this year, and it was 480,000 last year, it is likely that next year's stock will be closer to say 520,000 than to 400,000.

An alternative and more formal way of stating the above is as follows. At time $t-1$, y_t is unknown and can be considered as a random variable. The idea that previous observations might contain useful information concerning the value of y_t implies that the relevant object of interest is the distribution of y_t conditional on the information set $\mathcal{Y}_{t-1}$, denoted as $f(y_t|\mathcal{Y}_{t-1})$. The main objective of time series modeling and forecasting is to characterize this conditional distribution. In particular, we often focus on the first two conditional moments of y_t, that is, the conditional mean $\mathsf{E}[y_t|\mathcal{Y}_{t-1}]$ and the conditional variance $\mathsf{V}[y_t|\mathcal{Y}_{t-1}]$. One of the reasons for considering the conditional mean and variance only is that these are sufficient to completely describe the properties of y_t in case the conditional distribution $f(y_t|\mathcal{Y}_{t-1})$ is normal. The models discussed in this book amount to particular specifications of the conditional mean $\mathsf{E}[y_t|\mathcal{Y}_{t-1}]$, with the conditional variance assumed to be constant in most cases. Models for $\mathsf{V}[y_t|\mathcal{Y}_{t-1}]$ that are relevant in case of conditional heteroskedasticity are discussed in Chapter 7.

The focus on the distribution of y_t conditional on $\mathcal{Y}_{t-1}$ reflects the fact that for a time series it makes sense to try and exploit past information to forecast the future. It might be, however, that there is no such information. In that case, where the observations y_{t-k} for $k = 1, 2, \ldots$, are not informative for the value of this variable at time t, the conditional distribution $f(y_t|\mathcal{Y}_{t-1})$ is identical to the unconditional distribution $f(y_t)$. In particular, the conditional and unconditional means of y_t are the same. If in addition the unconditional mean is equal to zero, such a time series is called a white noise time series. In this book a white noise time series will be denoted by ε_t. A more formal definition is given by the following three properties:

$$\mathsf{E}[\varepsilon_t] = 0 \qquad t = 1, 2, \ldots, T, \tag{3.1}$$

$$\mathsf{E}[\varepsilon_t^2] = \sigma^2 \qquad t = 1, 2, \ldots, T, \tag{3.2}$$

$$\mathsf{E}[\varepsilon_s \varepsilon_t] = 0 \qquad s, t = 1, 2, \ldots, T \text{ and } s \neq t. \tag{3.3}$$

In words, the mean of ε_t equals zero, all observations ε_t have the same variance σ^2 and there is no (linear) correlation between any past, current and future ε_t observations.

For many empirical time series y_t it however occurs that the observation at time t *does* depend on observations at $t-1$, $t-2$, and so on. Because y_t and y_{t-k} are observed k periods apart, we usually refer to k as the *time lag* and say that y_t depends

on its *lagged values* $y_{t-1}, y_{t-2}, \ldots$, or briefly, on its *lags*. In order to keep notation simple, it is convenient to write this lagging of y_t in terms of the so-called *lag operator* L. This operator is defined by

$$L^k y_t \equiv y_{t-k} \qquad \text{for } k = \ldots, -2, -1, 0, 1, 2, \ldots \tag{3.4}$$

Hence, $L^{-2} y_t = y_{t+2}$ and $L^0 y_t = y_t$, for example. Given a time series y_t, we may for example create a new time series that consists of the first differences, i.e. $y_t - y_{t-1} = (1 - L)y_t$. We call $(1 - L)$ then a *filter*. The algebra of the lag operator L is discussed in Dhrymes (1981, pages 19-24), among others. The algebra of lag polynomials $R(L)$, where $R(L)$ is defined as the set of all finite linear combinations of elements of the set $\{L^k;\ k = \ldots, -2, -1, 0, 1, 2, \ldots\}$ is isomorphic to the algebra of polynomial functions $R(z)$. An implication of this formal statement is that we can use L in typical polynomial operations such as products, ratios, and additions. For example, when $0 < \alpha < 1$, we can write

$$(1 - \alpha L)^{-1} = 1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots \tag{3.5}$$

Another example is

$$(1 + L^2)(1 - L^2) = 1 - L^4. \tag{3.6}$$

3.1 Autoregressive moving average models

Autoregressive [AR] model

Suppose that the observation of the time series y_t depends on its p most recent lags, that is, the conditional distribution of y_t can be written as $f(y_t|\mathcal{Y}_{t-1}) = f(y_t|y_{t-1}, \ldots, y_{t-p})$. If in addition we assume that the dependence is such that the conditional mean of y_t is a linear function of $y_{t-1}, \ldots, y_{t-p}$ while the conditional variance of y_t is constant, it follows that y_t can described by the linear model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \qquad t = p + 1, p + 2, \ldots, T, \tag{3.7}$$

where $\phi_1, \phi_2, \ldots, \phi_p$ are unknown parameters and ε_t is a white noise series as defined by (3.1)–(3.3). The ε_t series is not observed and has to be estimated from the data, based on the presupposed model for y_t. In (3.7), y_t is described by a regression model that includes only its own lagged observations, and hence this model is usually called an autoregressive model of order p [AR(p)], or an autoregression of order p. Using

the lag operator L, the model in (3.7) can be written more compactly as

$$\phi_p(L)y_t = \varepsilon_t, \tag{3.8}$$

where

$$\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p, \tag{3.9}$$

which is called the AR polynomial in L of order p. The weights on the lags are the parameters ϕ_1 to ϕ_p, and these express to what extent y_t depends on its own past.

The statement made above, that in the AR(p) model (3.7) the observation y_t is related to the p previous observations $y_{t-1}, \ldots, y_{t-p}$, is somewhat misleading. As the observation y_{t-p} is related to $y_{t-p-1}, \ldots, y_{t-2p}$ in the same way, there actually is dependence between y_t and *all* its past observations.

In order to be able to construct a meaningful forecast for future observations $y_{T+1}, y_{T+2}, \ldots$, using the AR(p) model in (3.7), it should first of all hold that this dependence on the past is constant. Indeed, if this dependence would randomly vary over time, there is no point in trying to forecast y_{T+h} as for any value h the relevant forecast function may differ. Furthermore, in order to make sensible statements about y_{T+h}, it should hold that the immediate past is more important than less recent observations. In other words, the impact of the observation at time $t = 10$, y_{10}, should be smaller for, say, y_{80} than for y_{11}. Note that we do not yet address seasonal time series here, where y_t can look more similar to y_{t-12} (in case of a monthly series) than to y_{t-1}. Due to (3.7), y_{10} involves a white noise observation ε_{10}, or, as commonly said, a *shock* at time $t = 10$, and hence we can also state that the impact of the shock at $t = 10$ should be less important for the observation y_{80} than for y_{11}. Similarly, the observation y_{80} should depend more on the shock occurring at $t = 79$ than on the shock occurring at $t = 10$.

In terms of (3.7), this can be stated somewhat more formally by expressing y_t as a function of all past shocks, as

$$y_t = [\phi_p(L)]^{-1}\varepsilon_t = \sum_{i=0}^{t-1} \theta_i \varepsilon_{t-i} + y_0, \tag{3.10}$$

where y_0 is a function of pre-sample starting values and the autoregressive parameters ϕ_1 to ϕ_p. The requirement that the effect of the shocks ε_{t-i}, $i = 0, 1, 2, \ldots$, on y_t becomes smaller as i becomes larger suggests that the values of θ_i should converge towards zero with increasing i. More precisely, the effect of shocks dies out if the condition $\sum_{i=1}^{\infty} |\theta_i| < \infty$ is satisfied.

To attach some interpretation to this statement, consider the first order autoregression

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \qquad t = 1, 2, \ldots, T. \tag{3.11}$$

Since (3.11) implies that

$$
\begin{aligned}
y_1 &= \phi_1 y_0 + \varepsilon_1, \\
y_2 &= \phi_1 y_1 + \varepsilon_2 = \phi_1 \phi_1 y_0 + \phi_1 \varepsilon_1 + \varepsilon_2, \\
y_3 &= \phi_1 y_2 + \varepsilon_3 = \phi_1(\phi_1 \phi_1 y_0 + \phi_1 \varepsilon_1 + \varepsilon_2) + \varepsilon_3, \\
&\vdots \\
y_t &= (\phi_1)^t y_0 + (\phi_1)^{t-1} \varepsilon_1 + (\phi_1)^{t-2} \varepsilon_2 + \cdots + \varepsilon_t,
\end{aligned}
$$

where y_0 is a pre-sample starting value, the AR(1) model in (3.11) can be written as

$$
y_t = (\phi_1)^t y_0 + \sum_{i=0}^{t-1} (\phi_1)^i \varepsilon_{t-i}, \qquad t = 1, 2, \ldots, T. \tag{3.12}
$$

Clearly, the parameters $(\phi_1)^i$ in (3.12) for the shocks ε_{t-i} converge to zero with increasing i when $|\phi_1| < 1$. In that case it also holds that $\sum_{i=0}^{\infty} (\phi_1)^i < \infty$.

Exercise 3.1

When ϕ_1 exceeds 1, (3.12) shows that the effect of shocks ε_{t-i} on y_t increases with i. In that case the time series y_t is called explosive, which is a feature that does not often occur in practice. In this book, we exclude the explosive case.

When ϕ_1 is exactly equal to 1, (3.12) simplifies to

$$
y_t = y_0 + \sum_{i=0}^{t-1} \varepsilon_{t-i}, \qquad t = 1, 2, \ldots, T, \tag{3.13}
$$

for which it is clear that the effects of, say, ε_{10} on, say, y_{11} and y_{80} are the same and both equal to 1. In general, a shock ε_t now has the same impact on all observations y_{t+h}, $h = 0, 1, 2, \ldots$, and shocks are said to have permanent effects. This means that if by chance a certain shock ε_t is very large and all subsequent shocks are small, the level of the time series after this time t changes quite dramatically relative to the situation that ε_t is small. This also suggests that a time series like (3.13) is highly unpredictable, as each future shock can have similar dramatic effects on the development of y_t. An AR(1) model with $\phi_1 = 1$ is therefore called a *random walk* model.

We illustrate the impact of the value of ϕ_1 on the pattern of y_t by simulating three time series with $T = 200$ observations from the AR(1) model (3.11) with different values of the parameter ϕ_1. The series, shown in Figure 3.1, are generated as follows. First, 200 independent numbers are generated from a standard normal distribution, denoted as $N(0, 1)$. These are white noise observations since they are drawn independently from a distribution with mean zero and with the same variance $\sigma^2 = 1$. As such they will

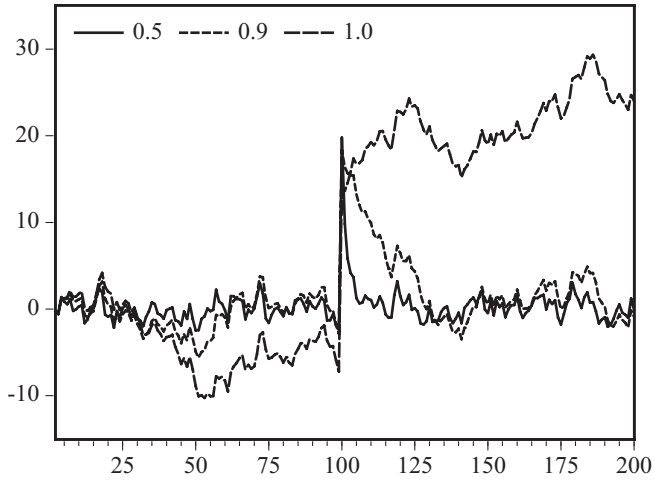


Figure 3.1: Simulated AR(1) time series with ϕ_1 equal to 0.5, 0.9 and 1.0.

be used as shocks ε_t. Next, we replace the observation at $t = 100$ ε_{100} by $\varepsilon_{100} + 20$, in order to illustrate the impact of a large shock on the time series y_t. Finally, y_1 is set equal to zero and the next 199 observations y_t for t running from 2 to 200 are generated by $y_t = \phi_1 y_{t-1} + \varepsilon_t$, where the AR-parameter ϕ_1 is set equal to 0.5, 0.9, and 1.0.

The solid line in Figure 3.1 corresponds with $\phi_1 = 0.5$. The effect of ε_{100} on y_t dies out fairly quickly, as can be observed from the fact that the time series quickly returns to its average level. This of course is not surprising given the representation of the AR(1) model in terms of lagged shocks in (3.12). This shows that the effect of the shock occurring at time t on y_{t+i} is equal to ϕ_1^i times ε_t, which declines to zero rapidly for $\phi_1 = 0.5$. The effect of the added value of 20 to ε_t at time 100 is 10 at $t = 101$, 5 at $t = 102$, and is negligible at, say, $t = 110$. This is not the case when $\phi_1 = 0.9$, as can be seen from the relevant time series in Figure 3.1. Indeed, it takes some time before $(0.9)^i$ times 20 becomes reasonably small. However, in both cases the time series display a tendency to go back to the mean level (also called *mean reversion*), or, in other words, shocks have only transitory effects. Finally, and most obvious from Figure 3.1, the time series does not return to its original average level after observation 100 when $\phi_1 = 1.0$. Hence, the effect of this shock seems to remain present in the time series, that is this shock seems to have a permanent effect. Again, this also follows from the representation of the AR(1) model given in (3.12). Of course, when large negative shocks occur after $t = 100$, the time series can go down again at a later stage.

In this chapter we limit the discussion to time series where shocks have only transitory effects. In case we suspect to have encountered a variable which displays permanent effects of shocks, we usually transform such a series to a time series with transitory effects by taking first differences $y_t - y_{t-1}$. This is closely related to the concept of stationarity, which will be discussed in detail in Chapter 4. We would then proceed with the analysis of $y_t - y_{t-1}$ instead of y_t. The motivation for taking first differences becomes clear from noting that in case $\phi_1 = 1$, the AR(1) model (3.11) is actually given by

$$y_t = y_{t-1} + \varepsilon_t. \tag{3.14}$$

From (3.14), it also follows immediately that the transformed time series $z_t = y_t - y_{t-1}$ can be described by the simple white noise model, that is

$$z_t = \varepsilon_t, \tag{3.15}$$

such that the past shocks ε_t have only transitory effects. In other words, while, say, ε_{100} can change the level of y_t after $t = 100$ permanently, it does not do so for the level of z_t in (3.15). Note again that when y_t is $\log(w_t)$, the z_t variable in (3.15) approximately equals the growth rate of w_t, and hence a single large shock may change the level of a series, but not so much the growth rate.

It sometimes may be the case that y_t needs to be differenced twice to obtain a similar result. Again, in order to keep notation simple, the differencing filter or differencing operator Δ_j that is often used, is defined by

$$\Delta_j^d = (1 - L^j)^d \qquad \text{for } d, j = \ldots, -2, -1, 0, 1, 2, \ldots \tag{3.16}$$

In practice we usually consider the cases $j = 1$ or S (with S being the number of seasons) and d equal to 0, 1, or 2. Notice that when d is 2, and $y_t = \log(w_t)$, the resultant second-order differenced series measures the change in the growth rate of w_t. In case a time series needs to be differenced d times, it is said to be integrated of order d, abbreviated as $I(d)$. When y_t is an $I(d)$ time series, and when it can be modeled with an AR(p) model after differencing it d times, the model for y_t can be written as

$$\Delta_1^d y_t - \phi_1 \Delta_1^d y_{t-1} - \cdots - \phi_p \Delta_1^d y_{t-p} = \varepsilon_t, \qquad t = p + d, p + d + 1, \ldots, T, \tag{3.17}$$

This model is usually abbreviated as an ARI(p, d) model.

Another terminology for permanent and transitory effects of shocks is based on the roots of the characteristic polynomial of an AR(p) model. The characteristic polynomial is nothing else than the lag polynomial $\phi_p(L)$ given in (3.9), but now considered as a function of z rather than the lag operator L, that is,

$$\phi_p(z) = 1 - \phi_1 z - \cdots - \phi_p z^p. \tag{3.18}$$

Its roots are the solutions to $\phi_p(z) = 0$. As an example, consider again the AR(1) model in (3.11). Its characteristic polynomial is given by

$$\phi_1(z) = 1 - \phi_1 z, \tag{3.19}$$

and its root is $z = (\phi_1)^{-1}$. When $\phi_1 = 1$, this solution equals 1, and in that case the AR(1) polynomial is said to have a unit root. When ϕ_1 is smaller than 1 in absolute value, the root of (3.19) exceeds 1. Since higher order AR(p) models may have complex roots, the solution to (3.19) is said to be outside the unit circle when $|\phi_1| < 1$.

A time series that can be described by an AR(p) model does not need to be differenced when all p solutions to its characteristic polynomial $\phi_p(z)$ as given in (3.18) are outside the unit circle, see Fuller (1976). In practice this can be hard to verify, especially because the parameters ϕ_i, $i = 1, 2, \ldots, p$, have to be estimated from the available data. Since we are mainly interested in knowing whether we need to difference or not, we usually check only whether one or more of the solutions of the characteristic polynomial are exactly equal to 1. For example, the characteristic polynomial of the ARI(p, d) model (3.17), as given by

$$(1 - z)^d - \phi_1(1 - z)^d z - \cdots - \phi_p(1 - z)^d z^p, \tag{3.20}$$

clearly has (at least) d roots $z = 1$. In practice an estimate for d needs to be obtained, and it appears that this is not easy. In Chapter 4, we discuss testing procedures for determining the number of unit roots.

Autoregressive moving average [ARMA] model

One of the crucial assumptions in the AR(p) model (3.7) is that the shocks ε_t are independent and identically distributed for all t, or at least that they are uncorrelated. In practice, this is usually checked by examining whether the residuals of the corresponding regression display the presumed white noise properties in (3.1)–(3.3), see Section 3.3 for a detailed discussion. It may well occur that the lag order p required to satisfy this requirement is rather large, that is, we need to include a large number of lags of y_t to fully capture the autocorrelation properties of the time series. When p increases, the number of unknown parameters to be estimated in (3.7) increases as well. Making use of the properties of the lag operator L as in (3.5) and (3.6), it is now possible to approximate a lengthy AR polynomial by a ratio of two polynomials which together involve fewer parameters. Stated otherwise, we may consider approximating the $\phi_p(L)$ polynomial in (3.8) by the ratio of a different $\phi_p(L)$ polynomial (with this p smaller than the previous one) and another polynomial $\theta_q(L)$. The resultant time series model is

$$\frac{\phi_p(L)}{\theta_q(L)} y_t = \varepsilon_t,$$

or

$$\phi_p(L)y_t = \theta_q(L)\varepsilon_t, \qquad t = p+1, p+2, \ldots, T, \tag{3.21}$$

with

$$\phi_p(L) = 1 - \phi_1 L - \cdots - \phi_p L^p,$$
$$\theta_q(L) = 1 + \theta_1 L + \cdots + \theta_q L^q,$$

where this notational convention is chosen such that the model in (3.21) is the regression model

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \tag{3.22}$$

This model is called an autoregressive moving average model of order (p, q) or, briefly, ARMA(p, q). When the y_t series needs to be differenced d times, such that y_t is replaced by $\Delta_1^d y_t$ in (3.22), we say that y_t is described by an autoregressive integrated moving average model of order (p, d, q) [ARIMA(p, d, q)]. It is exactly this class of univariate time series models that became very popular among practitioners through the seminal work of Box and Jenkins (1970). The label moving average is assigned to this model as the right-hand side of (3.21) mimics a moving average of ε_t terms. In addition to being a parsimonious approximation to a high-order AR(p) model, ARMA models also may arise through either temporal or cross-sectional aggregation of time series.

Exercise 3.2–3.4

It should be mentioned that for many practical purposes we consider the ARI model instead of the ARIMA model. The main reasons for this preference are that the parameters in ARI models can be easily estimated (see Section 3.3), that diagnostic measures can be easily computed (see also Section 3.3), and that such ARI(p, d) models can be easily extended to allow for seasonality, shifts in mean or trends, and non-linearity.

Moving average [MA] model

In some practical cases it may however be convenient to consider a simplified version of an ARMA(p, q) model, that is, the MA(q) model given by

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \tag{3.23}$$

or

$$y_t = \theta_q(L)\varepsilon_t, \tag{3.24}$$

with

$$\theta_q(L) = 1 + \theta_1 L + \cdots + \theta_q L^q. \tag{3.25}$$

An important feature of the MA(q) model (and hence also of the ARMA(p, q) model) is that the explanatory variables in (3.23), that is, ε_{t-1} to ε_{t-q}, are unobserved and have to be estimated using the available data. In order to avoid estimation problems that may arise because of this, the MA order q is usually kept quite small. In practice, this q is often set at 1 or 2 (or $S = 4$ or 12 in case of seasonal time series).

At first sight it may seem that y_t does not depend on its own past when an MA(q) model describes this variable. However, similar to (3.10), we may rewrite (3.24) as an AR model of infinite order,

$$[\theta_q(L)]^{-1} y_t = \varepsilon_t, \tag{3.26}$$

which shows that y_t in fact depends on all its previous values. For example, for the MA(1) model

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}, \tag{3.27}$$

it follows that (assuming that ε_0 is 0)

$$\begin{aligned}
y_1 &= \varepsilon_1, \\
y_2 &= \varepsilon_2 + \theta_1 \varepsilon_1 = \varepsilon_2 + \theta_1 y_1, \\
y_3 &= \varepsilon_3 + \theta_1 \varepsilon_2 = \varepsilon_3 + \theta_1 (y_2 - \theta_1 y_1), \\
y_4 &= \varepsilon_4 + \theta_1 \varepsilon_3 = \varepsilon_4 + \theta_1 (y_3 - \theta_1 (y_2 - \theta_1 y_1)), \\
&\vdots
\end{aligned}$$

or, in general,

$$y_t = \varepsilon_t + \sum_{i=1}^{t-1} (-1)^{i-1} \theta_1^i y_{t-i}. \tag{3.28}$$

This expression will become useful when estimating θ_1, as to be discussed in Section 3.3.

Similar to the concept of a unit root in the AR(p) polynomial (3.9), the MA(q) polynomial (3.25) may also contain one or more unit roots. In case of the MA(1) model (3.27), the characteristic polynomial given by

$$1 + \theta_1 z, \tag{3.29}$$

contains a unit root when $\theta_1 = -1$. If so, the MA(1) model is called non-invertible. Intuitively, invertibility of the MA(q) model means that the values of the shocks ε_t can be recovered from the observed time series y_t. In principle, this can be done by

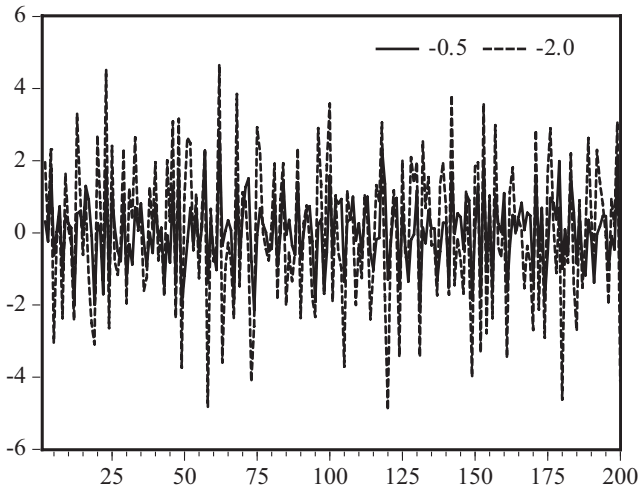


Figure 3.2: Simulated MA(1) time series with θ_1 equal to -0.5 and -2.0.

writing the model in the equivalent AR-form as in (3.26). A crucial requirement for invertibility however is that the resulting AR-coefficients converge towards zero or, stated more precisely, that $\sum_{i=0}^{\infty} |\phi_i| < \infty$. It turns out that this condition holds only if the roots of the characteristic MA-polynomial

$$\theta_q(z) = 1 + \theta_1 z + \cdots + \theta_q z^q, \tag{3.30}$$

are all outside the unit circle. For example, in the MA(1) model the coefficients in the equivalent AR-representation are given by $\phi_i = (-1)^i \theta_1^i$, which follows from (3.28). The sum of the absolute values of ϕ_i is finite only when $|\theta_1| < 1$.

Two simulated realizations of MA(1) processes with θ_1 equal to -0.5 and -2.0 are displayed in Figure 3.2, where the 200 observations on ε_t are drawn from the same $N(0,1)$ distribution.

Clearly, the impact of the values of the MA parameter on the time series pattern is less clear-cut as in case of AR models. The key difference between the MA model with $\theta_1 = -0.5$ and -2.0 is that the latter seems to display more reaction to large values of one period lagged values of ε_t, that is, its variance is larger. Given that the impact of shocks ε_t becomes zero after one period by construction (as the series are generated from an MA model of order 1), large shocks do not tend to change the level of the time series permanently. Comparing Figure 3.2 with Figure 3.1 clearly shows that the impact of shocks is a much less important topic to study for MA models.

It may still be of interest to examine if $\theta_1 = -1$, as the presence of a unit root in the MA polynomial usually is an indication of overdifferencing. For example, consider the ARMA(1,1) model with $\phi_1 = 1$,

$$y_t - y_{t-1} = \varepsilon_t + \theta_1 \varepsilon_{t-1}. \tag{3.31}$$

In case it holds that $\theta_1 = -1$, y_t has been erroneously differenced once too often as the polynomial (1-L) cancels from both sides. Formal tests for overdifferencing are derived in Breitung (1994), Franses (1995), and Tsay (1993), among others. In theory, we want to difference a time series until the resultant time series can be described by an ARMA(p, q) model with neither of the polynomials $\phi_p(L)$ and $\theta_q(L)$ containing the component $(1 - L)$.

Mean of time series and intercept in model

So far it has been assumed implicitly that the unconditional mean μ of y_t is equal to 0. In case of a known $\mu \neq 0$, y_t can be replaced by $y_t - \mu$ in the above expressions, which implies that its mean is subtracted before the analysis. For example, the AR(1) model (3.11) can then be written as

$$(y_t - \mu) = \phi_1(y_{t-1} - \mu) + \varepsilon_t. \tag{3.32}$$

In practice, however, the unconditional mean μ is unknown and needs to be estimated from the data. A simple method to achieve this is to rewrite for example (3.32) as

$$y_t = (1 - \phi_1)\mu + \phi_1 y_{t-1} + \varepsilon_t, \tag{3.33}$$

or

$$y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t, \tag{3.34}$$

where $\alpha = (1 - \phi_1)\mu$. Note that this essentially modifies the regression model (3.11) by including an intercept term α. For the general AR(p) model this regression becomes

$$\phi_p(L)y_t = \alpha + \varepsilon_t. \tag{3.35}$$

The unconditional mean μ of y_t can now be determined by multiplying (3.35) with $\phi_p(L)^{-1}$ on the left and right and taking expectations, which renders

$$\mu = [\phi_p(L)]^{-1}\alpha \tag{3.36}$$

$$= [\phi_p(1)]^{-1}\alpha \tag{3.37}$$

$$= \frac{\alpha}{1 - \phi_1 - \cdots - \phi_p}, \tag{3.38}$$

where replacing L by 1 in $\phi_p(L)$ is done because $L^k\alpha = \alpha$ for any value of k. The inclusion of an intercept in an AR regression is quite important in case of trending

variables, as we will see in the next chapter (and also in Chapter 9 where common trends will be discussed).

In case of an MA(q) model with an intercept, that is

$$y_t = \alpha + \theta_q(L)\varepsilon_t, \tag{3.39}$$

the intercept α corresponds with the mean μ as ε_t is a zero mean series.

When including an intercept term in an ARMA model, it may occur that the t-ratio of the estimated value of α suggests that it is insignificant and hence that α could be deleted from the model. The expression for the AR(1) model in (3.33) shows, however, that deleting α may not always be sensible. In fact, when μ is not zero, but $\alpha = (1 - \phi_1)\mu$ is imposed to be zero, this restriction assumes ϕ_1 is equal to 1, while it might not be. As a consequence, deleting an intercept biases estimates of ϕ_1 towards 1. With the estimation methods to be outlined in Section 3.3, we cane asily verify this by generating data with $\mu = 10$ and $\phi_1 = 0.5$, which yield an estimate $\hat{\phi}_1 \approx 1$ when the AR(1) regression does not include an intercept. For many practical applications, it is therefore better to include an intercept in the estimation model, even though it is not significantly different from zero.

Exercise 3.6

3.2 Autocorrelation and identification

The ARMA model discussed in the previous section has an important feature that makes it distinct from many other econometric models. This property is that the ability of an ARMA model to describe a certain time series can be "recognized" by specific features of the actual data. These features are the so-called autocorrelations and partial autocorrelations. The process of recognizing a possibly appropriate model is called *identification*, see Box and Jenkins (1970). The idea is that if a time series is best described by an ARMA(p,q) model, it should display autocorrelation properties that correspond to those of that particular model. In practice, the orders p and q of the ARMA model are unknown and have to be estimated from the data (see Section 3.4). This can be done by computing the empirical autocorrelations and partial autocorrelations to see whether these match certain patterns implied by different ARMA models. In this section, we discuss the autocorrelation and partial autocorrelation functions in detail, and illustrate how these can be used to identify some simple ARMA time series models.

Autocorrelation

The k-th order autocorrelation of a time series y_t is defined by

$$\rho_k = \gamma_k / \gamma_0, \tag{3.40}$$

where γ_k is the k-th order autocovariance of y_t, that is,

$$\gamma_k = \mathsf{E}[(y_t - \mathsf{E}[y_t])(y_{t-k} - \mathsf{E}[y_{t-k}])], \qquad k = \ldots, -2, -1, 0, 1, 2, \ldots \tag{3.41}$$

Given (3.41), it is clear that $\rho_0 = 1$, $\rho_{-k} = \rho_k$ and that $-1 < \rho_k < 1$ for all k. The collection of all autocorrelations ρ_k, $k = 0, 1, 2, \ldots$, is called the autocorrelation function [ACF].

This ACF can be useful to characterize ARMA time series models. A simple example is the white noise series ε_t for which $\rho_k = 0$ for all $k \neq 0$. As another example, consider the AR(1) model

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \varepsilon_t, \qquad t = 2, 3, \ldots, n. \tag{3.42}$$

As discussed in the previous section, when $|\phi_1| < 1$, μ in (3.42) is the unconditional mean of y_t, which can easily checked by taking expectations of the left- and right-hand sides of (3.42) and assuming that $\mathsf{E}[y_t] = \mathsf{E}[y_{t-1}]$.

Exercise 3.7

In order to calculate the ACF, we start with the variance

$$\gamma_0 = \mathsf{E}[(y_t - \mathsf{E}[y_t])(y_t - \mathsf{E}[y_t])]. \tag{3.43}$$

For the AR(1) model, the right-hand side of (3.43) is equal to

$$\begin{aligned}\mathsf{E}[(y_t - \mu)(y_t - \mu)] = \mathsf{E}[\phi_1(y_{t-1} - \mu)\phi_1(y_{t-1} - \mu)] + \mathsf{E}[\varepsilon_t^2] \\ + 2\mathsf{E}[\phi_1(y_{t-1} - \mu)\varepsilon_t].\end{aligned} \tag{3.44}$$

In order to solve (3.44) for the variance of y_t, we make use of a number of general results on covariances between the time series y_t and the shocks ε_t. First, consider again (3.10), where the AR(1) model is written as

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + y_0, \tag{3.45}$$

where the parameters are scaled by θ_0 (which is equal to 1 anyway). Its one-period lagged version is

$$y_{t-1} = \varepsilon_{t-1} + \theta_1\varepsilon_{t-2} + \theta_2\varepsilon_{t-3} + \cdots + y_0, \tag{3.46}$$

From (3.46), it is clear that $\mathsf{E}[y_{t-1}\varepsilon_t] = 0$. In fact, moving (3.46) even further back in time, it becomes evident that $\mathsf{E}[y_{t-j}\varepsilon_t] = 0$ for all $j > 0$. Second, from (3.45) we can see that $\mathsf{E}[y_t\varepsilon_t] = E[\varepsilon_t^2] = \sigma^2$ for all t. Third and finally, the covariance of μ with a time series is of course equal to zero. Combining these results, and making use of the fact that for a stationary time series the unconditional variance is constant over time, (3.44) becomes

$$\gamma_0 = (1 - \phi_1^2)^{-1}\sigma^2, \qquad \text{when } |\phi_1| < 1. \tag{3.47}$$

The first order autocovariance for an AR(1) time series is derived as:

$$\begin{aligned}\gamma_1 &= \mathsf{E}[(y_t - \mu)(y_{t-1} - \mu)] \\ &= \mathsf{E}[\phi_1(y_{t-1} - \mu)(y_{t-1} - \mu)] + \mathsf{E}[\varepsilon_t(y_{t-1} - \mu)] \\ &= \phi_1\gamma_0.\end{aligned} \tag{3.48}$$

Hence the first-order autocorrelation ρ_1 for the AR(1) model becomes

$$\rho_1 = \gamma_1/\gamma_0 = \phi_1. \tag{3.49}$$

To calculate ρ_k, it is convenient to consider the following relationship for an AR(1) model

$$\mathsf{E}[(y_t - \mu)(y_{t-k} - \mu)] = \mathsf{E}[\phi_1(y_{t-1} - \mu)(y_{t-k} - \mu)], \tag{3.50}$$

or $\gamma_k = \phi_1\gamma_{k-1}$, where we again used that $\mathsf{E}[y_{t-k}\varepsilon_t] = 0$. Dividing both sides of $\gamma_k = \phi_1\gamma_{k-1}$ by γ_0 results in a recursive relationship for the autocorrelations:

$$\rho_k = \phi_1\rho_{k-1} \qquad \text{for } k = 1, 2, 3, \ldots \tag{3.51}$$

The autocorrelations of an AR(1) model with $|\phi_1| < 1$ thus decline exponentially towards zero. For example, when $\phi_1 = 0.8$, the first four (theoretical) autocorrelations are 0.8, 0.64, 0.512 and 0.4096. In practice we can estimate such correlations for real data, see below. We can then examine whether the empirical autocorrelations display the theoretical pattern. If so, we can consider an AR(1) model for the time series at hand.

The derivations for the AR(1) model above are valid only when the condition $|\phi_1| < 1$ is satisfied. Earlier we saw that $\phi_1 = 1$ is a special case, as it implies random walk behavior of the time series y_t, in the sense that the AR(1) polynomial has a unit root and shocks ε_t have permanent effects. To demonstrate its consequences for the ACF, it is convenient to rewrite the random walk model

$$y_t = y_{t-1} + \varepsilon_t, \tag{3.52}$$

as

$$y_t = \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \cdots + \varepsilon_2 + \varepsilon_1 + y_0, \qquad t = 1, 2, \ldots, T. \tag{3.53}$$

As $\mathsf{E}[\varepsilon_t] = 0$ for all t, and $\mathsf{E}[\varepsilon_t \varepsilon_{t-k}] = 0$ for all $k \neq 0$, and assuming that y_0 is a fixed constant, which we set equal to 0 for convenience, it follows that the unconditional mean $\mathsf{E}[y_t] = 0$ for all t. By contrast, the unconditional variance and auto-covariances of y_t are no longer constant. From (3.53), it follows that

$$\gamma_{0,t} = \mathsf{E}[y_t^2] = t\sigma^2, \tag{3.54}$$

where the additional subscript t indicates that the value of the variance depends on t. In fact, the variance increases linearly with t. Comparing (3.53) with its one-period lagged version

$$y_{t-1} = \varepsilon_{t-1} + \varepsilon_{t-2} + \cdots \varepsilon_2 + \varepsilon_1 + y_0, \tag{3.55}$$

we find that

$$\gamma_{1,t} = \mathsf{E}[y_t y_{t-1}] = (t-1)\sigma^2, \tag{3.56}$$

which, together with (3.54), results in

$$\rho_{1,t} = \frac{t-1}{t}. \tag{3.57}$$

Similarly, it follows that in general $\rho_{k,t} = (t-k)/t$ for any $k > 0$. Results similar to (3.57) can be derived for any AR(p) model that has a factor $(1-L)$ in its AR polynomial $\phi_p(L)$ by replacing y_t in (3.52) by the filtered $\phi_{p-1}(L)y_t$ series, where $\phi_p(L) = \phi_{p-1}(L)(1-L)$.

Because of the time-varying nature of the autocorrelations, this ACF is not of much use to yield information on a possibly adequate ARMA model. In addition, note that all autocorrelations $\rho_{k,t}$ tend to the value 1 as t increases. In other words, the ACF is not interpretable for AR models with unit roots. Hence, from now on it is assumed in this chapter that if there is such an $(1-L)$ component, it has been removed by filtering the time series with the Δ_1 filter. To keep notation simple we will continue to use y_t for the appropriately transformed data. Also, and for similar reasons, we set μ equal to zero from now on (except when otherwise indicated).

In principle, determining the autocorrelations for higher order autoregressive models proceeds along similar lines as demonstrated for the AR(1) model. For example, consider the AR(2) model written conveniently as

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} = \varepsilon_t. \tag{3.58}$$

Multiplying both sides by y_{t-1}, taking expectations, and dividing by γ_0 results in

$$\rho_1 - \phi_1 \rho_0 - \phi_2 \rho_1 = 0. \tag{3.59}$$

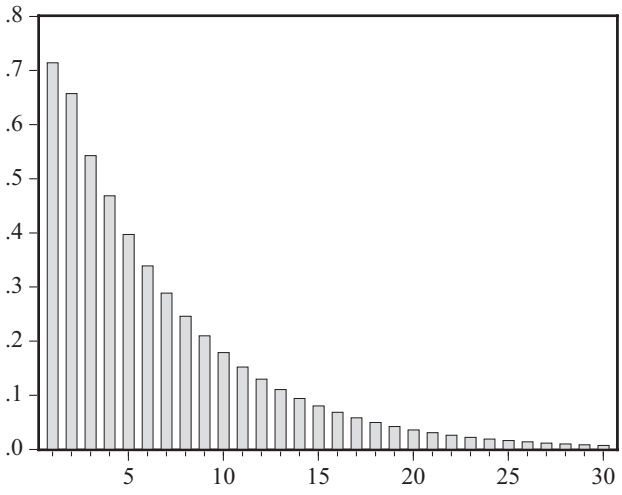


Figure 3.3: Theoretical autocorrelation function of an AR(2) process with $\phi_1 = 0.5$ and $\phi_2 = 0.3$.

As $\rho_0 = 1$, we first obtain that

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}. \tag{3.60}$$

To obtain an expression for ρ_2, analogous operations are carried out on (3.58) except for using y_{t-2} instead of y_{t-1}, yielding

$$\rho_2 - \phi_1\rho_1 - \phi_2\rho_0 = 0. \tag{3.61}$$

Substituting (3.60) into (3.61) gives

$$\rho_2 = \frac{\phi_1^2}{1 - \phi_2} + \phi_2. \tag{3.62}$$

Analogous to (3.59) and (3.61) we can derive that in general

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2}. \qquad \text{for } k = 2, 3, 4, \ldots \tag{3.63}$$

To find expressions for the ACF for AR(p) models with $p > 2$, we can use the same techniques as above.

In general it holds that the ACF of an AR process shows an exponentially decaying pattern. Consider for example the theoretical autocorrelation function of an AR(2) model as (3.58) with parameters $\phi_1 = 0.5$ and $\phi_2 = 0.3$ (in Figure 3.3) and $\phi_1 = 0.5$ and $\phi_2 = 0.49$ (in Figure 3.4). Both graphs show the ACF up to order $k = 30$. Figure 3.3 clearly shows the exponential decay in the values of the autocorrelations ρ_k when k

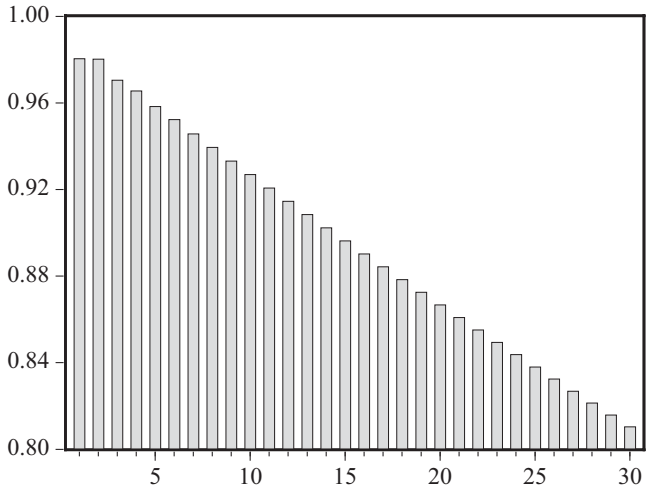


Figure 3.4: Theoretical autocorrelation function of an AR(2) process with $\phi_1 = 0.5$ and $\phi_2 = 0.49$.

increases, where ρ_{30} is indeed very close to zero. On the other hand, when ϕ_2 is increased from 0.3 to 0.49, the ACF is Figure 3.4 shows that the ACF does not decline quickly at all. In fact, ρ_{30} still exceeds 0.8 (note the scale on the vertical axis). This feature reflects the fact that the AR(2) polynomial $1 - 0.5z - 0.49z^2$ is almost equal to $(1 - z)(1 + 0.5z)$, with the latter containing the $(1 - z)$ unit root component. Recall that, in the presence of a unit root, the expression for $\rho_{k,t}$ given before suggests that their values are close to 1 for all k. Hence, a first tentative indication that a time series y_t can be described by an ARMA model of which the AR part contains the $(1 - L)$ component, which should be removed by applying the first differencing filter Δ_1, is given by a very slow decay of the ACF. Box and Jenkins (1970) in fact recommend to use this visual evidence as a tool to decide upon applying the Δ_1 filter or not. At present there exist statistical testing procedures that enable us to make a more formal decision about the presence of a unit root or not, see Chapter 4.

The roots of the characteristic polynomial $1 - \phi_1 z - \phi_2 z^2 = 0$ of an AR(2) model are given by $z_{1,2} = (\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2})/2\phi_2$. In case $\phi_1^2 + 4\phi_2 < 0$ these roots are complex, that is, $z_{1,2} = a \pm bi$, where i is the imaginary number defined by $i^2 = -1$. This results in a cyclical pattern of the ACF. The corresponding time series y_t displays a cyclical pattern with cycle length

$$c = 2\pi/[\tan^{-1}(b/a)], \tag{3.64}$$

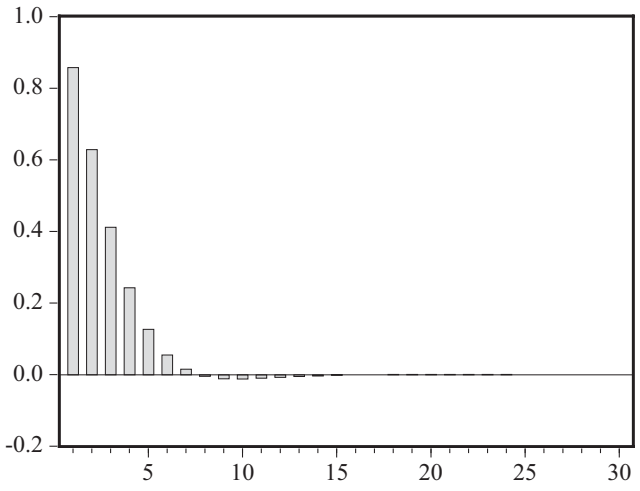


Figure 3.5: Theoretical autocorrelation function of an AR(2) process with $\phi_1 = 1.2$ and $\phi_2 = -0.4$.

which is a result that follows from standard differential calculus. This cyclical pattern can be illustrated by the ACF values of AR(2) models with $\phi_1 = 1.2$ and $\phi_2 = -0.4$ (in Figure 3.5) and with $\phi_1 = 1.0$ and $\phi_2 = -0.5$ (in Figure 3.6).

With (3.64) we can show that c is larger for the series with the ACF as in Figure 3.5 than for that in Figure 3.6. From Figure 3.6 it can be observed that positive and negative peaks in the ACF occur at lags 4, 8, 12, and so on. As the solutions to the corresponding characteristic polynomial $1 - z + 0.5z^2$ are equal to $z_{1,2} = 1 \pm i$, (3.64) indicates that c is indeed exactly equal to 8.

Figures 3.3 to 3.6 show that the values of ρ_k for $k = 3, 4, 5, \ldots$ can still be quite large for an AR(2) model. Hence, the ACF may not be particularly useful to identify whether an AR model of specific order is suitable. In fact, the ACF is more useful in case of MA(q) models. Consider for example the MA(2) model

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}, \tag{3.65}$$

and correspondingly

$$y_{t-k} = \varepsilon_{t-k} + \theta_1 \varepsilon_{t-k-1} + \theta_2 \varepsilon_{t-k-2}. \tag{3.66}$$

As all covariances between ε_t and its lags are equal to zero, the variance γ_0 equals

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)\sigma^2. \tag{3.67}$$

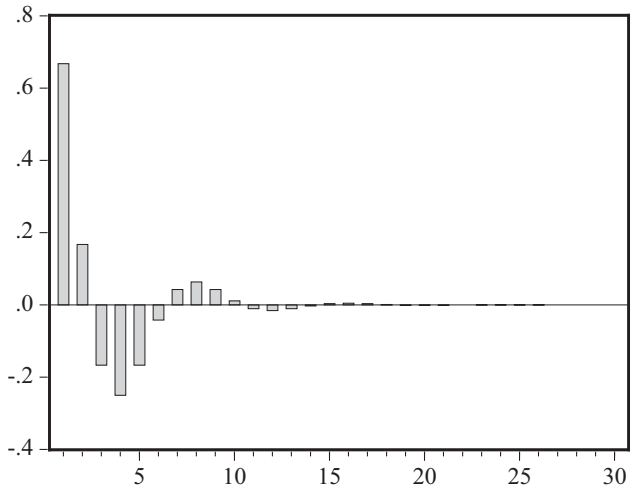


Figure 3.6: Theoretical autocorrelation function of an AR(2) process with $\phi_1 = 1.0$ and $\phi_2 = -0.5$.

With (3.65) and (3.66) we can see that

$$\gamma_1 = \mathsf{E}[y_t y_{t-1}] = (\theta_1 + \theta_1\theta_2)\sigma^2, \tag{3.68}$$

$$\gamma_2 = \mathsf{E}[y_t y_{t-2}] = \theta_2\sigma^2, \tag{3.69}$$

$$\gamma_k = 0 \qquad \text{for } k = 3, 4, \ldots, \tag{3.70}$$

and hence that $\rho_k = 0$ for all $k > 2$. In general, for an MA(q) model it holds that

$$\gamma_k = \begin{cases} \left[\sum_{i=0}^{q-k} \theta_i\theta_{i+k}\right]\sigma^2 & \text{for } k = 0, 1, \ldots, q, \\ 0 & \text{for } k > q, \end{cases} \tag{3.71}$$

with $\theta_0 = 1$.

The above implies that when in practice the empirical ACF [EACF] is available, and the values are zero after the q-th lag, we may decide to tentatively select an MA(q) model for y_t for further analysis.

For ARMA(p, q) models the pattern of the ACF is a mixture of the ACF patterns for pure AR and MA models. For example, consider the ARMA(1,1) model

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1\varepsilon_{t-1}, \tag{3.72}$$

for which we can derive (along similar lines as above) that

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \sigma^2 + \theta_1 \mathsf{E}[y_t\varepsilon_{t-1}] \\ &= \phi_1\gamma_1 + [1 + \theta_1(\phi_1 + \theta_1)]\sigma^2, \\ \gamma_1 &= \phi_1\gamma_0 + \theta_1\sigma^2, \\ \gamma_2 &= \phi_1\gamma_1, \\ \gamma_k &= \phi_1\gamma_{k-1} \qquad \text{for } k = 3, 4, 5, \ldots\end{aligned}$$

such that (after some algebra)

$$\rho_k = \frac{\phi_1^{k-1}(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{(1 + 2\phi_1\theta_1 + \theta_1^2)}, \qquad \text{for } k = 1, 2, 3, \ldots \tag{3.73}$$

From this expression it can be seen that ρ_k can take a wide variety of values for distinct choices of ϕ_1 and θ_1. This suggests that the identification of an ARMA time series model from the patterns of the ACF alone may be rather difficult. Notice that when $\phi_1 = -\theta_1$, the autocorrelations ρ_k become equal to 0 for all $k \neq 0$ as (3.72) collapses to a white noise model.

Exercise 3.8–3.9

Partial autocorrelation

The ACF is helpful to identify that an MA model of some order q is possibly useful to describe y_t because $\rho_k = 0$ for all $k > q$ as shown by (3.71). The ACF appears less useful to identify AR(MA) models. The AR(2) examples in Figures 3.3–3.6 suggest that the ACF can take a wide variety of patterns depending on the specific values of the parameters ϕ_1 and ϕ_2. In any case, the autocorrelations ρ_k do not become equal to zero after a specific lag order k. To understand the reason for this, consider the AR(1) model

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \tag{3.74}$$

which can be written as

$$y_t = \phi_1^2 y_{t-2} + \varepsilon_t + \phi_1\varepsilon_{t-1}. \tag{3.75}$$

This shows that the inclusion of y_{t-1} in a regression model for y_t implies that y_t also depends on y_{t-2} (though the dependence is slightly weaker as $|\phi_1^2| < |\phi_1|$ given that we require $|\phi_1| < 1$). This can of course also be observed from the expression for the autocorrelations $\rho_k = \phi_1^k$, which follows from (3.49) and (3.51). Now, what is helpful to identify an AR model is to notice that adding y_{t-2} to the regression

(3.74), which already includes y_{t-1}, would not help in explaining y_t, that is, the corresponding parameter should equal zero. Loosely speaking, the so-called k-th order partial autocorrelation measures the dependence or correlation between y_t and y_{t-k}, after their common dependence on the intermediate observations $y_{t-1}, \ldots, y_{t-k+1}$ has been removed. More precisely, the k-th order partial autocorrelation is defined as the coeffient ψ_k in the regression

$$y_t = \eta_1 y_{t-1} + \eta_2 y_{t-2} + \cdots + \eta_{k-1} y_{t-k+1} + \psi_k y_{t-k} + u_t. \tag{3.76}$$

Along these lines we can construct the complete partial autocorrelation function [PACF]. The first-order partial autocorrelation is given by the value of ψ_1 in

$$y_t = \psi_1 y_{t-1} + u_t, \tag{3.77}$$

where u_t would be a white noise error time series when the process generating y_t is indeed an AR(1). From (3.77) it follows that ψ_1 equals γ_1/γ_0, such that by construction we have that $\psi_1 = \rho_1$ for all time series models. The second partial autocorrelation ψ_2 can be obtained from the regression model

$$y_t = \eta_1 y_{t-1} + \psi_2 y_{t-2} + u_t. \tag{3.78}$$

Exercise 3.10

In case of an AR(1) time series, ψ_2 equals zero. In case of an AR(2) or higher, it is different from zero. Similarly, for an AR(2) series, it holds that $\psi_3 = 0$ in the regression

$$y_t = \eta_1 y_{t-1} + \eta_2 y_{t-2} + \psi_3 y_{t-3} + u_t. \tag{3.79}$$

Hence, in general when ψ_{p+1} equals zero, while ψ_p does not, we may wish to consider an AR model of order p. More formal derivations of the PACF, where it is also shown that the ψ_k can be written as functions of the ρ_k, can be found in Box and Jenkins (1970).

Overdifferencing

As shown in Chapter 2, many economic time series display trending behavior, and, as will become clear in Chapter 4, should be differenced using the Δ_1 filter to remove the $(1 - L)$ component in the AR polynomial. It may however be that we make a mistake and erroneously apply the Δ_1 filter once too often such that the resultant time series is overdifferenced. For example, when the white noise series $y_t = \varepsilon_t$ is differenced, the model for $y_t - y_{t-1}$ becomes

$$y_t - y_{t-1} = \varepsilon_t + \theta_1 \varepsilon_{t-1} \qquad \text{with } \theta_1 = -1. \tag{3.80}$$

Defining $z_t = \Delta_1 y_t$, the first order autocorrelation of z_t when $\theta = -1$ equals $\rho_1 = -0.5$, which follows from combining (3.71) with $k = q = 1$ with the fact that $\gamma_0 = 2\sigma^2$ in this case. In general, it can be shown that overdifferencing results in a typical pattern of the ACF. Suppose the autocovariances of the series y_t are denoted as γ_k and those of $\Delta_1 y_t$ as γ_k^*, then

$$\gamma_k^* = \mathsf{E}[(y_t - y_{t-1})(y_{t-k} - y_{t-k-1})] = 2\gamma_k - \gamma_{k-1} - \gamma_{k+1}. \tag{3.81}$$

Given this connection between γ_k^* and γ_j, $j = k-1, k, k+1$, it follows that

$$\sum_{i=1}^{\infty} \rho_i^* = \frac{\gamma_1 - \gamma_0}{2\gamma_0 - 2\gamma_1} = -0.5, \tag{3.82}$$

where ρ_i^* is the i-th order autocorrelation of $\Delta_1 y_t$. In sum, if we consider differencing the series y_t, for example because its ACF dies out only very slowly, it is useful to examine the ACF of the first-differenced series. In case its values sum up to about -0.5, we may take this as evidence that we have differenced once too often.

Exercise 3.11

Empirical (partial) autocorrelation functions

In practice, for a given economic or business time series y_t, the autocorrelation and partial autocorrelation functions have to be estimated. The k-th order autocorrelation can be estimated by means of

$$\hat{\rho}_k = \hat{\gamma}_k / \hat{\gamma}_0, \tag{3.83}$$

where $\hat{\gamma}_k$ is an estimate of the k-th order autocovariance, that is

$$\hat{\gamma}_k = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y}), \tag{3.84}$$

where $\bar{y}$ denotes the sample mean of y_t, $t = 1, 2, 3, \ldots, T$. The $\hat{\rho}_k$ for $k = 0, 1, 2 \ldots$ form the empirical ACF [EACF]. As an illustration, consider $\hat{\rho}_k$ for $k = 1, \ldots, 20$ for annual differences of the log monthly US industrial production for the period 1959–2012, as shown in Figure 3.7. It is clear from this graph that the EACF values dies out quite quickly.

The sample equivalents of ψ_k, which form the empirical partial ACF [EPACF], can be obtained by applying ordinary least squares [OLS] to

$$y_t - \bar{y} = \eta_1(y_{t-1} - \bar{y}) + \cdots + \eta_{k-1}(y_{t-k+1} - \bar{y}) + \psi_k(y_{t-k} - \bar{y}) + v_t, \tag{3.85}$$

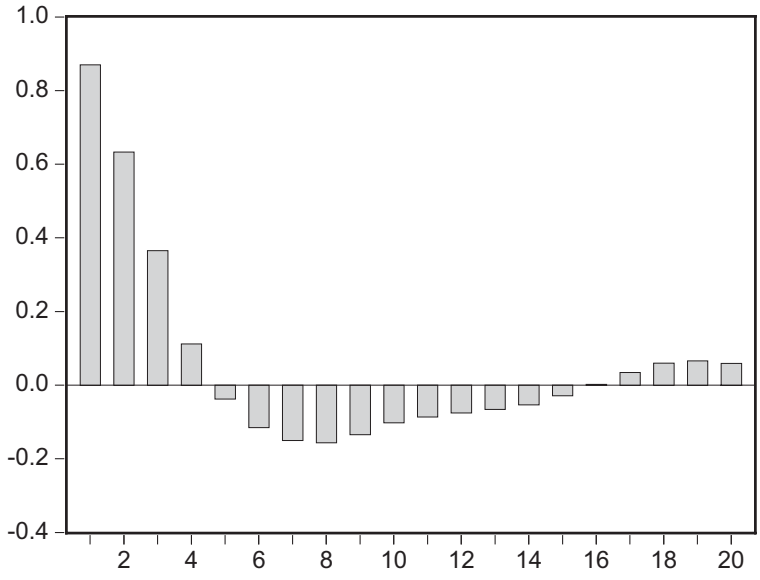


Figure 3.7: Empirical autocorrelation function of annual differences of the log monthly US industrial production (not seasonally adjusted), 1959–2012.

for any value of k, where v_t is not necessarily a white noise time series. Notice that (3.85) only renders an estimate of the k-th order partial autocorrelation $\hat{\psi}_k$ parameter. To obtain the complete EPACF, (3.85) should be estimated for $k = 1, 2, 3, \ldots$

In principle, we may consider the estimated t-statistics for the ψ_k in (3.85) to establish the significance of the EPACF values. For the EACF values we should consider the distribution of the $\hat{\rho}_k$. This distribution can be shown to depend on the underlying true model, see Box and Jenkins (1970), among others. In practice, we usually approximate the distribution of the $\hat{\rho}_k$ and $\hat{\psi}_k$ by setting their asymptotic standard errors to $1/\sqrt{T}$. Details of how good this approximation is can be found in, for example, Granger and Newbold (1986). In our book, we follow the usual approach by saying that ρ_k and ψ_k are significant at the 5 percent level in case for their empirical counterparts it holds that the intervals $(\hat{\rho}_k - 2/\sqrt{T}, \hat{\rho}_k + 2/\sqrt{T})$ and $(\hat{\psi}_k - 2/\sqrt{T}, \hat{\psi}_k + 2/\sqrt{T})$, respectively, do not include zero.

As an illustration of the EACF and EPACF, consider their first 12 values as these are given in Table 3.1 for the annual differences of log monthly revenue-passenger kilometres of European airlines, for the period 1994.1–2006.12. The E(P)ACFs are computed using all 156 observations in this period, and omitting the observations

Table 3.1: Empirical (partial) autocorrelation functions for annual differences of log monthly revenue-passenger kilometres of European airlines, 1994.1–2006.12

	All observations		Without 2001.9–2001.12	
Lag	EACF	EPACF	EACF	EPACF
1	0.803*	0.803*	0.713*	0.713*
2	0.598*	−0.131	0.478*	−0.064
3	0.429*	−0.023	0.299*	−0.038
4	0.301*	−0.012	0.090	−0.189*
5	0.269*	0.178*	−0.001	0.061
6	0.266*	0.038	−0.048	−0.013
7	0.272*	0.043	−0.045	0.059
8	0.298*	0.093	0.047	0.136
9	0.242*	−0.156	0.078	−0.043
10	0.152	−0.073	0.148	0.118
11	0.057	−0.067	0.187*	−0.001
12	−0.069	−0.162	0.109	−0.121

Note: An asterisk indicates significance at the 5% level. The estimated standard error for the full sample is 0.160, and 0.163 for the sample without the observations 2001.7–2001.12.

2001.9–2001.12. The first obvious feature of the EACF in Table 3.1 is that these four aberrant data points have a large effect on the EACF. In fact, $\hat{\rho}_1$ is 0.803 for the complete sample, while it is 0.71 for the sample less the four last months of 2001. Furthermore, the EACF declines much faster towards zero for the interrupted sample. For the full sample, the empirical autocorrelations stay significant until lag 9. Hence it is difficult to suggest a AR type of model as there is no exponential decrease. On the contrary, the EACF and EPACF patterns of the interrupted sample seem to suggest the possible adequacy of an AR(1).

In practice, we usually do not go through all possible models that are indicated as possibly useful by the EACF and EPACF. In fact, the key issues are often (i) whether the EACF values die out sufficiently quickly, where sufficiency here is not a formal concept

but merely a rule based on experience, (ii) whether the EACF signals overdifferencing, and (iii) whether the EACF and EPACF show any significant and easily interpretable peaks at certain lags, preferably at short horizons. The main reason for this less formal approach in practice is that each variant of an ARMA model implies certain properties of the ACF and PACF, but given the fact that these functions have to be estimated, a given set of EACF and EPACF values may suggest a wealth of possibly useful models. Hence, usually we select a seemingly reasonable set of tentative models, that is, we pick values for p and/or q, then estimate model parameters and we apply diagnostic checks to see whether the models capture the dynamics of the time series sufficiently well. If so, we may employ additional criteria to select a final model, as discussed in Section 3.4.

3.3 Estimation and diagnostic measures

A useful specification strategy for ARMA time series starts with an inspection of the EACF and EPACF values, to check which values are significant such that reasonably simple ARMA model structures can be hypothesized, to estimate the parameters of the various models, and to investigate whether the estimated residuals can be viewed approximately as white noise. This strategy amounts to a subtle interplay between identification, estimation and modification, and practical experience is needed to get some skill. Usually, it does not make much sense to start off with a very large ARMA model and to simplify it by deleting insignificant parameters. The reason for this is that we are likely to encounter a situation in which parts of the AR and MA components cancel out. Intuitively, if data are generated from an AR(1) model, but we estimate the parameters in the ARMA(2,1) model

$$(1 - \alpha_1 L)(1 - \alpha_2 L)y_t = (1 + \theta_1 L)\varepsilon_t, \tag{3.86}$$

the true parameter values are $\alpha_2 = \phi_1$, $\alpha_1 = 0$, and $\theta_1 = 0$ (where α_1 and α_2 could be interchanged, of course). However, (3.86) holds for any values of α_1 and θ_1 with $\alpha_1 = -\theta_1$ because then the model reduces to the true AR(1) specification. Hence, we may expect estimation problems for the parameters, and also problems with the distribution of the t-test statistics for α_1 and θ_1. Of course, when we only consider AR models, we may start off with an AR(p^*) with p^* large, and work downwards to an AR(p), where p is a smaller value than the initial p^*. Notice that this constitutes an additional advantage of AR models over ARMA models.

Once the parameters have been estimated, the residuals $\hat{\varepsilon}_t$ are usually inspected for the presence of some remaining autocorrelation. Again this is in contrast to regression models based on cross-sectional data, as in case of ARMA models the results of the diagnostic checks can provide clear-cut suggestions for modification of the model.

In this section, we discuss methods for estimating the unknown parameters in AR and ARMA models. Other estimation routines can be found in the advanced literature on ARMA models, see Hamilton (1994), among others. Furthermore, we consider two often applied tests for correlation in the $\hat{\varepsilon}_t$ time series. Usually several other diagnostic measures are applied to check the adequacy of the estimated model, including tests for the presence of aberrant observations, heteroskedasticity, and non-linearity, but these are discussed later in the relevant chapters.

Estimation of AR models

The parameters in the AR(p) model, given by

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \qquad t = p+1, p+2, \ldots, T, \tag{3.87}$$

can be estimated by ordinary least squares [OLS], where the observations y_1 to y_p are used as starting-values. It can be shown that the OLS estimators of the parameters α and $\phi_1, \ldots, \phi_p$ are consistent and asymptotically normal, and that standard t-statistics can be used to investigate their significance. The unconditional mean μ of y_t can be estimated using

$$\hat{\mu} = \frac{\hat{\alpha}}{1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots - \hat{\phi}_p}. \tag{3.88}$$

Again it should stressed that imposing α to be zero, while μ is not, forces the estimate $(1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots - \hat{\phi}_p)$ towards zero, and hence spuriously suggests the presence of a unit root, see (3.38).

Consider the following estimation results for an AR model of order 1 for $\Delta_1 2 y_t$ for log monthly revenue-passenger kilometres of European airlines, for the estimation sample 1994.1–2006.12, while omitting the observations for 2001.9–2001.12:

$$\begin{matrix} \Delta_{12} y_t = 0.019 + 0.711\ \Delta_{12} y_{t-1} + \hat{\varepsilon}_t, \\ \qquad\quad (0.004)\ \ (0.052) \end{matrix} \tag{3.89}$$

where estimated standard errors appear in parentheses. Clearly, $\hat{\phi}_1$ is significantly different from zero. The mean μ of $\Delta_{12} y_t$ is estimated as $0.019/(1 - 0.711) = 0.064$. In Figure 3.8, we present the graph of the time series, the fit from the regression (3.89) and the estimated residual time series. This figure illuminates a typical feature of the fit from AR time series models. It seems that the fitted line is approximately equal to the original time series, but one period lagged. Given the expression in (3.89), this is not surprising. Furthermore, it seems that the AR(1) model finds difficulties in fitting

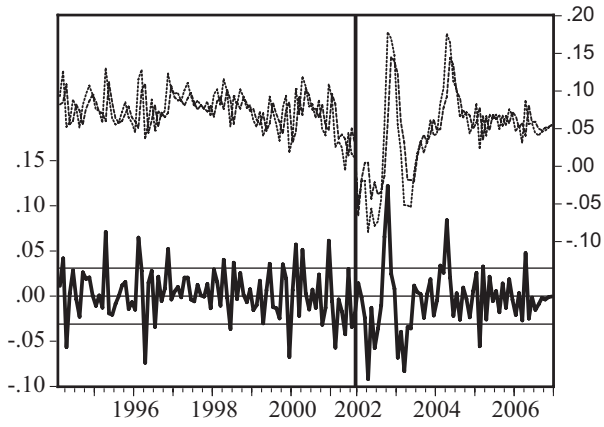


Figure 3.8: Typical fit of an AR time series model: estimation results of an AR(1) on $\Delta_{12} y_t$ with y_t the log monthly revenue-passenger revenue-passenger kilometres of European airlines, 1994.1–2001.8 and 2002.1–2006.12. The short-dashed and long-dashed lines correspond with the actual time series and fitted values, respectively. The solid line represents the residuals. The vertical line indicates the gap in the estimation sample due to the omission of the observations for 2001.9–2001.12.

the more extreme observations. In other words, these observations cannot be predicted well given the past (which, of course, we would not expect so).

Estimation of ARMA models

There exists a wide variety of estimation methods for ARMA models. The main reason for this is that the lagged ε_t variables in the MA part are unobserved, and hence have to be estimated as well. For example, for the ARMA(1,1) model

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \tag{3.90}$$

not only the parameters ϕ_1 and θ_1 are unknown, the ε_{t-1} variable is as well.

A simple estimation procedure based on the autocorrelation properties was proposed by Tuan (1979) and Galbraith and Zinde-Walsh (1994). From the general expression for the k-th order autocorrelation of an ARMA(1,1) model as given in (3.90), it follows

that

$$\rho_1 = \frac{(1+\phi_1\theta_1)(\phi_1+\theta_1)}{(1+2\phi_1\theta_1+\theta_1^2)}, \tag{3.91}$$

$$\rho_2 = \phi_1\rho_1. \tag{3.92}$$

We can rewrite (3.91) as a quadratic equation in the moving average parameter θ_1,

$$\theta_1^2 + b\theta_1 + 1 = 0 \quad \text{with} \quad b = \frac{\phi_1^2 + 1 - 2\rho_1\phi_1}{\phi_1 - \rho_1}. \tag{3.93}$$

Note that b is not well-defined if $\rho_1 = \phi_1$. It can be shown from (3.91) that this only occurs if $|\phi_1| = 1$ or $\theta_1 = 0$. Hence, we need to assume stationarity and a non-zero moving average coefficient to rule out both of these cases. Under this assumption, $|b| > 2$, and the quadratic equation has solutions given by

$$\theta_1 = \frac{-b \pm \sqrt{b^2-4}}{2}, \tag{3.94}$$

where one solution is less than 1 in absolute value, while the other solution is larger. Estimates of the parameters ϕ_1 and θ_1 now can be obtained as follows. We first estimate the AR-parameter as $\hat{\phi}_1 = \hat{\rho}_2/\hat{\rho}_1$ based on (3.92), where $\hat{\rho}_k$ denotes the k-th order empirical autocorrelation. We combine this with $\hat{\rho}_1$ to obtain an estimate $\hat{b}$ from (3.93). Finally, the MA-parameter is estimated by substituting $\hat{b}$ in (3.94), where we select the solution $|\hat{\theta}_1| < 1$ to obtain an invertible model.

An alternative estimation method that is frequently used starts by rewriting the ARMA(1,1) model as

$$(1+\theta_1 L)^{-1} y_t = \phi_1 (1+\theta_1 L)^{-1} y_{t-1} + \varepsilon_t. \tag{3.95}$$

Denoting $z_t = (1+\theta_1 L)^{-1} y_t$ such that

$$z_t = y_t - \theta_1 y_{t-1} + \theta_1^2 y_{t-2} - \theta_1^3 y_{t-3} + \cdots, \tag{3.96}$$

we can construct the z_t series for a given value of θ_1 and assuming that $y_0 = 0$, as

$$\begin{aligned} z_1 &= y_1, \\ z_2 &= y_2 - \theta_1 y_1, \\ z_3 &= y_3 - \theta_1 y_2 + \theta_1^2 y_1, \\ &\vdots, \end{aligned}$$

and so on. We can then estimate ϕ_1 via OLS applied to (3.95). This regression gives an estimated $\hat{\varepsilon}_t$ series, which can be used in (3.90) (setting $\varepsilon_1 = 0$), to obtain new estimates for both ϕ_1 and for θ_1 in a second step. Given these new estimates, the residuals from (3.90) provide a new $\hat{\varepsilon}_t$ series, which can be used again to obtain new

estimates for ϕ_1 and θ_1 from the ARMA(1,1) regression. This iterative procedure can be continued until convergence.

As an illustration, consider the first differences of the monthly log prices of silver, shown in levels in Figure 2.20, for the period February 1978–December 2012 (the January 1978 observation is lost due to first differencing). Inspecting the EACF for this series, only the first order autocorrelation appears significant at 0.253 (based on the asymptotic standard error $1/\sqrt{T}$ with $T = 419$), thereby suggesting an MA(1) model for this series. Using the iterative estimation method outlined above we obtain

$$\underset{}{\Delta_1 y_t} = \underset{(0.0053)}{0.0038} + \hat{\varepsilon}_t + \underset{(0.046)}{0.337}\,\hat{\varepsilon}_{t-1}, \tag{3.97}$$

with standard errors given in parentheses below the parameter estimates. Since this is an MA model, the mean of $\Delta_1 y_t$ is given by the intercept, which does not differ significantly from 0. The θ_1 parameter clearly is significant at the 5% level.

Diagnostic testing for residual autocorrelation

An obvious requirement for an ARMA time series model is that the time series of residuals is approximately white noise. In particular, the residuals should have insignificant autocorrelations at all lags. If this were not the case, we may have missed some dynamic structure in y_t that could have been incorporated in an ARMA model.

The k-th order autocorrelation of the estimated residuals can be computed as

$$r_k(\hat{\varepsilon}) = \frac{\sum_{t=k+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}, \tag{3.98}$$

for $k = 1, 2, 3 \ldots$. When the estimated model is adequate, the population equivalents of $r_k(\hat{\varepsilon})$ are asymptotically uncorrelated and have variances that can be approximated by $(T-k)/(T^2+2T) \approx T^{-1}$. Hence, under the additional assumption of normality, a rough check at the 5 percent significance level is to test whether the estimated residual autocorrelations lie within the $\pm 2/\sqrt{T}$ interval. Ljung and Box (1978) propose a joint test for the significance of the first m residual autocorrelations, which is given by

$$\mathrm{LB}(m) = T(T+2)\sum_{k=1}^{m}(T-k)^{-1} r_k^2(\hat{\varepsilon}). \tag{3.99}$$

The LB(m) statistic asymptotically follows a $\chi^2(m-p-q)$ distribution under the hypothesis of no residual autocorrelation provided that m/T is small and m is moderately large. Residual autocorrelation in an ARMA(p,q) model may not only be caused by the orders p and q being too low, but also by other types of model misspecification, such as non-linearity or neglected outliers, see Lumsdaine and Ng (1999). Hence, the LB test might be considered as a general test for any kind of dynamic misspecification

in the model, and therefore it is usually called a portmanteau test. A drawback of this portmanteau test is that it may be helpful to detect that the estimated model is inadequate, but it is not useful in indicating how the model should be modified. Furthermore, if for example the order m is set too large, the LB test lacks power against low order residual autocorrelation.

An alternative is to consider the nested hypotheses tests developed in Godfrey (1979), among others. As they are based on the Lagrange Multiplier (LM) principle, these tests are relatively easy to compute. For example, to test an AR(p) model against an AR($p+r$) or against an ARMA(p, r) model, the LM test is obtained by running the auxiliary regression

$$\hat{\varepsilon}_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \alpha_{p+1}\hat{\varepsilon}_{t-1} + \cdots + \alpha_{p+r}\hat{\varepsilon}_{t-r} + v_t, \quad (3.100)$$

where $\hat{\varepsilon}_t$ are the residuals of the AR(p) model, with $\hat{\varepsilon}_t$ set equal to zero when $t < p+1$. The test statistic is calculated as $T R^2$ where R^2 is the coefficient of determination from (3.100) and it is asymptotically $\chi^2(r)$ distributed under the null hypothesis that the AR(p) model is adequate. We denote the F-version of this LM test as $F_{\mathrm{AC},1-r}$. The simulation results in Hall and McAleer (1989) indicate that this LM test often has higher power than the LB test.

For the AR(1) model for the of European airlines, the auxiliary regression as (3.100) for residual autocorrelation of order 1 results in

$$\underset{}{\hat{\varepsilon}_t} = \underset{(0.0053)}{0.001} - \underset{(0.074)}{0.0215}\,\Delta_1 y_{t-1} + \underset{(0.110)}{0.064}\,\hat{\varepsilon}_{t-1} + \hat{v}_t, \quad (3.101)$$

for the sample period 1994.1–2001.8 and 2002.1–2006.12 (where $\hat{\varepsilon}_0$ and the interior missing values of $\hat{\varepsilon}_{t-1}$ for 2001.9–2001.12 both are set equal to zero) with an R^2 of 0.0024, where $T = 151$. The $F_{AC,1-1}$-test takes the value of 0.367, which is not significant at the 5% significance level of the $F(1, 148)$ distribution. Hence, the AR(1) model for this revenue-passenger kilometres series does not need to be enlarged by including additional lags of y_t or of ε_t.

In case of an MA(1) model (and this applies naturally to higher order MA models), it is necessary to take into account the fact that the regressor $\hat{\varepsilon}_{t-1}$ cannot be added to the model as that regressor is already included. In that case, we construct new time series

$$y_t^* = y_t + \hat{\theta}_1 y_{t-1}^*, \quad \text{with } y_0^* = 0,$$
$$\hat{\varepsilon}_t^* = \hat{\varepsilon}_t + \hat{\theta}_1 \hat{\varepsilon}_{t-1}^*, \quad \text{with } \hat{\varepsilon}_0^* = 0,$$

and perform the auxiliary regression

$$\hat{\varepsilon}_t = \hat{\alpha}_1 \hat{\varepsilon}_{t-1}^* + \beta_1 y_{t-1}^* + \cdots + \beta_r y_{t-r}^* + v_t, \quad (3.102)$$

in order to test the null hypothesis of the MA(1) model against an MA(1+r) or an ARMA(r,1) model as the alternative.

For the MA(1) model estimated for the first difference of the log silver prices, the resulting $F_{AC,1-1}$-test against an MA(2) or an ARMA(1,1) model takes a value of 3.16, which is not significant at the 5% level.

Diagnostic testing for normality of residuals

To facilitate the interpretation of, for example, parameter estimates and t-ratios, the residuals should be approximately normally distributed. A common approach to check this is by comparing the standardized third and fourth moments, that is, the skewness and kurtosis of the residuals with the values implied by the normal distribution. The skewness [SK] of $\hat{\varepsilon}_t$ can be computed as

$$\text{SK} = \hat{m}_3/\hat{m}_2^{3/2}, \tag{3.103}$$

and the kurtosis [K] as

$$\text{K} = \hat{m}_4/\hat{m}_2^2, \tag{3.104}$$

where $\hat{m}_j$ is the j-th moment of $\hat{\varepsilon}_t$, given by

$$\hat{m}_j = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^j, \quad j = 2, 3, 4. \tag{3.105}$$

Under the null hypothesis of normality (and given the absence of autocorrelation in $\hat{\varepsilon}_t$), we can construct the test statistics $\text{SK}^* = \sqrt{T/6}\,\text{SK}$ and $\text{K}^* = \sqrt{T/24}\,(\text{K}-3)$, which are independent and each have an asymptotic $N(0,1)$ distribution, see Lomnicki (1961). The well-known Jarque-Bera test is given by

$$\text{JB} = (\text{SK}^{*2} + \text{K}^{*2}) \sim \chi^2(2), \tag{3.106}$$

see Bera and Jarque (1982). Bai and Ng (2005) discuss the necessary modifications to the test statistics for time series data. Rejection of the null hypothesis of normality may indicate that there are some outlying observations, or that the error process is not homoskedastic. In Chapters 6 and 7 we deal with time series models that incorporate these features.

The 151 residuals from the AR(1) model for the annual differences of log montly revenue-passenger kilometres of European airlines have SK = 0.22 and $K = 4.84$. The JB takes the value of 22.66, which is significant at the 5 percent level. The 335 residuals of the MA(1) model for the differenced log silver prices have skewness and kurtosis equal to −0.45 and 11.72, respectively. Given that here there are much more observations, the JB test now attains the very large value of 1340.37. Hence, the time series of returns on silver seems to display one or more outlying observations.

Other diagnostic tests

For out-of-sample forecasting, it is important that the time series continues to behave similarly during the forecasting period ("out-of-sample") as it does within the estimation sample ("in-sample"). If this is so, there is confidence in the possible usefulness of the time series model for forecasting purposes. If the time series model suffers from structural breaks in-sample, these breaks should be taken into account when generating forecasts. Tests for structural breaks, as well as tests for specific other types of outliers will be discussed in Chapter 6.

Other relevant diagnostic checks, which can point toward possibly suitable modifications of the time series model, are presented in Chapters 7 and 8. In Chapter 7, we discuss diagnostic checks for the presence of conditional heteroskedasticity. In Chapter 8, we will focus on tests for specific forms of non-linearity.

3.4 Model selection

Identification, parameter estimation and the application of diagnostics can result in a set of tentatively useful models, in the sense that these models cannot be rejected using the above diagnostic measures. We may then want to select a final model using some additional criteria. One particular option is to consider all models for out-of-sample forecasting in order to decide which model performs best on some previously unseen data. This is discussed in more detail in the next section. Here we discuss several useful model selection criteria that are based on the performance of the models within the estimation sample.

A survey of model selection criteria is given in, for example, De Gooijer *et al.* (1985). It seems sensible to assume that no model is *a priori* preferable, and hence that the models should be treated symmetrically. This corroborates with the views expressed in Granger *et al.* (1995). In general, this implies that a final model is selected which optimizes the value of a certain criterion function.

Two popular criteria to select between time series models are the information criteria put forward by Akaike (1974) and Schwarz (1978). Both criteria evaluate the models based on their in-sample fit, while taking into account the number of estimated parameters, or the "parsimony" of the different models. When T now denotes the number of effective observations (which are the observations used to estimate the parameters), and when k denotes the number of ARMA parameters to be estimated, the Akaike Information Criterion [AIC] is given by

$$\mathrm{AIC}(k) = T \log \hat{\sigma}^2 + 2k, \tag{3.107}$$

where $\hat{\sigma}^2 = \sum_{t=1}^{T} \hat{\varepsilon}_t^2 / T$ is the estimated residual variance. The ARMA orders p and q that minimize AIC(k) are selected. The same decision rule applies for the Schwarz Information Criterion [SIC], which is given by

$$\text{SIC}(k) = T \log \hat{\sigma}^2 + k \log T. \tag{3.108}$$

Comparing the expressions for AIC and SIC, it is clear that when $T \geq 8$, the SIC criterion penalizes the inclusion of regressors (and thus of additional parameters) more than the AIC criterion does. This means that the model orders selected with SIC are usually smaller than the model orders selected with AIC.

3.5 Forecasting

Once one or more time series models have been selected we may consider forecasting future values of the time series at hand. Specifically, we may generate an h-step ahead forecast for y_t, where h denotes the so-called forecast horizon. Given that the sample that is used for specifying the time series model consists of T observations $y_1, y_2, \ldots, y_T$, the forecast concerns the observation y_{T+h} and is based on the set $\mathcal{Y}_T$. Three different, but related types of forecasts can be considered. First, a *point forecast* of y_{T+h}, denoted as $\hat{y}_{T+h|T}$, provides a specific value for this observation. In principle any number would provide a valid point forecast, but obviously the aim is to make forecasts that are as accurate as possible. Forecast accuracy is measured by means of a so-called *loss function*, which then also determines what the *optimal* h-step ahead point forecast is. The underlying idea is that any difference between the actual value y_{T+h} and the forecast $\hat{y}_{T+h|T}$ implies a certain loss for the forecast user. The best possible point forecast is the value that minimizes the expected value of this loss function. In this book, we assume that the forecast user has a quadratic loss function, that is

$$\text{Loss}_{T+h|T} = e_{T+h|T}^2, \tag{3.109}$$

where $e_{T+h|T}$ denotes the forecast error, that is $e_{T+h|T} = y_{T+h} - \hat{y}_{T+h|T}$. In that case, it can be shown that the optimal point forecast is the conditional expectation of y_{T+h}, that is

$$\hat{y}_{T+h|T} = \mathsf{E}[y_{T+h}|\mathcal{Y}_T]. \tag{3.110}$$

We note that by using the conditional expectation of y_{T+h} as the point forecast, the conditional mean of the forecast error $e_{T+h|T}$ is equal to zero, such that the quadratic loss function boils down to the forecast error variance. For that reason, in the examples below we provide explicit expressions for this variance.

We refer the interested reader to Christoffersen and Diebold (1996, 1997) for a characterization of optimal point forecasts under alternative loss functions, such as absolute loss $|e_{T+h|T}|$ (where the optimal point forecast is the median of y_{T+h}).

Second, an *interval forecast* consists of a lower bound $\widehat{L}_{T+h|T}$ and an upper bound $\widehat{U}_{T+h|T}$ such that interval $(\widehat{L}_{T+h|T}, \widehat{U}_{T+h|T})$ contains the actual value y_{T+h} with a certain probability. The subscripts on $\widehat{L}$ and $\widehat{U}$ again indicate that these are h-step ahead forecasts made at time $t = T$, that is, conditional on $\mathcal{Y}_T$. Obviously, many choices of $\widehat{L}_{T+h|T}$ and $\widehat{U}_{T+h|T}$ satisfy this requirement. It is common to construct interval forecasts in such a way that they are symmetric around the point forecast $\hat{y}_{T+h}$, that is, $(\widehat{L}_{T+h|T}, \widehat{U}_{T+h|T}) = (\hat{y}_{T+h|T} - c, \hat{y}_{T+h|T} + c)$ for a certain c.

Third, a *density forecast* concerns the conditional distribution of y_{T+h}, denoted as $f(y_{T+h}|\mathcal{Y}_T)$. A density forecast provides a complete characterization of the future observation y_{T+h}, in the sense that it can be used to construct any sort of point and interval forecast for this observation as well.

Forecasting with MA models

Constructing forecasts from ARMA models is quite straightforward, as will become clear from the next few examples. Consider the MA(2) model with zero mean

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}, \qquad t = 1, 2, \ldots, T, \tag{3.111}$$

and assume that our aim is to construct an h-step ahead forecast $\hat{y}_{T+h|T}$ using this model. Starting with the one-step ahead forecast for y_{T+1}, we note that (3.111) implies that for the true observation at $T + 1$ it holds that

$$y_{T+1} = \varepsilon_{T+1} + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}. \tag{3.112}$$

At time T, the value of ε_{T+1} is yet unknown. However, we know that its conditional expectation at time T, $\mathsf{E}[\varepsilon_{T+1}|\mathcal{Y}_T]$, is equal to zero. Hence, the optimal point forecast y_{T+1} (assuming quadratic loss) equals

$$\hat{y}_{T+1|T} = \mathsf{E}[y_{T+1}|\mathcal{Y}_T] = \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}. \tag{3.113}$$

In practice, the values of θ_1, θ_2, ε_T, and ε_{T-1} are of course unknown and should be estimated, but for convenience of notation here we assume that these are given.

Comparing the expressions for y_{T+1} and $\hat{y}_{T+1|T}$, the one-step ahead forecast error (or prediction error) $e_{T+1|T}$ is equal to the shock occurring at $t = n + 1$, that is,

$$e_{T+1|T} = y_{T+1} - \hat{y}_{T+1|T} = \varepsilon_{T+1}. \tag{3.114}$$

Hence, the variance of the forecast error $\mathsf{V}[e_{T+1|T}]$ is equal to σ^2, which is the variance of ε_t.

For two steps ahead, we have

$$e_{T+2|T} = y_{T+2} - \hat{y}_{T+2|T} = (\varepsilon_{T+2} + \theta_1\varepsilon_{T+1} + \theta_2\varepsilon_T) - (\theta_2\varepsilon_T), \tag{3.115}$$

as at time T, ε_{T+2} and ε_{T+1} are unknown, such that $\hat{y}_{T+2|T} = \mathsf{E}[y_{T+2}|\mathcal{Y}_T] = \theta_2\varepsilon_T$. The variance of the two-step ahead forecast error equals $(1 + \theta_1^2)\sigma^2$. For three steps ahead, we get

$$\hat{y}_{T+3|T} = \mathsf{E}[y_{T+3}|\mathcal{Y}_T] = \mathsf{E}[\varepsilon_{T+3} + \theta_1\varepsilon_{T+2} + \theta_2\varepsilon_{T+1}|\mathcal{Y}_T] = 0, \tag{3.116}$$

that is, there is no memory in the MA(2) model that can help to forecast y_{T+3}. Hence, the corresponding forecast error $e_{T+3|T}$ is equal to the actual observation y_{T+3},

$$e_{T+3|T} = y_{T+3} - \hat{y}_{T+3|T} = \varepsilon_{T+3} + \theta_1\varepsilon_{T+2} + \theta_2\varepsilon_{T+1}, \tag{3.117}$$

such that the forecast error variance equals $\mathsf{V}[e_{T+3|T}] = (1 + \theta_1^2 + \theta_2^2)\sigma^2$. In fact, it follows from (3.111) that for any horizon $h \geq 3$, the optimal point forecast $\hat{y}_{T+h|T}$ from the the MA(2) model is equal to 0, and the corresponding forecast error is equal to the actual observation with variance $\mathsf{V}[e_{T+h|T}] = (1 + \theta_1^2 + \theta_2^2)\sigma^2$.

In general, for an MA(q) model the h-step ahead forecast equals

$$\hat{y}_{T+h|T} = \sum_{i=0}^{q} \theta_{i+h}\varepsilon_{T-i}, \tag{3.118}$$

with $\theta_0 = 1$ and $\theta_{i+h} = 0$ for $i + h > q$. From (3.118) it follows that, for a zero mean time series that can be described by an MA(q) model, $\hat{y}_{T+h|T} = 0$ when $h > q$. The h-step ahead forecast error corresponding with (3.118) is equal to

$$e_{T+h|T} = y_{T+h} - \hat{y}_{T+h|T} = \sum_{i=0}^{h-1} \theta_i\varepsilon_{T+h-i}. \tag{3.119}$$

The white noise assumption on the shocks ε_t implies that

$$\mathsf{E}[e_{T+h|T}|\mathcal{Y}_T] = 0, \quad \text{and} \tag{3.120}$$

$$\mathsf{V}[e_{T+h|T}] = \mathsf{E}[e_{T+h}^2|\mathcal{Y}_T] = \sigma^2 \sum_{i=0}^{h-1} \theta_i^2. \tag{3.121}$$

Given that

$$y_{T+h} = \hat{y}_{T+h|T} + e_{T+h|T}, \tag{3.122}$$

and given that $\hat{y}_{T+h|T}$ is the conditional expectation of y_{T+h}, it follows that the conditional variance of y_{T+h} is equal to $\mathsf{V}[e_{T+h|T}]$. In fact, the conditional distribution of y_{T+h} at time T, or the density forecast, is equal to the distribution of $e_{T+h|T}$, except with mean $\hat{y}_{T+h|T}$ instead of zero. Assuming normality of ε_t, it follows from (3.119) that

$e_{T+h|T}$ and y_{T+h} are normally distributed. In that case, a 95 percent interval forecast for y_{T+h} is given by

$$(\hat{y}_{T+h|T} - 1.96\sqrt{\mathsf{V}[e_{T+h|T}]}, \hat{y}_{T+h|T} + 1.96\sqrt{\mathsf{V}[e_{T+h|T}]}). \tag{3.123}$$

Forecasting with AR models

For an AR(p) model it holds that y_t depends on all the previous observations, and therefore the h-step ahead forecasts have similar dependence. Consider for example the AR(2) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t, \tag{3.124}$$

and the one-step ahead forecast at time T, that is

$$\hat{y}_{T+1|T} = \phi_1 y_T + \phi_2 y_{T-1}. \tag{3.125}$$

As the true value at $t = n + 1$ is given by

$$y_{T+1} = \phi_1 y_T + \phi_2 y_{T-1} + \varepsilon_{T+1}, \tag{3.126}$$

the one-step ahead forecast error $e_{T+1|T}$ again is equal to ε_{T+1} with variance σ^2. For two steps ahead, we obtain

$$\begin{aligned}\hat{y}_{T+2|T} &= \mathsf{E}[\phi_1 y_{T+1} + \phi_2 y_T + \varepsilon_{T+2}|\mathcal{Y}_T] \\ &= \phi_1 \hat{y}_{T+1|T} + \phi_2 y_T \\ &= \phi_1(\phi_1 y_T + \phi_2 y_{T-1}) + \phi_2 y_T.\end{aligned} \tag{3.127}$$

As

$$\begin{aligned}y_{T+2} &= \phi_1 y_{T+1} + \phi_2 y_T + \varepsilon_{T+2} \\ &= \phi_1(\phi_1 y_T + \phi_2 y_{T-1} + \varepsilon_{T+1}) + \phi_2 y_T + \varepsilon_{T+2},\end{aligned} \tag{3.128}$$

it holds that $e_{T+2|T} = \varepsilon_{T+2} + \phi_1 \varepsilon_{T+1}$ and

$$\mathsf{V}[e_{T+2|T}] = (1 + \phi_1^2)\sigma^2. \tag{3.129}$$

For three steps ahead, we would have

$$\begin{aligned}\hat{y}_{T+3|T} &= \mathsf{E}[\phi_1 y_{T+2} + \phi_2 y_{T+1} + \varepsilon_{T+3}|\mathcal{Y}_T] \\ &= \phi_1 \hat{y}_{T+2|T} + \phi_2 \hat{y}_{T+1|T} \\ &= \phi_1(\phi_1(\phi_1 y_T + \phi_2 y_{T-1}) + \phi_2 y_T) + \phi_2(\phi_1 y_T + \phi_2 y_{T-1}),\end{aligned} \tag{3.130}$$

and as

$$\begin{aligned} y_{T+3} &= \phi_1 y_{T+2} + \phi_2 y_{T+1} + \varepsilon_{T+3} \\ &= \phi_1(\phi_1(\phi_1 y_T + \phi_2 y_{T-1} + \varepsilon_{T+1}) + \phi_2 y_T + \varepsilon_{T+2}) \\ &\quad + \phi_2(\phi_1 y_T + \phi_2 y_{T-1} + \varepsilon_{T+1}) + \varepsilon_{T+3}, \end{aligned} \tag{3.131}$$

the forecast error $e_{T+3|T} = \varepsilon_{T+3} + \phi_1\varepsilon_{T+2} + (\phi_1^2 + \phi_2)\varepsilon_{T+1}$ with variance

$$\mathsf{V}[e_{T+3|T}] = (1 + \phi_1^2 + \phi_2^2 + 2\phi_1^2\phi_2 + \phi_1^4)\sigma^2, \tag{3.132}$$

which shows that $\mathsf{V}[e_{T+3|T}] > \mathsf{V}[e_{T+2|T}]$.

Exercise 3.12

In general, for AR(p) models it holds that the variance of the forecast error increases with the forecast horizon, that is, $\mathsf{V}[e_{T+h|T}] > \mathsf{V}[e_{T+h-1|T}]$ for all $h > 1$. The expression in (3.132) clearly shows that the expression for the h-step ahead forecast error variance can be notationally cumbersome. It is then more useful to write an AR(p) model into MA format, and as such use similar formulas as (3.121). For example, the AR(2) model (3.124) can be written as

$$y_t = \varepsilon_t + \eta_1\varepsilon_{t-1} + \eta_2\varepsilon_{t-2} + \eta_3\varepsilon_{t-3} + \cdots, \tag{3.133}$$

for which it holds that

$$\mathsf{V}[e_{T+h|T}] = \mathsf{E}[e_{T+h}^2|\mathcal{Y}_T] = \sigma^2 \sum_{i=0}^{h-1} \eta_i^2, \qquad \text{with } \eta_0 = 1. \tag{3.134}$$

For the AR(2) model, it is easy to verify that $\eta_1 = \phi_1$ and $\eta_2 = \phi_1^2 + \phi_2$, such that for $h = 3$ (3.134) indeed is equal to (3.132).

Exercise 3.13

The h-step ahead forecasts for ARMA models are derived along similar lines as for AR and MA models. For example, for the ARMA(1,1) model we have

$$\hat{y}_{T+1|T} = \phi_1 y_T + \theta_1\varepsilon_T, \tag{3.135}$$

with obviously again $\mathsf{V}[e_{T+1|T}] = \sigma^2$. Further,

$$\begin{aligned} \hat{y}_{T+2|T} &= \mathsf{E}[\phi_1 y_{T+1} + \varepsilon_{T+2} + \theta_1\varepsilon_{T+1}|\mathcal{Y}_T] \\ &= \phi_1\hat{y}_{T+1} \\ &= \phi_1(\phi_1 y_T + \theta_1\varepsilon_T). \end{aligned} \tag{3.136}$$

The actual value y_{T+2} can be expressed as

$$\begin{aligned} y_{T+2} &= \phi_1 y_{T+1} + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} \\ &= \phi_1(\phi_1 y_T + \varepsilon_{T+1} + \theta_1 \varepsilon_T) + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1}, \end{aligned} \tag{3.137}$$

such that $e_{T+2|T} = \varepsilon_{T+2} + (\phi_1 + \theta_1)\varepsilon_{T+1}$ and $\mathsf{V}[e_{T+2|T}] = (1 + \theta_1^2 + \phi_1^2 + 2\phi_1\theta_1)\sigma^2$. This last expression can also be derived by writing the ARMA(1,1) model as

$$\begin{aligned} y_t &= (1 - \phi_1 L)^{-1}(1 + \theta_1 L)\varepsilon_t \\ &= \varepsilon_t + \eta_1 \varepsilon_{t-1} + \eta_2 \varepsilon_{t-2} + \eta_3 \varepsilon_{t-3} + \cdots, \end{aligned} \tag{3.138}$$

where $\eta_1 = \phi_1 + \theta_1$.

A final remark about constructing forecasts concerns forecasting the original time series w_t when a model has been made for $y_t = \log(w_t)$. Consider again the MA(2) model and its one-step ahead forecast

$$\hat{y}_{T+1|T} = \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1}. \tag{3.139}$$

If we then take the forecast for w_{T+1} as

$$\hat{w}_{T+1|T} = \exp(\hat{y}_{T+1|T}), \tag{3.140}$$

it is easy to show that $\hat{w}_{T+1|T}$ is biased for w_{T+1} as

$$\begin{aligned} \mathsf{E}[w_{T+1}] &= \mathsf{E}[\exp(\varepsilon_{T+1} + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1})] \\ &= \exp(\sigma^2/2)\mathsf{E}[\exp(\theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1})] \\ &= \exp(\sigma^2/2)\hat{w}_{T+1|T}, \end{aligned} \tag{3.141}$$

when normality of ε_t is assumed. Hence, in case of a model for logs an unbiased forecast of w_{T+1} is given by $\exp(\sigma^2/2)\hat{w}_{T+1|T}$, where $\hat{w}_{T+1|T}$ is called the "naive" forecast. For two steps ahead, we have that

$$\begin{aligned} \mathsf{E}[w_{T+2}] &= \mathsf{E}[\exp(\varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \varepsilon_T)] \\ &= \exp[(1 + \theta_1^2)\sigma^2/2]\mathsf{E}[\exp(\theta_2 \varepsilon_T)] \\ &= \exp[(1 + \theta_1^2)\sigma^2/2]\hat{w}_{T+2|T}. \end{aligned} \tag{3.142}$$

When ARMA models are written in MA format as (3.133), proper expressions for the correction factor of the naive forecasts for w_{T+h} can be derived, see Granger and Newbold (1976) for additional derivations.

Evaluating and comparing forecasts

Forecasting performance is a useful tool for evaluating and comparing time series models. A common practical procedure is to keep P observations apart in order

to be able to evaluate the h-step ahead forecasts $\hat{y}_{T+h+i|T+i} = \mathsf{E}[y_{T+h+i}|\mathcal{Y}_{T+i}]$ for $i = 0, \ldots, P-h$ from models which are constructed using the first T observations. Several choices need to be made when implementing the forecasting exercise. First, the forecast horizon h needs to be selected. Short-term forecasts often are of most interest, suggesting to take $h = 1$, but in other applications long-term forecasts may be relevant as well. Second, the available sample of size $T + P$ should be split into an initial part of T observations that is used number for model specification and parameter estimation, and a second part (or hold-out sample) of P observations for which forecasts are made and evaluated. Intuitively, we would like to set P as large as possible, in order to have a large number of forecasts to assess the accuracy of competing models. On the other hand, though, we would also like to set T sufficiently large, to help identifying potentially useful models and to obtain reasonably accurate parameter estimates. There is no strict rule that can guide the appropriate choice of T and P. Third, we should decide whether or not to re-estimate the model parameters during the forecast period. That is, when making the forecast $\hat{y}_{T+h+i|T+i} = \mathsf{E}[y_{T+h+i}|\mathcal{Y}_{T+i}]$ for a given $i = 0, \ldots, P-h$, we can either use the initial parameter estimates based on observations $y_1, \ldots, y_T$, or re-estimate the parameters based on the set $\mathcal{Y}_{T+i}$ containing observations up to y_{T+i}. If we decide to re-estimate the parameters, we should then choose between using an expanding window or a moving window for estimation. In the first case, for any given value of i we estimate the parameters using all available observations $y_1, y_2, \ldots, y_{T+i}$. In the second case, we delete the first i data points such that the estimation sample has the same size T for all the P forecasts.

The forecast accuracy of an individual time series model can be assessed in a variety of ways. For point forecasts $\hat{y}_{T+h+i|T+i}$, an obvious possibility is to consider their precision in terms of the loss function upon which they are based. For example, for point forecasts based on the quadratic loss function (3.109), a sensible forecast evaluation criterion is the mean squared prediction error [MSPE], which can be computed as

$$\begin{aligned} \text{MSPE}(h) &= \frac{1}{P-h+1} \sum_{i=0}^{P-h} (y_{T+h+i} - \hat{y}_{T+h+i|T+i})^2 \\ &= \frac{1}{P-h+1} \sum_{i=0}^{P-h} e^2_{T+h+i|T+i}. \end{aligned} \tag{3.143}$$

It is useful to note that the MSPE can be written as the sum of the forecast error variance and the squared bias, that is,

$$\text{MSPE}(h) = \hat{\sigma}^2_{e_h} + \bar{e}^2_h, \tag{3.144}$$

where $\hat{\sigma}^2_{e_h} = \frac{1}{P-h+1} \sum_{i=0}^{P-h} (e_{T+h+i|T+i} - \bar{e}_h)^2$ and $\bar{e}_h = \frac{1}{P-h+1} \sum_{i=0}^{P-h} e_{T+h+i|T+i}$. In practice, it is often desirable to have forecasts that are unbiased, that is average forecast

error $\bar{e}_h$, should be close to zero. If this is not the case, the model under- or over-estimates the conditional mean of the time series. Usually, this can be interpreted as saying that the deterministic component in the model such as mean and trend is not adequately specified. On the other hand, in some applications it turns out that slightly biased forecasts have considerably lower variance than unbiased ones, such that the MSPE may be lower.

Alternative forecast evaluation criteria include the mean absolute error [MAE],

$$\text{MAE}(h) = \frac{1}{P-h+1} \sum_{i=0}^{P-h} \left| y_{T+h+i} - \hat{y}_{T+h+i|T+i} \right|, \tag{3.145}$$

and the mean absolute percentage error [MAPE],

$$\text{MAPE}(h) = \frac{1}{P-h+1} \sum_{i=0}^{P-h} \left| \frac{y_{T+h+i} - \hat{y}_{T+h+i|T+i}}{y_{T+h+i}} \right|. \tag{3.146}$$

It should be mentioned that MAPE is not very useful when the time series y_t can take values very close to zero, as in the case of growth rates.

Comparison of different time series models in terms of out-of-sample forecast accuracy can be based on criteria such as the MSPE defined in (3.143), where obviously the model that gives the smallest MSPE value is the preferred one as it gives the most accurate forecasts. Diebold and Mariano (1995) consider several statistics for testing whether the difference in MSPE of two competing models, say A and B, is significant or not. The test statistic that has become most popular in practice is based on the so-called loss-differential d_t, defined a

$$d_t \equiv e^2_{A,t|t-h} - e^2_{B,t|t-h},$$

where $e_{A,t|t-h}$ and $e_{B,t|t-h}$ denote the forecast errors from models A and B, respectively. The null hypothesis to be tested is that the MSPEs are equal, which can be restated as $\mathsf{E}[d_t] = 0$. Given a sequence of P realizations d_t for $t = T+h, \ldots, T+h+P-1$, the sample mean loss differential $\bar{d} = \frac{1}{P}\sum_{i=0}^{P-1} d_{T+h+i}$ divided by its sample standard deviation has a standard normal distribution asymptotically, that is

$$\frac{\bar{d}}{\sqrt{\hat{\sigma}^2_{d_t}/P}} \sim N(0,1),$$

where $\hat{\sigma}^2_{d_t}$ is the variance of d_t, which can be computed as

$$\hat{\sigma}^2_{d_t} = \hat{\gamma}_0 + 2\sum_{j=0}^{h-1} \hat{\gamma}_j,$$

with $\widehat{\gamma}_j$ denoting the j-th order sample autocovariance

$$\hat{\gamma}_j = \frac{1}{P} \sum_{i=0}^{P-1-j} \left(d_t - \bar{d}\right)\left(d_{t-j} - \bar{d}\right).$$

The correction of the sample variance $\hat{\gamma}_0$ with the autocovariances $\hat{\gamma}_j$, $j = 1, \ldots, h-1$, is based on the fact that forecast errors for h-step ahead forecasts are serially correlated up to (at least) order $h-1$ by construction. We refer to Newbold and Harvey (2002) and West (2006) for detailed discussions of evaluation of point forecasts, and in particular statistics for comparing predictive accuracy.

For evaluating interval forecasts, an obvious possibility is to check whether indeed p percent of the forecasts indeed lie within the p percent forecast interval. If so, we gain confidence in the model. If not, the variance of the data is likely under- or overestimated. Formal test statistics for evaluating such interval forecasts are developed in Christoffersen (1998) and Wallis (2003), see also Clements (2005) for a review. Issues involved in evaluating density forecasts are discussed in Corradi and Swanson (2006).

Finally, we should note that here we have discussed only the most important ingredients for out-of-sample forecasting. More extensive treatments of forecasting can be found in Clements and Hendry (1998, 1999), while the chapters in the edited volumes by Clements and Hendry (2002) and Elliot *et al*. (2006) focus on certain specific issues.

CONCLUSION

In this chapter we have discussed some important concepts in univariate time series modeling and forecasting. These concepts should be a useful basis when analyzing time series with the typical features reviewed in Chapter 2. In many cases below we will confine ourselves to ARI(p, d) models. Such models have several advantages over ARIMA(p, d, q) models. For example, parameter estimation and diagnostic checking are quite straightforward, the longer memory makes AR models perhaps more useful for forecasting, and they can be extended to multivariate models fairly easily (as we will see in Chapter 9). Also, non-linearity and outliers are more easily handled within the ARI framework.

In the next chapter we start off with a discussion of trends. The adequacy of forecasts for economic and business time series can largely depend on the appropriate form of the trend in the model. The concept of unit roots in the AR polynomial will be shown to play a crucial role.

EXERCISES

3.1 Suppose we change the model in (3.11) to $y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t$. Express the starting value y_0 in terms of pre-sample observations, α and ϕ.

3.2 Derive the first four values π_1, π_2, π_3, and π_4 in the approximate polynomial

$$\pi(L) = 1 - \pi_1 L - \pi_2 L^2 - \pi_3 L^3 - \pi_4 L^4 \tag{3.147}$$

when it holds true that

$$\pi(L) = \frac{1 - \phi_1 L - \phi_2 L^2}{1 + \theta_1 L}. \tag{3.148}$$

This last ratio of polynomials corresponds with an ARMA(2,1) model.

3.3 Consider a bi-annual time series y_t, $t = 1, 2, \ldots, T$, which can be described by

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \tag{3.149}$$

where ε_t is a standard white noise variable with variance σ_ε^2. Suppose one aggregates the bi-annual data into annual observations X_a, with $a = 1, 2, \ldots, T/2$, that is,

$$X_1 = y_1 + y_2, \quad X_2 = y_3 + y_4, \quad X_3 = y_5 + y_6, \ldots$$

and in general

$$X_a = y_{2a-1} + y_{2a}.$$

Show that the ARMA model for X_a is

$$X_a = \alpha X_{a-1} + V_a + \theta V_{a-1}, \tag{3.150}$$

where α and θ are functions of ϕ_1 and σ_ε^2.

3.4 Consider the variables y_t and x_t, which can be described by

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

and

$$x_t = \beta x_{t-4} + u_t + \theta u_{t-1}.$$

Show that the variable z_t defined by $z_t = y_t + x_t$ can be described by an ARMA(5,4) model.

3.5 Consider a time series y_t, which can be described by

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \tag{3.151}$$

where ε_t is a standard white noise variable with variance σ_ε^2. Unfortunately, it turns out that y_t is only observed with measurement error, that is, one observes

z_t, given by

$$z_t = y_t + u_t, \tag{3.152}$$

instead of y_t, where u_t is also a standard white noise variable with variance σ_u^2. It is known that ε_t and u_t are mutually uncorrelated at all lags.

a. Derive the ARMA model for z_t
b. Is it possible to retrieve the parameters in the model for y_t from the parameter estimates for this ARMA(1,1) model for z_t? And, when it is assumed that σ_u^2 is equal to σ_ε^2?

3.6 Generate 100 artificial time series of $T = 200$ observations from the AR(1) model given in (3.33) $\mu = 10$ and $\phi_1 = 0.5$. Set the starting value $y_0 = \mu = 10$ for all series, but use different shocks ε_t, $t = 1, \ldots, 200$ with $\varepsilon_t \sim N(0, 1)$. For each of those series, estimate the parameter in an AR(1) regression that does not include an intercept, that is, $y_t = \phi y_{t-1} + \varepsilon_t$. Examine the properties of the least squares estimates. In particular, how does the mean of $\hat{\phi}_1$ relate to the true value $\phi_1 = 0.5$?

3.7 Show that the unconditional mean of a time series y_t that can be described by the AR(1) model (3.42) is equal to μ when $|\phi_1| < 1$.

3.8 Show that the k-th order autocorrelation of an ARMA(1,1) model is given by (3.73).

3.9 Consider the following ARMA(1,2) model for a time series y_t, that is,

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \tag{3.153}$$

where ε_t is a standard white noise process with variance σ^2, and with $|\phi_1| < 1$, and where ϕ_1, θ_1 and θ_2 are unequal to zero. Give expressions for the first three autocorrelations of y_t.

3.10 Show that it follows from (3.77) that the first-order partial autocorrelation $\psi_1 = \gamma_1/\gamma_0$.

3.11 Show that (3.81) implies that the sum of the autocorrelations of an overdifferenced time series is equal to -0.5, as suggested by (3.82).

3.12 Prove that the 2- and 3-steps ahead forecast error variance for an AR(2) model are given by (3.129) and (3.132), respectively.

3.13 Verify that for an AR(2) model $\eta_1 = \phi_1$ and $\eta_2 = \phi_1^2 + \phi_2$ in (3.134), such that for $h = 3$ the resulting expression of the forecast error variance is equal to (3.132).

4 Trends

The examples in Chapter 2 demonstrate that many time series in economics and business have a trending pattern, where for macroeconomic time series such trends typically move upwards. Although many practitioners would be able to indicate roughly what a trend is ("a general tendency for a variable to increase or decrease over time"), a formal definition of a trend cannot be given otherwise than in the context of a model. Put differently, only after we have agreed upon a time series model that (presumably) describes the data at hand, we can define a trend within this framework and discuss the 'trend properties' of the time series in a meaningful way. In this book, the focus is on ARMA-type time series models, and hence the current chapter deals with trends within this model class.

For several reasons it is important to investigate the appropriate formulation of the trend in a time series prior to putting effort into modeling other data features and forecasting. First and foremost, the trend will dominate long-run out-of-sample forecasts, although also short-run forecasts can be affected, as we will see in Section 4.4. An inadequate specification of the trend possibly leads to forecasts that are biased or inaccurate in other ways. Second, a time series that displays a trend is nonstationary, in the sense that it does not have a constant mean. In addition, depending on the nature of the trend, the unconditional variance of the series is not constant over time but increases with every new observation. This implies that the autocorrelation function can also vary over time, simply because it depends on the variance. For any given time series we can compute the sample mean, variance and autocorrelations. For trending time series these estimates are not meaningful, however, as they do not converge to specific values when the number of observations increases. In other words, for summary statistics like mean, variance and autocorrelations to be interpretable, these three measures should be constant over time, which typically is not the case for trending series. Hence, considerable care should be exercised when analyzing such time series, and various aspects of such analysis will be discussed in this chapter.

A more formal definition of stationarity of a time series y_t is that the following three properties should hold:

$$\mathsf{E}[y_t] = \mu \qquad \text{for all } t = 1, 2, \ldots, T, \tag{4.1}$$

$$\mathsf{E}[(y_t - \mu)^2] = \gamma_0 \qquad \text{for all } t = 1, 2, \ldots, T, \tag{4.2}$$

$$\mathsf{E}[(y_t - \mu)(y_{t-k} - \mu)] = \gamma_k \qquad \text{for all } t = 1, 2, \ldots, T, \text{ and } k = \ldots, -2, -1, 0, 1, 2, \ldots \tag{4.3}$$

where μ, γ_0 and γ_k are all finite-valued numbers. For a given time series it is difficult to verify whether these three conditions are satisfied at the same time. Intuitively, to verify (4.1) with a certain test statistic, we need an estimate of the variance of y_t, which in turn should obey (4.2), which in turn depends on the validity of (4.1). Hence, for practical diagnostic purposes (4.1)–(4.3) are not readily useful.

One way to overcome the practical limitations of (4.1)–(4.3) while investigating the stationarity or trending behavior of a time series is to consider stationarity and trends within the framework of an AR(p) time series model, possibly with deterministic components. That is, it is possible to examine for which values of the autoregressive parameters or for which components in the deterministic part of an AR(p) model the conditions (4.1)–(4.3) do not hold. Obviously, when the deterministic component includes the variable t, $t = 1, 2, \ldots, T$, the time series has a deterministic trend. For example, suppose an useful model is $y_t = \mu + \delta t + \varepsilon_t$, with $\delta \neq 0$ and ε_t a white noise series with variance σ^2. In this case the unconditional mean of y_t is equal to $\mu + \delta t$ and thus varies over time, but its unconditional variance equals σ^2 and is constant. On the other hand, when the AR(p) part of the model contains the component $(1 - L)$, we say that the y_t series has a stochastic trend, as desribed below. The unconditional mean of a series with a stochastic trend may be constant or time-varying depending on the specification of the deterministic component, but in any case its unconditional variance increases with time, as we will demonstrate in Section 4.1. In sum, different trend specifications imply different time series properties.

In practice we proceed as follows to determine the appropriate trend specification for a given time series. An AR(p) model is fitted to the time series y_t. Next we test whether the trend variable t contributes to the explanation of y_t, or whether a $(1 - L)$ component can be separated out from the AR(p) part, or perhaps even both. Having a $(1 - L)$ component in the AR-polynomial is equivalent to having a solution equal to 1 for the corresponding characteristic equation. Hence, when examining the relevance of a stochastic trend in a given time series we say that we investigate the presence of a unit root.

In this chapter, we review the deterministic and stochastic trend specifications in AR(p) models in Section 4.1. Methods to check whether a time series is stationary or not, and hence to make the necessary choice between the two trend characterizations,

are discussed in the following two sections, focusing on the Dickey and Fuller (1979) method in Section 4.2 and on the Kwiatkowski *et al.* (1992) method in Section 4.3. The literature on this topic has expanded enormously in the last few decades, mainly because the statistical methods involved are non-standard. Of course, it is virtually impossible to treat all issues here. The interested reader should consult more extensive surveys, which appear in Hamilton (1994), Hansen (1996), and Phillips and Xiao (1998), amongst many others. The main aim of this chapter is to show how trends can be modeled in the context of AR models, and how the various decisions that need to be made can affect out-of-sample forecasting. The latter topic is discussed in Section 4.4. In Chapter 9 we will see that a decision on trends in univariate time series has an impact on how to proceed with modeling a set of time series in a multivariate model.

4.1 Modeling trends

To discuss the various possible representations of trends within the context of an autoregressive time series model, consider the AR(1) model where the time series y_t is considered in deviation of a linear deterministic trend, that is

$$y_t - \mu - \delta t = \phi_1(y_{t-1} - \mu - \delta(t-1)) + \varepsilon_t, \quad t = 1, 2, \ldots, T. \tag{4.4}$$

When $|\phi_1| < 1$, the time series y_t mean-reverts to $\mu + \delta t$, see below. This can be rewritten in regression format as

$$y_t = (1-\phi_1)\mu + \phi_1\delta + (1-\phi_1)\delta t + \phi_1 y_{t-1} + \varepsilon_t, \tag{4.5}$$

or more compactly as

$$y_t = \alpha + \beta t + \phi_1 y_{t-1} + \varepsilon_t, \tag{4.6}$$

which corresponds with the model that we would consider for the estimation of μ, δ and ϕ_1, see also Chapter 3 for the case without deterministic trend.

Exercise 4.1

Defining $z_t = y_t - \mu - \delta t$, we can express z_t in terms of the current and lagged shocks ε_{t-i}, $i = 0, 1, \ldots, t-1$ by recursively substituting lagged z_t values in (4.4) as

$$z_t = (\phi_1)^t z_0 + \sum_{i=1}^{t} (\phi_1)^{t-i} \varepsilon_i, \tag{4.7}$$

where z_0 is a pre-sample starting value of z_t. This of course resembles (3.12). As discussed in Chapter 3, when $|\phi_1| < 1$, (4.7) indicates that more recent shocks ε_i have a larger impact on z_t than less recent ones. In fact, the effect of such shocks dies out in

the long run or, in other words, such shocks are transitory. It follows from (4.7) and the analysis in the previous chapter that, while assuming that $|\phi_1| < 1$, the unconditional mean of z_t is equal to z_0, which we typically set equal to 0 for convenience. Hence, the unconditional mean of y_t is equal to $\mu + \delta t$ for all $t = 1, 2, \ldots$. Likewise, it follows that the unconditional variance of y_t is equal to $\sigma^2/(1 - \phi_1^2)$.

As (4.4) can be written as $z_t = \phi_1 z_{t-1} + \varepsilon_t$, or equivalently as

$$\Delta_1 z_t = (\phi_1 - 1) z_{t-1} + \varepsilon_t, \tag{4.8}$$

it can be observed that when $|\phi_1| < 1$, positive values of z_{t-1} will on average lead to a negative value of $\Delta_1 z_t$ and hence to a decrease in z_t. Similarly, negative values tend to increase z_t. Given that positive and negative values of z_t correspond with y_t being larger or smaller than its unconditional mean $\mu + \delta t$, we say that when $|\phi_1| < 1$, the time series y_t displays mean-reverting behavior to the trend $\mu + \delta t$, which is also called trend-reverting behavior. The results in Chapter 3 on the ACF and PACF of an AR(1) model show that these functions take constant values when $|\phi_1|$ is smaller than 1. As the deterministic trend variable t is included in (4.4), when $|\phi_1| < 1$ we say that y_t is a trend-stationary time series and that it can be described by a deterministic trend model [DT model].

When $\phi_1 = 1$ in (4.4), the time series y_t does not show mean-reverting behavior as in that case (4.8) reduces to

$$\Delta_1 z_t = \varepsilon_t, \tag{4.9}$$

or in terms of y_t,

$$y_t = \delta + y_{t-1} + \varepsilon_t. \tag{4.10}$$

This particular model is called a random walk with drift, where the drift equals δ. With $\phi_1 = 1$, it also holds that (4.9) can be written as

$$z_t = z_0 + \sum_{i=1}^{t} \varepsilon_i, \tag{4.11}$$

or again in terms of y_t as

$$y_t = y_0 + \delta t + \sum_{i=1}^{t} \varepsilon_i, \tag{4.12}$$

where y_0 is some function of the pre-sample observations and μ. The partial sum time series $S_t = \sum_{i=1}^{t} \varepsilon_i$ that appears in (4.12) is called a stochastic trend. Hence, when the time series y_t can be described by the random walk with drift model (4.10) with $\delta \neq 0$, (4.12) shows that it then has both a deterministic trend and a stochastic trend. In order to avoid confusion, when $\phi_1 = 1$, y_t is said to be described by a stochastic trend model [ST model], irrespective of whether $\delta = 0$ or not. The key property of an

ST model is that shocks ε_{t-i} have a permanent effect on the time series y_t, see (4.12), as the weights on ε_{t-i} are all equal to 1.

Exercise 4.2

The AR(1) polynomial for (4.10) equals $(1 - L)$, and the solution to the corresponding characteristic equation $(1 - z) = 0$ equals 1. Therefore, the ST model corresponds with an AR model with a unity solution to the characteristic equation, or a so-called unit root. Notice that (4.10) implies that the first differences $y_t - y_{t-1}$ equal a white noise series, albeit with a non-zero mean δ, and this series is stationary by definition. In general it holds that an ST time series can be made stationary by applying the Δ_1 differencing filter. Therefore, we sometimes call y_t a difference-stationary time series.

When ε_t in (4.4) is replaced by $\eta_t = [\phi_{p-1}(L)]^{-1}\varepsilon_t$, where $\phi_{p-1}(L)$ does not contain the component $(1 - L)$, all the above results continue to hold. Hence, when an AR(p) polynomial $\phi_p(L)$ can be decomposed as $\phi_{p-1}(L)(1 - L)$, the time series y_t has a stochastic trend. When $\phi_p(L)$ does not contain $(1 - L)$ and the deterministic part of the model includes a trend term t, y_t has a deterministic trend. This already suggests that a simple way to choose between a DT or an ST model amounts to looking for the presence of a $(1 - L)$ component in the AR-polynomial, that is, for a unit root.

To obtain an idea of the differences between DT and ST models, consider the data depicted in Figure 4.1, which are generated from

$$\text{DT} : y_t = 0.2t + \varepsilon_t, \quad t = 1, 2, 3, \ldots, 200, \tag{4.13}$$

$$\text{ST} : y_t = 0.2 + y_{t-1} + \varepsilon_t, \quad t = 1, 2, 3 \ldots, 200 \text{ and } y_0 = 0, \tag{4.14}$$

where the observations for the shocks ε_t are the same in both equations, and drawn from a $N(0, 0.25)$ distribution. Clearly, the upward trend in both time series is similar, as can be expected given (4.6) and (4.12). The key difference between the series is that the ST time series can deviate substantially from this trend for prolonged periods of time due to the stochastic trend $S_t = \sum_{i=1}^{t} \varepsilon_i$ also shown in the figure. This demonstrates the lack of mean-reverting forces for ST time series. For values of δ that are small relative to the variance of the ε_t observations, we can in fact generate data which never seem to revert to some mean, see also Figure 3.1. The ST time series thus displays features different from the DT time series because of the partial sum series S_t.

Exercise 4.3–4.4

As an empirical example to illustrate that the patterns in Figure 4.1 can reflect the behavior of truly observed time series, consider the results from the auxiliary regression

$$y_t = \hat{\alpha} + \hat{\beta}t + \hat{u}_t, \tag{4.15}$$

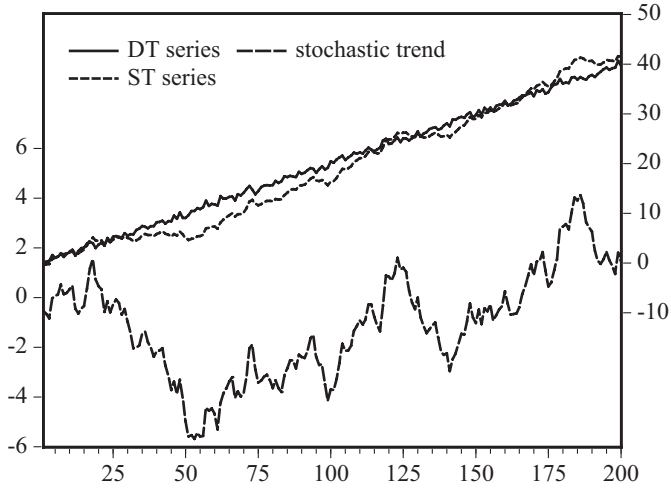


Figure 4.1: Simulated time series from deterministic trend and stochastic trend models.

which was also used in Chapter 2, for the quarterly US industrial production series (seasonally adjusted) for the sample period 1960.1–2012.4 in Figure 4.2. This graph displays the actual values of the time series y_t, the fitted values $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$ from (4.15) and the residuals $\hat{u}_t$. Clearly, there is an upward trend in US industrial production. Additionally, the residuals $\hat{u}_t$ seem to mimic the patterns of the stochastic trend S_t in Figure 4.1, and hence it may be that this variable can best be described by an ST model.

The selection between ST and DT models for y_t may be important from an economic point of view. For example, it can be useful to know whether shocks to a certain time series have permanent effects or not, as in the ST and DT models, respectively. If a certain policy rule creates a large value of ε_t, in the ST model its effect lasts indefinitely, while in the DT model its effect is not noticeable anymore soon. From a forecasting perspective it is also important to determine the most appropriate trend specification. Consider for example the random walk model with drift

$$y_t = y_{t-1} + \delta + \varepsilon_t. \tag{4.16}$$

The two-step ahead forecast at time T implied by this model is

$$\hat{y}_{T+2|T} = \hat{y}_{T+1|T} + \delta = y_T + 2\delta. \tag{4.17}$$

As the true value y_{T+2} equals

$$y_{T+2} = y_{T+1} + \delta + \varepsilon_{T+2} = y_T + 2\delta + \varepsilon_{T+2} + \varepsilon_{T+1}, \tag{4.18}$$

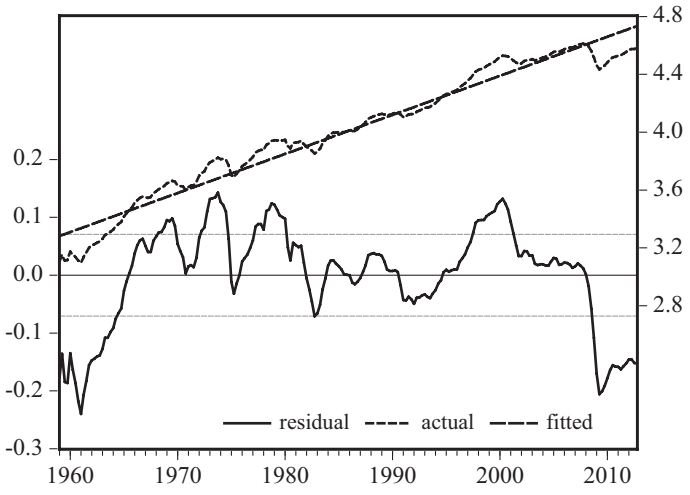


Figure 4.2: Results of a regression of US industrial production on a constant and a linear deterministic trend.

the variance of the corresponding forecast error $\mathsf{V}[e_{T+2|T}]$ is equal to $2\sigma^2$. On the other hand, for the DT model

$$y_t = \alpha + \beta t + \phi_1 y_{t-1} + \varepsilon_t, \tag{4.19}$$

with $|\phi_1| < 1$, it is easily derived that $\mathsf{V}[e_{T+2|T}] = (1 + \phi_1^2)\sigma^2$. Obviously, when ϕ_1 is less than 1 in absolute value, the forecast error variance for the DT model is smaller than for the random walk with drift model. It can also be shown that the difference in $\mathsf{V}[e_{T+h|T}]$ between the two models increases with the forecast horizon h. Hence, the out-of-sample forecasts from the ST model are less certain than those from the DT model. The higher level of uncertainty for the ST model, which is reflected by the fact that shocks have a permanent effect and thus can change the level of y_t permanently, is in turn reflected by wider interval forecasts relative to the DT model. We discuss the issue of trends and forecasting in more detail in Section 4.4.

A wide variety of methods is available for selecting between an ST and a DT model for a given empirical time series y_t. These methods either examine the possible presence of a component $(1 - L)$ in the AR-polynomial of the AR(p) model for y_t, or the relative importance of the stochastic trend component $S_t = \sum_{i=1}^{t} \varepsilon_i$. The first set of methods are called unit root tests, and a method that is often applied in practice will be discussed in Section 4.2. The second set of methods are called stationarity tests, and one of these will be discussed in Section 4.3.

Integration

A time series y_t that requires the first differencing filter Δ_1 to remove a stochastic trend is called integrated of order 1 [I(1)]. There are also time series that even after first differencing still contain a stochastic trend. An example is given by

$$y_t = 2y_{t-1} - y_{t-2} + \delta + \varepsilon_t, \tag{4.20}$$

which can be written as

$$z_t = z_{t-1} + \delta + \varepsilon_t, \tag{4.21}$$

with $z_t = y_t - y_{t-1}$, such that $\Delta_1^2 y_t = \delta + \varepsilon_t$. In this case, y_t is an I(2) time series as it needs the Δ_1 filter twice to be rendered stationary.

One way to understand the possible practical relevance of I(2) processes, which typically seem to occur for nominal monetary aggregates and price levels of rapidly growing economies, is by means of the representation

$$\Delta_1 y_t = \delta_t + \varepsilon_t, \tag{4.22}$$

$$\delta_t = \delta_{t-1} + \eta_t, \tag{4.23}$$

where η_t is a white noise series with variance σ_η^2, and the variance of ε_t is denoted σ^2 as before. That is, the random walk y_t has a time-varying drift δ_t (or, equivalently, the growth rate of y_t has a time-varying mean), which is again a random walk process, see Harvey (1989). When the variance of η_t equals zero, the mean of $\Delta_1 y_t$ is constant, and hence $\Delta_1 y_t$ shows mean-reverting behavior such that y_t is I(1). The expressions in (4.22) and (4.23) can be combined into

$$\Delta_1^2 y_t = v_t, \tag{4.24}$$

where $v_t = \eta_t + (1 - L)\varepsilon_t$ is an MA(1) process. The variance of the v_t series is $\sigma_\eta^2 + 2\sigma^2$, its first-order autocovariance is equal to $\gamma_1 = -\sigma^2$, while all higher-order autocovariances are equal to 0. Hence, the first-order autocorrelation of v_t, that is, $\rho_1 = -\sigma^2/(\sigma_\eta^2 + 2\sigma^2)$ is bounded between -0.5 and 0. When σ_η^2 is very small relative to σ^2, ρ_1 approximates -0.5, and hence it seems that differencing y_t twice is once too often. Put differently, the $\Delta_1^2 y_t$ series may easily seem overdifferenced. In general, the model in (4.24) can be rewritten as

$$\Delta_1^2 y_t = (1 + \theta_1 L)v_t, \tag{4.25}$$

with v_t a white noise error series and $\theta_1 < 0$, which follows from the fact that the first-order autocorrelation in an MA(1) model is equal to $\theta_1/(1 + \theta_1^2)$. Hence, in case σ_η^2 is relatively small such that δ_t only changes slightly, the MA component in (4.25) will be large in the sense that $\theta_1 \to -1$ in order to get ρ_1 close to -0.5. This could

make it difficult to decide whether y_t should be differenced one or two times, as the Δ_1 polynomial almost cancels out from both sides.

Exercise 4.5

An I(2) time series has a growth rate that fluctuates randomly over time and hence has a double stochastic trend. This can be observed from solving (4.21) as

$$z_t = z_0 + \delta_t + \sum_{i=1}^{t} \varepsilon_i,$$

such that

$$\begin{aligned}
y_1 &= z_1 = z_0 + \delta + \varepsilon_1, \\
y_2 &= z_1 + z_2 = z_0 + \delta + \varepsilon_1 + z_0 + 2\delta + \varepsilon_1 + \varepsilon_2, \\
y_3 &= z_0 + \delta + \varepsilon_1 + z_0 + 2\delta + \varepsilon_1 + \varepsilon_2 + z_0 + 3\delta + \varepsilon_1 + \varepsilon_2 + \varepsilon_3, \\
&\vdots \\
y_t &= y_0 + z_0 t + \delta t(t+1)/2 + \sum_{i=1}^{t} \sum_{j=1}^{i} \varepsilon_j. \qquad (4.26)
\end{aligned}$$

This result shows that when δ is different from zero, we should be able to detect an I(2) time series from its graph as it displays explosive growth through the component $t(t+1)/2$. Furthermore, the double stochastic trend appears in the $\sum_{i=1}^{t} \sum_{j=1}^{i} \varepsilon_j$ component, as it needs double differencing to be removed.

An empirical example of a possibly I(2) time series is given in Figure 4.3, showing the stock of motorcycles in the Netherlands for the period 1946–2012, together with the fitted values $\hat{y}_t$ and residuals $\hat{u}_t$ from the regression on a constant, a linear trend, and a quadratic trend, that is,

$$y_t = \hat{\alpha} + \hat{\beta} t + \hat{\gamma} t(t+1)/2 + \hat{u}_t, \qquad (4.27)$$

compare (4.26). Obviously, the residuals do not look anything like white noise and in fact, they might display the smooth behavior of a double stochastic trend $\sum_{i=1}^{t} \sum_{j=1}^{i} \varepsilon_j$.

Common stochastic trends

As mentioned earlier, it is important to analyze the nature of the trends in univariate time series also for the purpose of multivariate modeling, that is, simultaneously analysing two or more time series. The main reason is that there are occasions where several time series have stochastic trends in common. As an example, it is unlikely that the trend

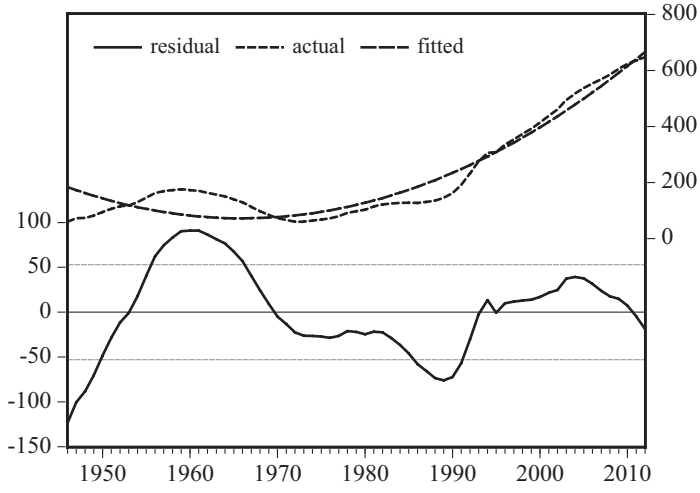


Figure 4.3: Results of a regression of stock of motorcycles on a quadratic deterministic trend.

in real disposable income is different from that in real consumption, simply because if this were the case we would persistently consume too much or save too much. In the long run, income and consumption should be in equilibrium. We may expect that consuming too much now (relative to the current level of income, that is), will lower future consumption and hence that there will be some form of correction towards equilibrium. Later in Chapter 9, we will see that common trends and this mechanism of equilibrium correction are closely linked in the sense that one implies the other, and vice versa.

Exercise 4.6

As an illustrative example of the possible presence of common stochastic trends, consider the estimated partial sums of the residuals of the regression in (4.15) for the five Latin American GDP per capita series in Figure 4.4. Given that these five countries are closely related, not only geographically but also in economic terms, it is not unexpected that the S_t series seem to display common patterns. Hence, it may be that certain linear combinations of these S_t variables do not contain stochastic trend components. In fact, it seems from Figure 4.4 that if there are any stochastic trends in Argentina, Brazil, and Chile, they may have a similar pattern. We will return to these data in Chapter 9, when dealing with the concept of common trends in terms of its companion concept of common integratedness, or briefly, cointegration.

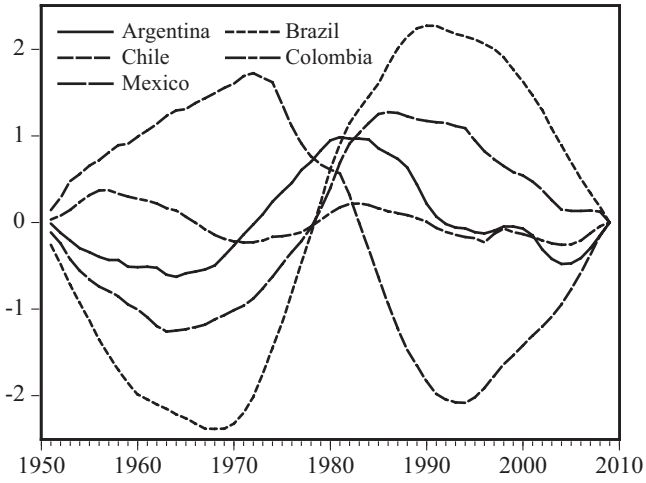


Figure 4.4: Partial sums of residuals from regression (4.15) for Latin American GDP per capita series, 1951–2010.

Growth curves

In principle, trends in an ARMA model imply possibly unbounded behavior of the time series. For many macroeconomic data this may be sensible, but for business economic data, we may sometimes expect that there is some upper bound to the time series. An example of such nonlinear trends is a growth curve, describing product sales or adoption over time. In principle, the ST and DT models allow the y_t series to be unbounded. This may imply that these models cannot be useful for variables such as the unemployment rate, because this variable is bounded by the values 0 and 100. However, in small samples we may still find that ST and DT models do yield approximately adequate data descriptions, although we should then exercise care when forecasting many periods ahead.

For many marketing economic time series, it is conceivable that the time series converge to a certain maximum or minimum as time goes by. The market penetration of durable consumer goods such as personal computers or mobile phones may be close to 100 percent in the long-run. Sales may have been small initially, while they rise through the adoption of the new product by the majority, and ultimately sales may decline such that penetration converges to some saturation level. Hence, for time series variables that characterize some product or market life-cycle, we may wish to modify the above trend models to allow for a saturation level.

An overview of such so-called growth curves is given in Mahajan *et al.* (1993), see also Meade and Islam (1995). Two frequently used growth curves in practice are the

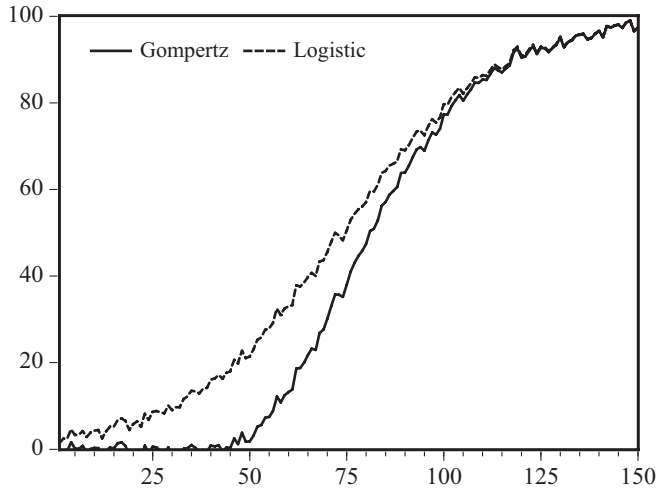


Figure 4.5: Example of a Gompertz curve (4.28) and a logistic curve (4.29) with $\alpha = 100$, $\beta = 4$ and $\gamma = 0.05$.

Gompertz growth curve, given by

$$y_t = \alpha \exp[-\beta \exp(-\gamma t)], \tag{4.28}$$

and the logistic growth curve, given by

$$y_t = \alpha/[1 + \beta \exp(-\gamma t)], \tag{4.29}$$

where α is the saturation level, and α, β and γ are all positive parameters. An example of a typical growth curve pattern is given in Figure 4.5, showing two time series y_t of length $T = 150$, where one series is generated using a Gompertz curve (4.28) with $\alpha = 100$, $\beta = 40$, and $\gamma = 0.05$, while the other is generated using a logistic curve (4.29) with the same parameter values. Standard normal shocks $\varepsilon_t \sim N(0, 1)$ are added to both series.

The expressions in (4.28) and (4.29) show that as time proceeds (that is, when t increases), the y_t value approaches α. The key difference between the two growth curves is the rate of increase towards this saturation level. This can be understood most easily by considering the point of inflection, say, τ, where growth is fastest. It is easily shown that for both the Gompertz curve and for the logistic curve it holds that $\partial^2 y_t / \partial^2 t = 0$ at time $\tau = \log \beta / \gamma$. The level of y_t at the inflexion point differs though. For the Gompertz curve, it holds that $y_\tau = \alpha/e$, while for the logistic curve $y_\tau = \alpha/2$. Hence, growth is slower (faster) before (after) the inflexion point for the Gompertz curve than for the logistic curve, see also Figure 4.5.

A growth curve model that is frequently applied in marketing is the Bass model, see Bass (1969). The model assumes a population of m potential adopters of a new product. For each adopter, the time to adoption is a random variable with a distribution $F(\tau)$ and density $f(\tau)$, such that the hazard rate satisfies

$$\frac{f(\tau)}{1-F(\tau)} = p + qF(\tau). \tag{4.30}$$

The parameters p and q are associated with innovation and imitation. The cumulative number of adopters at time τ, denoted as $N(\tau)$, is a random variable with mean

$$\overline{N}(\tau) = \mathsf{E}[N(\tau)] = mF(\tau). \tag{4.31}$$

The function $\overline{N}(\tau)$ thus satisfies the differential equation

$$\overline{n}(\tau) = \frac{d\overline{N}(\tau)}{d\tau} = p[m - \overline{N}(\tau)] + \frac{q}{m}\overline{N}(\tau)[m - \overline{N}(\tau)]. \tag{4.32}$$

The solution of this differential equation for cumulative adoption is

$$\overline{N}(\tau) = mF(\tau) = m\left[\frac{1-e^{-(p+q)\tau}}{1+\frac{q}{p}e^{-(p+q)\tau}}\right] \tag{4.33}$$

and for adoption itself it is

$$\overline{n}(\tau) = mf(\tau) = m\left[\frac{p(p+q)^2e^{-(p+q)\tau}}{(p+qe^{-(p+q)\tau})^2}\right]. \tag{4.34}$$

Like the logistic and Gompertz curve, $\overline{N}(\tau)$ has a sigmoid pattern and $\overline{n}(\tau)$ has a hump-shaped pattern.

In practice one has discretely observed data, like per year or per quarter. Denote X_t as sales and denote N_t as cumulative sales, where t corresponds to the discretely observed data. Bass (1969) proposes the regression model

$$\begin{aligned} X_t &= pm + (q-p)N_{t-1} - \frac{q}{m}N_{t-1}^2 + \varepsilon_t \\ &= \alpha_1 + \alpha_2 N_{t-1} + \alpha_3 N_{t-1}^2 + \varepsilon_t, \end{aligned} \tag{4.35}$$

where ε_t is white noise with variance σ^2. Bass (1969) recommends use OLS to estimate the parameters in (4.35), where non-linear least squares [NLS] is needed to estimate the standard errors of $\hat{p}$, $\hat{q}$ and $\hat{m}$. Of course, for forecasting only estimates for $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$ are required and OLS can be used.

Boswijk and Franses (2005) have proposed an alternative to the expression in (4.35), which is based on the notion that $\overline{N}(\tau)$ resembles an equilibrium path around which the actual cumulative adoptions fluctuate. The stochastic features of the diffusion process originate from the tendency of the data to revert to that equilibrium path in

an error-correction-type of way. The model to be fitted to actual data would then become

$$X_t = \alpha_1 + \alpha_2 N_{t-1} + \alpha_3 N_{t-1}^2 + \alpha_4 X_{t-1} + \varepsilon_t \tag{4.36}$$

Boswijk and Franses (2005) also propose to make the error process heteroskedastic, so that uncertainty around the diffusion path is largest around the sales peak.

Another convenient empirical version of the Bass model is proposed in Srinivasan and Mason (1986). These authors propose to apply NLS to

$$X_t = m[F_t(p, q) - F_{t-1}(p, q)] + \varepsilon_t, \tag{4.37}$$

where

$$F_t(p, q) = \left[\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}} \right]. \tag{4.38}$$

The Bass model is very often used to forecast future sales data. It is important to recognize Van den Bulte and Lilien (1997) that with data available only before the inflection point reliable estimates of p and q cannot be obtained. Practical sales forecasting of durable products thus usually proceeds along other lines; see Lilien *et al.* (2000) for a detailed account. First, one considers the sales patterns of products that are similar, think of sales data for the PlayStation3 to predict the patterns of PlayStation4. For these similar data one obtains initial values of p and q. The expected value of the total sales is usually set by the manager. What is then helpful are cross-country comparison studies, where the innovation and imitation characteristics of countries are summarized, see for example Chandrasekaran and Tellis (2008), Talukdar *et al.* (2002) and Tellis *et al.* (2003). One can make the characteristics of the diffusion process a function of characteristics of countries.

To make point forecasts for sales one can use one of the above expressions, like in (4.35), (4.36) or (4.37). The model in (4.37) seems easiest to construct forecasts. When $t = n$ is the forecasting origin, and h is the horizon, one can simply use

$$X_{n+h} = \hat{m}[F_{n+h}(\hat{p}, \hat{q}) - F_{n+k-1}(\hat{p}, \hat{q})] + \varepsilon_t \tag{4.39}$$

When the error term is an ARMA type process, straightforward modifications of (4.39) can be made.

Fractional integration

Now we turn back to the models without and with unit roots. The difference between the DT and ST models as discussed above is quite large, in the sense that in a DT model shocks have temporary effects that decay exponentially fast, while in an ST model shocks have permanent effects that do not die out at all. For many economic

time series, it seems that shocks have long-lasting but nevertheless temporary effects. A typical example concerns the effects of the oil crises in the 1970s on macro-economic variables such as output and inflation. Hence, there might be a need for time series models allowing for shocks to have temporary effects that decline at a rate that is slower than exponential. This can be achieved by means of so-called fractional differencing or fractional integration, by allowing d in the filter $(1-L)^d$ to take non-integer values.

The concept of fractional integration within the context of ARIMA models was first put forward by Granger and Joyeux (1980) and Hosking (1981). A fractionally integrated model appears useful to describe time series with very long cycles for which it is difficult to estimate their mean. Such series may also concern variables that experience occasional level shifts, see Chapter 6. Typical applications of such a model include inflation rates, foreign exchange rates, and volatility in financial markets, see, for example, Hassler and Wolters (1995), Cheung (1993), and Andersen *et al.* (2003). Baillie (1996) provides an extensive survey.

The simplest fractionally integrated time series model is given by

$$(1-L)^d y_t = \varepsilon_t, \qquad \text{with } 0 < d < 1, \tag{4.40}$$

where the fractional differencing operator $(1-L)^d$ is defined by the binomial expansion

$$(1-L)^d = 1 - dL - \frac{d(1-d)L^2}{2!} - \frac{d(1-d)(2-d)L^3}{3!} - \cdots - \frac{d(1-d)(2-d)\cdots((j-1)-1-d)L^j}{j!} - \cdots. \tag{4.41}$$

which obviously becomes equal to 1 for $d = 0$ and equal to $(1-L)$ for $d = 1$. The expansion in (4.41) shows that y_t can be described by an AR model of infinite order with a specific structure imposed on the autoregressive coefficients. Similarly, y_t can be written as an MA model of infinite order. In this representation, the coefficient for ε_{t-k} is proportional to k^{d-1} for large k, which demonstrates that the effects of a shock declines only at a hyperbolic rate. Furthermore, it can be shown that the ACF of y_t does not decline towards zero at the familiar exponential rate but rather at a much slower hyperbolic rate, where ρ_k is proportional to k^{2d-1} when $d < 0.5$. For that reason, fractionally integrated time series often are said to have "long memory." When $0 < d < 0.5$, the time series is stationary, in the sense that the sum of the absolute autocorrelations is still finite. When $0.5 < d < 1$, y_t is non-stationary.

The so-called fractional white noise model (4.40) may not be sufficient to adequately describe the dynamics of the y_t series. It can be augmented with additional autoregressive and moving average terms though, which leads to the autoregressive fractionally

integrated moving average model of orders p, d, and q [ARFIMA(p,d,q)], that is,

$$\phi_p(L)(1-L)^d y_t = \theta_q(L)\varepsilon_t, \tag{4.42}$$

where $\phi_p(L)$ and $\theta_q(L)$ are lag polynomials of orders p and q, respectively, as defined in the previous chapter. The conditions for stationarity of this model are $d < 0.5$ together with the requirement that all roots of the AR-polynomial are outside the unit circle. Similarly, invertibility requires that $d > -0.5$ and that all solutions to $\theta_q(z) = 0$ are outside the unit circle.

Several estimation methods for the parameters in ARFIMA(p,d,q) models are available, including exact maximum likelihood developed by Sowell (1992). In practice, the interest often centers on the value of the long memory parameter d, for which the semi-parametric estimator proposed by Geweke and Porter-Hudak (1983) is very popular. Beran (1995) proposes an approximate maximum likelihood (AML) estimator for invertible and possibly non-stationary ARFIMA models based on least squares. The AML estimator amounts to minimizing the sum of squared residuals

$$Q_T(\theta) = \sum_{t=1}^{T} e_t^2(\theta), \tag{4.43}$$

where $\theta = (\phi, d, \theta, \mu)$ and the residuals $e_t(\theta)$ are computed as

$$e_t(\theta) = (y_t - \mu) - \sum_{j=1}^{t+p-1} \pi_j (y_{t-j} - \mu),$$

where the π_j's are the autoregressive coefficients in the infinite order AR representation

$$(y_t - \mu) - \pi_1(y_{t-1} - \mu) - \pi_2(y_{t-2} - \mu) - \cdots = \varepsilon_t,$$

or $\pi(L)(y_t - \mu) = \varepsilon_t$ with $\pi(L) = \theta_q^{-1}(L)\phi_p(L)(1-L)^d$. The AML estimator is asymptotically efficient if the errors ε_t are normally distributed. When normality of ε_t does not hold, it is still consistent and asymptotically normal. We refer to Doornik and Ooms (2003, 2004) for a useful review and comparison of alternative estimation methods.

As mentioned above, long memory is often found in the volatility of financial asset returns. Figure 4.6 shows the first 250 empirical autocorrelations for absolute first differences of the log gold price series, over the period January 1, 1978–December 31, 2012 (9131 observations). As discussed in Chapter 2, first differences of logs are approximately equal to returns, but absolute returns often are considered as a measure of volatility. While the first-order autocorrelation is not particularly large at $\hat{\rho}_1 = 0.251$, Figure 4.6 suggests that the decay of the EACF is very slow. Given the fact that the standard error of $\hat{\rho}_k$ is equal to 0.01, even $\hat{\rho}_{200} = 0.074$ is still significantly positive! We use the AML method to estimate the parameters in an ARFIMA(0,d,1),

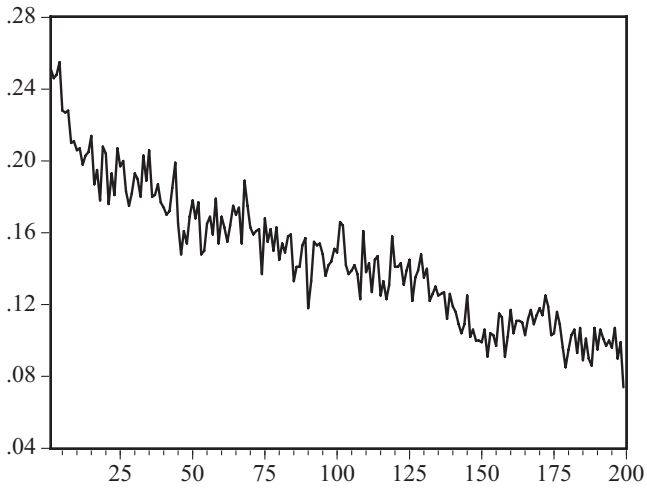


Figure 4.6: Empirical autocorrelation function for absolute daily gold returns, 1/1/1978–31/12/2012.

where the MA(1) component is included to handle residual autocorrelation. This gives the following estimation results, with standard errors of the estimated parameters in parentheses:

$$\underset{}{(1-L)^{\hat{d}}(y_t - 0.877)} = \underset{(0.122)}{(1 + 0.256L)\hat{\varepsilon}_t} \quad \text{with } \underset{(0.019)}{\hat{d} = 0.340} \tag{4.44}$$

The estimate of d is significantly positive but also significantly less than 0.5, suggesting that volatility of gold returns indeed is fractionally integrated but stationary.

Although often an ARFIMA model may provide an improvement over (possibly lengthy) AR models in terms of in-sample fit, the evidence on the usefulness of long memory models for out-of-sample forecasting is mixed. On the one hand, Crato and Ray (1996) show that simple AR models outperform ARFIMA models in out-of-sample forecasting, also because selecting the appropriate AR and MA orders p and q is difficult and because estimation of d can be quite complicated. By contrast, Brodsky and Hurvich (1999) provide simulation evidence that a fractionally integrated model provides substantially more accurate forecasts than an approximating ARMA model, especially at long forecast horizons. Bhardwaj and Swanson (2006) conduct an extensive empirical analysis of a large number of macroeconomic and financial time series, also finding that often the forecasts from ARFIMA models are more accurate than those from a variety of AR, MA, and ARMA models.

4.2 Unit root tests

As already indicated before, one of the specific features of a time series that is governed by a stochastic trend is that its AR representation (or the AR part of an ARMA model) contains the component $(1 - L)$. In other words, the AR polynomial can then be decomposed as

$$\begin{aligned}\phi_p(L) &= 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p \\ &= (1 - \phi_1^* L - \phi_2^* L^2 - \cdots - \phi_{p-1}^* L^{p-1})(1 - L),\end{aligned} \tag{4.45}$$

or written more compactly

$$\phi_p(L) = \phi_{p-1}(L)(1 - L). \tag{4.46}$$

From (4.46), it is easy to see that having a component $(1 - L)$ in the AR-polynomial $\phi_p(L)$ is equivalent to saying that $\phi_p(L)$ contains a unit root, that is, $z = 1$ is a solution of the characteristic equation

$$\phi_p(z) = 0.$$

From (4.45) it follows that in that case the AR-parameters $\phi_1, \ldots, \phi_p$ sum to one, because

$$\phi_p(1) = 1 - \phi_1 - \phi_2 - \cdots - \phi_p = 0. \tag{4.47}$$

For an I(1) series, $\phi_p(L)$ contains a single unit root, such that the polynomial $\phi_{p-1}(L)$ in (4.46) has all roots outside the unit circle. For an I(2) time series, it holds that the AR-polynomial contains two unit roots, that is $z = 1$ also is a solution to $\phi_{p-1}(z) = 0$. Equivalently, when a time series is I(2), it holds true that

$$\phi_p(L) = \phi_{p-2}(L)(1 - L)^2, \tag{4.48}$$

where all the roots of $\phi_{p-2}(L)$ are outside the unit circle.

A single unit root

In order to test for a single unit root in an AR(p) model, we may wish to test the restriction (4.47). A simple approach to this testing problem is proposed in Dickey and Fuller (1979). It is based on the fact that we can always write

$$\phi_p(L) = (1 - \phi_1 - \phi_2 - \cdots - \phi_p)L + \phi_{p-1}^*(L)(1 - L), \tag{4.49}$$

where $\phi^*_{p-1}(L) = 1 - \phi^*_1 L - \cdots - \phi^*_{p-1} L^{p-1}$ with $\phi^*_j = -\sum_{i=j+1}^{p} \phi_i$. For example, consider the AR(2) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t, \tag{4.50}$$

for which (4.49) yields

$$1 - \phi_1 L - \phi_2 L^2 = (1 - \phi_1 - \phi_2)L + \phi^*_1(L)(1 - L), \tag{4.51}$$

where $\phi^*_1(L) = (1 - \phi^*_1 L)$ with $\phi^*_1 = -\phi_2$. The expression in (4.51) can be used to rewrite any AR(2) model into

$$\Delta_1 y_t = (\phi_1 + \phi_2 - 1) y_{t-1} + \phi^*_1 \Delta_1 y_{t-1} + \varepsilon_t. \tag{4.52}$$

Exercise 4.7

In general, based on (4.49), any AR(p) model can be written as

$$\Delta_1 y_t = \rho y_{t-1} + \phi^*_1 \Delta_1 y_{t-1} + \cdots + \phi^*_{p-1} \Delta_1 y_{t-(p-1)} + \varepsilon_t, \tag{4.53}$$

where $\rho = \phi_1 + \phi_2 + \cdots + \phi_p - 1$. When ρ equals zero, (4.53) collapses to an AR(p-1) model for $\Delta_1 y_t$, that is, the AR(p) model containing a unit root becomes an ARI(p-1,1) model. Thus, testing the null hypothesis of a unit root obviously corresponds to testing whether $\rho = 0$ in (4.53). Indeed, the test for a unit root in an AR(p) polynomial proposed by Dickey and Fuller (1979) concerns the t-statistic for the parameter ρ in this auxiliary regression. The alternative hypothesis usually is taken to be stationarity of the time series y_t, that is all roots of the AR-polynomial are outside the unit circle. This requires that the sum of the AR-parameters $\phi_1, \ldots, \phi_p$ is less than 1. Hence the relevant alternative is $\rho < 0$. Values of $\phi_1 + \cdots + \phi_p$ exceeding 1 correspond with an explosive time series, for which a casual glance at the graph already indicates that it cannot be described by a DT model. Hence, the test is one-sided. The asymptotic distribution of the t-statistic for ρ [$t(\hat{\rho})$] is non-standard, see also Phillips (1987). Intuitively, under the null hypothesis of a unit root, the y_t series contains a stochastic trend, and hence its variance and autocovariances depend on time. The denominator of the t-statistic includes a function of such variances, and therefore the distribution of $t(\hat{\rho})$ is not normal. Even though the variance is time-dependent, Phillips (1987) shows that the asymptotic distribution of $t(\hat{\rho})$ exists. In the first panel of Table 4.1, we display the critical values for $t(\hat{\rho})$ in the so-called augmented Dickey-Fuller [ADF] regression (4.53), which were obtained using Monte Carlo simulation. The null hypothesis of a unit root is rejected when $t(\hat{\rho})$ value is below (to the left of) the critical value. If the null hypothesis cannot be rejected, the y_t series should be first differenced prior to any further analysis.

Table 4.1: Critical values for tests to select between deterministic trend and stochastic trend models, where the auxiliary regression may contain a constant and a trend

Auxiliary regression	Sample size	Critical value			
		10%	5%	2.5%	1%
Testing for unit roots with the Dickey-Fuller method (t-test)					
No constant, no trend	25	−1.60	−1.95	−2.26	−2.66
	50	−1.61	−1.95	−2.25	−2.62
	100	−1.61	−1.95	−2.24	−2.60
	250	−1.62	−1.95	−2.23	−2.58
	500	−1.62	−1.95	−2.23	−2.58
	∞	−1.62	−1.95	−2.23	−2.58
Constant, no trend	25	−2.63	−3.00	−3.33	−3.75
	50	−2.60	−2.93	−3.22	−3.58
	100	−2.58	−2.89	−3.17	−3.51
	250	−2.57	−2.88	−3.14	−3.46
	500	−2.57	−2.87	−3.13	−3.44
	∞	−2.57	−2.86	−3.12	−3.43
Constant and trend	25	−3.24	−3.60	−3.95	−4.38
	50	−3.18	−3.50	−3.80	−4.15
	100	−3.15	−3.45	−3.73	−4.04
	250	−3.13	−3.43	−3.69	−3.99
	500	−3.13	−3.42	−3.68	−3.98
	∞	−3.12	−3.41	−3.66	−3.96
Testing for stationarity with the KPSS test (accumulated partial sums)					
Constant, no trend	∞	0.347	0.463	0.574	0.739
Constant and trend	∞	0.119	0.146	0.176	0.216

Source: Fuller (1976) and Kwiatkowski *et al.* (1992).

Lag order selection

The asymptotic distribution of the t-statistic of ρ in the ADF regression (4.53) does not depend on the AR-order p or, in other words, on the number of lagged first differences that is included, under the assumption that this is fixed in advance. Hall (1994) shows that when the lag order p is selected by means of t-tests on the parameters $\phi_1^*, \ldots, \phi_{p-1}^*$ of the lagged first differences $\Delta_1 y_{t-i}$ or via application of the AIC or SIC, the asymptotic distribution remains the same, such that the critical values from the appropriate rows in Table 4.1 can still be used. Intuitively, this can be understood from the fact that (i) in case of a single unit root, the $\Delta_1 y_t$ series does not have a stochastic trend component, while (ii) in case it is already stationary the Δ_1 filter implies overdifferencing. Note that the latter situation leads to a non-invertible MA-component, but these are stationary by definition. In both cases the t-statistics of the AR parameters have an asymptotic standard normal distribution.

In finite samples, the distribution of the ADF test does depend on the fact that the lag order p needs to be selected, and upon the manner in which this is done. Cheung and Lai (1995) provide so-called response surfaces, which can be used to compute appropriate critical values for any sample size T and fixed lag order p. For example, in finite samples, the distribution of the ADF test shifts to the left such that the appropriate critical values become more negative, see also Table 4.1. Put differently, using asymptotic critical values would lead to overrejections of the unit root null hypothesis when it is in fact true. We refer to Schwert (1989), Agiakloglou and Newbold (1992) and Lopez (1997) for simulation evidence on the effects of the choice of p on the size and power properties of the ADF test. Ng and Perron (1995) compare different methods of selecting the AR order, including information criteria such as AIC and SIC, and a so-called general-to-specific methodology, which starts by setting p in (4.53) at a fairly large value and then sequentially lowers it by eliminating the highest-order lagged first difference until it is significant at a pre-specified level. The latter method is favored over those based on information criteria, as it shows less size distortions and has comparable power. Ng and Perron (2001) developed a set of modified information criteria for the purpose of lag selection in unit root test regressions, which alleviate these problems to a large extent.

One of the reasons why the lag order is important and can severely affect the properties of the ADF test is that the AR model in the auxiliary test regression (4.53) might in reality be an approximation to an ARMA(p,q) model. Said and Dickey (1984) show that the ADF test is valid only when the lag order in the ADF regression satisfies certain lower and upper bound conditions. In addition, for such an ARMA model, the MA component should not be approximately similar to the AR component. For example, we may expect difficulties in testing the null hypothesis $\rho = 0$ in the model

$$y_t = (\rho + 1)y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \tag{4.54}$$

when θ_1 is very close to -1, as in that case the AR and MA polynomial are both approximately equal to $(1 - L)$ under the null hypothesis. Schwert (1989) and Ng and Perron (1995) provide simulation evidence showing that indeed the ADF test suffers from severe size distortions under such circumstances.

Deterministic components

The ADF test regression in (4.53) does not include deterministic components such as a constant and a trend. This means that under the null hypothesis the time series y_t has a stochastic trend without drift, while it is a stationary series with zero mean under the alternative. For empirical applications this often is unrealistic, because many time series in economics and business have a non-zero mean or an upward trend. It is fairly straightforward to modify the ADF test regression to incorporate these features. For example, by including an intercept, the model under the null hypothesis becomes a random walk with drift. A further consideration then is that the models under the null and the alternative hypotheses should be 'competitive', in the sense that both are plausible descriptions of the time series under investigation. For example, if only an intercept is included in (4.53), the random walk with drift under the null is tested against the alternative of a stationary AR(p) around a constant mean, which is a test that can almost be done by simply looking at the graph of the time series. Clearly these models imply rather different time series properties, as the model under the null suggests an upward sloping trend (assuming the intercept is positive), while the model under the alternative does not include a (deterministic) trend at all. In sum, we should be careful in treating deterministic components in the ADF test. For illustration, consider again the AR(1) model in (4.4),

$$y_t - \mu - \delta t = \phi_1(y_{t-1} - \mu - \delta(t-1)) + \varepsilon_t, \tag{4.55}$$

which can be rewritten as

$$y_t = (1 - \phi_1)\mu + \phi_1\delta + (1 - \phi_1)\delta t + \phi_1 y_{t-1} + \varepsilon_t. \tag{4.56}$$

When ϕ_1 is unequal to 1, (4.56) can be written as the deterministic trend model

$$y_t = \alpha + \beta t + \phi_1 y_{t-1} + \varepsilon_t, \tag{4.57}$$

with both α and β unequal to zero. However, when $\phi_1 = 1$, (4.56) reduces to the random walk with drift

$$y_t = \delta + y_{t-1} + \varepsilon_t. \tag{4.58}$$

This implies that in the ADF test regression

$$\Delta_1 y_t = \alpha + \beta t + \rho y_{t-1} + \phi_1^* \Delta_1 y_{t-1} + \cdots + \phi_{p-1}^* \Delta_1 y_{t-(p-1)} + \varepsilon_t, \tag{4.59}$$

under the null hypothesis not only ρ equals zero, but also $\beta = 0$. Dickey and Fuller (1981) propose a joint F-test of the hypothesis $\rho = \beta = 0$, and in case of no clear trend, $\rho = \alpha = 0$. In the latter case, the term βt is not included in the regression (4.59), such that we effectively test a random walk without drift against stationarity around a constant, possibly non-zero mean.

A common practical procedure in case we want to consider (4.59) is to use the critical values in panels 2 and 3 of Table 4.1, depending on whether the linear deterministic trend is included in the test regression or not. These critical values are generated using a data generating process [DGP] of a random walk with no drift, and using (4.59) with p set equal to 1. Clearly, the distribution of the ADF test shifts to the left when deterministic components are included. Intuitively, if the data are generated by a random walk model, the inclusion of a trend biases the estimate for ρ away from zero, and hence even larger values of the test statistic are required to be able to reject the null hypothesis.

Exercise 4.8

It should be mentioned that when the DGP really is a random walk with drift μ, the asymptotic distribution of $t(\hat{\rho})$ in (4.59) approaches the standard normal distribution with increasing values of μ, see Hylleberg and Mizon (1989). In practice, however, we never know for sure how big μ is (before knowing whether ρ equals 0 or not), and the regression (4.59) is simply estimated in all cases where the data show an upward or downward trending pattern. Campbell and Perron (1991) show that erroneously neglecting deterministic terms is worse than including redundant variables, which supports this practice.

Two unit roots

Notice that when the y_t series really is I(2), the $\Delta_1 y_t$ variable still has a stochastic trend. Haldrup (1994) shows that in that case the critical values to test one unit root versus no unit root in Table 4.1 cannot be used. If we suspect the possible presence of two unit roots in the AR(p) part of the model, we better first investigate the null hypothesis of two unit roots versus the alternative of a single unit root. In case this null is rejected, we may proceed with testing one versus zero unit roots. In this sequential testing procedure, the relevant ADF test regression in the first round becomes

$$\Delta_1^2 y_t = \rho^* \Delta_1 y_{t-1} + \phi_1^{**} \Delta_1^2 y_{t-1} + \cdots + \phi_{p-2}^{**} \Delta_1^2 y_{t-(p-2)} + \varepsilon_t. \qquad (4.60)$$

Dickey and Pantula (1987) show that critical values for $t(\hat{\rho}^*)$ are again as given in Table 4.1.

Empirical applications

Given its relevance for modeling trends, there have appeared many studies on the size and power properties of the Dickey-Fuller test procedure. As discussed before, when an AR model is the DGP, it appears that the size of the test in small samples is reasonably accurate. In contrast, when an ARMA model is the DGP, with the MA component having a root close to the unit circle, the empirical size of the test becomes too large, see Schwert (1989), Agiakloglou and Newbold (1992), and Lopez (1997). Unfortunately, in small samples the power of the ADF test is not very high for time series that have autoregressive roots quite close to one, which seems to be the relevant case for many economic variables. As expected, it is not easy to distinguish between $\phi_p(1) = 0$ and $\phi_p(1) = 0.05$. However, when the sample size increases, the power appears to increase quite rapidly. For practical application, a simple guideline is to evaluate the ADF test not only at the 5 percent level, but also at, say, the 10 percent (or even the 20 percent) level. The overall conclusion is that care should be exercised when evaluating the ADF test results. In case of doubt, we may even be better off assuming the possible adequacy of both the DT and ST model, and to see which of the two does a better job in out-of-sample forecasting. Further confidence in the empirical outcomes is also obtained when the ADF test results appear robust to changes in the sample size, to outliers, to additional lags and to the inclusion or exclusion of deterministic components.

We illustrate the ADF test procedure by testing for the presence of unit roots in AR models for various empirical time series discussed previously. The ADF test is implemented using (4.59), where the value of p is determined by a 'general-to-specific' procedure starting with lag order $p_{\max}$ and sequentially lowering it by eliminating the highest-order lagged first difference until its coefficient is significant at the 10% significance level. The test results appear in Table 4.2.

The first two series in Table 4.2 concern the daily gold and silver prices covering the period 1978–2002. Note that we exclude the later years as Figure 2.20 indicate a huge trend from 2002 until 2012. The results are mixed. The auxiliary regressions for the ADF test do not include a deterministic linear trend, as Figure 2.20 shows that no such trends seem present in these series. Hence, a deterministic trend model does not seem a reasonable data description. A constant is included, though, to allow for a non-zero mean under the DT alternative. For silver, the ADF test equals -1.74, indicating that the unit root null cannot be rejected. By contrast, we find support for a DT model for gold given that ADF$= -3.156$, which allows rejection of the unit root hypothesis at the 5% significance level. Perhaps this reflects that these time series contain several aberrant observations at the end of the 1970s, which may influence the test for a unit root, as we will see in Chapter 6.

The Latin American GDP per capita series obviously show trending patterns, see Figure 2.1 and hence the auxiliary ADF regressions include a constant and a trend. For

Table 4.2: Testing for unit roots: some empirical examples

Variable	Linear trend?	T	p_{max}	p	ADF
Gold	no	6249	13	4	−3.156*
Silver	no	6249	13	3	−1.742
GDP per capita in					
Argentina	yes	60	4	1	−1.569
Brazil	yes	60	4	1	−1.342
Chile	yes	60	4	1	−1.292
Colombia	yes	60	4	2	−2.565
Mexico	yes	60	4	1	−1.433
US IP (SA)	yes	160	4	2	−3.594*
UK Con (SA)	yes	160	4	1	−1.363

Notes: * indicates significance at the 5% level. "Linear trend?" indicates whether a linear deterministic trend is included in the auxiliary regression (4.59) or not. T denotes the effective number of observations, p is the lag order of the AR(p) model where the possible $(1 - L)$ component is still included, and ADF is the value of the t-statistic for ρ in the augmented Dickey-Fuller regression (4.59). The value of p is determined by a 'general-to-specific' procedure starting with lag order p_{max} in (4.59) and sequentially lowering it by eliminating the highest-order lagged first difference until its coefficient is significant at the 10% significance level.

all five countries, we cannot reject the presence of unit root even at the 10 percent level, indicating that shocks to output occurring in these economies have permanent effects. Also note that for all series, the sequential AR-order selection procedure indicates that almost no lagged first differences need to be included in the test regression.

Finally, using the estimation sample 1961.1 to 2000.4 (where the first few observations are used as pre-sample starting values and the remainder of the time series is saved for forecast evaluation later on) for the quarterly US industrial production and UK (log) consumption series (seasonally adjusted), we find that the ADF test statistic takes the value −3.594 and −1.363 respectively. It seems that shocks to the AR(2) model for the production variable are not permanent, which may not be in line with

the graphical impression from Figure 4.2. For the UK consumption series, the value clearly indicates the existence of a stochastic trend.

As an example of the possible presence of two unit roots, consider again the stock of motorcycles in The Netherlands for the period 1948–1991 as shown in Figure 2.2. An AR(2) model for this series (when no logs are taken) results in

$$\underset{}{y_t} = \underset{(2.463)}{3.618} + \underset{(0.084)}{1.928}\, y_{t-1} - \underset{(0.083)}{0.954}\, y_{t-2} + \hat{\varepsilon}_t. \tag{4.61}$$

Clearly, the value of 2 is included in the 95 percent confidence interval for $\hat{\phi}_1$ and -1 is included in the similar interval for $\hat{\phi}_2$, suggesting that the AR(2) polynomial can in fact be written as $(1 - 2L + L^2) = (1 - L)^2$. As mentioned above, the proper procedure is now to first consider the auxiliary regression (4.60), which for this time series gives

$$\Delta_1^2 y_t = \underset{}{0.514} - \underset{(0.084)}{0.055}\, \Delta_1 y_{t-1} + \hat{\varepsilon}_t, \tag{4.62}$$

which results in a t-ratio of -0.660. In sum, this variable seems to be I(2).

4.3 Stationarity tests

The unit root testing procedure discussed in the previous section compares the ST model under the null hypothesis with the DT model under the alternative. When the unit root null hypothesis cannot be rejected, the ST model is preferred over the DT model. The reason for taking the ST model as the most important model (as the null hypothesis is only rejected if the test finds strong evidence against it) is that the ST model assumes permanent effects of shocks, which can be important from an economic policy perspective. Given that the power of unit root tests against relevant alternatives is not particularly high, it may also be of interest to apply tests of the null hypothesis of stationarity (so, treat the two models the other way around), to prevent us from possibly erroneously concluding that a given time series has a unit root due to the statistical properties of the ADF test.

A test that takes stationarity as the null hypothesis is developed in Kwiatkowski *et al.* (1992) [KPSS]. It is based on the idea of decomposing a time series into the sum of a deterministic trend δt, a random walk S_t, and a stationary error process u_t, that is,

$$y_t = \delta t + S_t + u_t, \tag{4.63}$$

where S_t is the random walk or stochastic trend

$$S_t = \sum_{i=1}^{t} \varepsilon_i = S_{t-1} + \varepsilon_t, \tag{4.64}$$

with $S_0 = 0$. Notice that this is a generalization of the representation of a random walk with drift, as given in (4.12). When the variance of ε_t, denoted as σ^2 as usual, is equal to zero, the random walk component S_t disappears from (4.63) such that the time series y_t is stationarity. Hence, the null hypothesis of stationarity to be tested is given by $\sigma_\varepsilon^2 = 0$. As shown in Kwiatkowski *et al.* (1992), the appropriate test statistic is computed as

$$\hat{\eta} = \frac{1}{T^2 s^2(l)} \sum_{t=1}^{T} \hat{S}_t^2, \tag{4.65}$$

where $\hat{S}_t = \sum_{i=1}^{t} \hat{e}_i$ is the partial sum series of the residuals $\hat{e}_t$ from the auxiliary regression

$$y_t = \hat{\tau} + \hat{\delta} t + \hat{e}_t, \tag{4.66}$$

and $s^2(l)$ is an estimate of the so-called long-run variance $\sigma^2 = \lim_{T\to\infty} T^{-1}\mathsf{E}[S_T^2]$. Following Phillips (1987) and Phillips and Perron (1988), we can estimate this by

$$s^2(l) = \frac{1}{T}\sum_{t=1}^{T} \hat{e}_t^2 + \frac{2}{T}\left[\sum_{j=1}^{l} w(j,l) \sum_{t=j+1}^{T} \hat{e}_t \hat{e}_{t-j}\right], \tag{4.67}$$

where the weights $w(j, l)$ can be set equal to

$$w(j,l) = 1 - j/(l+1), \tag{4.68}$$

see Newey and West (1987), although also other weights are possible. The lag length l is usually set proportional to $T^{\frac{1}{3}}$, based on Newey and West (1994).

The asymptotic distribution of the test statistic $\hat{\eta}$ in (4.65) is derived in Kwiatkowski *et al.* (1992). For further use, we present critical values in the last panel of Table 4.1. In the first row of that panel, we give the critical values in case (4.66) does not contain a linear trend, which is the relevant statistic for testing the null hypothesis of stationarity around a constant mean. The null hypothesis of (trend-)stationarity is rejected when $\hat{\eta}$ exceeds the critical value. Notice that as $\hat{\eta}$ can take positive values only, this test procedure is also one-sided.

As an illustration, consider again the (seasonally adjusted) UK consumption series, for the sample period 1961.1 to 2000.4. We use the regression (4.66) including a trend and we set the lag length l in (4.67) equal to 10 based upon the Newey and West (1994) procedure. The resulting KPSS test statistic η takes the value 0.203, which implies that the null hypothesis of stationarity is rejected at the 5% significance level. Hence,

the ADF and the KPSS test suggest that the ST model is to be preferred over the DT model as a representation of the trend in UK consumption.

4.4 Forecasting

An important issue when selecting a model for the trend in economic data is the impact of this choice on out-of-sample forecasts, and in particular on their accuracy. Adding to the remarks made in Section 4.1, here we discuss if and how point forecasts can differ across DT and ST models, and illustrate that interval forecasts become wider when a unit root is assumed.

Consider the case where a time series y_t can be described by an AR(1) model with a non-zero unconditional mean μ,

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \varepsilon_t. \tag{4.69}$$

Using recursive substitution, it follows that for any $h \geq 1$ we can write

$$\begin{aligned} y_{t+h} - \mu &= \phi_1^h(y_t - \mu) + \varepsilon_{t+h} + \phi_1\varepsilon_{t+h-1} + \cdots + \phi_1^{h-1}\varepsilon_{t+1} \\ &= \phi_1^h(y_t - \mu) + \sum_{i=1}^{h} \phi_1^{h-i}\varepsilon_{t+i}. \end{aligned} \tag{4.70}$$

Hence, the optimal h-step ahead forecast of y_{T+h} made at time T is given by

$$\hat{y}_{T+h|T} \equiv \mathsf{E}[y_{T+h}|\mathcal{Y}_T] = \mu + \phi_1^h(y_t - \mu), \tag{4.71}$$

and the corresponding forecast error is equal to $e_{T+h|T} = \sum_{i=1}^{h} \phi_1^{h-i}\varepsilon_{T+i}$ with variance $\mathsf{V}[e_{T+h|T}] = \sigma^2 \sum_{i=1}^{h} \phi_1^{2(h-i)}$. Specifically, for $h = 1$, 2, and 3 we have

$$\begin{aligned} \mathsf{V}[e_{T+1|T}] &= \sigma^2, \\ \mathsf{V}[e_{T+2|T}] &= (1 + \phi_1^2)\sigma^2, \\ \mathsf{V}[e_{T+3|T}] &= (1 + \phi_1^2 + \phi_1^4)\sigma^2. \end{aligned}$$

Assuming normality of ε_t, we may construct a 95% interval forecast for y_{T+h} as

$$(\hat{y}_{T+h|T} - 1.96\sqrt{\mathsf{V}[e_{T+h|T}]},\ \hat{y}_{T+h|T} + 1.96\sqrt{\mathsf{V}[e_{T+h|T}]}),$$

as discussed in Section 3.5.

Now suppose that we decide to use an AR(1) model for forecasting y_t but impose a unit root, that is, we set $\phi_1 = 1$ in (4.69). Hence, effectively the model reduces to a random walk without drift $y_t = y_{t-1} + \varepsilon_t$. This leads to a number of interesting insights concerning both point and interval forecasts. First, from (4.71) it follows that for all horizons h the point forecast $\hat{y}_{T+h|T}$ becomes equal to the current value y_T.

Hence, the difference between the point forecasts obtained from the stationary AR(1) model (4.69) and the random walk is equal to

$$(\phi_1^h - 1)(y_T - \mu),$$

which becomes large when ϕ_1 is far from one or the current level of the time series is far from its unconditional mean μ.

Second, the variance of the h-step ahead forecast error becomes equal to $\mathsf{V}[e_{T+h|T}] = h\sigma^2$ under the unit root assumption, which clearly exceeds that of the stationary AR(1) model. In other words, the nonstationarity of a random walk time series is reflected by wider interval forecasts or, put differently, less confidence in point forecasts.

To illustrate this, consider again the stock of motorcycles in the Netherlands. We obtain h-step ahead forecasts for this series using an unrestricted AR(2) model that includes a quadratic deterministic trend, an ARI(1,1) model with a linear deterministic trend, and an I(2) model including an intercept. The latter model is selected based on the Dickey-Fuller test as discussed before in Section 4.2. The parameters in the three models are estimated using the sample period 1948–1991, and out-of-sample forecasts are generated for 1992 to 2011, that is, the forecast horizon h ranges from 1 to 20. Table 4.3 shows the standard deviation of the h-step ahead forecast errors. Obviously, the more unit roots are imposed, the larger these standard deviations are. Furthermore, the differences increase substantially when h gets larger. Given that larger standard deviations of the forecast errors lead to wider interval forecasts, at first sight it may seem odd that the actual observations for 2008–2011 are outside the 95% interval forecast for the ARI(1,1) model, while they are not for the AR(2) model. This can be explained by the fact that the point forecasts $\hat{y}_{T+h|T}$, which determines the center (or 'location') of the interval forecast, also differs across models.

Another illustration is provided by the US industrial production series, for which we consider the unrestricted AR(2) model

$$\underset{}{4y_t} = \underset{(0.058)}{0.157} + \underset{(0.136)}{(0.287(t/1000)} + \underset{(0.076)}{1.500y_{t-1}} - \underset{(0.075)}{0.544y_{t-2}} + \hat{\varepsilon}_t. \qquad (4.72)$$

and the ARI(1,1) model

$$\underset{}{\Delta y_t} = \underset{(0.0015)}{0.0037} + \underset{(0.076)}{0.538\ \Delta y_{t-1}} + \hat{\varepsilon}_t. \qquad (4.73)$$

where the parameters in both models are estimated for the sample period 1960.3 to 1990.4. The h-step out-of-sample point forecasts generated from both models for 1991.1–2005.4, as well as the lower and upper bounds of the 95% interval forecasts, are displayed in Figures 4.7 and 4.8.

The point forecasts from the ARI(1,1) model in Figure 4.8 seem somewhat closer to the actual observations in the out-of-sample period than the forecasts from the AR(2)

Table 4.3: Standard errors of *h*-step ahead forecasts of the level of the stock of motorcycles

	Model		
Year	AR(2)	ARI(1,1)	I(2)
1992	5.23 (o)	5.32	4.81 (o)
1993	10.86 (o)	11.70	10.76 (o)
1994	16.92	19.28	18.01
1995	23.10	27.82	26.37
1996	29.21	37.13	35.70
1997	35.15	47.10	45.92
1998	40.87	57.63	56.96
1999	46.37	68.65	68.76
2000	51.68	80.10	81.27
2001	56.84	91.93	94.46
2002	61.91	104.09	108.29
2003	66.94	116.57	122.74
2004	71.99	129.32	137.77
2005	77.11	142.32	153.38
2006	82.35	155.56	169.52
2007	87.73	169.02	186.20
2008	93.29	182.68 (o)	203.40
2009	99.04	196.53 (o)	221.09
2010	105.00	210.56 (o)	239.26
2011	111.17	224.77 (o)	257.91

Notes: The table reports standard errors of *h*-step ahead forecast errors of level of the stock of motorcycles in the Netherlands for the years 1992-2011, using models specified and estimated using observations for the period 1948–1991. An "o" in parentheses indicates that the true observation lies outside the 95% interval forecast.

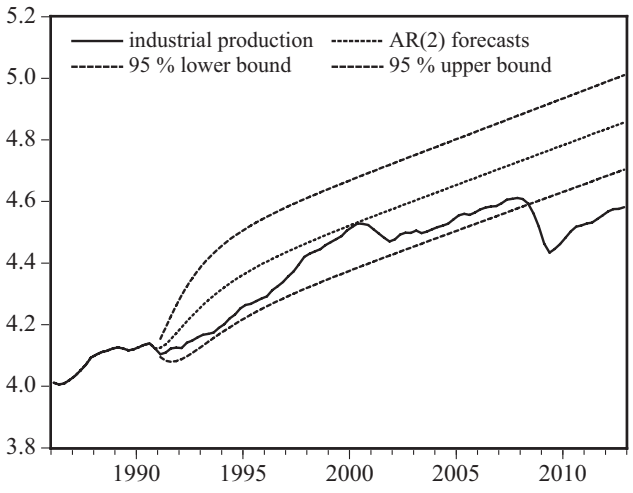


Figure 4.7: Point forecasts and 95% interval forecasts from the AR(2) (4.72) for US industrial production, 1991.1–2012.4.

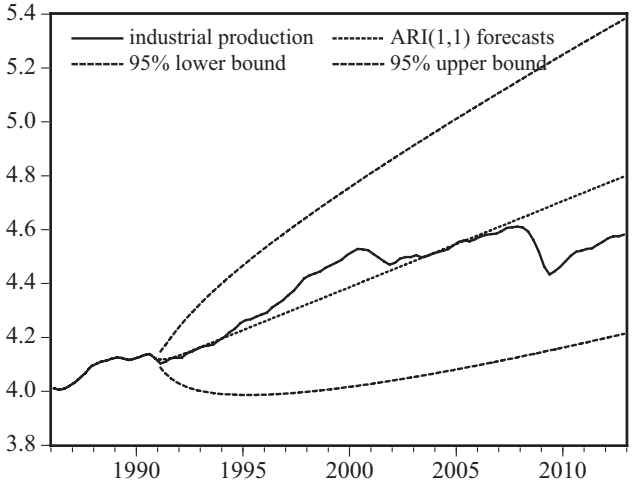


Figure 4.8: Point forecasts and 95% interval forecasts from the ARI(1,1) model (4.73) for US industrial production, 1991.1–2012.4.

model in Figure 4.7. This comes at the cost of much wider 95% interval forecasts though or, put differently, larger uncertainty in the point forecasts.

The discussion of forecasting with an AR(1) model at the start of this section may suggest that we should always use unrestricted models and never impose unit roots, as this leads to large uncertainty in point forecasts or, equivalently, wider interval forecasts. Several additional issues play a role, however, that qualify this conclusion. First, it may be that the autoregressive representation of a given time series y_t truly does contain a unit root. If that is the case, of course it is not harmful to impose it. In fact, it is beneficial to do so, as it reduces the number of parameters to be estimated. This leads us to the second issue, namely that in the analysis above we have implicitly assumed that the parameters ϕ_1 and μ in the AR(1) model (4.69) are known. In practice, they of course need to estimated based on observations $y_1, \ldots, y_T$. Denoting these estimates as $\hat{\phi}_1$ and $\hat{\mu}$, the h-step ahead point forecast from the AR(1) model is $\hat{y}_{T+h|T} = \hat{\mu} + \hat{\phi}_1^h(y_t - \hat{\mu})$ instead of (4.71). Using (4.70), the corresponding forecast error can be written as

$$e_{T+h|T} = \sum_{i=1}^{h} \phi_1^{h-i} \varepsilon_{T+i} + (\hat{\mu} - \mu)(\hat{\phi}_1 - 1) + (\phi_1^h - \hat{\phi}_1^h)(y_T - \mu). \quad (4.74)$$

The forecast error variance $\mathsf{V}[e_{T+h|T}]$, given by the expected value of the square of (4.74), thus increases due to the estimation uncertainty in both μ and ϕ_1. In fact, for autoregressive parameters close to 1, it may become larger than the forecast error variance from the random walk model (which is $h\sigma^2$ as shown above), such that imposing a unit root in that case leads to tighter interval forecasts. We refer to Elliott (2006) for elaborate discussion of this issue.

Third, and perhaps most important, in reality we often are not certain whether a time series contains a unit root or not. Furthermore, even if the time series does not have a stochastic trend, we do not know how 'close' it is to being unit root nonstationary, and how large the impact of estimation error is. As a practical device, we may first use the Dickey-Fuller (or any other unit root) test to examine whether the time series at hand contains a unit root, and to select between a DT model or an ST model. Diebold and Kilian (2000) provide simulation evidence showing that such a so-called 'pre-test' procedure is useful, in the sense that it generally leads to more accurate forecasts relative to always imposing a unit root or never doing so. Stock and Watson (1999) conduct a large-scale empirical forecasting experiment for US macro-economic time series that supports this finding.

CONCLUSION

In this chapter we have reviewed a selection of methods that can help to decide on the most useful representation of a trend in economic time series. For practical purposes

it seems sensible to evaluate various models with different trend specifications and to select the most appropriate model based on forecasting the hold-out data. In the next chapter, we will see that the analysis of trends can become more involved when the time series also displays seasonality.

EXERCISES

4.1 Suppose you have estimated the regression model (4.6) for a given time series y_t, $t = 1, \ldots, T$, which has produced estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\phi}_1$. How would you proceed to obtain estimates of μ and δ in (4.4)?

4.2 Show that the AR(1) model with deterministic linear trend as given in (4.4) becomes equal to the random walk with drift model in (4.9) in case $\phi_1 = 1$.

4.3 Simulate time series y_t, $t = 1, \ldots, T$ with $T = 200$ from the DT and ST models given in (4.13) and (4.14) for values of the drift δ equal to 0.01, 0.05, 0.10, 0.50 and 1. Does the difference between the DT and ST series depend on the value of δ?

4.4 Simulate time series from the DT and ST models

$$\text{DT} : y_t = 0.1t + \phi_1(y_{t-1} - 0.1(t-1)) + \varepsilon_t, \quad t = 1, 2, 3, \ldots, 200, \tag{4.75}$$

$$\text{ST} : y_t = 0.1 + y_{t-1} + \varepsilon_t, \qquad t = 1, 2, 3 \ldots, 200 \text{ and } y_0 = 0, \tag{4.76}$$

where the shocks ε_t are the same in both equations, and drawn from a $N(0, 1)$ distribution. Consider values of the AR(1) parameter ϕ_1 equal to 0.5, 0.7, 0.9, 0.95 and 0.98. How does the difference between the DT and ST series depend on the value of ϕ_1?

4.5 Show that the expressions in (4.22) and (4.23) can be combined into (4.24).

4.6 This idea is due to the famous econometrician Prof. C.W.J. Granger, for which he received the Nobel prize in 2003. Have a look at http://www.nobelprize.com to read the main reasons why he received this important award.

4.7 Consider the AR(3) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \varepsilon_t. \tag{4.77}$$

Show that when $\phi_1 + \phi_2 + \phi_3 = 1$, this can be written as an AR(2) model for $(1 - L)y_t$. Rewrite the AR(3) model such that this parameter restriction can be tested using a $t-$test.

4.8 Use simulation to demonstrate that the distribution of the Dickey-Fuller t-test for $\rho = 0$ in (4.59) with $p = 0$ shifts to the left when the linear deterministic trend βt is included, for time series generated using a random walk with drift.

5 Seasonality

In this chapter the focus is on seasonal fluctuations in business and economic time series. The graphs in Section 2.2 suggest that seasonal fluctuations can be the dominant source of variation in a quarterly or monthly observed time series, once we have dealt with the trend in the series. For example, the results of the auxiliary regression (2.3) for the first differences of quarterly UK consumption indicates that more than 90 percent of the variation in this series can be assigned to seasonal movements. Similar numbers are routinely found for a host of other (macro-)economic time series, see Miron (1996), among others. A second observation from the graphs in Section 2.2 is that often the seasonal patterns do not appear to be stable over time. Such evolving patterns may emerge because the time series behavior in one or a few seasons changes, while it may also be that total seasonal variation changes as time passes by.

There are two different, and in fact opposing, attitudes towards seasonality. The first is to consider seasonality as a 'nuisance', which is not of much interest in itself and only complicates the analysis of other relevant time series features, such as the trend and nonlinearity. Put differently, seasonality is regarded as a form of data contamination, suggesting that seasonal fluctuations should be removed prior to any further analysis of the time series. This is the rationale behind so-called seasonal adjustment procedures. The second attitude towards seasonality is to consider it as an inherent feature of a time series, which should be incorporated in a model that is used for description and forecasting. The latter is the approach taken in this chapter.

Explicitly modeling seasonality in a time series seems preferable to removing it *a priori* for at least three reasons. First, in many situations the seasonal fluctuations and forecasts thereof *are* of interest in their own right. For example, in marketing and tourism applications, forecasts on possible changes in seasonality can be very useful, additional to forecasts of the long-run trend. Similarly, information on the nature of seasonal movements in variables such as production, investment, and inventory building may be of interest, both for individual firms and for policymakers, see Carpenter and Levy (1998). Second, seasonal adjustment procedures essentially assume that an observed time series y_t can be considered as the sum (or product) of a non-seasonal

component y_t^{ns} and a seasonal component y_t^s. The objective of these procedures is to remove the latter component, such that the resulting seasonally adjusted series can then be analyzed, for example, for the presence of deterministic or stochastic trends using the methods discussed in the previous chapter. Essentially, this presupposes that seasonal fluctuations are independent from other features of a time series, such as its trend and cyclical behavior. This assumption may not always be realistic in practice, see Miron and Beaulieu (1996), Cecchetti *et al.* (1997), and Krane and Wascher (1999), among others. It also goes against theoretical models that explicitly describe economic decision-making in a seasonal context, see Ghysels (1994) and Miron (1996) for reviews. Third, the seasonal component y_t^s is unobserved and needs to be estimated from the data before it can be removed. It has been well-documented that this may distort other time series features, including trends, structural breaks, and nonlinearity, see Ghysels and Perron (1993, 1996), Ghysels *et al.* (1996), and Franses and Paap (1999), among others. Hence, when analyzing the trend characteristics of a time series, we may obtain false conclusions due to the initial seasonal adjustment. We refer to Fok *et al.* (2006) for an assessment of the performance of the popular Census X12-ARIMA and TRAMO/SEATS seasonal adjustment procedures.

Seasonality can be incorporated in autoregressive moving average [ARMA] models in several different ways. The two most common approaches will be discussed in this chapter. In Section 5.1, we consider deterministic and (nonstationary) stochastic seasonality, resembling the discussion of deterministic and stochastic trends in the previous chapter. It is shown that the selection between these alternative representations of seasonality can be based on tests for seasonal unit roots, which subsequently are reviewed in Section 5.2. A third approach to modeling seasonality concerns so-called periodic autoregressive [PAR] models, in which the AR parameters are allowed to vary across the seasons. We refer to Franses and Paap (2004) for a detailed discussion of this approach. Forecasting seasonal time series is the focus of Section 5.3. We discuss the properties of out-of-sample forecasts that are obtained with different models for seasonality, and review the empirical evidence on the forecast accuracy of the alternative approaches.

Notation

The notation in this chapter is as follows. The time series y_t, $t = 1, 2, \ldots, T$, is observed during S seasons per year, where S may take such values as 2, 4, 6, 12, or 13. We should note that the methods discussed in this chapter are not limited in their use to seasonal patterns that occur within a year. They are equally applicable for analysing, for example, day-of-the-week effects, which sometimes is relevant for financial time series. For notational convenience, but without loss of generality, we typically assume that the first observation y_1 is made in the first season $s = 1$ of the year, and that $T/S = N$ is integer, that is, observations y_t are available for all seasons in each of the

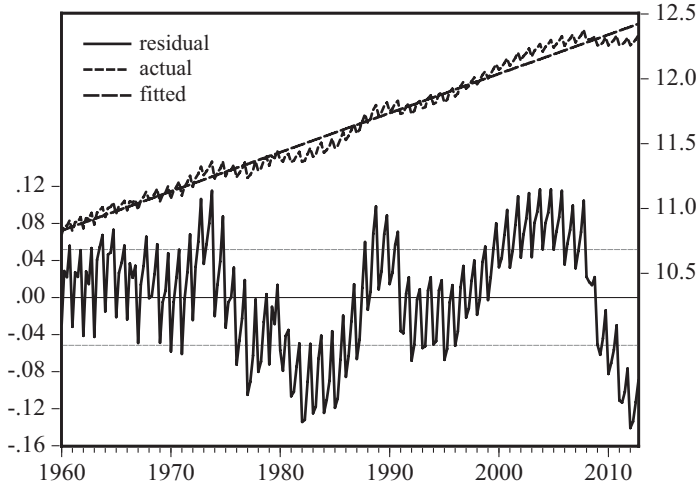


Figure 5.1: Results of a regression of quarterly UK household consumption on an intercept and a linear deterministic trend.

N years. Sometimes we write $t = S(n-1)+s$, for certain $n = 1, \ldots, N$, to indicate that the t-th observation of the time series corresponds with the s-th season in year n. We make extensive use of so-called seasonal dummy variables, which are denoted $D_{s,t}$, $s = 1, 2, \ldots, S$. These $D_{s,t}$ variables take a value of 1 in season s and a value of 0 in all other seasons. Note that in a regression with S seasonal dummies, it is not possible to include a constant term because of perfect multicollinearity.

Exercise 5.1

5.1 Modeling seasonality

We first illustrate some of the issues that play a role in modeling seasonality by means of the quarterly time series on UK household final consumption expenditures over the period 1960.1–2012.4, as shown in Figure 2.8. The time series contains a pronounced upward trend, and an immediate first question to address is whether this trend is deterministic or stochastic. We therefore apply the Augmented Dickey-Fuller test, as discussed in Section 4.2, including an intercept and a linear deterministic trend, and selecting $p = 8$ based on testing for the significance of the lagged first differences in the auxiliary regression (4.59). The resulting test statistic takes the value of -2.525, which exceeds the 10 percent critical value shown in Table 4.1. The exact finite sample p-value is 0.316. Hence, it is clear that the trend in this series is best represented

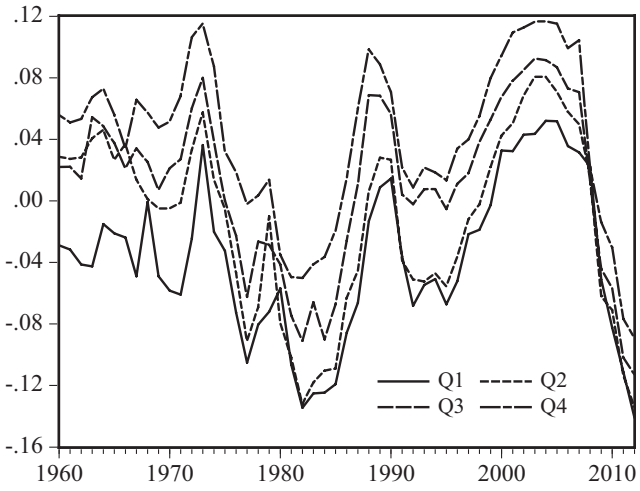


Figure 5.2: Vector-of-quarters plot of deviations from linear trend of UK household consumption.

as being stochastic. This can also be seen from Figure 5.1, which plots the residuals from regressing y_t on a constant and a linear trend, as in (2.1). The deviations from the deterministic trend take in general quite long to disappear. Only during 1960.1–1970.1 it remains unclear whether the deviations should be considered as permanent, or whether trend-reversion is a more appropriate characterization.

Another feature that is apparent in Figure 5.1, as well as in the first differences $\Delta_1 y_t$ in Figure 2.9, is that seasonal fluctuations are a dominant source of variation in the consumption series, irrespective of how the trend is modeled. In that respect, recall that the regression of the first differences on quarterly dummies, as in (2.3), has an R^2 of more than 90 percent. What might not be completely evident from these graphs is whether the seasonal pattern is stable over time or whether it changes. To shed more light on this issue, consider Figures 5.2 and 2.10, showing the vector-of-quarter plots, as discussed in Section 2.2, for the deviations from a linear trend and the first differences, respectively. From Figure 5.2, it appears that the deviation from the linear trend is always highest in the fourth quarter and lowest in the first, with the difference between the two being roughly constant over time. The deviation in the third quarter is just below that in the fourth, where again the difference appears quite stable. This is rather different for the second quarter though, for which the deviation from the trend sometimes is relatively high, while at other times it is almost as low as during the first quarter. A similar conclusion can be reached from Figure 2.10, which suggests that average consumption growth in the second quarter has become considerably lower after 1980, while average growth in the other quarters is relatively stable throughout

the entire sample period. As will be made clear in the following, the stability of the seasonal pattern in a time series, or the lack thereof, is the essential ingredient in the distinction between deterministic and stochastic seasonality. In addition, this is closely related to the representation of the trend as deterministic or stochastic.

Deterministic seasonality

A non-trending seasonal time series has different means in different seasons of the year. An obvious way to incorporate this property in an ARMA(p,q) model for such a time series is by including seasonal dummies, that is,

$$\phi_p(L)(y_t - \mu_1 D_{1,t} - \mu_2 D_{2,t} - \cdots - \mu_S D_{S,t}) = \theta_q(L)\varepsilon_t, \tag{5.1}$$

where $\phi_p(L)$ and $\theta_q(L)$ are the familiar lag polynomials. Assuming that all p roots of $\phi_p(L)$ are outside the unit circle, it follows that the unconditional mean of y_t in season s is equal to $\mathsf{E}[y_t|t = S(n-1)+s] = \mu_s$, for $s = 1, \ldots, S$. The overall unconditional mean is $\mathsf{E}[y_t] = \mu = 1/S \sum_{s=1}^{S} \mu_s$.

Exercise 5.2

When all the roots of the AR-polynomial $\phi_p(L)$ are real-valued, the seasonal pattern in y_t is only captured by means of the seasonal dummy variables $D_{s,t}$ or, equivalently, the season-specific means $\mu_s, s = 1, \ldots, S$. As discussed in Section 3.2, the characteristic equation $\phi_p(z) = 0$ may also have complex solutions of the form $z = a \pm bi$ when $p \geq 2$, where a and b are functions of the AR parameters. In that case, the time series y_t displays a cycle with length $c = 2\pi/(\tan^{-1}(b/a)]$, see (3.64). It may well be that c is such that the resulting cyclical pattern bears close resemblance to a seasonal cycle, for example when $c = 4$ for quarterly observed time series. In that case, we say that y_t has stationary stochastic seasonality.

For trending series such as UK consumption, (5.1) may be extended by including a linear deterministic trend term. Consider for example the AR(1) model,

$$y_t - \sum_{s=1}^{S} \mu_s D_{s,t} - \delta t = \phi_1(y_{t-1} - \sum_{s=1}^{S} \mu_s D_{s,t-1} - \delta(t-1)) + \varepsilon_t, \tag{5.2}$$

where we have restricted the trend growth rate δ to be the same across seasons. Allowing for different trend growth rates δ_s does not seem realistic for most economic time series, but we do return to this point below.

As discussed extensively in the previous chapter, a non-seasonal time series y_t that can be described by an AR(1) model as in (5.2) (but with $\mu_s = \mu$ for all $s = 1, \ldots, S$, see (4.4)) has a deterministic trend when $|\phi_1| < 1$, while the trend is stochastic when

$\phi_1 = 1$. This continues to hold when the seasonal intercepts μ_s are included. Defining $z_t = y_t - \sum_{s=1}^{S} \mu_s D_{s,t} - \delta t$, we can express z_t by recursive substitution for lagged z_t values in (5.2) as

$$z_t = (\phi_1)^t z_0 + \sum_{i=1}^{t} (\phi_1)^{t-i} \varepsilon_i, \tag{5.3}$$

or, in terms of the original time series y_t, as

$$y_t = (\phi_1)^t y_0 + \sum_{s=1}^{S} \mu_s (D_{s,t} - (\phi_1)^t D_{s,0}) + \delta t + \sum_{i=1}^{t} (\phi_1)^{t-i} \varepsilon_i. \tag{5.4}$$

When $|\phi_1| < 1$, the shocks ε_i have transitory effects as $(\phi_1)^{t-i}$ tends to zero when t increases. In addition, setting $y_0 = \sum_{s=1}^{S} \mu_s D_{s,0}$ for convenience, the unconditional mean of y_t is equal to $\mu_s + \delta t$ when $t = S(n-1) + s$, and the unconditional variance of y_t is equal to $\sigma^2/(1 - \phi_1^2)$. In sum, the series y_t is trend-stationary with a deterministic trend growth rate δ and seasonal pattern due to the intercepts μ_s, $s = 1, \ldots, S$. Note that the seasonal pattern is not only deterministic, but in addition it is stable over time.

Setting $\phi_1 = 1$ in (5.4) results in

$$y_t = y_0 + \sum_{s=1}^{S} \mu_s (D_{s,t} - D_{s,0}) + \delta t + \sum_{i=1}^{t} \varepsilon_i. \tag{5.5}$$

Comparing this expression with (4.12), the only difference appears to be the term $\sum_{s=1}^{S} \mu_s (D_{s,t} - D_{s,0})$. This is just a constant, the value of which depends on the season s that corresponds to observation t. In particular, assuming that $t = S(n-1) + s$, $\sum_{s=1}^{S} \mu_s (D_{s,t} - D_{s,0}) = \mu_s - \mu_1$. This leads to the following two conclusions. First, also in case the time series y_t is seasonal, it contains a deterministic trend δt and a stochastic trend $S_t = \sum_{i=1}^{t} \varepsilon_i$ when $\phi_1 = 1$ in (5.2). Second, also when the AR(1) model contains a unit root, the seasonality is captured in a deterministic way by means of the term $\sum_{s=1}^{S} \mu_s (D_{s,t} - D_{s,0})$ in (5.4) that varies across the seasons. Note that also in the presence of a stochastic trend as in (5.5), the seasonal pattern is stable over time.

Before proceeding, another crucial feature of time series with pronounced seasonality that should be pointed out is that observations in different years n and n^* but for the same season s appear much more closely related than observations within the same year n but for different seasons s and s^*. For example, Table 5.1, columns 2 and 4, show the EACFs for the first differences $\Delta_1 y_t$ and for the deviations from a linear deterministic trend (y_t^{DT}) of UK consumption. Clearly, the autocorrelations at the seasonal lags 4, 8, 12, and 16 are much larger (in absolute value) than even the first three autocorrelations. The strong dependence of observations in the same season across different years can to some extent be captured by means of the dummy variables $D_{s,t}$ in (5.2), but often not completely. For example, for the UK consumption series

Table 5.1: Empirical autocorrelation functions of UK consumption, 1960.1–2012.4

Lag	$\Delta_1 y_t$	$(\Delta_1 y_t)^{SD}$	y_t^{DT}	$\Delta_4 y_t$	$\Delta_1\Delta_4 y_t$
1	−0.344	−0.339	0.671	0.780	−0.209
2	−0.239	0.141	0.561	0.653	0.040
3	−0.333	−0.222	0.606	0.509	0.172
4	0.922	0.536	0.867	0.294	−0.473
5	−0.319	−0.246	0.535	0.280	0.143
6	−0.248	0.106	0.405	0.205	−0.050
7	−0.336	−0.268	0.437	0.151	−0.018
8	0.902	0.504	0.689	0.106	0.024
9	−0.310	−0.287	0.363	0.053	−0.106
10	−0.248	0.111	0.232	0.044	0.072
11	−0.335	−0.292	0.265	0.006	−0.082
12	0.877	0.436	0.513	−0.001	0.051
13	−0.285	−0.215	0.199	−0.028	0.023
14	−0.263	0.036	0.062	−0.069	−0.104
15	−0.327	−0.235	0.096	−0.072	0.045
16	0.846	0.327	0.345	−0.089	−0.117

Note: $(\Delta_1 y_t)^{SD}$ indicates the residuals from the regression of $\Delta_1 y_t$ on four quarterly dummies. y_t^{DT} indicates the residuals from the regression of y_t on a constant and a linear trend. The asymptotic standard error of the empirical autocorrelations is 0.069.

the residuals from a regression of the first differences on quarterly dummies still has the largest and significant autocorrelations at the seasonal lags 4, 8, 12, and 16, see Table 5.1, column 3. For this reason, we often consider autoregressive models of order (at least) S for seasonal time series to incorporate this correlation more directly.

Consider the special case where the autoregressive polynomial $\phi_S(L) = 1 - \phi_S L^S$, that is we only include the seasonal lag y_{t-S}. Also including seasonal dummies and a

linear trend term results in the model

$$y_t - \sum_{s=1}^{S} \mu_s D_{s,t} - \delta t = \phi_S(y_{t-S} - \sum_{s=1}^{S} \mu_s D_{s,t-S} - \delta(t-S)) + \varepsilon_t. \quad (5.6)$$

When the parameter $|\phi_S| < 1$ such that all S solutions of the characteristic equation $\phi_S(z) = 1 - \phi_S z^S = 0$ are outside the unit circle, the model in (5.6) is nothing else than a (restricted) stationary AR(S) model with a deterministic trend and deterministic seasonality. The situation that is of more interest is when $\phi_S = 1$, such that the AR-polynomial $\phi_S(L)$ reduces to $1 - L^S$. In that case, all S solutions of the corresponding characteristic equation are on the unit circle. This gives rise to the occurrence of (nonstationary) stochastic seasonality, which we discuss next.

Stochastic seasonality

Setting $\phi_S = 1$ in (5.6), the model reduces to

$$y_t = y_{t-S} + \delta S + \varepsilon_t. \quad (5.7)$$

From (5.7), it follows that the 'skip-sampled' series $\ldots, y_{t-2S}, y_{t-S}, y_t, y_{t+S}, y_{t+2S}, \ldots$, consisting of all observations related to the same season as y_t, is a random walk with drift equal to δS. When $t = S(n-1) + s$, by recursive substitution for y_{t-S} in (5.7), we can express y_t as

$$y_t = y_s + \delta S(n-1) + \sum_{i=1}^{n} \varepsilon_{t-S(i-1)}. \quad (5.8)$$

Similarly, for y_{t+1} we obtain

$$y_{t+1} = y_{s+1} + \delta S(n-1) + \sum_{i=1}^{n} \varepsilon_{t+1-S(i-1)}. \quad (5.9)$$

Comparing (5.8) and (5.9), the consecutive observations y_t and y_{t+1} are determined by different starting values y_s and y_{s+1} and by a different set of shocks. Hence, they have nothing in common and are uncorrelated. This in fact holds for any combination of y_t with y_{t+s} for $s = 1, \ldots, S-1$. In sum, the time series y_t, $t = 1, 2, \ldots$ considered as a whole is composed of S independent random walks, one for each season of the year. For this reason, y_t is usually referred to as a seasonal random walk (with drift in case $\delta \neq 0$).

Exercise 5.3–5.4

We can develop some more intuitive understanding of the properties of a seasonal random walk by examining the case of quarterly time series in more detail. Setting

$S = 4$ in (5.7) yields $y_t = y_{t-4} + \delta^* + \varepsilon_t$, where $\delta^* = 4\delta$, or

$$\Delta_4 y_t = (1 - L^4) y_t = \delta^* + \varepsilon_t. \tag{5.10}$$

The lag polynomial $1 - L^4$ can be decomposed as

$$\begin{aligned}(1 - L^4) &= (1 - L)(1 + L)(1 + L^2) \\ &= (1 - L)(1 + L)(1 - iL)(1 + iL),\end{aligned} \tag{5.11}$$

showing immediately that its roots equal $1, -1$, and $\pm i$. In the general case, the solutions to $(1 - z^S) = 0$ are given by $\cos(2\pi k/S) + i \sin(2\pi k/S)$ for $k = 0, 1, 2, \ldots, S - 1$, which all lie on the unit circle. The solution for $k = 0$ is equal to 1, and is called the nonseasonal unit root, while the $S - 1$ other solutions are called seasonal unit roots, see Hylleberg *et al.* (1990).

Exercise 5.5–5.6

Consider now the following three transformations of the time series y_t,

$$y_{1,t} = (1 + L + L^2 + L^3) y_t, \tag{5.12}$$

$$y_{2,t} = (1 - L + L^2 - L^3) y_t, \tag{5.13}$$

$$y_{3,t} = (1 - L^2) y_t. \tag{5.14}$$

Each of the filters leading to $y_{1,t}$, $y_{2,t}$ and $y_{3,t}$ impose all but one (or two in case of $(1 - L^2)$) of the unit roots, which follows from the fact that $(1 - L^4)$ can be decomposed in different ways, in particular as $(1 - L^4) = (1 + L + L^2 + L^3)(1 - L)$, or $(1 - L^4) = (1 - L + L^2 - L^3)(1 + L)$, or $(1 - L^4) = (1 - L^2)(1 + L^2)$. Hence, given that the quarterly time series y_t follows a seasonal random walk $\Delta_4 y_t = \delta^* + \varepsilon_t$, it follows that

$$(1 - L) y_{1,t} = \delta^* + \varepsilon_t, \tag{5.15}$$

$$(1 + L) y_{2,t} = \delta^* + \varepsilon_t, \tag{5.16}$$

$$(1 + L^2) y_{3,t} = \delta^* + \varepsilon_t. \tag{5.17}$$

The process $y_{1,t}$ has the non-seasonal unit root 1, and is in fact a conventional random walk with drift. Due to (5.16), we can write $y_{2,t}$ as

$$y_{2,t} = -y_{2,t-1} + \delta^* + \varepsilon_t, \tag{5.18}$$

which demonstrates that this process has a cycle with a length of two quarters. For that reason, the associated unit root -1 is called the semi-annual root. Finally, the roots $\pm i$ imply cycles of four periods, and hence are called the annual roots. This follows from observing that (5.17) gives

$$y_{3,t} = -y_{3,t-2} + \delta^* + \varepsilon_t. \tag{5.19}$$

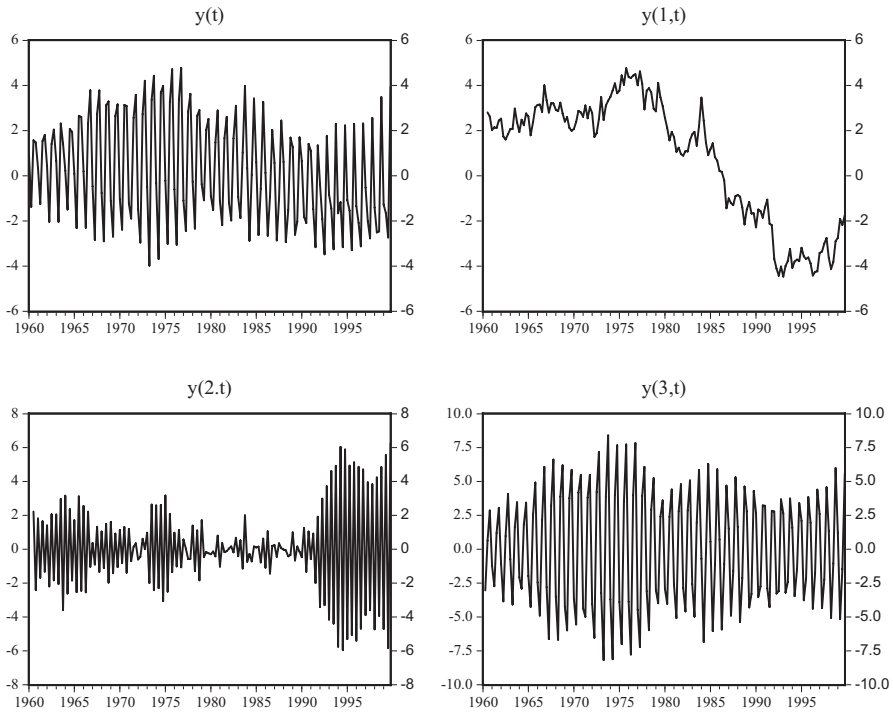


Figure 5.3: Simulated quarterly time series y_t from seasonal random walk and transformations $y_{1,t}$, $y_{2,t}$ and $y_{3,t}$ computed according to (5.12)–(5.14).

Exercise 5.7–5.8

Figure 5.3 displays a simulated example where a quarterly time series y_t is generated for the period 1960.1–1999.4 using the seasonal random walk model (5.10) with $\delta^* = 0$ and $\varepsilon_t \sim N(0, 0.25)$. The pre-sample starting values $y_{-3}, \ldots, y_0$ are set equal to 2, 0, -2 and 1, respectively. In addition to the series y_t in the upper-left panel, Figure 5.3 also shows the three transformed series obtained from (5.12)–(5.14). The $y_{1,t}$ series shows the typical behavior of a random walk without drift, with no tendency to exhibit mean-reversion to a constant level. The series $y_{2,t}$ contains a cycle of two quarters, but with constantly changing characteristics. Finally, $y_{3,t}$ shows an annual cycle, which at first sight appears to be quite stable, although its amplitude is somewhat time-varying. This first impression is misleading though. As the vector-of-quarters plot of $y_{3,t}$ in Figure 5.4 reveals, the cycle does change substantially after 1980. In particular, the average level of the observations in the first and third quarters changes such that these seasons appear to be "trading places".

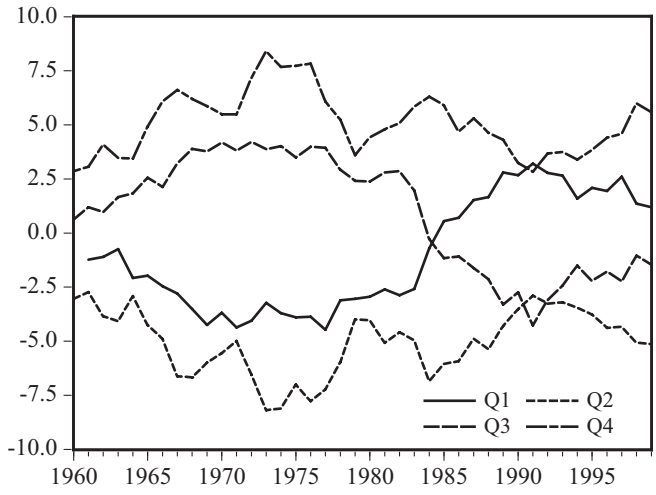


Figure 5.4: Vector-of-quarters plot of $y_{3,t} = (1 - L^2)y_t$ for simulated seasonal random walk.

The simulated example of a seasonal random walk suggests that such a series may display quite erratic seasonal behavior. The remaining question to be addressed therefore is what are exactly the implications of the presence of the S unit roots in the restricted AR(S) model (5.6) for the seasonality properties of y_t. Note that $\phi_S = 1$ implies that we need to apply the seasonal differencing filter $\Delta_S = (1 - L^S)$ to the time series y_t in order to make it stationary. This also can be seen from rewriting (5.7) as

$$y_t - y_{t-S} = \Delta_S y_t = \delta S + \varepsilon_t. \tag{5.20}$$

The seasonal dummies $D_{s,t}$ do not appear in (5.9) as these are annihilated by the seasonal differencing. Nevertheless, the seasonal random walk model in (5.20) is very well capable of describing seasonal patterns in the time series y_t. As noted above, the observations in the different seasons are driven by different sets of shocks that have permanent effects. Hence, we may observe that observations during a certain season are systematically above observations in another season, simply because at some point in the past a large shock occurred in that particular season. Another implication is that the seasonal pattern may experience substantial changes over time. A change in the level of y_t due to a large value of ε_t need not be accompanied by a similar change in y_{t+1}, but it will affect the level of y_{t+S}, y_{t+2S}, ... in the same way as it affects y_t. This sometimes is referred to by the phrase that "winter may become summer", see Figure 5.4. In fact, given that the effect of shocks does not disappear in (5.20), such changes in the seasonal pattern are permanent, until they are offset by other shocks. The time series y_t is therefore said to display nonstationary stochastic seasonality, as opposed

to stationary stochastic seasonality caused by complex roots in the AR-polynomial, as discussed before. In the remainder of this chapter, we are mainly concerned with the distinction between deterministic seasonality and nonstationary stochastic seasonality. For that reason, we refer to the latter simply as stochastic seasonality.

In the restricted AR(S) model with $\phi_S(L) = 1 - \phi_S L^S$ considered above, either all roots or none of the roots are on the unit circle. In the general case, where $\phi_S(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_S L^S$, we may have some of the roots on the unit circle and others outside. The same of course applies to higher order AR(p) models with $p \geq S$. We would then like to test for the presence of different seasonal unit roots, as a way of distinguishing between deterministic and stochastic seasonality. In addition, the unit roots that are present determine the appropriate differencing filter that is required to make the time series stationary. For quarterly time series, for example, when only the nonseasonal root 1 and the seasonal root -1 are present, the AR polynomial can be written as $\phi_p(L) = (1 - L)(1 + L)\phi_{p-2}(L)$, with $\phi_{p-2}(L)$ having all roots outside the unit circle. In that case, the filter $(1 - L)(1 + L) = (1 - L^2)$ is sufficient in the sense that $(1 - L^2)y_t = y_t - y_{t-2}$ is stationary. Applying the seasonal differencing filter $(1 - L^4)$ would imply overdifferencing.

Exercise 5.10

Several testing methods have been developed to test for the presence of seasonal unit roots and thereby to formally investigate the most appropriate differencing filter for y_t. These methods are based on extensions of the Dickey-Fuller method discussed in Chapter 4. The next section reviews the most popular of these methods.

Seasonal ARIMA models

The strong dependence between observations in the same season s in different years n and n^* was already noticed by Box and Jenkins (1970). This led them to consider the class of seasonal ARIMA [SARIMA] models, which makes the correlation of y_t with its seasonal lags explicit by augmenting an ARIMA model with the seasonal differencing filter and seasonal AR and MA components, as follows. The SARIMA(p,d,q)(P,D,Q) model is given by

$$\phi_p(L)\Phi_P(L)\Delta_1^d\Delta_S^D(y_t - \mu_t) = \theta_q(L)\Theta_Q(L)\varepsilon_t, \tag{5.21}$$

where $\phi_p(L)$ and $\theta_q(L)$ are the familiar AR and MA polynomials, while $\Phi_P(L)$ and $\Theta_Q(L)$ are seasonal AR and MA polynomials, defined for quarterly time series as

$$\Phi_P(L) = 1 - \Phi_1 L^4 - \Phi_2 L^8 - \cdots - \Phi_P L^{4P},$$
$$\Theta_Q(L) = 1 + \Theta_1 L^4 + \Theta_2 L^8 + \cdots + \Theta_Q L^{4Q}.$$

The term μ_t in (5.21) represents the unconditional mean of y_t at time t, which may contain deterministic components such as seasonal dummies and a linear deterministic trend, that is

$$\mu_t = \mu_1 D_{1,t} + \mu_2 D_{2,t} + \cdots + \mu_S D_{S,t} + \delta t.$$

As an example, for $p = P = 1$, $d = D = 0$ and $q = Q = 1$ and setting $\mu_t = 0$, the resulting model is

$$(1 - \phi_1 L)(1 - \Phi_1 L^4) y_t = (1 + \theta_1 L)(1 + \Theta_1 L^4)\varepsilon_t,$$

which amounts to the restricted ARMA(5,5) model

$$y_t = \phi_1 y_{t-1} + \Phi_1 y_{t-4} + \phi_1 \Phi_1 y_{t-5} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \Theta_1 \varepsilon_{t-4} + \theta_1 \Theta_1 \varepsilon_{t-5}. \tag{5.22}$$

The Box and Jenkins (1970) approach to selecting the orders (p,d,q) and (P,D,Q) and the contents of μ_t in the SARIMA model (5.21) amounts to applying various transformations to y_t and investigating the EACFs (and EPACFs) of the transformed series, see, for example, Abraham and Ledolter (1983) and Granger and Newbold (1986). Typically, we consider the EACFs of y_t, $\Delta_1 y_t$, $(\Delta_1 y_t)^{SD}$, y_t^{DT}, $\Delta_S y_t$, and $\Delta_1 \Delta_S y_t$. Note that the double differencing filter $\Delta_1 \Delta_S$ amounts to the transformation $y_t - y_{t-1} - y_{t-S} + y_{t-S-1}$. The purpose is to find a suitable transformation of the time series such that the EACF and EPACF show easily interpretable patterns, which can be used to identify an appropriate SARIMA model. For many seasonal time series it appears that the EACF and EPACF of the doubly transformed series $\Delta_1 \Delta_S y_t$ suggest parsimonious model structures. That is, only a few (partial) autocorrelations are significantly different from zero, indicating that the orders p, P, q and Q in (5.21) can be set to low values when $d = D = 1$, such that the resulting model contains only a small number of parameters to estimate. In fact, in practice for many series it appears that $p = P = 0$, $q = Q = 1$ are suitable values to capture the remaining autocorrelation in the doubly-differenced series. This leads to the so-called "airline model"

$$\Delta_1 \Delta_S y_t = (1 + \theta_1 L)(1 + \Theta_1 L^S)\varepsilon_t, \tag{5.23}$$

which derives its name from its successful first application to a monthly time series of airline passengers reported in Box and Jenkins (1970). Notice that the MA part of this model contains only two unknown parameters θ_1 and Θ_1.

For the quarterly UK consumption series, the largest empirical autocorrelations for $\Delta_1 \Delta_4 y_t$ occur at lags 1 and 4, with smaller but still significant (and positive) values at lags 3 and 5, see the last column of Table 5.1. This matches perfectly with the theoretical ACF of the doubly-differenced series implied by the airline model, which

is equal to

$$\rho_1 = \frac{\theta_1 + \theta_1\Theta_1^2}{1 + \theta_1^2 + \Theta_1^2 + \theta_1^2\Theta_1^2} = \frac{\theta_1}{1 + \theta_1^2}, \tag{5.24}$$

$$\rho_S = \frac{\Theta_1}{1 + \Theta_1^2}, \tag{5.25}$$

$$\rho_{S-1} = \rho_{S+1} = \frac{\theta_1\Theta_1}{(1 + \theta_1^2)(1 + \Theta_1^2)}, \tag{5.26}$$

$$\rho_k = 0 \qquad \text{for all nonnegative } k \neq 0,\ 1,\ S-1,\ S,\ S+1. \tag{5.27}$$

We estimate an 'unrestricted' airline model, with MA terms at lags 1, 4 and 5, for this series, which yields the results

$$\Delta_1\Delta_4 y_t = \underset{(0.00020)}{-0.00013} + \hat{\varepsilon}_t - \underset{(0.049)}{0.134}\hat{\varepsilon}_{t-1} - \underset{(0.049)}{0.672}\hat{\varepsilon}_{t-4}, \tag{5.28}$$

with standard errors in parentheses, and where we have deleted the insignificant MA(5) component. This specification seems reasonably adequate, as the residuals $\hat{\varepsilon}_t$ do not have significant autocorrelations, although the Jarque-Bera statistic suggests that normality of the residuals should be rejected. The dynamic pattern of y_t seems to be characterized by only two MA parameters and a double differencing filter.

The $\Delta_1\Delta_4$ differencing filter that is used in the airline model for quarterly time series assumes the presence of two nonseasonal unit roots 1 and three seasonal unit roots -1 and $\pm i$, see also (5.11). The double nonseasonal unit root implies that the original time series y_t is I(2), which may not be realistic for many economic time series. In addition, the seasonal unit roots are imposed without any prior testing of their relevance. In fact, in many empirical applications the $\Delta_1\Delta_S y_t$ series appears to be overdifferenced. Indications for overdifferencing can be obtained in several different ways, including the EACF of $\Delta_1\Delta_S y_t$ and the roots of the MA-polynomial in an estimated airline model.

Exercise 5.11–5.12

A typical feature of many doubly differenced seasonal time series is that the EACF at lags 1 and S can take values which are close to -0.5. For example, $\hat{\rho}_4$ in the last column of Table 5.1 equals -0.488. From the expressions of the theoretical autocorrelations of the airline model, it is seen that the seasonal MA-parameter Θ_1 should be close to -1 to achieve such large negative values of ρ_S, see (5.25). This may then suggest that the $\Delta_1\Delta_S y_t$ series is in fact overdifferenced, as the $(1 - L^S)$ factor will appear on the right hand side of the airline model (5.23), making it redundant in the $\Delta_1\Delta_S$ filter. Similarly, when $(1 - L)$ is redundant, the theoretical value of ρ_1 would equal -0.5.

Overdifferencing leads to unit roots in the MA polynomial and, hence, non-invertibility of the model, as discussed in Chapter 3. Additional information on the possibility that the double differencing filter may not be needed can therefore be obtained from solving the characteristic equation $\theta_q(z)\Theta_q(z) = 0$ using the estimated MA parameters. To illustrate this, consider the characteristic equation corresponding to the MA polynomial in (5.28), that is

$$(1 - 0.134z - 0.672z^4) = 0.$$

The solutions to this polynomial are 0.94, $0.03 \pm 0.90i$ and -0.87, where $i^2 = -1$. The first root obviously is close to 1, suggesting that two nonseasonal unit roots is one too many for this series. Hence, the $\Delta_1\Delta_4$ filter may be reduced to simple seasonal differencing Δ_4. In addition, the pair of complex roots appears to be quite close to the unit circle as well. It is however difficult to tell from just their values if they are not significantly different from the annual unit roots $\pm i$, which would simplify the required differencing filter for y_t even further. This requires proper statistical testing procedures, which are discussed in the next section.

5.2 Seasonal unit root tests

If the seasonal differencing filter Δ_S is required to transform y_t to stationarity, a time series is said to be seasonally integrated. As discussed in the previous section, the use of a certain differencing filter amounts to an assumption on the number of seasonal and nonseasonal unit roots in a time series. Any such unit roots need to be removed prior to further analysis by differencing the time series to make it stationary. Hence, we do not want to difference the series "too little", for example, by only applying the $(1 - L)$ filter when unit roots 1 and -1 are present. At the same time we want to avoid overdifferencing, and it therefore does not seem advisable to routinely apply the Δ_S filter to any given seasonal series y_t. Instead, we may want to test whether the Δ_S filter indeed amounts to an adequate data transformation or whether a nested filter such as Δ_1 is sufficient, thereby selecting between the deterministic and stochastic seasonality models. Hylleberg *et al.* (1990) [HEGY hereafter] develop a testing procedure for seasonal and non-seasonal unit roots in a quarterly time series, and hence to test the adequacy of the $(1 - L^4)$ filter versus its nested variants like $(1 - L)$ or $(1 + L)$.

Recall the decomposition of the seasonal differencing filter $(1 - L^4)$ discussed in the previous section

$$(1 - L^4) = (1 - L)(1 + L)(1 + L^2), \tag{5.29}$$

and the transformations $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$ as defined in (5.12)–(5.14). As the filters leading to $y_{1,t}$ and $y_{2,t}$ each contain three of the four possible unit roots, an obvious

idea would be to test for the presence of the remaining one. Similarly, the filter leading to $y_{3,t}$ contains the roots 1 and -1, and we may test for the presence of the complex pair $\pm i$. As the roots -1 and $\pm i$ are imposed in the transformation that renders $y_{1,t}$, we may test for the non-seasonal unit root 1 in this series by testing the null hypothesis $\phi_1 = 1$ in

$$y_{1,t} = \phi_1 y_{1,t-1} + \varepsilon_t, \tag{5.30}$$

against the alternative $\phi_1 < 1$, where for the moment we leave deterministic terms out of consideration. This might be done using the conventional Dickey-Fuller test discussed in Section 4.2. Note that rewriting (5.30) in terms of a regression of the first difference of $y_{1,t}$ on its lagged level gives

$$(1 - L)y_{1,t} = \Delta_4 y_t = \pi_1 y_{1,t-1} + \varepsilon_t, \tag{5.31}$$

where $\pi_1 = \phi_1 - 1$.

Along the same lines, we may test for the presence of a unit root equal to -1 in $y_{2,t}$, which already assumes the roots 1 and $\pm i$, by testing $\phi_2 = 1$ against $\phi_2 < 1$ in

$$y_{2,t} = -\phi_2 y_{2,t-1} + \varepsilon_t. \tag{5.32}$$

It is convenient to rewrite (5.32) as

$$(1 + L)y_{2,t} = \Delta_4 y_t = -\pi_2 y_{2,t-1} + \varepsilon_t, \tag{5.33}$$

where $\pi_2 = \phi_2 - 1$. We can then implement the unit root test by regressing $\Delta_4 y_t$ on $-y_{2,t-1}$ and testing whether the coefficient π_2 is zero against the alternative that it is negative based on its t-statistic.

Finally, the real roots 1 and -1 are imposed to obtain $y_{3,t}$, such that we should test for the presence of the pair of complex roots $\pm i$. This is the case if $\phi_3 = 1$ and $\phi_4 = 0$ in the AR(2) representation

$$y_{3,t} = -\phi_4 y_{3,t-1} - \phi_3 y_{3,t-2} + \varepsilon_t. \tag{5.34}$$

Defining $\pi_3 = \phi_3 - 1$ and $\pi_4 = \phi_4$, (5.32) can be written as

$$(1 + L^2)y_{3,t} = \Delta_4 y_t = -\pi_3 y_{3,t-2} - \pi_4 y_{3,t-1} + \varepsilon_t, \tag{5.35}$$

which allows testing the relevant null hypothesis $\pi_3 = \pi_4 = 0$ straightforwardly by means of an F-test.

As shown in HEGY, the transformations $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$ as given in (5.12)–(5.14) are asymptotically uncorrelated. Hence, all three unit root tests discussed above can in fact be conducted using the joint auxiliary regression

$$\Delta_4 y_t = \pi_1 y_{1,t-1} - \pi_2 y_{2,t-1} - \pi_3 y_{3,t-2} - \pi_4 y_{3,t-1} + \varepsilon_t. \tag{5.36}$$

The t-tests for $\pi_1 = 0$ and $\pi_2 = 0$ are denoted as $t(\pi_1)$ and $t(\pi_2)$. As the relevant alternative hypotheses for the unit roots 1 and -1 are that the roots are outside the unit circle, the t-tests are one-sided tests, similar to the standard Dickey-Fuller tests discussed in the previous chapter. The significance of π_3 and π_4 is evaluated through the joint F-test, denoted as $F(\pi_3, \pi_4)$. Additionally, we may consider F-tests for jointly testing the restrictions on π_2, π_3 and π_4, or on all 4 π_i parameters, see Ghysels *et al.* (1994).

Exercise 5.13

The HEGY tests for seasonal unit roots essentially were developed starting from the restricted AR(4) model $(1 - \phi_4 L)y_t = \varepsilon_t$. In practice we may consider more general AR(p) specifications $\phi_p(L)y_t = \varepsilon_t$, where in case y_t displays seasonal fluctuations the autoregressive order p is likely to exceed the number of seasons S. HEGY demonstrate that their approach to testing for unit roots remains valid in that case, based on the expansion

$$\phi_p(L) = -\pi_1 L\phi_1(L) + \pi_2 L\phi_2(L) + (\pi_3 L + \pi_4)L\phi_3(L) + \phi^*_{p-4}(L)(1 - L^4), \tag{5.37}$$

where the polynomials $\phi_i(L)$, $i = 1, 2, 3$, are defined by

$$\begin{aligned}
\phi_1(L) &= (1 + L + L^2 + L^3),\\
\phi_2(L) &= (1 - L)(1 + L^2) = (1 - L + L^2 - L^3),\\
\phi_3(L) &= (1 - L^2),
\end{aligned}$$

that is, these are the filters leading to the transformed series $y_{1,t}$, $y_{2,t}$ and $y_{3,t}$, as defined in (5.12)–(5.14). The decomposition in (5.37) leads to the following 'augmented' auxiliary regression

$$\Delta_4 y_t = \pi_1 y_{1,t-1} - \pi_2 y_{2,t-1} - \pi_3 y_{3,t-2} - \pi_4 y_{3,t-1} + \sum_{i=1}^{p-4} \phi^*_i \Delta_4 y_{t-i} + \varepsilon_t, \tag{5.38}$$

which can be used for testing for seasonal unit roots in AR(p) models for a quarterly time series y_t for any value of $p \geq 4$.

Exercise 5.14

The joint null hypothesis in the HEGY test procedure for quarterly data is that the $(1 - L^4)$ filter is the appropriate filter to remove unit roots. Hence, the series y_t is nonstationary under the null, implying that the asymptotics for the various t- and F-tests are nonstandard. Discussions of the relevant asymptotic distributions are given in Engle, Granger, Hylleberg and Lee (1993) and HEGY. Similar to the standard Dickey-Fuller tests discussed in Chapter 4, these asymptotic distributions do not depend on the AR-order p, but do depend on deterministic terms that may be added to the auxiliary regression (5.38), which we therefore discuss next.

Deterministic components

In the previous chapter it was noted that, when testing for a unit root, it is important that the models under the null and the alternative hypotheses are 'competitive', in the sense that both should be reasonable descriptions of the trend properties of the time series under investigation. In the present case of testing for seasonal unit roots using the auxiliary regression (5.38), the series y_t displays (nonstationary) stochastic seasonality under the joint null hypothesis $\pi_1 = \cdots = \pi_4 = 0$. Under the alternative that all π_i, $i = 1, \ldots, 4$, coefficients are nonzero, y_t can be described by a stationary AR(4) model. Although this may contain stationary stochastic seasonality, this is of a rather different character than the patterns implied by the seasonal unit roots under the null. Hence, it is important to augment (5.38) with deterministic terms that make the seasonality in the model under the alternative more competitive to that of the null model. In addition, we may want to include a linear deterministic trend, in case this seems to be present in the time series y_t.

In general, we may extend (5.38) to

$$\Delta_4 y_t = \mu_t + \pi_1 y_{1,t-1} - \pi_2 y_{2,t-1} - \pi_3 y_{3,t-2} - \pi_4 y_{3,t-1} + \sum_{i=1}^{p-4} \phi_i^* \Delta_4 y_{t-i} + \varepsilon_t, \tag{5.39}$$

where μ_t contains the deterministic components. HEGY consider five different specifications for μ_t, which can all be nested in

$$\mu_t = \mu_1 D_{1,t} + \mu_2 D_{2,t} + \mu_3 D_{3,t} + \mu_4 D_{4,t} + \delta t. \tag{5.40}$$

The five specific cases are (i) no constant, no dummies, no trend: $\mu_1 = \cdots = \mu_4 = \delta = 0$; (ii) constant, no dummies, no trend: $\mu_1 = \cdots = \mu_4 = \mu$ for some $\mu \neq 0$ and $\delta = 0$; (iii) constant, no dummies, trend: $\mu_1 = \cdots = \mu_4 = \mu$ for some $\mu \neq 0$; (iv) constant,

dummies, no trend: $\delta = 0$; and (v) constant, dummies, and trend (no restrictions on (5.40)). For practical purposes, the latter two choices for μ_t are the most relevant. Smith and Taylor (1998) propose a more general specification for μ_t including seasonal linear trends, replacing the linear trend term δt in (5.40) with $\sum_{s=1}^{4} \delta_s D_{s,t} t$. Although the asymptotic distributions of the HEGY test statistics typically depend on the specification chosen for the deterministic component, they are sometimes invariant to the choice of μ_t. In particular, for the two specifications of μ_t considered here, the asymptotic distributions of the $t(\pi_2)$ and $F(\pi_3, \pi_4)$ statistics are each the same.

Tables with critical values for the $t(\pi_1)$, $t(\pi_2)$ and $F(\pi_3, \pi_4)$ statistics are given in HEGY and Franses and Hobijn (1997) for several finite sample sizes. Harvey and van Dijk (2006) conduct a so-called response surface analysis, which enables straightforward computation of critical values for any number of observations. Critical values for the $t(\pi_1)$, $t(\pi_2)$, $F(\pi_3, \pi_4)$ test statistics are displayed in Table 5.2 for samples of 10, 20, 30, 40, and 50 years of quarterly observations.

In HEGY it is shown that the asymptotic distribution of the $t(\pi_1)$ test is the same as that of the standard Dickey-Fuller test for a non-seasonal unit root in non-seasonal time series. Comparing the critical values in Table 5.2 with those in Table 4.1 seems to substantiate this asymptotic result. In HEGY it is also shown that the $t(\pi_2)$ test has the same asymptotic distribution as the $t(\pi_1)$ test. As mentioned above though, the distribution of $t(\pi_2)$ is the same, irrespective of whether a deterministic trend is included in (5.40) or not, while the distribution of $t(\pi_1)$ is different in those cases.

A final remark about the distributions of the HEGY test statistics concerns the lag order p in (5.38). It was mentioned above that the asymptotic distributions of the $t(\pi_1)$, $t(\pi_2)$, and $F(\pi_3, \pi_4)$ statistics do not depend on the value of p. This continues to hold in case the value of p is unknown and is determined using one of the procedures discussed in Section 4.2 in the context of the Dickey-Fuller test for a nonseasonal unit root. In finite samples, however, the distributions are sensitive to the choice of p, and to the way this value is selected. Harvey and van Dijk (2006) provide response surfaces that enable computation of accurate finite sample critical values of the HEGY tests, allowing for the lag order to be determined endogenously, using commonly applied selection methods. Burridge and Taylor (2004) suggest an alternative approach by employing bootstrap techniques to determine appropriate critical values for the particular sample size and lag order at hand.

The HEGY test approach has been evaluated in various simulation exercises, see Hylleberg (1995) and Ghysels *et al.* (1994), among others. The results in the latter study indicate that the size of the tests deteriorates when the DGP is a seasonal MA series with a parameter close to the unit circle. For practical applications, it is therefore again necessary to thoroughly check the EACF of the error series of the auxiliary test regression. If this EACF does not die out at seasonal lags, we should be cautious with the interpretation of HEGY test outcomes. Furthermore, and similar to the standard

Table 5.2: Critical values for HEGY one-sided $t(\pi_1)$ and $t(\pi_2)$-tests and for the $F(\pi_3, \pi_4)$-test in quarterly time series

		No trend			Trend		
Test	**Years**	**10%**	**5%**	**1%**	**10%**	**5%**	**1%**
$t(\pi_1)$	10	−2.45	−2.76	−3.41	−3.01	−3.33	−4.00
	20	−2.51	−2.81	−3.40	−3.06	−3.36	−3.95
	30	−2.53	−2.82	−3.41	−3.09	−3.38	−3.95
	40	−2.54	−2.83	−3.41	−3.10	−3.38	−3.95
	50	−2.54	−2.84	−3.42	−3.10	−3.39	−3.95
$t(\pi_2)$	10	−2.45	−2.76	−3.41	−2.43	−2.75	−3.40
	20	−2.51	−2.81	−3.40	−2.50	−2.80	−3.39
	30	−2.53	−2.82	−3.41	−2.52	−2.82	−3.40
	40	−2.54	−2.83	−3.41	−2.53	−2.83	−3.41
	50	−2.54	−2.84	−3.42	−2.54	−2.84	−3.41
$F(\pi_3, \pi_4)$	10	5.43	6.61	9.36	5.37	6.54	9.28
	20	5.52	6.59	8.98	5.49	6.56	8.95
	30	5.55	6.60	8.90	5.53	6.58	8.89
	40	5.57	6.61	8.87	5.56	6.60	8.86
	50	5.58	6.62	8.86	5.57	6.61	8.85

Note: The auxiliary test regression contains a constant, seasonal dummies and possibly a trend.
Source: Harvey and van Dijk (2006).

Table 5.3: Testing for seasonal unit roots: some empirical examples

Variable	T	$p_{\max}$	p	$t(\pi_1)$	$t(\pi_2)$	$F(\pi_3, \pi_4)$
US industrial production	120	12	8	−3.228*	−1.519	18.419***
US unemployment rate	120	12	7	−2.113	−2.623*	15.751***
UK consumption	120	12	12	−2.030	−1.513	3.890

Notes: ***, ** and * indicate significance at the 1%, 5%, and 10% levels, respectively. T denotes the effective number of observations, p is the lag order of the AR(p) model where the possible $(1 - L^4)$ component is still included. The value of p is determined by a 'general-to-specific' procedure starting with lag order $p_{\max}$ in (5.39) and sequentially lowering it by eliminating the highest-order lagged first difference until its coefficient is significant at the 10% significance level.

Dickey-Fuller tests, the power of the tests for seasonal unit roots is not high when the DGP is close to the null hypothesis.

The HEGY tests are applied to the quarterly US industrial production, US unemployment rate, and UK consumption series, for the sample period 1963.1 to 1992.4 (where the first three years of observations are used as pre-sample starting values and the remainder of the time series is saved for forecast evaluation later on). As both the industrial production and consumption series display a clear upward trend and pronounced seasonality, we include a constant, seasonal dummies and a linear trend in the auxiliary regression (5.39). A trend is not included in the test regression for the US unemployment rate. The lag order is determined using sequential testing for significance of the coefficient of the highest-order lagged seasonal difference at the 10% level. Starting with $p_{\max} = 12$, this results in AR orders equal to 8, 7, and 12 for the US industrial production, US unemployment rate, and the UK consumption series, respectively. The HEGY test results are shown in Table 5.3. For all three series, we cannot reject the hypotheses that $\pi_1 = 0$ and $\pi_2 = 0$ at the 5% significance level. In addition, we note that for US industrial production and US unemployment rate the $t(\pi_1)$ and $t(\pi_2)$ statistic respectively does exceed the 10% critical value, suggesting that the nature of the trend in this series is somewhat uncertain. For UK consumption we can neither reject that $\pi_3 = \pi_4 = 0$. From these results we therefore conclude, at least for now, that the Δ_4 filter is most appropriate for this series. For US industrial production and US unemployment we convincingly reject $\pi_3 = \pi_4 = 0$ at conventional significance levels, with the $F(\pi_3, \pi_4)$ statistic taking the values 18.42 and 15.75. Hence, these series do not appear to contain the pair of complex unit roots $\pm i$, suggesting that the $(1 - L^2)$ filter is appropriate for both industrial production and unemployment.

5.3 Forecasting

The various models discussed in this chapter contain rather different descriptions of the seasonal patterns of the time series y_t, with different properties. For example, while models with deterministic seasonality imply a seasonal pattern that is stable over time, models with seasonal unit roots suggest the presence of changing, and in fact nonstationary, seasonal patterns. Intuitively it seems clear that selecting a model with an appropriate characterization of the seasonal properties of a given time series is important for the accuracy of out-of-sample forecasts, especially for short horizons up to a year, say.

Constructing point forecasts from the models discussed in this chapter is relatively straightforward, given that all these models are linear. A few remarks concerning the properties of the point forecasts and their uncertainty or the properties of the associate forecast errors are useful here. First, consider the seasonal random walk model

$$y_t = y_{t-S} + \delta + \varepsilon_t.$$

For horizons h smaller than the number of seasons S, the h-step ahead point forecast at time T is given by

$$\hat{y}_{T+h|T} = \mathsf{E}[y_{T+h}|\mathcal{Y}_T] = y_{T+h-S} + \delta, \tag{5.41}$$

while for horizons $h > S$, we have the relationship

$$\hat{y}_{T+h|T} = \hat{y}_{T+h-S|T} + \delta, \tag{5.42}$$

such that forecasts at horizons beyond a year can be obtained recursively. Notice that, although the seasonal random walk model allows for changes in the seasonal pattern of y_t, the out-of-sample point forecasts essentially are 'no-change' forecasts and repeat the pattern observed during the last year before the forecast origin.

For all horizons h up to and including a year, (5.41) implies that the associated forecast error is equal to

$$e_{T+h|T} = y_{T+h} - \hat{y}_{T+h|T} = \varepsilon_{T+h},$$

with variance $\mathsf{V}[e_{T+h|T}] = \sigma^2$. Similarly, for all forecast horizons between one and two years, we find that $e_{T+h|T} = \varepsilon_{T+h} + \varepsilon_{T+h-S}$ with variance $\mathsf{V}[e_{T+h|T}] = 2\sigma^2$. In general, when forecasting for a horizon h in the k-th year beyond the current time T, the forecast error variance is equal to $k\sigma^2$. Consequently, the width of a 95% interval forecast for y_{T+h}, which may be computed as

$$(\hat{y}_{T+h|T} - 1.96\sqrt{\mathsf{V}[e_{T+h|T}]}, \hat{y}_{T+h|T} + 1.96\sqrt{\mathsf{V}[e_{T+h|T}]}),$$

resembles a step-function, which is constant within a year and increases with $2 \times 1.96\sigma$ when going to the next year.

Second, consider the airline model for a quarterly time series

$$\Delta_1 \Delta_4 y_t = (1 + \theta_1 L)(1 + \Theta_1 L^4)\varepsilon_t, \tag{5.43}$$

or

$$y_t = y_{t-1} + y_{t-4} - y_{t-5} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \Theta_1 \varepsilon_{t-4} + \theta_1 \Theta_1 \varepsilon_{t-5}. \tag{5.44}$$

Assuming the parameters θ_1 and Θ_1 are known, the out-of-sample forecasts from this model for horizons $h = 1, \ldots, 5$ are given by

$$\begin{aligned}
\hat{y}_{T+1|T} &= y_T + y_{T-3} - y_{T-4} + \theta_1 \varepsilon_T + \Theta_1 \varepsilon_{T-3} + \theta_1 \Theta_1 \varepsilon_{T-4}, \\
\hat{y}_{T+2|T} &= \hat{y}_{T+1|T} + y_{T-2} - y_{T-3} + \Theta_1 \varepsilon_{T-2} + \theta_1 \Theta_1 \varepsilon_{T-3}, \\
\hat{y}_{T+3|T} &= \hat{y}_{T+2|T} + y_{T-1} - y_{T-2} + \Theta_1 \varepsilon_{T-1} + \theta_1 \Theta_1 \varepsilon_{T-2}, \\
\hat{y}_{T+4|T} &= \hat{y}_{T+3|T} + y_T - y_{T-1} + \Theta_1 \varepsilon_T + \theta_1 \Theta_1 \varepsilon_{T-1}, \\
\hat{y}_{T+5|T} &= \hat{y}_{T+4|T} + \hat{y}_{T+1|T} - y_T + \theta_1 \Theta_1 \varepsilon_T,
\end{aligned}$$

while for longer forecast horizons, we have

$$\hat{y}_{T+h|T} = \hat{y}_{T+h-1|T} + \hat{y}_{T+h-4|T} - \hat{y}_{T+h-5|T}, \qquad \text{for } h = 6, 7, 8, \ldots \tag{5.45}$$

Note that (5.45) indicates that the forecasts are a deterministic function after $h = 5$. In fact, we can write (5.45) as $\Delta_1 \hat{y}_{T+h|T} = \Delta_1 \hat{y}_{T+h-4|T}$ and as $\Delta_4 \hat{y}_{T+h|T} = \Delta_4 \hat{y}_{T+h-1|T}$. In words, the quarterly change in the forecasts for y is the same as in the corresponding quarter of the previous year, while the annual change in the forecasts is the same for all quarters. The airline model therefore is sometimes said to deliver 'same change' forecasts.

There are relatively few studies that have considered the relative forecasting performance of different models for seasonality. Osborn *et al.* (1999) find that seasonal unit root models produce more accurate point forecasts than models with deterministic seasonality. Paap *et al.* (1997), on the other hand, demonstrate that this may be reversed when structural breaks in deterministic seasonality are allowed for, see Chapter 6 for details. Finally, Franses and van Dijk (2005) examine the forecasting performance of various models for seasonality and nonlinearity for quarterly industrial production series of 18 OECD countries. In general, linear models with fairly simple descriptions of seasonality are found to outperform at short forecast horizons, whereas nonlinear models with more elaborate seasonal components dominate at longer horizons.

The fact that the results from these empirical studies do not suggest one particular representation of seasonality that gives the most accurate forecasts in general should

not be interpreted as saying that seasonality and the way it is incorporated in a time series model is not all that important for forecasting. Rather, it suggests that the type of seasonality varies across time series. The relevance of modelling seasonality in an appropriate way for forecasting can be understood by considering the consequences of forecasting with misspecified models. This is done in Ghysels *et al.* (2006) in detail, here we only review the most important results.

First, we examine the consequences of using a seasonal random walk or a SARIMA model for forecasting, when the seasonality in a time series y_t is in fact deterministic. To focus on the effects of wrongly imposing seasonal unit roots, we assume that the series does contain the nonseasonal root 1. In particular, consider the AR(1) model with deterministic seasonality as given in (5.2). Imposing the nonseasonal unit root by setting $\phi_1 = 1$ and setting $\delta = 0$ for convenience, we can write this model as

$$y_t = y_{t-1} + \sum_{s=1}^{S} \mu_s^* D_{s,t} + \varepsilon_t, \tag{5.46}$$

where $\mu_s^* = \mu_s - \mu_{s-1}$, with $\mu_0 = \mu_S$. Hence, the optimal one-step ahead forecast at time T is given by

$$\hat{y}_{T+1|T} = y_T + \sum_{s=1}^{S} \mu_s^* D_{s,T+1},$$

for which the associated forecast error is equal to ε_{T+1} with variance σ^2. When we use a seasonal random walk model $\Delta_S y_t = v_t$ for forecasting instead, the one-step ahead forecast at time T is given by $\hat{y}^{SRW}_{T+1|T} = y_{T-S+1}$, with the forecast error being equal to the annual difference $e^{SRW}_{T+1|T} = y_{T+1} - y_{T-S+1}$. We can express this in terms of the shocks ε_t by recursively substituting for lagged y_t in (5.46), which allows this to be written as

$$\begin{aligned}
y_t &= y_{t-2} + \sum_{s=1}^{S} \mu_s^* (D_{s,t} + D_{s,t-1}) + \varepsilon_t + \varepsilon_{t-1} \\
&= y_{t-3} + \sum_{s=1}^{S} \mu_s^* (D_{s,t} + D_{s,t-1} + D_{s,t-2}) + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} \\
&\;\;\vdots \\
&= y_{t-S} + \sum_{s=1}^{S} \mu_s^* (D_{s,t} + D_{s,t-1} + D_{s,t-2} + \cdots + D_{s,t-S+1}) \qquad (5.47) \\
&\qquad\qquad + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \cdots + \varepsilon_{t-S+1} \\
&= y_{t-S} + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \cdots + \varepsilon_{t-S+1}, \qquad (5.48)
\end{aligned}$$

where the last equality follows from the fact that the sum of the dummies $D_{s,t}, \ldots, D_{s,t-S+1}$ is equal to a constant, while $\sum_{s=1}^{S} \mu_s^* = 0$. Hence, it follows that $e_{T+1|T}^{SRW}$ is equal to the sum $\varepsilon_{T+1} + \varepsilon_T + \cdots + \varepsilon_{T-S+2}$, with variance equal to $\mathsf{V}[e_{T+1|T}^{SRW}] = S\sigma^2$. Concluding, we find that wrongly imposing seasonal unit roots leads to an increase of the MSPE proportional to the periodicity of the data.

We may expect that things would be even worse when a SARIMA-type model (5.21) with $d = D = 1$ were used, but this turns out not to be the case. Assume that we use the model $\Delta_1 \Delta_S y_t = u_t$ for forecasting y_t, such that the one-step ahead forecast at time T is given by $\hat{y}_{T+1|T}^{SARIMA} = y_{T-S+1} - y_T + y_{T-S} = y_{T-S+1} - \Delta_S y_T$. The resulting forecast error is equal to $y_{T+1} - y_{T-S+1} + \Delta_S y_T = \Delta_S y_{T+1} + \Delta_S y_T$. From (5.48) it follows that this is equal to $\varepsilon_{T+1} - \varepsilon_{T-S+1}$, which has variance $2\sigma^2$, independent of the value of S.

As an empirical example, consider the quarterly US unemployment rate. The analysis of this time series so far suggests that a nonseasonal unit root may be present in the data, while seasonality may best be characterized as deterministic. Using the sample period 1963.1–1992.4 for specification and estimation, it appears that an AR(6) model with quarterly dummies yields an adequate description of the $\Delta_1 y_t$ series. Using this model to compute one-step ahead forecasts for the level of the unemployment rate y_t for the period 1993.1–2012.4 gives a root MSPE equal to 0.264. When the seasonal unit roots -1 and $\pm i$ are imposed as well, we find that an AR(2) model is adequate for the $\Delta_4 y_t$ series. The resulting one-step ahead forecasts for y_t over the period 1993.1–2005.5 have an RMSPE equal to 0.308, more than 10 percent larger compared to the forecasts based on the AR model for $\Delta_1 y_t$. The EACF of the doubly-differenced series $\Delta_1 \Delta_4 y_t$ has significant values at lags 1–5, suggesting an AR(5) may be suitable. Estimating this model gives the results

$$\begin{aligned} \Delta_1 \Delta_4 y_t = {} & \underset{}{0.0565} + \underset{(0.309)}{1.284 \Delta_1 \Delta_4 y_{t-1}} - \underset{(0.087)}{0.777 \Delta_1 \Delta_4 y_{t-2}} \\ & - \underset{(0.151)}{0.687 \Delta_1 \Delta_4 y_{t-3}} - \underset{(0.147)}{0.042 \Delta_1 \Delta_4 y_{t-4}} - \underset{(0.086)}{0.349 \Delta_1 \Delta_4 y_{t-5}} + \hat{\varepsilon}_t \end{aligned} \tag{5.49}$$

The roots of the estimated AR-polynomial equal $0.91 \pm 0.25i$, $-0.03 \pm 0.91i$, and 0.48, indicating that both nonseasonal unit roots and the complex pair $\pm i$ are wrongfully imposed through double-differencing.

If we ignore this and proceed with forecasting y_t using this model, we find an RMSPE equal to 0.496, again considerably higher than the value found for the model based on the first differences $\Delta_1 y_t$.

Second, we consider the reverse situation, that is, the consequences of using an AR model with deterministic seasonality for forecasting, when the seasonality in a time series y_t is in fact stochastic. At first sight, this appears to be less damaging, for the following reason. For simplicity, we assume that the series y_t follows a seasonal random

walk $\Delta_S y_t = \varepsilon_t$. Again, to focus on the effects of misspecification of the seasonality in y_t, we assume that the nonseasonal unit root is imposed and consider the properties of forecasts obtained from an AR(p) model with deterministic seasonality for $\Delta_1 y_t$, that is

$$\Delta_1 y_t = \sum_{s=1}^{S} \alpha_s D_{s,t} + \phi_1 \Delta_1 y_{t-1} + \phi_2 \Delta_1 y_{t-2} + \cdots + \phi_p \Delta_1 y_{t-p} + v_t. \tag{5.50}$$

Notice that the seasonal random walk $\Delta_S y_t = \varepsilon_t$ can be written as

$$\Delta_1 y_t = -\Delta_1 y_{t-1} - \Delta_1 y_{t-2} - \cdots - \Delta_1 y_{t-(S-1)} + \varepsilon_t.$$

From the properties of OLS, it follows that as long as the AR-order p in (5.50) is equal to or larger than $S-1$, the estimates of the parameters in this model are consistent and converge to their true values $\alpha_s = 0$, $s = 1, \ldots, S$, $\phi_i = -1$, $i = 1, \ldots, S-1$, and $\phi_i = 0$, $i \geq S$, when the sample size T becomes large. Assuming that this occurs, the one-step ahead forecast error from (5.50) is equal to ε_{T+1} with variance σ^2, identical to the correct optimal one-step ahead forecast from the seasonal random walk model itself. Hence, it does not seem to be harmful at all to mistake stochastic seasonality for deterministic seasonality.

In practice, things are more complicated. First, an important requirement for this result is that $p \geq S-1$, while in practice a low-order AR-model may appear to be adequate because a substantial part of the seasonality in $\Delta_1 y_t$ can be accounted for by the seasonal dummies in (5.50). Second, for small sample sizes T, estimation uncertainty in the parameters in the AR(p) model for $\Delta_1 y_t$ inflates the forecast error variance. Hence, in practice it is likely to be worthwhile to impose seasonal unit roots when they are indeed present.

Also in this case we should avoid overdifferencing though. As shown before, when the SARIMA model with double-differencing $\Delta_1 \Delta_S y_t = u_t$ is used for forecasting, the one-step ahead forecast error is given by $y_{T+1} - y_{T-S+1} + \Delta_S y_T = \Delta_S y_{T+1} + \Delta_S y_T$, which is equal to $\varepsilon_{T+1} + \varepsilon_T$ for a seasonal random walk y_t. So again, the one-step ahead forecast error variance is equal to $2\sigma^2$.

The above analysis suggests that the assumptions made about the presence of seasonal unit roots, or, in other words, the choice of differencing filter can be important for the accuracy of out-of-sample forecasts. In practice, the true nature of the seasonality in a given time series is of course unknown and needs to be determined from the data. It is therefore recommended to first assess the presence of seasonal unit roots by applying the HEGY tests discussed in Section 5.2. This suggests which differencing filter should be applied to y_t to make it tationary. The conventional time series tools discussed in Chapter 3 can then be used to specify an appropriate model for this transformed series, which subsequently can be used for forecasting purposes.

To illustrate this, consider the quarterly US industrial production series. The HEGY test results in Table 5.3 suggest that the real roots 1 and -1 may be present in this series, such that the filter $(1 - L^2)$ may be an adequate transformation to render the series stationary, that is, we should consider modelling $y_t - y_{t-2}$. Using the observations for the period 1963.1–1992.4 for model specification, we arrive at an AR(5) model for this series. One-step ahead forecasts for the level series y_t for the period 1993.1–2012.4 have an RMSPE equal to 1.246. If we only imposed the nonseasonal unit root 1 and considered modelling $\Delta_1 y_t$, we end up with using an AR(5) model for this series as well (but now with quarterly dummies included), which produces an RMSPE of 1.355. On the other hand, we may also impose the seasonal unit roots $\pm i$ and consider modelling the annual differences $\Delta_4 y_t$. Model selection criteria and tests for residual autocorrelation suggest using an AR(4) model for this series, which produces an RMSPE of 1.438 for the one-step ahead forecasts of y_t. Hence, both under- and overdifferencing relative to the outcome of the HEGY tests produces less accurate out-of-sample forecasts.

CONCLUSION

In this chapter we have reviewed two different possibilities for incorporating seasonality in ARIMA models. First, deterministic seasonality can be accounted for by means of seasonal dummy variables, which presupposes a seasonal pattern that is stable over time. Second, stochastic seasonality or the presence of changing seasonal patterns leads to the adoption of models that contain seasonal unit roots. Statistical testing procedures are available for distinguishing between deterministic and stochastic seasonality, although in practice this may be difficult. Evidently, there are many other issues in seasonality that have not been discussed here. We refer to Ghysels and Osborn (2001) for a more comprehensive coverage of econometric analysis of seasonal time series, including seasonal adjustment procedures and nonlinear seasonal models. Finally, Osborn (2002) and Ghysels *et al.* (2006) review issues that are relevant in forecasting seasonal time series, see also Franses and Paap (2002) for an extensive discussion of forecasting with periodic models.

EXERCISES

5.1 Consider the first differences of quarterly log US industrial production, over the period 1963.1–2002.4. Create 4 quarterly dummies $D_{s,t}$ and estimate the model

$$y_t = \alpha + \mu_1 D_{1,t} + \mu_2 D_{2,t} + \mu_3 D_{3,t} + \mu_4 D_{4,t} + \varepsilon_t.$$

What happens? Why? What happens if α is restricted to 0? What happens if μ_1 is restricted to 0? How can the parameters be interpreted in these two cases?

5.2 Generate a quarterly time series from an AR(1) model with seasonal means, that is, (5.1) with $p = 1, q = 0$. Set $\mu_1 = 1, \mu_2 = -1, \mu_3 = -3, \mu_4 = -1, \phi_1 = 0.7$, and $\sigma^2 = 1$. Set the starting value y_0 equal to -1. Estimate the model

$$y_t = \alpha_1 D_{1,t} + \alpha_2 D_{2,t} + \alpha_3 D_{3,t} + \alpha_4 D_{4,t} + \phi_1 y_{t-1} + \varepsilon_t.$$

How do you find estimates of the seasonal means μ_s, $s = 1, 2, 3, 4$?

5.3 Show that the restricted AR(S) model in (5.6) reduces to the seasonal random walk in (5.7) when $\phi_S = 1$.

5.4 Simulate a quarterly random walk according to (5.7) with $S = 4$, and $\delta = 0$, for sample size $T = 200$, setting the starting values $y_1 = 1, y_2 = -1, y_3 = -3, y_4 = -1$. Construct the vector-of-quarter plot for this series. What do you observe? Relate this to Figure 5.2.

5.5 Show that $\pm i$ are solutions to the characteristic equation $\phi_4(z) = (1 - z^4) = 0$.

5.6 What are the solutions to the characteristic equation $\phi_{12}(z) = (1 - z^{12}) = 0$, corresponding to a monthly seasonal random walk?

5.7 Show that the seasonal differencing filter $(1 - L^4)$ can be decomposed as $(1 - L^4) = (1 + L + L^2 + L^3)(1 - L)$, as $(1 - L^4) = (1 - L + L^2 - L^3)(1 + L)$, and as $(1 - L^4) = (1 - L^2)(1 + L^2)$.

5.8 Show that the complex pair of unit roots $\pm i$ in the case of a quarterly seasonal random walk $y_t = y_{t-4} + \varepsilon_t$ imply cycles of four periods.

5.9 Consider an AR(3) model for a quarterly series y_t, that is

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \varepsilon_t,$$

where ε_t is a white noise series with variance σ^2. Derive the parameter restrictions for $\{\phi_1, \phi_2, \phi_3\}$ that should hold in case y_t has (a) a single nonseasonal unit root, (b) two nonseasonal unit roots, and (c) three seasonal unit roots: -1, and $\pm i$.

5.10 Suppose the data generating process for a time series y_t is $(1 - L^2)y_t = \varepsilon_t$, where ε_t is a white noise process. What are the autocorrelation properties of the seasonally differenced series $u_t = (1 - L^4)y_t$?

5.11 Show that the $\Delta_1\Delta_4$ differencing filter that is used in the airline model for quarterly time series, that is, (5.23) with $S = 4$, assumes the presence of two nonseasonal unit roots 1 and three seasonal unit roots -1 and $\pm i$.

5.12 How can we detect overdifferencing from the ACF of the MA part of a SARIMA model?

5.13 Use simulation to verify that the transformations $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$ as given in (5.12)–(5.14) are asymptotically uncorrelated in case the series y_t follows a quarterly seasonal random walk.

5.14 Consider the decomposition of the AR(p) polynomial in (5.37). Derive expressions for π_1, π_2, π_3, π_4, and $\phi^*_{p-4}(L) = 1 - \phi^*_1 L - \phi^*_2 L^2 - \cdots - \phi^*_{p-4} L^{p-4}$ in terms of the original coefficients $\phi_1, \ldots, \phi_p$ for the case $p = 5$.

5.15 Consider the UK industrial production series in the EViews workfile UKIP.wf1. (Note that the data have already been transformed to logarithms)

a. Examine the unit root properties of this series using the sample period 1960.1–1989.4.

b. Specify and estimate a model for $\Delta_4 y_t = (1 - L^4)y_t$ using the sample period 1960.1–1989.4, and use this model to compute 1-step ahead forecasts of y_t for the period 1990.1–2004.4.

c. Specify and estimate a model for $\Delta_2 y_t = (1 - L^2)y_t$ using the sample period 1960.1–1989.4, and use this model to compute 1-step ahead forecasts of y_t for the period 1990.1–2004.4.

d. What is your conclusion about the most appropriate model / unit root specification, taking into account all the evidence above?

6 Aberrant observations

For many economic time series variables it can happen that one or more observations are markedly different from the other observations. This often is due to the occurrence of exceptional and usually unpredictable events. Such outliers occur rarely, are (often) unforecastable, and are (assumed to be) caused by exogenous influences. An illustrative example is given by the May 1968 uproar in France, which caused important macro-economic variables such as industrial production to have a much lower value than usual, see Figures 6.1 and 6.2.

As another example, the stock market crash on Monday October 19, 1987, may be considered as an extraordinary event, which gave rise to a return that is markedly different from the bulk of the data, see Figures 6.3 and 6.4. Of course, stock market crashes do happen once in a while, but the fact that automated trading programs were one of the main reasons for the 'Black Monday' crash to be that dramatic makes this return observation quite exceptional.

Sometimes aberrant observations are part of the process which you want to model, that is, they are in fact the most interesting observations in a time series. For example, the effect of substantial price discounts on product sales makes the low price data very informative. Other effects that can lead to outliers are more difficult to capture with a time series forecasting model. For example, again price discounts but now by competitors generally are impossible to predict by the own company. It is generally understood that when aberrant observations are neglected completely, they can be very harmful, in the sense that they may have a large effect on parameter estimates and on forecasting, which we will demonstrate below. In sum, aberrant observations often are very influential, and hence we somehow have to deal with them.

Outliers and the model

To deepen the discussion, it is important to recognize that aberrant data points are different from the other observations within the context of a model. So, again a time series feature is defined in terms of a model (like a trend and seasonality). To illustrate,

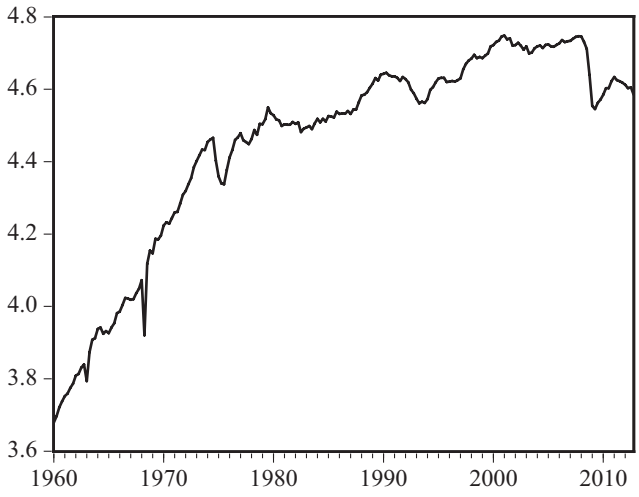


Figure 6.1: Quarterly log industrial production France (1960Q1–2004Q4).

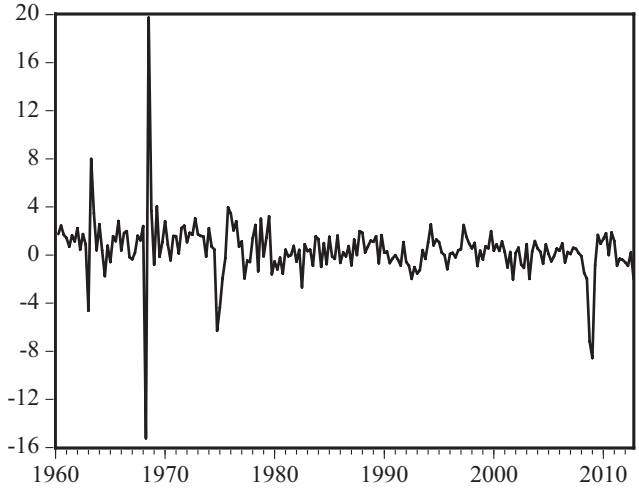


Figure 6.2: Quarterly growth rates of industrial production France (1960Q1–2004Q4).

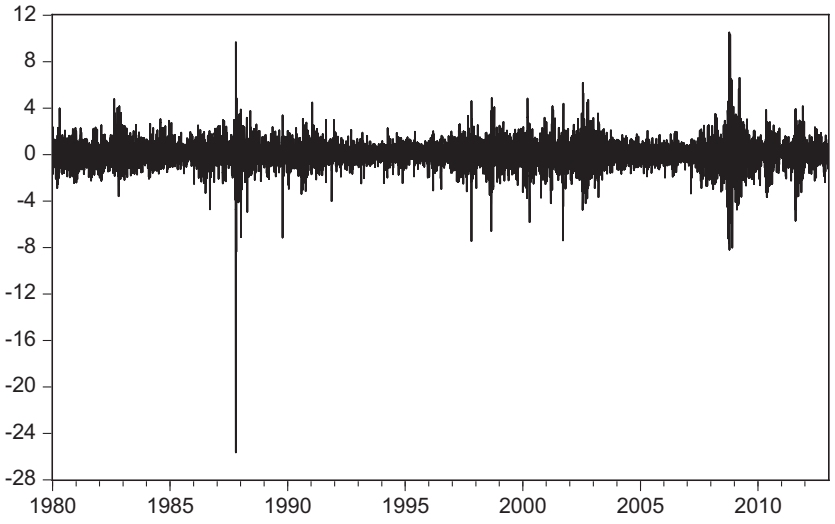


Figure 6.3: Daily returns on the Dow Jones index.

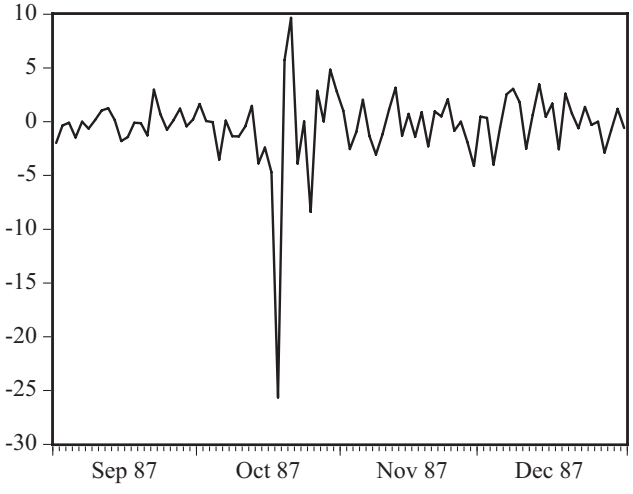


Figure 6.4: Daily returns on the Dow Jones index, September 1–December 31, 1987.

suppose we have ten observations on y_t, and these are

$$9.23,\ \ 11.67,\ \ 8.93,\ \ 12.01,\ \ 10.73,\ \ 100.98,\ \ 8.45,\ \ 9.79,\ \ 10.15,\ \ 10.90.$$

If the model would be $y_t = \mu + \varepsilon_t$ with $y_t \sim N(\mu, \sigma^2)$, the sixth observation can be considered an outlier.

Exercise 6.1

However, if the model would be

$$y_t = \mu_1 + \mu_2 z_t + \varepsilon_t,$$

where $z_t = 1$ with probability 0.1 and $z_t = 0$ otherwise, the sixth observation can be considered as a regular observation.

As another illustration, consider now a time series model. For example, suppose that a time series y_t is generated by the AR(1) model

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \qquad t = 1, 2, \ldots, T, \tag{6.1}$$

where $|\phi_1| < 1$ and ε_t is assumed to be drawn from a normal distribution $N(0, \sigma^2)$. Assume furthermore that $\sigma^2 = 1$ and y_0 is set equal to -1. Given that the resulting time series y_t is stationary with unconditional mean equal to zero and unconditional variance $1/(1-\phi_1^2)$, it is highly unlikely that the observation at $t = 20$ is equal to 10. Suppose however, that the observations on y_t are recorded by a statistical agency, and that, when downloading these numbers, somehow a mistake is made and y_{20} erroneously becomes 11.5 instead of the true value of 1.5, say. In that case, we observe a new variable y_t^*, which is y_t as in (6.1) with one observation changed. For this y_t^* variable, we can easily conclude that the value of y_{20}^* does not correspond to the general pattern of the series, and also that is shall be hard to predict using y_{19}^*. In fact, for most data points it approximately holds that $y_t^* = \phi_1 y_{t-1}^*$, while for y_{20}^* (and y_{21}^*, as will become clear below) this evidently is not the case. Hence, y_{20}^* can be called an outlier. Note that as there are various types of outliers or even sequences of outliers, we use the more general phrase of aberrant observations instead of outliers in this chapter.

What could go wrong?

The above example already shows that it can be important to account for outliers when modeling and forecasting economic time series. In fact, neglecting such aberrant observations can have at least three major effects. The first of these is that parameter estimators can be biased when outliers are neglected. For example, when the y_t^* series (with the outlier at $t = 20$) is used to estimate the autoregressive parameter ϕ_1, we may expect that this $\hat{\phi}_1$ can be quite different from the true value. Below, we will

demonstrate that for the so-called additive outlier [AO], the least squares estimate $\hat{\phi}_1$ is biased towards zero, while for a so-called level shift (where the mean of the series changes permanently) it is biased towards 1. Hence, when we forecast y_{T+1}, using the estimate $\hat{\phi}_1$ obtained from the contaminated y_t^* series, we can expect some forecasting bias.

The second effect of neglecting aberrant observations, especially those which are close to the forecasting origin T, is that forecasts can be very inaccurate. For example, suppose that the above y_t^* series ends at $t = 20$, and we wish to forecast y_{21}^*. In that case, the value of 11.5 for y_{20}^* will have a dramatic effect on this forecast.

The third effect is that when an AR model as in (6.1) is estimated for the y_t^* series, we may expect that the estimate of the residual variance $\hat{\sigma}^2$ for these contaminated data is much larger than the true value σ^2. This, in turn, implies that the corresponding interval forecasts will be much too wide.

How to look at outliers?

There are various ways to look at aberrant observations and how they should be dealt with when modeling and forecasting economic variables. One approach assumes that there is (almost) no *a priori* knowledge about the observation(s) that could possibly be aberrant. For the statistical analysis of a time series, we may then assume that apparently irregular data are generated by a distribution that is different from most other data points. For example, we can assume that most returns on the Dow Jones index are normally distributed with mean 0 and variance σ^2, while for only a few it holds that the data have been generated by a normal distribution with variance ω^2 with ω^2 being much larger than σ^2. This means that the data originate from a mixture of distributions.

An alternative view is to replace the normality assumption of the error process as in (6.1) by the assumption of (for example) a Student-t distribution. This distribution has fatter tails than the normal distribution and may generate more extreme values of ε_t, so that the corresponding y_t data can also take more extreme values.

Yet another approach, which focuses more on estimating ϕ_1 in (6.1), concerns replacing OLS by so-called robust estimation methods, see Denby and Martin (1979) and Bustos and Yohai (1986). Although these approaches incorporate the possibility to examine particular observations with respect to their possible influence, see for example Lucas (1996), in general these methods focus on reducing the impact of aberrant data on estimation and forecasting without paying specific attention to the influential observations themselves. We will discuss this view in Section 6.3 below.

In that Section 6.3 we also pay attention to a fourth view which assumes that the modeler has some *a priori* knowledge about the likely location and relevance of such outliers. An outlier detection technique can then be used to allocate them. Once found,

regressors to describe these outliers can be included on the right-hand side of the model.

From a modeling and forecasting point of view it is important to have knowledge of the occurrence of aberrant observations, in particular their timing and nature. For example, a one-time promotion (like "buy three for the price of two") in week j can generate a large increase in sales in week j, maybe a little decrease in week $j+1$, and then its effect may fade out. Removal of this observation would not make it possible to study the effect of promotion on sales. See Leone (1995) for an approach that explicitly allows for outliers and level shifts to study the possible long-run impact of promotional activities. In this chapter, we include (models for) aberrant data within the class of ARMA type time series models. Again, all these models are regression-based. In general, it seems sensible to include as much information as possible on aberrant data in our model, so that we can decide on their impact on modeling and forecasting in a next step.

The outline of this chapter is as follows. In Section 6.1, we give some models that allow us to describe several forms of aberrant data and to understand how these may emerge in economic variables. Such models can be useful to test their impact, as well as to exploit their description in a forecasting model. In Section 6.2 we focus on the consequences of neglecting outliers. In Section 6.3 we describe methods to deal with outliers. Finally, in Section 6.4 we discuss the issue of unit root testing in the presence of aberrant data.

6.1 Modeling aberrant observations

In this section we discuss how aberrant observations can be represented in the context of a model. As such, we describe various types of aberrant observations.

Suppose we are interested in a time series x_t, and assume that it can be described by the AR(1) model,

$$x_t = \phi_1 x_{t-1} + \varepsilon_t, \tag{6.2}$$

or

$$(1 - \phi_1 L)x_t = \varepsilon_t,$$

where $|\phi_1| < 1$, L is the usual lag operator defined as $Ly_t \equiv y_{t-1}$, and $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$.

Suppose further that instead of x_t we actually observe the contaminated series y_t, where

$$y_t = x_t + \zeta_t d_t, \tag{6.3}$$

where

$$d_t = \begin{cases} +1 \text{ or } -1, & \text{if an outlier occurs,} \\ 0, & \text{otherwise.} \end{cases} \tag{6.4}$$

For example, we may have $P(d_t = 1) = P(d_t = -1) = \frac{\pi}{2}$ and $P(d_t = 0) = 1 - \pi$, with $0 < \pi < 1$ and where π is small. Note that this means that d_t determines if and when an outlier occurs. The ζ_t is the so-called contamination process, which determines the characteristics of the outlier.

Additive Outlier [AO]

An additive outlier [AO] can be viewed as an observation which is the original observation plus or minus some value. An AO only affects the observation at time, say, τ:

$$\ldots, y_{\tau-1} = x_{\tau-1}, \quad y_\tau = x_\tau + \zeta, \quad y_{\tau+1} = x_{\tau+1}, \ldots$$

where $\zeta_t \equiv \zeta \neq 0$, such that

$$y_t = x_t + \zeta d_t \tag{6.5}$$

Note that from (6.5) when AR(1) dynamics are imposed one gets

$$(1 - \phi_1 L)y_t = (1 - \phi_1 L)x_t + (1 - \phi_1 L)\zeta d_t,$$

or

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + (1 - \phi_1 L)\zeta d_t.$$

Figure 6.5 depicts an example of an additive outlier [AO]. The series x_t is generated according to an AR(1) model $x_t = \phi_1 x_{t-1} + \varepsilon_t$, with $\phi_1 = 0.7$ and $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma = 1$. An AO of size 10σ occurs at $t = 50$.

Suppose that x_t is generated by an AR(1) model with parameter ϕ_1, such that $|\phi_1| < 1$. In case we do not observe x_t but y_t as in (6.5), at time τ there is an AO of magnitude ζ which appears on the left-hand side of the AR(1) model, while at time $\tau + 1$, this AO observation moves towards the right-hand side of the model in the y_{t-1} part. In other words, when we make a scatter plot of y_t versus y_{t-1}, and this quite important, a (neglected) AO gives rise to two irregular data points. In Figure 6.6, we give such a scatter plot for a simulated time series with $T = 100$, $\phi_1 = 0.7$, $\sigma = 1$, $\zeta = 10\sigma$ and $\tau = 50$. Clearly, we observe the two data points (y_t, y_{t-1}) which do not seem to correspond to the general cloud of observations.

As is obvious from Figure 6.6, an AO is reflected in two pairs of (y_t, y_{t-1}) in case of an AR(1) time series x_t. Hence, neglecting such an AO while estimating the parameter in an AR(1) model for y_t using OLS would result in two large residuals. In fact, as y_τ

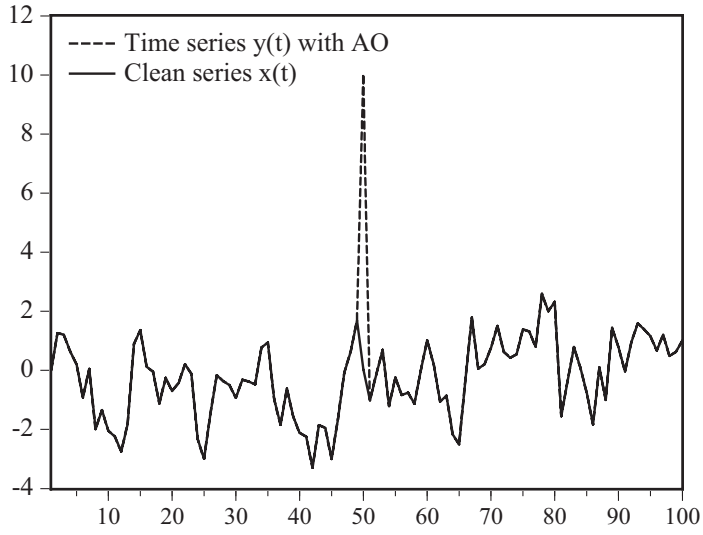


Figure 6.5: Example of an additive outlier [AO]. The series x_t is generated according to an AR(1) model $x_t = \phi_1 x_{t-1} + \varepsilon_t$, with $\phi_1 = 0.7$ and $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma = 1$. A single AO of size 10σ occurs at $t = \tau = 50$.

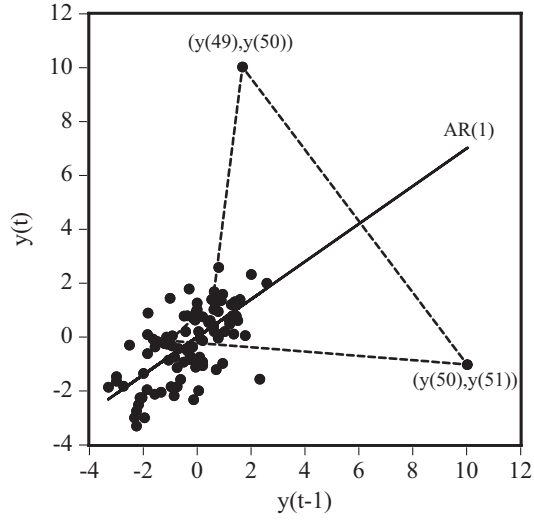


Figure 6.6: Example of an additive outlier [AO]. The series x_t is generated according to an AR(1) model $x_t = \phi_1 x_{t-1} + \varepsilon_t$, with $\phi_1 = 0.7$ and $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma = 1$. A single AO of size 10σ occurs at $t = \tau = 50$. The solid black line indicates the true AR(1) regression line, $y_t = \phi_1 y_{t-1}$.

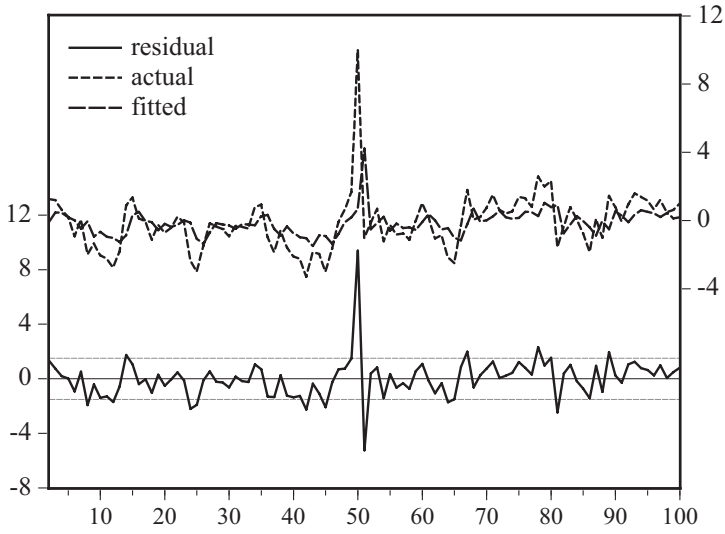


Figure 6.7: Effect of neglecting a single additive outlier on residuals of AR(1) model.

is an extraordinary data point, its forecast $\hat{y}_\tau$ given $y_{\tau-1}$ and some estimate of $\hat{\phi}_1$ can be quite different from the true y_τ value. Additionally, y_τ itself is a biased predictor variable for $y_{\tau+1}$. In sum, when an AO in an AR(1) series is neglected, we find two large errors, that is, $\hat{\varepsilon}_\tau$ and $\hat{\varepsilon}_{\tau+1}$. This is illustrated by Figure 6.7, showing the same simulated time series as before, the fit from an AR(1) model (based on OLS estimates), and the estimated residual series. Clearly, there are two large residuals.

Given that there are two observations for an AR(1) series that do not correspond with the cloud of (y_t, y_{t-1}) points, as visualized in Figure 6.7, we may expect that neglecting AOs can have a serious impact on the OLS estimator of ϕ_1. This will be discussed in more detail in the next section.

Innovation Outlier [IO]

An innovation outlier [IO] also affects future observations at $t = \tau + 1, \tau + 2, \ldots$, but its effect disappears in the same way as "regular" shocks ε_t do. This can be seen as follows. Compare the series x_t and y_t, given by

$$x_t = \phi_1 x_{t-1} + \varepsilon_t,$$
$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \zeta d_t,$$

and assume that $d_\tau = 1$, while $d_t = 0$ for all $t \neq \tau$. Hence, y_t and x_t are identical up to $t = \tau$, at which point y_t experiences an unusually (possibly) large shock $\varepsilon_t + \zeta d_t$, that is,

$$\begin{aligned} y_{\tau-1} &= x_{\tau-1}, \\ y_\tau &= x_\tau + \zeta, \\ y_{\tau+1} &= x_{\tau+1} + \phi_1 \zeta, \\ y_{\tau+2} &= x_{\tau+2} + \phi_1^2 \zeta, \quad \text{etc} \end{aligned}$$

An IO in y_t can thus be interpreted as an AO in the error process ε_t, or as an "unusual shock" (think of an oil crises) (see, for example, Figure 6.8)

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \zeta d_t. \tag{6.6}$$

Exercise 6.2

We can write an IO in the general format above, that is, $y_t = x_t + \zeta_t d_t$ as follows. From (6.6), we have

$$\begin{aligned} (1 - \phi_1 L) y_t &= \varepsilon_t + \zeta d_t \\ &= (1 - \phi_1 L) x_t + \zeta d_t, \end{aligned}$$

such that dividing by $(1 - \phi_1 L)$ gives

$$y_t = x_t + \frac{\zeta}{(1 - \phi_1 L)} d_t = x_t + \zeta_t d_t, \tag{6.7}$$

where $\zeta_t \equiv \frac{\zeta}{1-\phi_1 L}$.

Exercise 6.3

As all observations after time τ obey the AR(1) model, we can expect that the corresponding pairs of $(y_t, y_{\tau-1})$ (with $t > \tau$) lie on the regression line with slope ϕ_1. When ζ is very large, the graph in Figure 6.9 suggests that neglecting an IO leads to a biased estimate of the possible intercept (if this is added to (6.6)), and that $\hat{\phi}_1$ may show virtually no bias. When such an IO is neglected for an AR(1) series, we will have only a single extraordinarily large estimated residual, due to the fact that $\hat{\phi}_1 y_{\tau-1}$ is a biased predictor for y_τ. This effect is visualized in Figure 6.10 for a simulated AR(1) time series with $\zeta = 10\sigma$, $\tau = 50$, $\phi_1 = 0.7$, and $\sigma = 1$.

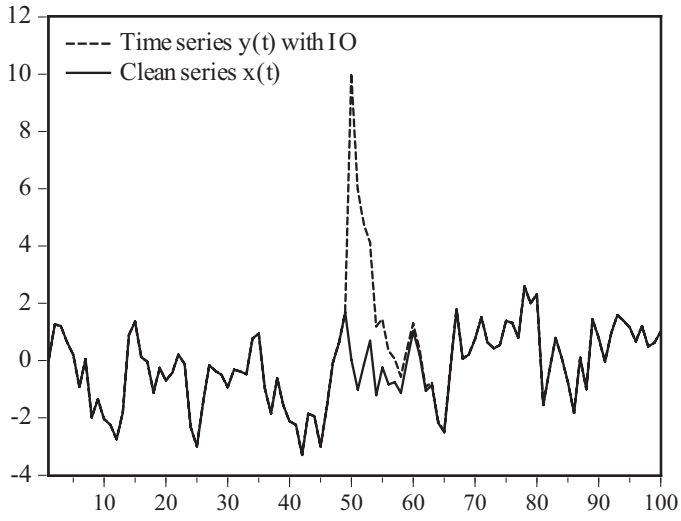


Figure 6.8: Example of an innovation outlier [IO]. The series x_t is generated according to an AR(1) model $x_t = \phi_1 x_{t-1} + \varepsilon_t$, with $\phi_1 = 0.7$ and $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma = 1$. A single IO of size 10σ occurs at $t = \tau = 50$.

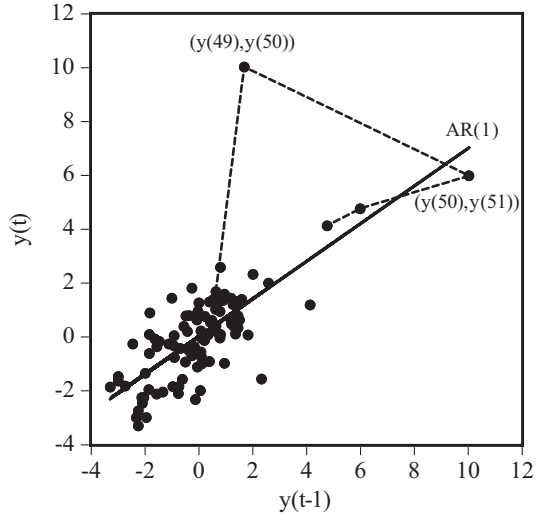


Figure 6.9: Example of an innovation outlier [IO]. The series x_t is generated according to an AR(1) model $x_t = \phi_1 x_{t-1} + \varepsilon_t$, with $\phi_1 = 0.7$ and $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma = 1$. A single IO of size 10σ occurs at $t = \tau = 50$. The solid black line indicates the AR(1) regression line, $y_t = \phi_1 y_{t-1}$.

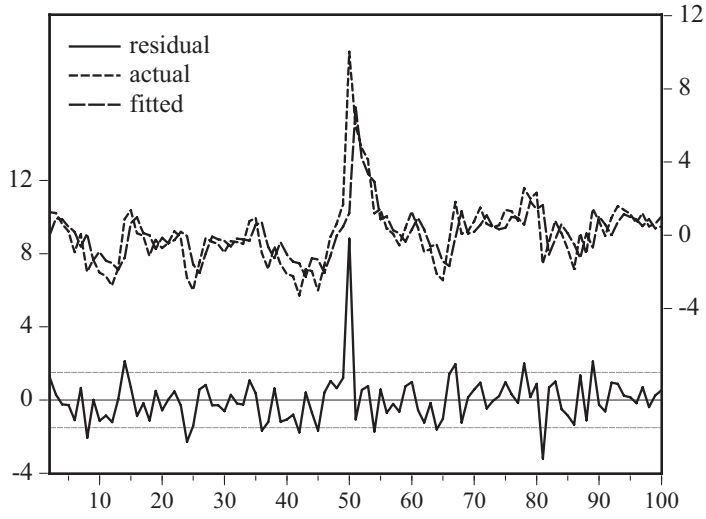


Figure 6.10: Effect of neglecting a single innovation outlier on residuals of AR(1) model.

Transient change [TC]

A so-called transient change [TC] also affects future observations at $t = \tau + 1, \tau + 2, \ldots,$ but its effect disappears in a different way than "regular" shocks ε_t. This can be represented as

$$\begin{aligned} y_{\tau-1} &= x_{\tau-1}, \\ y_{\tau} &= x_{\tau} + \zeta, \\ y_{\tau+1} &= x_{\tau+1} + \delta\zeta, \\ y_{\tau+2} &= x_{\tau+2} + \delta^2\zeta, \quad \text{and so on,} \end{aligned}$$

where δ is not equal to the AR(1) parameter ϕ_1. In other words, $\zeta_t \equiv \frac{\zeta}{1-\delta L}$ for a certain $0 < \delta < 1$, such that

$$y_t = x_t + \frac{\zeta}{(1-\delta L)} d_t. \tag{6.8}$$

To arrive at a time series model specification for the observed series y_t, note that because of this equality it follows that

$$\begin{aligned} (1-\phi_1 L) y_t &= (1-\phi_1 L) x_t + (1-\phi_1 L) \frac{\zeta}{(1-\delta L)} d_t \\ &= \varepsilon_t + \zeta \frac{(1-\phi_1 L)}{(1-\delta L)} d_t, \end{aligned}$$

or

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \zeta \frac{(1-\phi_1 L)}{(1-\delta L)} d_t. \tag{6.9}$$

Exercise 6.4

Level Shift [LS]

A level shift [LS] also affects future observations at $t = \tau + 1, \tau + 2, \ldots$, and here its effect does not disappear at all. This can be put into equations like

$$\begin{aligned} y_{\tau-1} &= x_{\tau-1}, \\ y_\tau &= x_\tau + \zeta, \\ y_{\tau+1} &= x_{\tau+1} + \zeta, \\ y_{\tau+2} &= x_{\tau+2} + \zeta, \quad \text{etc} \end{aligned}$$

In general, for $t \geq \tau$ we have $y_t = x_t + \zeta d_\tau = x_t + \zeta L^{t-\tau} d_t$.

In practice, of course we do not know the timing of possible level shifts. Hence, we may write

$$\begin{aligned} y_t &= x_t + \zeta d_t + \zeta d_{t-1} + \zeta d_{t-2} + \cdots \\ &= x_t + \zeta(1 + L + L^2 + \cdots) d_t = x_t + \frac{\zeta}{(1-L)} d_t, \end{aligned} \tag{6.10}$$

such that $y_t = x_t + \zeta_t d_t$ with $\zeta_t \equiv \frac{\zeta}{1-L}$.

To arrive at a time series model specification for the observed series y_t, note that because

$$y_t = x_t + \frac{\zeta}{(1-L)} d_t,$$

it follows that

$$\begin{aligned} (1-\phi_1 L) y_t &= (1-\phi_1 L) x_t + (1-\phi_1 L) \frac{\zeta}{(1-L)} d_t \\ &= \varepsilon_t + \zeta \frac{(1-\phi_1 L)}{(1-L)} d_t, \end{aligned}$$

or

$$y_t = \phi_1 y_{t-1} + \varepsilon_t + \zeta \frac{(1-\phi_1 L)}{(1-L)} d_t. \tag{6.11}$$

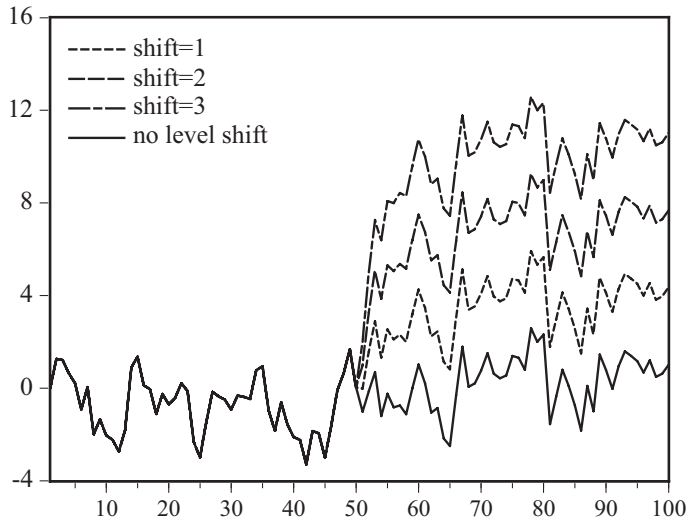


Figure 6.11: Example of a level shift [LS]. The series x_t is generated according to an AR(1) model $x_t = \phi_1 x_{t-1} + \varepsilon_t$, with $\phi_1 = 0.7$ and $\varepsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma = 1$. Level shifts of size 1σ, 2σ and 3σ occur at $t = \tau = 50$.

In this section we have seen that four types of (sets of) aberrant observations can be captured in a regression framework. This shall become important when one wants to test the data for the presence of such observations. We will return to this in Section 6.3 below. First, we use these expression to study the consequences of neglecting aberrant observations.

6.2 What happens if we neglect outliers?

Neglecting outliers (may) have effects on OLS parameter estimates, out-of-sample forecasts and properties of residuals, whereas the latter in turn can be useful to identify outlier.

To illustrate these three effects, consider again the AR(1) model for a series that contains a single outlier at $t = \tau$, that is,

$$x_t = \phi_1 x_{t-1} + \varepsilon_t,$$

$$y_t = x_t + \zeta_t d_t,$$

$$d_t = 1 \text{ when } t = \tau \text{ and } 0 \text{ otherwise,}$$

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

Effects on OLS estimates

Additive outliers cause the OLS estimate of ϕ_1 to be biased and inconsistent. If a single AO of magnitude ζ occurs in a sample of size T, it can be shown that

$$\mathsf{E}[\hat{\phi}_1] \approx \phi_1 - \frac{1}{T-1}\phi_1(1-\phi_1^2)(\zeta/\sigma)^2, \tag{6.12}$$

see Lucas (1996), for example. Given this expression, it is clear that $|\mathsf{E}[\hat{\phi}_1]| < |\phi_1|$, which means that the OLS estimate is biased towards 0. This is exemplified in Figure 6.6.

If Innovation Outliers occur, the OLS estimate of ϕ_1 is unbiased and consistent. In fact, IOs often increase the precision of OLS estimates as they are very informative observations. This can be learned from the graph in Figure 6.9.

Effects on forecasts

The effects of neglected outliers on forecasts are small if the outliers are somewhere in the middle of the sample, but of course, when they occur close to or even at the forecast origin, their effect can be substantial. Note that if the outlier occurs precisely at the forecast origin, it is impossible to identify whether it is an AO or an IO. This is due to the fact that we can only learn about the character of an outlier after observing its effect (or not) on subsequent observations, after time τ.

Furthermore, as neglected outliers can inflate the estimate of the residual variance, the effects on the size of the forecast interval can be large.

Effects on residuals

Consider again the representations of AOs and IOs in an AR(1) model

$$\begin{aligned} \text{AO:} \quad & y_t = \phi_1 y_{t-1} + \varepsilon_t + \zeta(1-\phi_1 L)d_t \\ \text{IO:} \quad & y_t = \phi_1 y_{t-1} + \varepsilon_t + \zeta d_t, \end{aligned}$$

where the key difference between the two models is in the multiplier of d_t.

Define the residuals as $e_t = y_t - \phi_1 y_{t-1}$, such that

$$\text{AO:} \quad e_t = \varepsilon_t + \zeta(d_t - \phi_1 d_{t-1}) \tag{6.13}$$

$$\text{IO:} \quad e_t = \varepsilon_t + \zeta d_t, \tag{6.14}$$

In words, an AO at $t = \tau$ affects residuals at $t = \tau$ and $t = \tau + 1$ ($e_\tau = \varepsilon_\tau + \zeta$, $e_{\tau+1} = \varepsilon_{\tau+1} - \phi_1\zeta$), while an IO at $t = \tau$ only affects the residual at $t = \tau$ ($e_\tau = \varepsilon_\tau + \zeta$). These two aspects are visualized in Figure 6.7 and 6.10.

Because $e_t \neq \varepsilon_t$ even if ϕ_1 is known, the properties of residuals are affected. For example, it is easy to appreciate that AOs and IOs lead to excess kurtosis (and thus to non-normality). Also, AOs give rise to autocorrelation in the (squared) residuals.

Exercise 6.5–6.6

6.3 What to do about outliers?

There are various options to handle outliers, and in this chapter we choose to address two approaches. First, one can use an outlier detection method, and second, one can resort to robust estimation methods.

As concerning outlier detection methods (like the ones proposed in Tsay (1988); Chen and Liu (1993)), the typical procedure is the following. First, one estimates a model using the observed time series x_t. One then inspects residuals for typical outlier patterns. Third, one eliminates the most convincing outlier by means of dummy variables (also called an intervention model). Then, steps one, two and three are repeated until no more outliers are detected.

6.3.1 Outlier detection methods

In this section we describe how outlier detection methods work. Assume that we suspect that an outlier has occurred at $t = \tau$. How can we test whether this really is the case?

One way to understand how outlier detection methods work, is to interpret (6.13)–(6.14) as a regression model for e_t, where we again use the AR(1) model for illustration, that is,

$$e_t = \zeta z_{it} + \varepsilon_t, \qquad t = 1, \ldots, T, \quad i = 1, 2, \tag{6.15}$$

with

$$\text{AO:} \quad z_{1t} = \begin{cases} 1 & \text{for } t = \tau \\ -\phi_1 & \text{for } t = \tau + 1 \\ 0 & \text{for } t > \tau + 1 \end{cases}$$

$$\text{IO:} \quad z_{2t} = \begin{cases} 1 & \text{for } t = \tau \\ 0 & \text{for } t > \tau \end{cases}$$

and $z_{it} = 0$, $i = 1, 2$, for all $t < \tau$. For each outlier type, one can obtain an estimate of ζ by regressing e_t on z_{it} resulting in

$$\hat{\zeta}_i(\tau) = \frac{\sum_{t=1}^{T} e_t z_{it}}{\sum_{t=1}^{T} z_{it}^2} = \frac{\sum_{t=\tau}^{T} e_t z_{it}}{\sum_{t=\tau}^{T} z_{it}^2}.$$

The significance of $\hat{\zeta}_i$ can be assessed by considering the associated t-statistic

$$\hat{\lambda}_i(\tau) = \frac{\hat{\zeta}_i(\tau)}{\hat{\sigma}_\varepsilon \left(\sum_{t=\tau}^{T} z_{it}^2\right)^{-1/2}},$$

where $\hat{\sigma}_\varepsilon$ is the OLS estimate of the standard deviation of ε_t.

Exercise 6.7

If the location of the outlier is somehow known, that is, there is a value of τ that is of interest, then $\hat{\lambda}_i(\tau)$ approximately follows a standard normal distribution.

Of course, in practice the location of an outlier is unknown. A solution is then to compute $\hat{\lambda}_i(\tau)$ for all $\tau = 1, 2, \ldots, T$ and to consider the supremum test statistic

$$\hat{\lambda}_i = \max_{1 \le \tau \le T} |\hat{\lambda}_i(\tau)|. \tag{6.16}$$

As the point-wise statistics $\hat{\lambda}_i(\tau_1)$ and $\hat{\lambda}_i(\tau_2)$ are not independent, $\hat{\lambda}_i$ follows a non-standard distribution, for which unfortunately no closed-form expression exists. Hence, standard critical values cannot be used, but have to be determined by simulation. In practice, a critical value of $C = 3$ or 3.5 is often used.

The value of τ for which the maximum in (6.16) occurs can be used as an estimate of the location of the outlier. To determine the type of the outlier, one can consider the maximum of $\hat{\lambda}_i$, $i = 1, 2$. And, if an outlier is detected and its type and location have been determined, its effects on the time series x_t can be removed using the dummy variables, after which the analysis is repeated with the corrected series.

6.3.2 Robust estimation methods

How robust estimation methods work is best outlined using the example of the ten observations before. Suppose again that we have 10 observations on x_t, and these are 9.23, 11.67, 8.93, 12.01, 10.73, 10.98, 8.45, 9.79, 10.15, 10.90.

Assume we want to estimate the parameter μ in the model $y_t \sim N(\mu, \sigma^2)$, where of course we are not certain whether that is the best model, but we simply asume it is. We can estimate μ by, for example, the sample mean $\hat{\mu} = 1/10 \sum_{t=1}^{10} x_t = 10.28$, and also by the sample median $\hat{\mu} = \text{med}(x_t, t = 1, \ldots, 10) = 10.44$

Suppose now that instead of x_t we observe y_t, where $y_t = x_t$ for $t = 1, \ldots, 10$ and $t \neq 6$, and $y_6 = x_6 + 90 = 100.98$. Assume again that we estimate the parameter μ in the model $N(\mu, \sigma^2)$ for the observed variable y_t. With the sample mean we get that $\hat{\mu} = 1/10 \sum_{t=1}^{10} y_t = 19.28$, while with the sample median we get $\hat{\mu} = \text{med}(y_t, t = 1, \ldots, 10) = 10.44$.

In other words, the sample mean is heavily influenced even by a single outlier. Or, the sample mean is not robust to outliers. The sample median on the other hand is *not* affected. So, the sample median is a *robust* estimator.

In a regression model similar results hold. For the AR(1) model, the OLS estimator implies that

$$\hat{\phi}_{1\,OLS} = \text{argmin} \sum_{t=1}^{T} (y_t - \phi_1 y_{t-1})^2, \tag{6.17}$$

that is, OLS minimizes the sum (which is of course T times the mean) of squared residuals. So, OLS is not robust to outliers. A robust alternative would be, for example, to minimize the *median* of the squared residuals.

More sophisticated robust methods use weighted least squares, that is,

$$\hat{\phi}_{1\,WLS} = \text{argmin} \sum_{t=1}^{T} w_t (y_t - \phi_1 y_{t-1})^2, \tag{6.18}$$

where $0 \leq w_t \leq 1$. The key idea is to take the weights w_t such that regular observations get large weights, and outliers get smaller weights. If an observation has a small weight w_t, it then does not influence the estimate of ϕ_1 very much.

As learned from graphs such as Figure 6.6, outliers are characterized by large values of the residual $y_t - \phi_1 y_{t-1}$, while especially those due to extreme values of y_{t-1} have considerable influence on parameter estimators.For this reason, robust estimation methods typically make w_t a function of y_{t-1} and $y_t - \phi_1 y_{t-1}$, such that $w_t \to 0$ if one of the two is very different from average values. Then, outliers become less influential.

Two often used examples of weight functions are trimmed least squares, that is,

$$w_t = \begin{cases} 1 & \text{if } |e_t/\sigma_e| < c \\ 0 & \text{otherwise} \end{cases}$$

and the Huber weights given by

$$w_t = \begin{cases} -c/(e_t/\sigma_e) & \text{if } e_t/\sigma_e < -c \\ 1 & \text{if } -c \leq e_t/\sigma_e < c \\ c/(e_t/\sigma_e) & \text{if } c \leq e_t/\sigma_e \end{cases}$$

As ϕ_1, the AR(1) parameter, is unknown and the weights w_t typically depend on ϕ_1, an iterative procedure is required. Given a starting value $\phi_1^{(0)}$, compute weights $w_t(\phi_1^{(0)})$

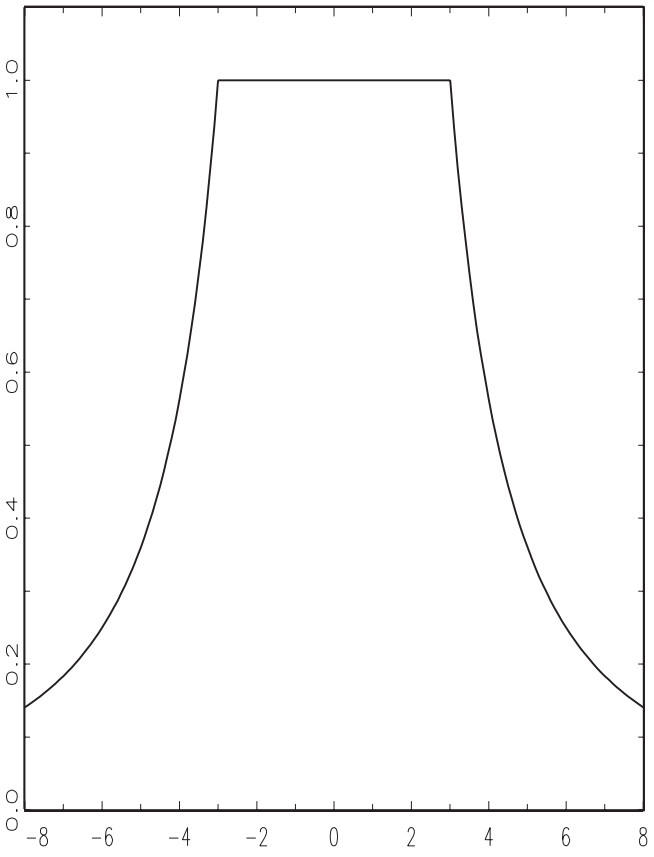


Figure 6.12: Huber weight function with $c = 3$.

Then, do WLS regression using $w_t(\phi_1^{(0)})$ to obtain estimate $\phi_1^{(1)}$. Next, compute weights $w_t(\phi_1^{(1)})$. Then, do WLS regression using $w_t(\phi_1^{(1)})$ to obtain estimate $\phi_1^{(2)}$. Finally, repeat this until estimates converge, that is, $\phi_1^{(i+1)} \approx \phi_1^{(i)}$.

The advantages of robust methods over the outlier detection methods are that no subjective judgement is required and that, after estimation, weights w_t can be used to determine which observations are to be considered to be outliers.

6.3.3 Some illustrations

An empirical illustration of what a neglected level shift can do is given by an analysis of the first part of the radio advertising expenditures series, see Figure 2.12. The expansion of broadcasting minutes was effectuated in 1982.01. However, as this change must have

been known beforehand, and given that contracts may have been settled earlier than this date, we set τ equal to 1981.13. Neglecting this obvious mean shift results in

$$\underset{(0.133)}{y_t = 0.180} + \underset{(0.017)}{0.978 y_{t-1}} + \hat{\varepsilon}_t, \tag{6.19}$$

for 142 effective observations, and clearly the estimate for ϕ_1 is close to 1. In fact, the Dickey-Fuller test gives the insignificant value of -1.321. The $F_{AC,1-1}$ and $F_{AC,1-13}$ diagnostics (see Chapter 3) do not indicate misspecification, but the JB test for normality with a value of 23.511 does, which could be seen as a sign that the data may suffer from outliers. Adding a level shift dummy variable for all observations after and including 1981.13, as well as single dummy variables for 1982.01, 1982.02 (to accommodate for outliers and possibly higher order AR effects) and additional lags for the same reason, and reducing the initial model by excluding all insignificant terms, we finally obtain

$$\underset{(0.314)}{y_t = 2.306} + \underset{(0.056)}{0.552 y_{t-1}} + \underset{(0.032)}{0.144 y_{t-13}} + \underset{(0.026)}{0.193 I_t[t \geq 1981.13]} + \hat{\varepsilon}_t. \tag{6.20}$$

The normality test now obtains a value of 4.911, which is not significant at the 5% level. This estimated model shows that the 60% increase in broadcasting time leads to only an 8.4% increase in advertising spending. Notice that (6.20) now also includes the y_{t-13} variables, which was not found necessary for (6.19).

To illustrate the analysis of a time series for which there may be AOs, Figure 6.13 shows quarterly observations on US manufacturers' new orders for non-defense capital goods over the period 1965.1–2010.4. The ADF regression, which includes a intercept and a trend, results in an ADS test value of -4.77. Hence y_t does not seem to have a stochastic trend at the 1% level. After removing the deterministic trend, the EACF of the residuals suggests that an AR(3) may be useful to describe the detrended log-orders series. Particularly large residuals are found in 1968, 1975, 1991 and 2009, see Figure 6.14 which shows the actual fit and residuals of the AR(3) model. In this illustration, we focus on the 1968 period. We observe a large positive residual in 1968.1 followed by a large negative residual for 1968.2. The residuals after these quarters are in the 95% confidence level. This pattern is consistent with the occurrence of an additive outlier in an autoregressive model with a large positive first-order AR coefficients and small higher-order AR coefficients (which is the case here). Note that a similar pattern occurs around 1991.2–1991.3.

We test for an possible AO or IO by running (6.15) (adapted to an AR(3) model) for all τ. The suprema of all values of $\hat{\lambda}_{AO}(\tau)$ and $\hat{\lambda}_{IO}(\tau)$ are equal to 6.15 and 4.38 respectively, both at $\tau = 1968.1$. Although both values exceed the critical value of 3.5, it is not immediately clear-cut given these values which type of outlier suits most adequately the aberrant observation of 1968.1. We therefore specificate the following

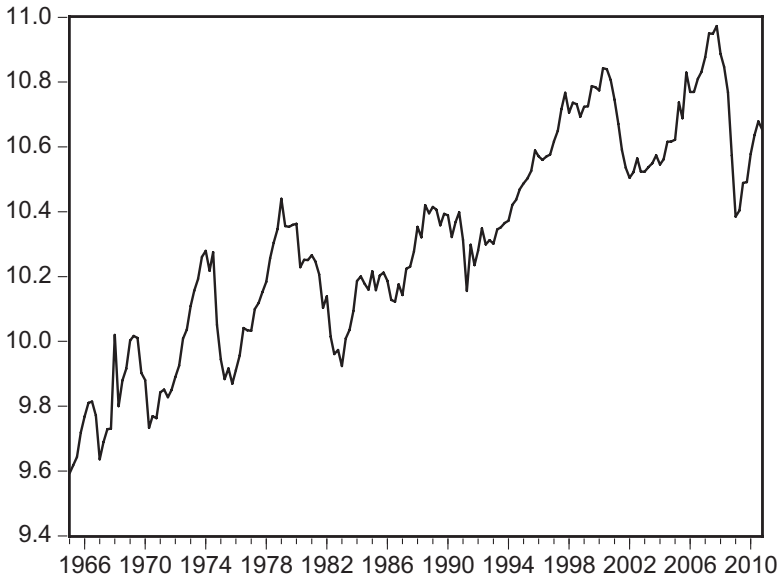


Figure 6.13: Quarterly (log) US manufacturers' new orders for non-defense capital goods.

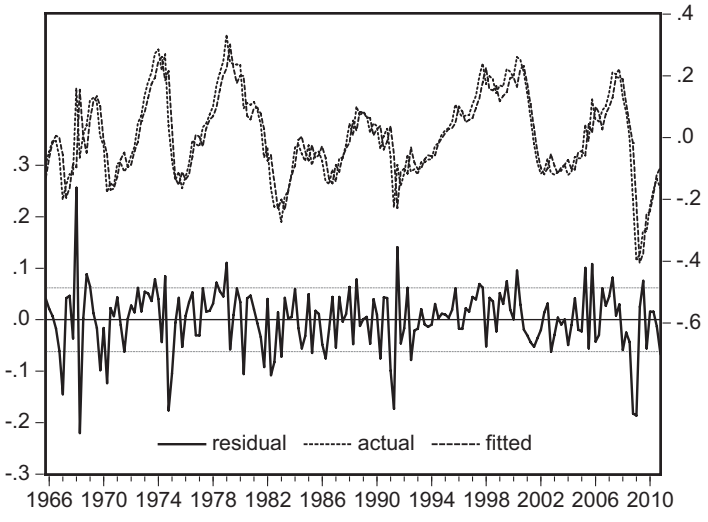


Figure 6.14: Actual, fitted en residuals from an AR(3) model applied to detrended quarterly (log) US manufacturers' new orders for non-defense capital goods.

AR(3) models, treating this observation as an AO and IO:

$$\phi_3 L(x_t - \mu - \delta_1 d_{t,1968.1}) = \varepsilon_t$$
$$\phi_3 L(x_t - \mu) = \delta_1^* d_{t,1968.1} + \varepsilon_t,$$

where $\phi_3 L = 1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3$ and $d_{t,1968.1}$ is a dummy variable that takes the value one the first quarter of 1968. After estimating the parameters, we inspect the implied residuals $\hat{\varepsilon}_t$. It seems that 1968.1 is an AO, as the corresponding residuals around 1968 are in the 95% confidence level, while there is still a peak at 1968.2 in the residuals of the AR(3) model treating 1968.1 as an IO.

For some of the examples above, it seems straightforward how to proceed. In many practical occasions, however, this is much more complicated. It may be quite difficult, as will be shown below, to distinguish a level shift model from a unit root model with an IO. Hence, it often seems wise to test for unit roots first using the methods described in the next section. If a unit root is found, we can impose it, and proceed with the techniques in this section. Notice however that overdifferencing a stationary AR(1) time series with a single IO may result in a series that seems to have an AO as the weight ζ of the IO will appear as $-\zeta$ in the next period. Additionally, inspection of the EACF may then suggest that one needs an MA type of model.

Exercise 6.8

When a time series seems to have many aberrant data, it can be that a univariate *linear* time series model such as an ARMA model does not yield a good description of the data. For example, nonlinear time series data can be characterized by regular regime switches (as will be shown in Chapter 8), and a linear model for these data may result in many large residuals. Furthermore, outliers may reflect the fact that a multivariate time series model (such as those discussed in Chapter 9) or an AR model with exogenous variables is more appropriate.

6.4 Outliers and unit root tests

In this last section of this chapter we deal with a practically very relevant issue, which is the interaction between features of time series when modeling data. Many interactions are possible, but the link between outliers and potential unit roots in the data is an interesting one.

In general, it is not easy to decide whether a time series has a unit root or not in case there are level shifts, sets of AOs, or changing deterministic trends, see Perron (2006) for a recent survey. For example, neglecting level shifts or breaking trends leads to spurious unit roots, see Perron (1989, 1990), and neglected AOs lead to a spurious

finding of stationarity, see Franses and Haldrup (1994). Given that knowledge of the presence of unit roots is important for forecasting and for multivariate modeling, it seems sensible to start any analysis with an examination of unit roots while allowing for possible aberrant observations. Again we assume that the break date is (approximately) known. If not, the rolling sample or recursive methods developed in, for example Banerjee and Stock (1992), or the minimum DF test approach in for example Zivot and Andrews (2002) and Perron and Vogelsang (1992) can be used. Alternatively, one may rely on the outlier robust tests for unit roots propagated in Lucas (1995, 1996), where alternative assumptions for the error process are considered as well as estimation methods alternative to OLS. In the first part of this section we focus on nonseasonal time series. Next, we briefly discuss seasonal time series.

Nonseasonal series, additive outliers

The results in Section 6.2, especially in (6.12), indicate that the estimate of ϕ_1 is downward biased in case of a neglected AO. Hence, even when the true parameter is 1, the $\hat{\phi}_1$ will be below 1 when there are neglected AOs. In other words, we would find a unit root less often in case the data are contaminated by one or more AOs. Hence, the ρ parameter in the Dickey-Fuller regression becomes large and negative. Depending on the number of observations and the variance of the ε_t process, this may lead to a significant $t(\hat{\rho})$-test value. In other words, the asymptotic distribution of the ADF (or DF) test becomes more skewed to the left when there are neglected additive outliers.

Franses and Haldrup (1994) recommend to follow the strategy in Chen and Liu (1993) to detect AOs, and to apply the DF test to an auxiliary regression which includes dummy variables for the detected AOs. It can be shown that the DF test based on this enlarged regression asymptotically follows the DF distribution tabulated in Table 4.1.

Nonseasonal series, level shifts

In contrast to AOs, (permanent) level shifts cause the DF test to reject the null hypothesis *not* often enough, that is, we find a spurious unit root when such a shift is not incorporated in the model for a stationary time series. Perron (1989, 1990) shows that the larger the shift is, that is, the larger is ζ in (6.10), the more $\hat{\phi}_1$ in an AR(1) model is biased towards unity. If we wish to allow for such a possible level shift, one should include it in the DF regression like

$$\Delta_1 y_t = \rho y_{t-1} + \omega I_t[t \geq \tau] + \lambda_1 I_t[t = \tau] + \lambda_2 I_t[t = \tau + 1] + \varepsilon_t, \quad (6.21)$$

where the dummy variables at τ and $\tau + 1$ ensure that there is a gradual shift, see Perron and Vogelsang (1992). When $\lambda = \tau/T$, i.e., λ measures the location of the level shift in the sample, Perron (1990) shows that the asymptotic distribution of $t(\rho)$ depends on λ only.

Table 6.1: Asymptotic critical values of Dickey-Fuller t-test in the presence of level shifts and breaking trends at a known date

	λ								
Size	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Level shift									
0.01	−3.67	−3.80	−3.88	−3.92	−3.90	−3.92	−3.88	−3.80	−3.67
0.05	−3.10	−3.23	−3.30	−3.35	−3.34	−3.35	−3.30	−3.23	−3.10
0.10	−2.78	−2.92	−2.99	−3.05	−3.04	−3.05	−2.99	−2.92	−2.78
Breaking trend									
0.01	−4.38	−4.65	−4.78	−4.81	−4.90	−4.88	−4.75	−4.70	−4.41
0.05	−3.75	−3.99	−4.17	−4.22	−4.24	−4.24	−4.18	−4.04	−3.80
0.10	−3.45	−3.66	−3.87	−3.95	−3.96	−3.95	−3.86	−3.69	−3.46

Note: $\lambda = \tau/T$, where τ is the observation for which there is a shift. The test regression for the level shift is given in (6.21) and the test regression for the breaking trend in (6.22).
Sources: Perron (1990) (Table 4) and Perron (1989) (Table VI.B), respectively.

In the first panel of Table 6.1, we give some critical values of this unit root test for various values of λ. Comparing these with those in Table 4.1, one observes that the critical values based on regression (6.21) have shifted to the left. Intuitively, this is because we include information in the regression model which favors the alternative hypothesis, and hence we should find, say, extra-large negative values for $t(\hat{\rho})$ to be able to reject the null hypothesis of a unit root.

A useful time series model for the first eleven years of radio advertising data is given in (6.20). The ADF value corresponding to this regression is −6.481. Comparing this value with the critical values in Table 6.1 shows that this series does not have a unit root. Note however that if one would have neglected the level shift in 1982.01, one would have decided that this series has a unit root, see (6.19).

Nonseasonal series, breaking trends

If additional to a level shift there is a changing trend, we may expect even more difficulty to detect stationarity using a standard unit root test. Hence, in that case it

seems sensible to test the null hypothesis of a unit root in a regression like

$$\Delta_1 y_t = \rho y_{t-1} + \mu + \nu t + \omega I_t[t \geq \tau]t + \lambda_l I_t[t = \tau] + \lambda_2 I_t[t \geq \tau] + \varepsilon_t. \tag{6.22}$$

Exercise 6.9

Perron (1989) derives the asymptotic distribution of $t(\hat{\rho})$, and again it turns out that this distribution only depends on $\lambda = \tau/T$. In the second panel of Table 6.1, we give some asymptotic critical values for the $t(\hat{\rho})$ test for regression models as (6.22). Given that this regression includes yet another set of regressors that favor the alternative hypothesis, it is no surprise that the critical values shift even further to the left as compared to the level shift model.

It should be mentioned here that the asymptotic critical values of the $t(\rho)$-test are different in case the location of the break λ is unknown. For example, for model (6.22), Zivot and Andrews (2002) show that the 5% critical value for the minimum value of $t(\rho)$, which is now calculated for all possible break points, is -5.08. For practical purposes where sometimes no precise information on the break date is available, one may use the following rule of thumb (for the 5% significance level). If a unit root test value is below -5.08, we can safely conclude that there is no unit root. If this value is above -3.75, there is a unit root. Any value in between can be viewed as corresponding to a possibly inconclusive region.

Seasonal time series and seasonal level shifts

Similar to the above arguments of the effects of level shifts on tests for unit roots in a nonseasonal time series, we may expect that level shifts in one or more seasons have an effect on tests for seasonal unit roots. As such seasonal unit roots amount to ever widening interval forecasts for out-of-sample forecasts as well, it is also important to examine seasonal unit roots while allowing for possible seasonal mean shifts. Simulation results in Paap *et al.* (1997) show that neglecting seasonal mean shifts leads to substantial forecasting errors. Franses and Vogelsang (1998) propose to enlarge the HEGY regression model in (5.39) by including dummy variables for the mean shifts. For the case of no additional lags, this model for quarterly data is

$$\begin{aligned}\Delta_4 y_t = &\sum_{s=1}^{4} \delta_s D_{s,t} + \sum_{s=1}^{4} \delta_s^* D_{s,t} I_t[t \geq \tau] + \sum_{j=1}^{4} \kappa_j I_t[t = \tau - 1 + j] \\ &+ \pi_1(1 + L + L^2 + L^3)y_{t-1} + \pi_2(-1 + L - L^2 + L^3)y_{t-1} \\ &+ (\pi_3 L + \pi_4)(1 - L^2)y_{t-1} + \varepsilon_t,\end{aligned} \tag{6.23}$$

Table 6.2: Asymptotic critical values of HEGY test statistics in the presence of seasonal level shifts at known break date

	λ								
Size	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
$t(\pi_2)$									
0.01	−3.68	−3.81	−3.87	−3.92	−3.94	−3.95	−3.90	−3.83	−3.67
0.05	−3.08	−3.22	−3.29	−3.34	−3.35	−3.35	−3.31	−3.22	−3.34
0.10	−2.77	−2.91	−2.99	−3.04	−3.05	−3.04	−3.00	−2.91	−3.77
$F(\pi_3, \pi_4)$									
0.01	6.37	7.02	7.52	7.80	7.93	7.87	7.56	7.01	6.35
0.05	7.47	8.20	8.74	9.07	9.16	9.11	8.81	8.25	7.50
0.10	9.88	10.80	11.38	11.74	11.68	11.72	11.35	10.84	9.86

Note: $\lambda = \tau/T$, where τ is the observation for which there is a shift.
Source: Franses and Vogelsang (1998).

where the four κ_j parameters in this regression are needed to allow for four possible mean shifts from τ onwards. Franses and Vogelsang (1998) show that the asymptotic distributions of the t-tests for π_1 and π_2 and the joint F-test for π_3 and π_4 only depend on λ. As the distribution of $t(\pi_1)$ is the same as that of $t(\rho)$. Table 6.2 only reports some critical values for the $t(\pi_2)$ and $F(\pi_3, \pi_4)$ test.

When these critical values are compared with those in Table 5.2 it is clear that the critical values of $t(\pi_2)$ shift to the left, while those of the $F(\pi_3, \pi_4)$ shift to the right. Notice that the values in Table 6.2 can also be used in case (6.23) contains a single nonbreaking deterministic trend variable as the regressors in the auxiliary regression models are orthogonal.

CONCLUSION

In this chapter we have assumed that a modeler is somehow interested in describing a few aberrant observations in a time series before generating out-of-sample forecasts. Usually this interest is motivated by prior knowledge of some specific events. However, if we do not have such information, the material in this chapter can still be useful as a

general check for model adequacy. When many observations are aberrant, we may want to reconsider the model, as it may be a sign of misspecification. Indeed, a multivariate model or a nonlinear model could have been better to describe and forecast the data more accurately. On the other hand, we can try to exploit the occurrence of sequences of outliers by modeling these clusters of aberrant data. The analysis of these clusters is the focus of the so-called ARCH model, which will be considered in the next chapter.

EXERCISES

6.1 Generate 50 observations from a standard normal distribution using Eviews. Change the first observation into 10. Make a histogram and compute the skewness and excess kurtosis. What do you observe? Now change the first observation into -10, and repeat the exercise. What do you observe now? Suppose you generate 5000 observations, and repeat both exercises. What do you observe and why?

6.2 Generate 200 observations from a AR(1) model with $\phi = 0.9$ and a standard normal error term using Eviews. Repeat this exercise after changing the error observation at time 100 into 20. Make graphs of the two series. What do you osberve? Repeat this exercise for $\phi = 0.99$ and then for $\phi = 1$. What do you observe now?

6.3 Show that this results is due to the fact that we can write $\frac{1}{1-\phi_1 L} = 1 + \phi_1 L + \phi_1^2 L^2 + \phi_1^3 L^3 + \cdots$

6.4 Find a way using Eviews to create 200 observations on an AR(1) process where $\phi = 0.9$ and $\delta = 0.6$, and $\tau = 100$ and $\zeta = 20$. Change δ into -0.6, what do you observe?

6.5 How do the results above for the AR(1) model carry over to the AR(p) model? Show for an AR(4) model that a neglected AO leads to five special residuals.

6.6 Create 200 observations for an AR(3) model with parameters $\phi_1 = 1.3$, $\phi_2 = -0.8$ and $\phi_3 = 0.3$. Introduce an AO at observation 100 of size 10σ. Create an autoregressive model for the contaminated series and save the residuals. Create a new variable which is the square of these residuals, and compute the autocorrelation function of this series. What do you observe?

6.7 Use the definitions of z_{1t} and z_{2t} to show that $\hat{\zeta}_1(\tau) = (e_\tau - \phi_1 e_{\tau+1})/(1 + \phi_1^2)$ and $\hat{\zeta}_2(\tau) = e_\tau$

6.8 Show this with simulated data.

6.9 Explain the inclusion of all regressors in this model.

7 Conditional Heteroskedasticity

The volatility of asset returns, which is viewed as a measure of uncertainty or risk, plays a crucial role in financial decision problems such as risk management, option pricing, and portfolio management. One of the most prominent features of asset return volatility is that it changes over time. In particular, periods of hectic movements in prices alternate with periods during which prices hardly change, see Section 2.4. This characteristic feature commonly is referred to as *volatility clustering*. In this chapter, we discuss time series models that can be used to describe this feature. In particular, we discuss (extensions of) the class of (Generalized) AutoRegressive Conditional Heteroskedasticity [(G)ARCH] models, introduced by Engle (1982) and Bollerslev (1986).

The outline of this chapter is as follows. In Section 7.1, we discuss representations of the basic GARCH model. Several extensions are briefly reviewed in Section 7.2. We emphasize which of the stylized facts of returns on financial assets can and cannot be captured by the various models. Several aspects that are relevant for implementing GARCH models in practice are discussed in some detail in Section 7.3. This includes testing for conditional heteroskedasticity, parameter estimation, and diagnostic checking. In Section 7.4 we focus on out-of-sample forecasting. Both the consequences for forecasting the conditional mean in the presence of ARCH, as well as forecasting volatility itself are discussed.

We should remark that the aim of this chapter is not to provide a complete account of the vast literature on GARCH models, but rather to provide an introduction to this area. For topics not covered in this chapter, the interested reader should consult one of the many surveys on GARCH models which have appeared in recent years. Bollerslev *et al.* (1992) provide a comprehensive overview of empirical applications of GARCH models to financial time series. Bollerslev *et al.* (1994) focus on the theoretical aspects of GARCH models. Gourieroux (1997) discusses in great detail how GARCH models can be incorporated in financial decision problems such as asset pricing and portfolio management. Additional reviews of GARCH and related models can be found in Diebold and Lopez (1995), Pagan (1996), Palm (1996), and Shephard (1996).

We first consider the time series of daily returns on the Dow Jones Index over the period January 1, 1990–December 31, 2005, to illustrate some of the stylized facts of

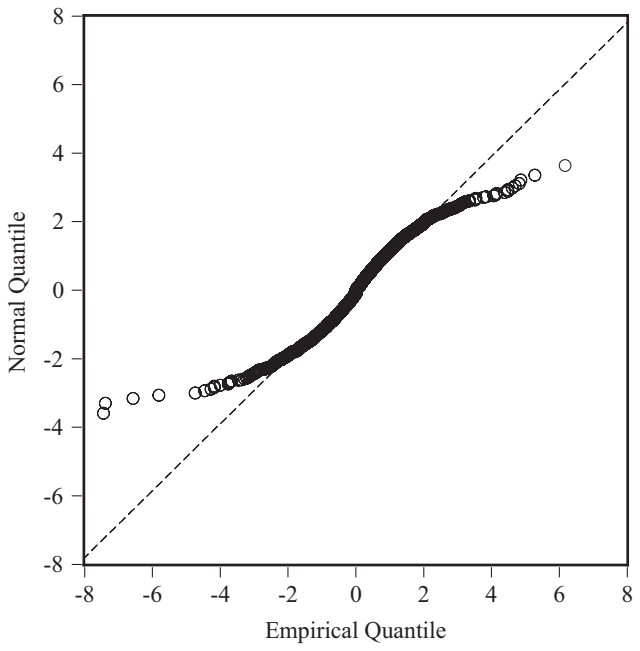


Figure 7.1: QQ-plot of daily returns on the Dow Jones index, January 1, 1990–December 31, 2005).

financial asset returns, and to motivate the use of models for time-varying conditional volatility discussed in this chapter. First, the daily stock index returns appear to be non-normally distributed. In particular, both positive and negative large returns occur more frequently than expected under normality, as indicated by the kurtosis being equal to 7.81. This also appears from the QQ-plot in Figure 7.1, which shows the quantiles of the empirical distribution against the quantiles of a normal distribution. If the series were normally distributed, these would form a 45-degree line. Clearly, in both the left and right tails of the distribution, the empirical quantiles are considerably larger. It also appears that small returns occur more frequently than under normality, which follows from the fact that around 0, the QQ-plot is steeper than the 45-degree line. In sum, the empirical distribution of daily returns is fat-tailed and peaked. The QQ-plot also suggests that large negative returns occur more often than large positive ones. This is confirmed by the skewness, which is equal to -0.22.

In the graph of the daily returns series in Figure 2.18, it appears that relatively volatile periods, characterized by large price changes and, hence, large returns, alternate with more tranquil periods in which prices are more stable and returns are, consequently,

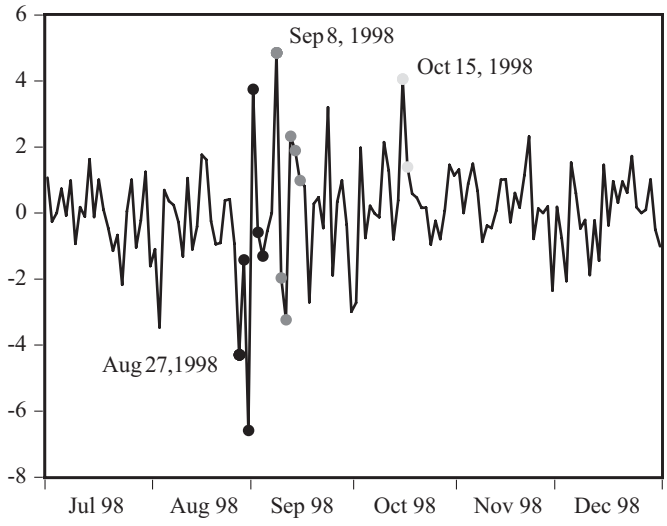


Figure 7.2: Daily returns on the Dow Jones index, July 1, 1998–December 31, 1998.

relatively small. In other words, large returns seem to occur in clusters, suggesting that the volatility of returns is varying over time.

As noted before in Chapter 2, volatility in financial markets is considered to be driven by the arrival of news. Hence, during periods with much news volatility is higher than during periods with less new information. To see how specific news facts affect volatility and lead to clusters of large returns, consider Figure 7.2, which shows the daily Dow Jones index returns for the second half of 1998. Three sets of daily observations are marked, corresponding to different news events. First, on August 27, 1998, the financial crisis in Russia reached a climax, with the Russian central bank cancelling foreign currency trading and the failure of the auction of Russian government bonds held during that day. The US stock market reacted with a return of -4.3%, and returns on subsequent days remained large due to the unrest this news caused. A large positive return of 5% is observed for Monday September 8, following a speech of Federal Reserve chairman Alan Greenspan on the Friday before in which he hinted at the possibility of lowering interest rates in the near future. Again, this led to a cluster of large returns lasting for several days, or, in other words, to an increase in volatility. Not all news events have such an extended effect on volatility. For example, on October 15, the Federal Reserve indeed decided to lower interest rates at an unscheduled meeting of its Federal Open Market Committee. Although this decision surprised the financial market and gave rise to a large return of $+4\%$, this was

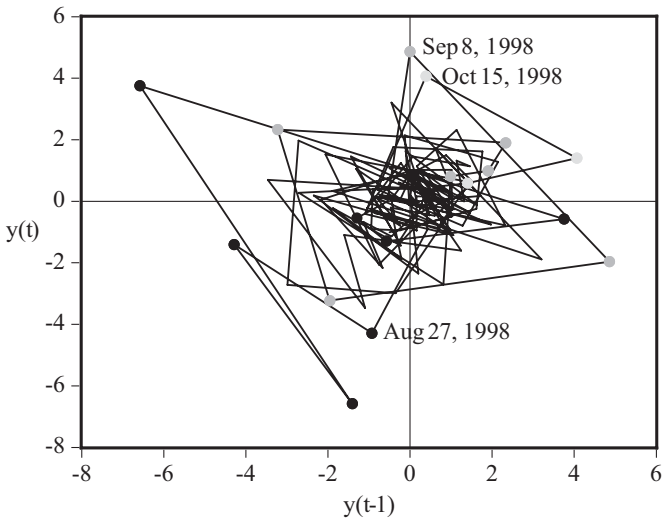


Figure 7.3: Daily returns on the Dow Jones index, July 1, 1998–December 31, 1998.

not followed by large returns during the following days. In general, it is believed that negative news has a larger impact on volatility than positive news.

The feature of volatility clustering also becomes apparent when inspecting connected scatterplots of the return of day t, denoted y_t, against the return of day $t-1$, as shown in Figure 7.3 for the second half of 1998.

A final feature of the daily Dow Jones returns series that is of interest here concerns the empirical autocorrelations, which are shown in Figure 7.4 for the returns, the squared returns and the absolute returns. The returns themselves have no significant autocorrelations at all at low orders. By contrast, for the absolute and squared returns, the autocorrelations start off at a moderate level between 0.15 and 0.20 but remain (significantly) positive for a substantial number of lags. Given that the absolute and squared returns are measures of volatility, this suggests that volatility is rather persistent.

7.1 Models for conditional heteroskedasticity

As discussed in Chapter 3, the main objective of time series modeling is to characterize the conditional distribution $f(y_t|\mathcal{Y}_{t-1})$, where $\mathcal{Y}_{t-1} \equiv \{y_1, y_2, \ldots, y_{t-1}\}$ denotes the

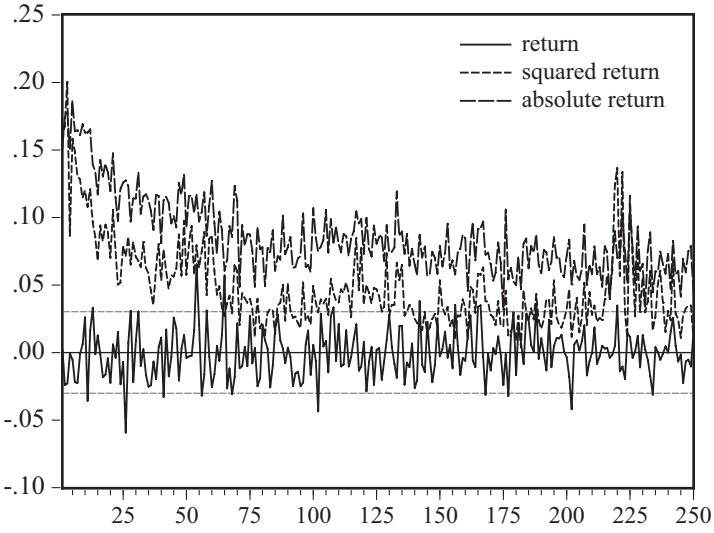


Figure 7.4: Empirical autocorrelation function of daily returns, squared returns, and absolute returns on the Dow Jones index, January 1, 1990–December 31, 2005. The dashed horizontal lines are bounds of the 95% asymptotic confidence interval.

set of all past time series observations. In particular, we often focus on the first two conditional moments of y_t, that is, the conditional mean $\mathsf{E}[y_t|\mathcal{Y}_{t-1}]$ and the conditional variance $\mathsf{V}[y_t|\mathcal{Y}_{t-1}]$. In previous chapters we have concentrated on different aspects of time series models for the conditional mean $\mathsf{E}[y_t|\mathcal{Y}_{t-1}]$, while simply assuming that the conditional variance $\mathsf{V}[y_t|\mathcal{Y}_{t-1}]$ is constant. Effectively this was done by imposing that the shock ε_t satisfies the white noise properties (3.1)–(3.3). In particular, ε_t was taken to be both unconditionally and conditionally homoskedastic, that is, $\mathsf{E}[\varepsilon_t^2] = \mathsf{E}[\varepsilon_t^2|\mathcal{Y}_{t-1}] = \sigma^2$ for all t. Here we relax part of this assumption and allow the *conditional variance* of ε_t to vary over time, that is, $\mathsf{E}[\varepsilon_t^2|\mathcal{Y}_{t-1}] = h_t$ for some non-negative function $h_t \equiv h_t(\mathcal{Y}_{t-1})$. Put differently, ε_t is *conditionally heteroskedastic*. A convenient way to express this in general is

$$\varepsilon_t = z_t\sqrt{h_t}, \tag{7.1}$$

where z_t is independent and identically distributed with zero mean and unit variance. In particular, we assume that z_t has a standard normal distribution throughout this chapter. Some remarks on this assumption are made at the end of this section.

From (7.1) and the properties of z_t it follows that the distribution of ε_t conditional upon the history $\mathcal{Y}_{t-1}$ is normal with mean zero and variance h_t. Also note that the

unconditional variance of ε_t is still assumed to be constant. Using the law of iterated expectations,

$$\sigma^2 \equiv \mathsf{E}[\varepsilon_t^2] = \mathsf{E}[\mathsf{E}[\varepsilon_t^2|\mathcal{Y}_{t-1}]] = \mathsf{E}[h_t]. \tag{7.2}$$

In other words, we assume that the unconditional expectation of h_t is constant.

To understand why allowing ε_t to be conditionally heteroskedastic is useful, note that the time series y_t can be written as

$$y_t = \mathsf{E}[y_t|\mathcal{Y}_{t-1}] + \varepsilon_t, \tag{7.3}$$

where the conditional mean $\mathsf{E}[y_t|\mathcal{Y}_{t-1}]$ may be given by an ARMA(p,q) model, for example. From (7.3) it follows that the conditional distribution $f(y_t|\mathcal{Y}_{t-1})$ is the same as the conditional distribution of ε_t, but with mean equal to $\mathsf{E}[y_t|\mathcal{Y}_{t-1}]$. The conditional variance of y_t is exactly the same as the conditional variance of ε_t though, as

$$\mathsf{V}[y_t|\mathcal{Y}_{t-1}] = \mathsf{E}[(y_t - \mathsf{E}[y_t|\mathcal{Y}_{t-1}])^2] = \mathsf{E}[\varepsilon_t^2|\mathcal{Y}_{t-1}] = h_t.$$

Thus, whenever we discuss the properties of the conditional variance of ε_t in the following, it should be understood that effectively we are talking about the conditional variance of the time series of interest y_t itself.

Obviously, we need to specify how the conditional variance of ε_t evolves over time in order to complete the model and make it useful for practical purposes. In the remainder of this section, we discuss various time series models for h_t. The properties of the resultant time series ε_t are used to see whether these models can capture (some of) the stylized facts of time series such as the daily Dow Jones index returns described above.

The ARCH model

Engle (1982) introduced the class of AutoRegressive Conditionally Heteroskedastic [ARCH] models to capture the volatility clustering of financial time series (even though the first empirical applications did not deal with high-frequency financial data). In the basic ARCH model, the conditional variance of the shock ε_t is a linear function of the squares of past shocks. For example, in the ARCH model of order 1, h_t is specified as

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2. \tag{7.4}$$

Obviously, the h_t should be non-negative, given that it represents a (conditional) variance. In order to guarantee that this is the case for the ARCH(1) model, the parameters in (7.4) have to satisfy the conditions $\omega > 0$ and $\alpha_1 \geq 0$. In case $\alpha_1 = 0$, h_t is constant and, hence, the series ε_t is conditionally homoskedastic.

It can be understood intuitively that the ARCH(1) model is able to describe volatility clustering by observing that (7.1) with (7.4) states that the conditional variance of ε_t

is an increasing function of the square of the shock that occurred in the previous time period. Therefore, if ε_{t-1} is large (in absolute value), ε_t is expected to be large (in absolute value) as well. In other words, large (small) shocks tend to be followed by large (small) shocks, of either sign.

An alternative way to see the same thing is to note that the ARCH(1) model can be rewritten as an AR(1) model for ε_t^2. Adding ε_t^2 to (7.4) and subtracting h_t from both sides gives

$$\varepsilon_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \nu_t, \tag{7.5}$$

where $\nu_t \equiv \varepsilon_t^2 - h_t = h_t(z_t^2 - 1)$. Notice that $\mathsf{E}[\nu_t|\mathcal{Y}_{t-1}] = 0$. Using the theory for AR models summarized in Chapter 3, it follows that (7.5) is covariance stationary if $\alpha_1 < 1$. In that case the unconditional mean of ε_t^2, or the unconditional variance of ε_t, can be obtained as

$$\sigma^2 \equiv \mathsf{E}[\varepsilon_t^2] = \frac{\omega}{1-\alpha_1}. \tag{7.6}$$

Furthermore, (7.5) can be rewritten as

$$\begin{aligned}\varepsilon_t^2 &= (1-\alpha_1)\frac{\omega}{1-\alpha_1} + \alpha_1 \varepsilon_{t-1}^2 + \nu_t \\ &= (1-\alpha_1)\sigma^2 + \alpha_1 \varepsilon_{t-1}^2 + \nu_t \\ &= \sigma^2 + \alpha_1(\varepsilon_{t-1}^2 - \sigma^2) + \nu_t. \end{aligned} \tag{7.7}$$

Assuming that $0 \leq \alpha_1 < 1$, (7.7) shows that if ε_{t-1}^2 is larger (smaller) than its unconditional expected value σ^2, ε_t^2 is expected to be larger (smaller) than σ^2 as well.

In the discussion of the time series of daily returns on the Dow Jones index, it was noted that volatility clustering results in positive empirical autocorrelations for the squared time series. The k-th order autocorrelation of ε_t^2 is defined as $\rho_k = \mathsf{E}[(\varepsilon_t^2 - \mathsf{E}(\varepsilon_t^2))(\varepsilon_{t-k}^2 - \mathsf{E}(\varepsilon_{t-k}^2)]/\mathsf{E}[\varepsilon_t^2 - \mathsf{E}(\varepsilon_t^2)]$ which is only well-defined if the unconditional fourth moment $\mathsf{E}[\varepsilon_t^4]$ exists and is constant. For the ARCH(1) model with normally distributed z_t, it follows from (7.1) and (7.4) that

$$\begin{aligned}\mathsf{E}[\varepsilon_t^4] &= \mathsf{E}[z_t^4 h_t^2] = 3\mathsf{E}[(\omega + \alpha_1 \varepsilon_{t-1}^2)^2] \\ &= 3\mathsf{E}[\omega^2 + 2\alpha_1 \varepsilon_{t-1}^2 + \alpha_1^2 \varepsilon_{t-1}^4] \\ &= 3\omega^2 + \frac{6\alpha_1 \omega^2}{(1-\alpha_1)} + 3\alpha_1^2 \mathsf{E}[\varepsilon_{t-1}^4] \\ &= \frac{3\omega^2(1+\alpha_1)}{(1-\alpha_1)} + 3\alpha_1^2 \mathsf{E}[\varepsilon_{t-1}^4]. \end{aligned}$$

Setting $\mathsf{E}[\varepsilon_t^4] = \mathsf{E}[\varepsilon_{t-1}^4]$, the unconditional fourth moment of ε_t is finite if $3\alpha_1^2 < 1$, in which case

$$\mathsf{E}[\varepsilon_t^4] = \frac{3\omega^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

Hence, it also follows that the kurtosis K_ε is equal to

$$\mathrm{K}_\varepsilon = \frac{\mathsf{E}[\varepsilon_t^4]}{\mathsf{E}[\varepsilon_t^2]^2} = \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2}. \tag{7.8}$$

Note that K_ε is always larger than the normal value of 3, implying that the ARCH(1) model can also capture excess kurtosis, which is another feature commonly observed in financial time series. It is useful to note that excess kurtosis is in fact a general property of models for conditional heteroskedasticity. From (7.1) it can be seen that the kurtosis of ε_t always exceeds the kurtosis of z_t, as

$$\mathsf{E}[\varepsilon_t^4] = \mathsf{E}[z_t^4]\mathsf{E}[h_t^2] \geq \mathsf{E}[z_t^4]\mathsf{E}[h_t]^2 = \mathsf{E}[z_t^4]\mathsf{E}[\varepsilon_t^2]^2, \tag{7.9}$$

which follows from Jensen's inequality.

When $3\alpha_1^2 < 1$, it follows from the AR(1) representation of the ARCH(1) model in (7.5) that the k-th order autocorrelation of ε_t^2 is equal to α_1^k. Given that α_1 is required to be positive, ρ_k is positive at all lags $k \geq 1$, corresponding with the empirical autocorrelations. Recall Figure 7.4, which shows that the first-order autocorrelation of the squared Dow Jones index returns is quite small, while the subsequent decay is very slow. It turns out that the ARCH(1) model cannot accommodate this pattern in the autocorrelation function. Matching the small first-order autocorrelation requires a small value of α_1, but this in turn would imply that the autocorrelations would become close to zero rather quickly. On the other hand, the slow decay of the EACF suggests a large value of α_1, but this would imply a large value of the first-order autocorrelation. In sum, it appears that the ARCH(1) model cannot describe the two characteristic features of the empirical autocorrelations of time series of squared returns simultaneously.

To remedy this deficiency, we may consider generalizations of the model. One possibility to allow for more persistent autocorrelations is including additional lagged squared shocks in the conditional variance function. The general ARCH(q) model is given by

$$h_t = \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_q\varepsilon_{t-q}^2. \tag{7.10}$$

To guarantee non-negativity of the conditional variance, it is required that $\omega > 0$ and $\alpha_i \geq 0$ for all $i = 1, \ldots, q$. The ARCH(q) model can be rewritten as an AR(q) model for ε_t^2 in exactly the same fashion as writing (7.4) as (7.5), that is,

$$\varepsilon_t^2 = \omega + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_q\varepsilon_{t-q}^2 + \nu_t. \tag{7.11}$$

It follows that the unconditional variance of ε_t is equal to

$$\sigma^2 = \frac{\omega}{1 - \alpha_1 - \cdots - \alpha_q}, \tag{7.12}$$

while the ARCH(q) model is covariance stationary if all roots of the lag polynomial $1 - \alpha_1 L - \cdots - \alpha_q L^q$ are outside the unit circle.

The GARCH model

To capture the dynamic patterns in conditional volatility adequately by means of an ARCH(q) model, q often needs to be taken quite large. It turns out that it can be quite cumbersome to estimate the parameters in such a model, because of the restrictions that need to be imposed to guarantee non-negativity and stationarity. To reduce the computational problems, it is common to impose some structure on the parameters in the ARCH(q) model, such as $\alpha_i = \alpha(q + 1 - i)/(q(q + 1)/2)$, $i = 1, \ldots, q$, which implies that the parameters of the lagged squared shocks decline linearly and sum to α, see Engle (1982, 1983). As an alternative solution, Bollerslev (1986) suggested to add lagged conditional variances to the ARCH model instead. For example, adding h_{t-1} to the ARCH(1) model (7.4) results in the Generalized ARCH [GARCH] model of order (1,1)

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}. \tag{7.13}$$

The parameters in this model should satisfy $\omega > 0$, $\alpha_1 > 0$ and $\beta_1 \geq 0$ to guarantee that $h_t \geq 0$, while α_1 must be strictly positive for β_1 to be identified, see also (7.16).

To see why the lagged conditional variance avoids the necessity of adding many lagged squared residual terms to the model, notice that (7.13) can be rewritten as

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1(\omega + \alpha_1 \varepsilon_{t-2}^2 + \beta_1 h_{t-2}), \tag{7.14}$$

or, by continuing the recursive substitution, as

$$h_t = \sum_{i=1}^{\infty} \beta_1^i \omega + \alpha_1 \sum_{i=1}^{\infty} \beta_1^{i-1} \varepsilon_{t-i}^2. \tag{7.15}$$

This shows that the GARCH(1,1) model corresponds to an ARCH(∞) model with a particular structure for the parameters of the lagged ε_t^2 terms.

Alternatively, the GARCH(1,1) model can be rewritten as an ARMA(1,1) model for ε_t^2 as

$$\varepsilon_t^2 = \omega + (\alpha_1 + \beta_1)\varepsilon_{t-1}^2 + \nu_t - \beta_1 \nu_{t-1}, \tag{7.16}$$

where again $\nu_t = \varepsilon_t^2 - h_t$. Using the theory for ARMA models discussed in Chapter 3, it follows that the GARCH(1,1) model is covariance stationary if and only if

$\alpha_1 + \beta_1 < 1$. In that case the unconditional mean of ε_t^2 or, equivalently, the unconditional variance of ε_t is equal to

$$\sigma^2 = \frac{\omega}{1 - \alpha_1 - \beta_1}. \tag{7.17}$$

Exercise 7.1

The ARMA(1,1) representation in (7.16) also makes clear why α_1 needs to be strictly positive for identification of β_1. If $\alpha_1 = 0$, the AR and MA polynomials both are equal to $1 - \beta_1 L$. Rewriting the ARMA(1,1) model for ε_t^2 as an MA(∞), these polynomials cancel out,

$$\varepsilon_t^2 = \frac{1 - \beta_1 L}{1 - \beta_1 L} \nu_t = \nu_t, \tag{7.18}$$

which shows that β_1 then is not identified.

As shown by Bollerslev (1986), the unconditional fourth moment of ε_t is finite if $(\alpha_1 + \beta_1)^2 + 2\alpha_1^2 < 1$. If in addition the z_t are assumed to be normally distributed, the kurtosis of ε_t is given by

$$\mathrm{K}_\varepsilon = \frac{3[1 - (\alpha_1 + \beta_1)^2]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2}, \tag{7.19}$$

which again is always larger than the normal value of 3. Notice that if $\beta_1 = 0$, (7.19) reduces to (7.8).

Again using the ARMA(1,1) representation in (7.16), the autocorrelations of ε_t^2 are obtained directly from (3.73). Substituting $\phi_1 = \alpha_1 + \beta_1$ and $\theta_1 = -\beta_1$, these can be expressed in the coefficients of the GARCH(1,1) model as

$$\rho_1 = \alpha_1 + \frac{\alpha_1^2 \beta_1}{1 - 2\alpha_1 \beta_1 - \beta_1^2}, \tag{7.20}$$

$$\rho_k = (\alpha_1 + \beta_1)^{k-1} \rho_1 \qquad \text{for } k = 2, 3, \ldots, \tag{7.21}$$

see Bollerslev (1988). Even though the autocorrelations still decline exponentially, the decay factor in this case is $\alpha_1 + \beta_1$. If this sum is close to 1, the autocorrelations will decrease only very gradually. If in addition α_1 is small, $\rho_1 \approx \alpha_1$. Hence, it seems that the GARCH(1,1) model can describe the two salient features of the EACF of ε_t^2, that is, a small first-order autocorrelation and a slow decay towards zero, simultaneously.

The GARCH(1,1) model may be generalized by including additional lagged squared ε_t's and lagged h_t's. The general GARCH(p,q) model is given by

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 h_{t-1} + \cdots + \beta_p h_{t-p}, \tag{7.22}$$

or

$$h_t = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)h_t, \tag{7.23}$$

where $\alpha(L) = \alpha_1 L + \cdots + \alpha_q L^q$ and $\beta(L) = \beta_1 L + \cdots + \beta_p L^p$. Assuming that all the roots of $1 - \beta(L)$ are outside the unit circle, the model can be rewritten as an infinite order ARCH model

$$h_t = \frac{\omega}{1 - \beta(1)} + \frac{\alpha(L)}{1 - \beta(L)}\varepsilon_t^2 \tag{7.24}$$

$$= \frac{\omega}{1 - \beta_1 - \cdots - \beta_p} + \sum_{i=1}^{\infty} \delta_i \varepsilon_{t-i}^2. \tag{7.25}$$

For non-negativity of the conditional variance it is required that all δ_i in (7.25) are non-negative. Nelson and Cao (1992) discuss the conditions this implies for the parameters α_i, $i = 1, \ldots, q$, and β_i, $i = 1 \ldots, q$, in the original model (7.22).

Alternatively, the GARCH(p,q) can be interpreted as an ARMA(m, p) model for ε_t^2 given by

$$\varepsilon_t^2 = \omega + \sum_{i=1}^{m} (\alpha_i + \beta_i)\varepsilon_{t-i}^2 - \sum_{i=1}^{p} \beta_i \nu_{t-i} + \nu_t, \tag{7.26}$$

where $m = \max(p, q)$, $\alpha_i \equiv 0$ for $i > q$ and $\beta_i \equiv 0$ for $i > p$. It follows that the GARCH(p,q) model is covariance stationary if all the roots of $1 - \alpha(L) - \beta(L)$ are outside the unit circle. We refer to He and Teräsvirta (1999) and Ling and McAleer (2002a) for additional results on the properties of the GARCH(p,q) model.

Even though the general GARCH(p,q) model might be of theoretical interest, the GARCH(1,1) model often appears adequate in practice, see also Bollerslev *et al.* (1992). Furthermore, many nonlinear extensions of the GARCH models, including the models be discussed in the next section, have only been considered for the GARCH(1,1) case.

Exercise 7.2

7.2 Various extensions

In this section we discuss various extensions of the basic GARCH(1,1) model with normally distributed standardized shocks z_t, as given by (7.1) with (7.13).

IGARCH

In applications of the GARCH(1,1) model (7.13) to high-frequency financial time series, it is often found that the estimates of α_1 and β_1 are such that their sum is close or equal to 1. Following Engle and Bollerslev (1986), the model that results when $\alpha_1 + \beta_1 = 1$ is commonly referred to as Integrated GARCH [IGARCH]. The reason for this is that the restriction $\alpha_1 + \beta_1 = 1$ implies a unit root in the ARMA(1,1) model for ε_t^2 given in (7.16), which then can be written as

$$(1 - L)\varepsilon_t^2 = \omega + \nu_t - \beta_1 \nu_{t-1}. \tag{7.27}$$

The analogy with a unit root in an ARMA model for the conditional mean of a time series is however rather subtle. For example, from (7.17) it is seen that the unconditional variance of ε_t is not finite in this case. Therefore, the IGARCH model is not covariance stationary. However, the IGARCH(1,1) model may still be strictly stationary, as shown by Nelson (1990). This can be illustrated by rewriting (7.13) as

$$\begin{aligned} h_t &= \omega + (\alpha_1 z_{t-1}^2 + \beta_1) h_{t-1} \\ &= \omega + (\alpha_1 z_{t-1}^2 + \beta_1)(\omega + (\alpha_1 z_{t-2}^2 + \beta_1) h_{t-2}) \\ &= \omega(1 + (\alpha_1 z_{t-1}^2 + \beta_1)) + (\alpha_1 z_{t-1}^2 + \beta_1)(\alpha_1 z_{t-2}^2 + \beta_1) h_{t-2}, \end{aligned}$$

and continuing the substitution for h_{t-i}, it follows that

$$h_t = \omega \left(1 + \sum_{i=1}^{t-1} \prod_{j=1}^{i} (\alpha_1 z_{t-j}^2 + \beta_1) \right) + \prod_{i=1}^{t} (\alpha_1 z_{t-i}^2 + \beta_1) h_0. \tag{7.28}$$

As shown by Nelson (1990), a necessary condition for strict stationarity of the GARCH(1,1) model is $\mathsf{E}[\ln(\alpha_1 z_{t-i}^2 + \beta_1)] < 0$. If this condition is satisfied, the impact of h_0 disappears asymptotically.

As expected, the autocorrelations of ε_t^2 for an IGARCH model are not defined properly. However, Ding and Granger (1996) show that the approximate autocorrelations are given by

$$\rho_k = \frac{1}{3}(1 + 2\alpha_1])(1 + 2\alpha_1^2)^{-k/2}. \tag{7.29}$$

Hence, the autocorrelations still decay exponentially. This is in sharp contrast with the autocorrelations for a random walk model, for which the autocorrelations are approximately equal to 1, see (3.57).

GARCH in mean

Many financial theories postulate a direct relationship between the return and risk of financial assets. For example, in the CAPM the excess return on a risky asset is

proportional to its non-diversifiable risk, which is measured by the covariance with the market portfolio. The GARCH in mean [GARCH-M] model introduced by Engle *et al.* (1987) was explicitly designed to capture such direct relationships between return and possibly time-varying risk (as measured by the conditional variance). This is established by including (a function) of the conditional variance h_t in the model for the conditional mean of the variable of interest y_t. For example, in the case of an AR(p) model, we may have

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \delta g(h_t) + \varepsilon_t, \tag{7.30}$$

where $g(h_t)$ is some function of the conditional variance of ε_t, h_t, which is assumed to follow a GARCH process. In most applications, $g(h_t)$ is taken to be the identity function or square root function, that is, $g(h_t) = h_t$ or $g(h_t) = \sqrt{h_t}$. The additional term $\delta g(h_t)$ in (7.30) often is interpreted as some sort of risk premium. As h_t varies over time, so does this risk premium.

To gain some intuition for the properties of y_t as implied by the GARCH-M model, consider (7.30) with $p = 0$ and $g(h_t) = h_t$ and assume that h_t follows an ARCH(1) process

$$y_t = \delta h_t + \varepsilon_t, \tag{7.31}$$

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2. \tag{7.32}$$

Substituting (7.32) in (7.31) and using the fact that $\mathsf{E}[\varepsilon_{t-1}^2] = \omega/(1-\alpha_1)$, see (7.6), it follows that the unconditional expectation of y_t is equal to

$$\mathsf{E}[y_t] = \delta\omega\left(1 + \frac{\alpha_1}{1-\alpha_1}\right).$$

Similarly, it can be shown that the unconditional variance of y_t is equal to

$$\mathsf{V}[y_t] = \frac{\omega}{1-\alpha_1} + \frac{(\delta\alpha_1)^2 2\omega^2}{(1-\alpha_1)^2(1-3\alpha_1^2)},$$

which is larger than the unconditional variance of y_t in the absence of the GARCH-M effect, as in that case $\sigma_y^2 = \frac{\omega}{1-\alpha_1}$. Another consequence of the presence of h_t as regressor in the conditional mean equation (7.31) is that y_t is serially correlated. As shown by Hong (1991),

$$\rho_1 = \frac{2\alpha_1^3\delta^2\omega}{2\alpha_1^2\delta^2\omega + (1-\alpha_1)(1-3\alpha_1^2)} \tag{7.33}$$

$$\rho_k = \alpha_1^{k-1}\rho_1 \qquad k = 2, 3, \ldots. \tag{7.34}$$

An overview of applications of GARCH-M models to stock returns, interest rates and exchange rates can be found in Bollerslev *et al.* (1992).

Nonlinear GARCH models and the news impact curve

As discussed in the previous section, for financial time series such as daily stock returns it appears to be the case that volatile periods often are initiated by a large negative shock, which suggests that positive and negative shocks may have an asymmetric impact on the conditional volatility of subsequent observations. This was recognized already by Black (1976), who suggested that a possible explanation for this finding might be the way firms are financed. When the value of (the stock of) a firm falls, the debt-to-equity ratio increases, which in turn leads to an increase in the volatility of the returns on equity. As the debt-to-equity ratio is also known as the leverage of the firm, this phenomenon commonly is referred to as the leverage effect.

The GARCH(1,1) model as given in (7.13) cannot capture such asymmetric effects of positive and negative shocks. As the conditional variance h_t only depends on the square of ε_{t-1}, positive and negative shocks of the same magnitude have the same effect on the conditional volatility, that is, the sign of the shock is not important. Over the years, many nonlinear extensions of the GARCH model have been developed to allow for different effects of positive and negative shocks or other types of asymmetries. Below, we discuss the two nonlinear GARCH models that have become most popular in practice. For a more complete overview, the interested reader is referred to Franses and van Dijk (2000).

Most nonlinear GARCH models are motivated by the desire to capture the different effects of positive and negative shocks on conditional volatility or other types of asymmetry. A natural question to ask then is whether all these models are indeed different from each other, or whether they are more or less similar. A convenient way to compare different GARCH models is by means of the so-called *News Impact Curve* [NIC], introduced by Pagan and Schwert (1990) and popularized by Engle and Ng (1993). The news impact curve measures how new information is incorporated into volatility. To be more precise, the NIC shows the relationship between the current shock or news ε_t and conditional volatility one period ahead h_{t+1}, holding constant all other past and current information. In the basic GARCH(1,1) model and nonlinear variants thereof, the only relevant information from the past is the current conditional variance h_t. Thus, the news impact curve for the GARCH(1,1) model (7.13) is given by

$$\mathrm{NIC}(\varepsilon_t | h_t = h) = \omega + \alpha_1 \varepsilon_t^2 + \beta_1 h = A + \alpha_1 \varepsilon_t^2, \tag{7.35}$$

where $A = \omega + \beta_1 h$. Hence, the news impact curve is a quadratic function centered on $\varepsilon_t = 0$. As the value of the lagged conditional variance h_t only affects the constant A in (7.35), it only shifts the NIC vertically, but does not change its basic shape. In practice, it is customary to take h_t equal to the unconditional variance σ^2.

Exponential GARCH

The earliest variant of the GARCH model which allows for asymmetric effects is the Exponential GARCH [EGARCH] model, introduced by Nelson (1991). The EGARCH(1,1) model is given by

$$\log(h_t) = \omega + \alpha_1 z_{t-1} + \gamma_1(|z_{t-1}| - \mathsf{E}[|z_{t-1}|]) + \beta_1 \log(h_{t-1}). \tag{7.36}$$

As the EGARCH model (7.36) describes the relation between past shocks and the *logarithm* of the conditional variance no restrictions on the parameters α_1, γ_1 and β_1 have to be imposed to ensure that h_t is non-negative. Using the properties of z_t – an independent and identically distributed variable with zero mean and unit variance – it follows that $g(z_t) \equiv \alpha_1 z_t + \gamma_1(|z_t| - \mathsf{E}[|z_t|])$ has mean zero and is uncorrelated. The function $g(z_t)$ is piecewise linear in z_t, as it can be rewritten as

$$g(z_t) = (\alpha_1 + \gamma_1) z_t I[z_t > 0] + (\alpha_1 - \gamma_1) z_t I[z_t < 0] - \gamma_1 \mathsf{E}[|z_t|].$$

Thus, negative shocks have an impact $\alpha_1 - \gamma_1$ on the log of the conditional variance, while for positive shocks the impact is $\alpha_1 + \gamma_1$. This property of the function $g(z_t)$ leads to an asymmetric news impact curve. In particular, the news impact curve for the EGARCH(1,1) model (7.36) is given by

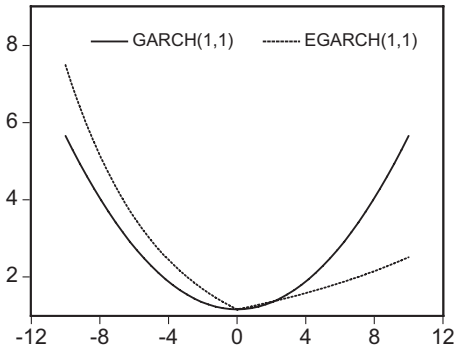
$$\mathrm{NIC}(\varepsilon_t | h_t = \sigma^2) = \begin{cases} A \exp\left(\frac{\alpha_1+\gamma_1}{\sigma}\varepsilon_t\right) & \text{for } \varepsilon_t > 0, \\ A \exp\left(\frac{\alpha_1-\gamma_1}{\sigma}\varepsilon_t\right) & \text{for } \varepsilon_t < 0, \end{cases} \tag{7.37}$$

with $A = \sigma^{2\beta_1} \exp(\omega - \gamma_1\sqrt{2/\pi})$.

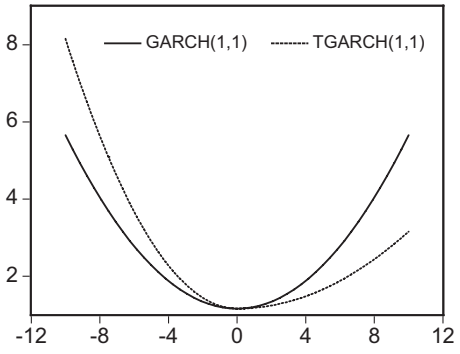
Typical news impact curves for the GARCH(1,1) and EGARCH(1,1) models are shown in panel (a) of Figure 7.5. The parameters in the models have been chosen such that the constants A in (7.37) and (7.35) are the same and, hence, the news impact curves are equal when $\varepsilon_t = 0$. The shape of the NIC of the EGARCH model is typical for parameterizations with $\alpha_1 < 0$, $0 \leq \gamma_1 < 1$ and $\gamma_1 + \beta_1 < 1$. For such parameter configurations, negative shocks have a larger effect on the conditional variance than positive shocks of the same size. For the range of ε_t for which the news impact curve is plotted in Figure 7.5, it also appears that negative shocks in the EGARCH model have a larger effect on the conditional variance than shocks in the GARCH model, while the reverse holds for positive shocks. However, as ε_t increases, the impact on h_t will eventually become larger in the EGARCH model for positive shocks as well, as the exponential function in (7.37) dominates the quadratic in (7.35) for large values of ε_t.

Threshold GARCH

The so-called Threshold GARCH [TGARCH] model introduced by Glosten *et al.* (1993) offers an alternative method to allow for asymmetric effects of positive and



(a)



(b)

Figure 7.5: News impact curves from the GARCH(1,1), EGARCH(1,1) and TGARCH(1,1) models, based on (7.35), (7.37) and (7.39).

negative shocks on volatility. The model is obtained from the GARCH(1,1) model (7.13) by assuming that the parameter of ε_{t-1}^2 depends on the sign of the shock, that is,

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \varepsilon_{t-1}^2 I[\varepsilon_{t-1} < 0] + \beta_1 h_{t-1}, \tag{7.38}$$

where as usual $I[\cdot]$ is an indicator function. The conditions for non-negativity of the conditional variance are $\omega > 0$, $\alpha_1 + \gamma_1/2 \geq 0$ and $\beta_1 > 0$. The condition for covariance stationarity is $\alpha_1 + \gamma_1/2 + \beta_1 < 1$, see Ling and McAleer (2002b). If this condition is satisfied, the unconditional variance of ε_t is $\sigma^2 = \omega/(1 - \alpha_1 - \gamma_1/2 - \beta_1)$. The news impact curve for the TGARCH model follows directly from (7.38) and is

equal to

$$\mathrm{NIC}(\varepsilon_t|h_t=\sigma^2) = A + \begin{cases} \alpha_1\varepsilon_t^2 & \text{if } \varepsilon_t > 0, \\ (\alpha_1+\gamma_1)\varepsilon_t^2 & \text{if } \varepsilon_t < 0. \end{cases} \tag{7.39}$$

where $A = \omega + \beta_1\sigma^2$. The news impact curve of the TGARCH model is a quadratic function centered on $\varepsilon_t = 0$, similar to the news impact curve of the basic GARCH model. However, the slopes of the TGARCH news impact curve are allowed to be different for positive and negative shocks. An example of the TGARCH news impact curve is shown in panel (b) of Figure 7.5, where we have set α_1 and γ_1 such that $\gamma_1 > 0$ while $\alpha_1 + \gamma_1/2$ is equal to the value of α_1 in the GARCH(1,1) model. In this case, the news impact curve is steeper than the GARCH news impact curve for negative news and less steep for positive news, which is the expected situation if negative shocks have a larger effect on the conditional variance than positive shocks. Comparing the NIC's of the EGARCH and TGARCH models as shown in panels (a) and (b) of Figure 7.5 shows that they are rather similar. Hence, the TGARCH model and the EGARCH model may be considered as alternative models for the same series.

Exercise 7.3–7.4

Alternative error distributions

So far we have assumed that the innovations z_t in (7.1) are normally distributed, which is equivalent to stating that the conditional distribution of ε_t is normal with mean zero and variance h_t. The *unconditional* distribution of a series ε_t for which the conditional variance follows a GARCH model is non-normal in this case. In particular, as the kurtosis of ε_t is larger than the normal value of 3, the unconditional distribution has fatter tails than the normal distribution. However, in many applications of the standard GARCH(1,1) model (7.13) to high-frequency financial time series it is found that the implied kurtosis of ε_t given in (7.19) is much smaller than the kurtosis of the observed time series.

The kurtosis of ε_t is an increasing function of the kurtosis of z_t, see (7.9), and, hence, K_ε can be increased by assuming a leptokurtic distribution for z_t. Following Bollerslev (1987), a popular choice has become the standardized Student t distribution with η degrees of freedom, that is,

$$f(z_t) = \frac{\Gamma((\eta+1)/2)}{\sqrt{\pi(\eta-2)}\Gamma(\eta/2)}\left(1+\frac{z_t^2}{\eta-2}\right)^{-(\eta+1)/2}, \tag{7.40}$$

where $\Gamma(\cdot)$ is the Gamma function. The Student t distribution is symmetric around zero (and thus $\mathsf{E}[z_t] = 0$), while it converges to the normal distribution as the number of degrees of freedom η becomes larger. A further characteristic of the Student t

distribution is that only moments up to order η exist. Hence, for $\eta > 4$, the fourth moment of z_t exists and is equal to $3(\eta - 2)/(\eta - 4)$. As this is larger than the normal value of 3, the kurtosis of ε_t will also be larger than in case z_t followed a normal distribution. The number of degrees of freedom of the Student t distribution need not be specified in advance. Rather, η can be treated as a parameter and can be estimated along with the other parameters in the model, as discussed in the next section.

7.3 Specification, estimation and evaluation

In this section we discuss some aspects that are relevant for implementing GARCH models in practice. This includes testing for conditional heteroskedasticity, parameter estimation, and diagnostic checking.

Testing for conditional heteroskedasticity

Even though it appears obvious from summary statistics and graphs such as Figure 2.18 that the conditional volatility of high-frequency financial time series changes over time, for other time series this may be less clear. In that case, it is useful to test for conditional heteroskedasticity prior to actually estimating a GARCH model. Engle (1982) developed such a test in the context of ARCH models based on the Lagrange Multiplier [LM] principle. The conditional variance h_t in the ARCH(q) model in (7.10) is constant if the parameters corresponding to the lagged squared shocks ε_{t-i}^2, $i = 1, \ldots, q$, are equal to zero. Therefore, the null hypothesis of conditional homoskedasticity can be formulated as H_0: $\alpha_1 = \cdots = \alpha_q$. The corresponding LM test can be implemented using the AR(q) representation in (7.11), and is computed as TR^2, where T is the sample size and the R^2 is obtained from the regression

$$\hat{\varepsilon}_t^2 = \omega + \alpha_1 \hat{\varepsilon}_{t-1}^2 + \cdots + \alpha_q \hat{\varepsilon}_{t-q}^2 + u_t, \tag{7.41}$$

where the residuals $\hat{\varepsilon}_t$ are obtained by estimating the model for the conditional mean of the observed time series y_t under the null hypothesis. The LM test statistic has an asymptotic $\chi^2(q)$ distribution under the null hypothesis. Note the resemblance of (7.41) with the regression that is used for constructing the LM test for autocorrelation in $\hat{\varepsilon}_t$ given in (3.100). In fact, the test for ARCH can be interpreted as a test for autocorrelation in the squared residuals. Lee (1991) shows that the LM test against this GARCH(p, q) alternative is the same as the LM test against the alternative of ARCH(q) errors.

The small sample properties of the LM test for linear (G)ARCH have been investigated quite extensively. In particular, it has been found that rejection of the null hypothesis of homoskedasticity might be due to other types of model misspecification, such as residual autocorrelation, outliers, nonlinearity, and omitted variables in the

model for the conditional mean, see Lumsdaine and Ng (1999) and van Dijk *et al.* (1999) for detailed discussion. In other words, if we find that for a given time series y_t the null hypothesis of conditional homoskedasticity should be rejected, this does not necessarily imply that the conditional variance is indeed changing over time. The LM test for conditional heteroskedasticity may therefore also be used as a general test for misspecification of an ARMA model for some time series.

Estimation

The parameters in GARCH models can be estimated by means of maximum likelihood [ML], together with the parameters in the model for the conditional mean $\mathsf{E}[y_t|\mathcal{Y}_{t-1}]$. We discuss this method in some detail here for the AR(p) model,

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + y_{t-p} + \varepsilon_t = G(x_t;\xi) + \varepsilon_t, \tag{7.42}$$

where $x_t = (1, y_{t-1}, \ldots, y_{t-p})'$, $\xi = (\alpha, \phi_1, \ldots, \phi_p)'$, and where $\varepsilon_t = z_t\sqrt{h_t}$ with z_t independent and identically distributed with zero mean and unit variance. The conditional variance h_t of ε_t is assumed to follow a possibly nonlinear GARCH model with parameters ψ. For example, in case a TGARCH(1,1) model (7.38) is specified for h_t, $\psi = (\omega, \alpha_1, \gamma_1, \beta_1)'$. The parameters in the models for the conditional mean and conditional variance are collected in the vector $\theta \equiv (\xi', \psi')'$. The true parameter values are denoted $\theta_0 = (\xi_0', \psi_0')'$. The conditional log likelihood for the t-th observation is equal to

$$l_t(\theta) = \ln f(\varepsilon_t/\sqrt{h_t}) - \ln\sqrt{h_t}, \tag{7.43}$$

where $f(\cdot)$ denotes the density of the shocks z_t. For example, if z_t is assumed to be normally distributed,

$$l_t(\theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln h_t - \frac{\varepsilon_t^2}{2h_t}. \tag{7.44}$$

Exercise 7.5

The maximum likelihood estimate [ML] for θ, which we denote as $\hat{\theta}_{\mathrm{ML}}$ is found by maximizing the log likelihood function $\mathcal{L}(\theta)$ for the full sample, which is the sum of the conditional log likelihoods as given in (7.43), that is, $\mathcal{L}(\theta) = \sum_{t=1}^{T} l_t(\theta)$. The MLE solves the first order condition

$$\sum_{t=1}^{T} \frac{\partial l_t(\theta)}{\partial \theta} = 0. \tag{7.45}$$

The vector of derivatives of the log likelihood with respect to the parameters is usually referred to as the score $s_t(\theta) \equiv \partial l_t(\theta)/\partial\theta$. The score can be decomposed as $s_t(\theta) = (\partial l_t(\theta)/\partial\xi', \partial l_t(\theta)/\partial\psi')$, where

$$\frac{\partial l_t(\theta)}{\partial \xi} = \frac{\varepsilon_t}{h_t}\frac{\partial G(x_t;\xi)}{\partial \xi} + \frac{1}{2h_t}\left(\frac{\varepsilon_t^2}{h_t} - 1\right)\frac{\partial h_t}{\partial \xi}, \tag{7.46}$$

$$\frac{\partial l_t(\theta)}{\partial \psi} = \frac{1}{2h_t}\left(\frac{\varepsilon_t^2}{h_t} - 1\right)\frac{\partial h_t}{\partial \psi}. \tag{7.47}$$

The second term on the right-hand side of (7.46) arises because the conditional variance h_t in general depends on ε_{t-1}, and thus on the parameters in the conditional mean for y_t, as $\varepsilon_{t-1} = y_{t-1} - G(x_{t-1};\xi)$.

As the first order conditions in (7.45) are nonlinear in the parameters, an iterative optimization procedure has to be used to obtain the MLE $\hat{\theta}_{\mathrm{ML}}$. If the conditional distribution $f(\cdot)$ is correctly specified, the resulting estimates are consistent and asymptotically normal. The asymptotic covariance matrix of $\sqrt{T}(\hat{\theta}_{\mathrm{ML}} - \theta_0)$ then is equal to A_0^{-1}, the inverse of the information matrix evaluated at the true parameter vector θ_0,

$$A_0 = -\frac{1}{T}\sum_{t=1}^{T}\mathsf{E}\left[\frac{\partial^2 l_t(\theta_0)}{\partial\theta\partial\theta'}\right] = \frac{1}{T}\sum_{t=1}^{T}\mathsf{E}[H_t(\theta_0)]. \tag{7.48}$$

The negative of the matrix of second-order partial derivatives of the log likelihood with respect to the parameters, $H_t(\theta) \equiv -\partial^2 l_t(\theta)/\partial\theta\partial\theta'$, is called the Hessian. The matrix A_0 can be consistently estimated by its sample analogue

$$A_T(\hat{\theta}_{\mathrm{ML}}) = -\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial^2 l_t(\hat{\theta}_{\mathrm{ML}})}{\partial\theta\partial\theta'}\right). \tag{7.49}$$

As argued in Section 7.2, conditional normality of ε_t often is not a very realistic assumption for high-frequency financial time series, as the resulting model fails to capture the kurtosis in the data. Instead, it is sometimes assumed that z_t is drawn from a (standardized) Student t distribution given in (7.40) or some other distribution. The parameters in the GARCH models then can be estimated by maximizing the log likelihood corresponding with this particular distribution. As we can never be sure that the specified distribution of z_t is the correct one, an alternative approach is to ignore the problem and base the likelihood on the normal distribution as in (7.44). This method usually is referred to as quasi maximum likelihood [QML] estimation. In general, the resulting estimates still are consistent and asymptotically normal, provided that the models for the conditional mean and conditional variance are correctly specified, see Bollerslev and Wooldridge (1992), Lee and Hansen (1994) and Lumsdaine (1996), among others.

Interestingly, consistency and asymptotic normality of the QML estimates do not require that the parameters in the GARCH(1,1) model satisfy the covariance stationarity condition $\alpha_1 + \beta_1 < 1$, but they continue to hold for the IGARCH(1,1) model. This is another difference with unit root models for the conditional mean. Recall from Chapter 4 that the properties of the estimates of, for example, autoregressive parameters change dramatically in case the model contains a unit root.

As the true distribution of z_t is not assumed to be the same as the normal distribution which is used to construct the likelihood function, the standard errors of the parameters have to be adjusted accordingly. In particular, the asymptotic covariance matrix of $\sqrt{T}(\hat{\theta} - \theta_0)$ is equal to $A_0^{-1} B_0 A_0^{-1}$, where A_0 is the information matrix (7.48) and B_0 is the expected value of the outer-product of the gradient matrix,

$$B_0 = \frac{1}{T} \sum_{t=1}^{T} \mathsf{E}\left[\frac{\partial l_t(\theta_0)}{\partial \theta} \frac{\partial l_t(\theta_0)}{\partial \theta'} \right] = \frac{1}{T} \sum_{t=1}^{T} \mathsf{E}[s_t(\theta_0) s_t(\theta_0)']. \tag{7.50}$$

The asymptotic covariance matrix can be estimated consistently by using the sample analogues for both A_0, as given in (7.49), and B_0, given by

$$B_T(\hat{\theta}_{\mathrm{ML}}) = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\partial l_t(\hat{\theta}_{\mathrm{ML}})}{\partial \theta} \frac{\partial l_t(\hat{\theta}_{\mathrm{ML}})}{\partial \theta'} \right) = \frac{1}{T} \sum_{t=1}^{T} s_t(\hat{\theta}_{\mathrm{ML}}) s_t(\hat{\theta}_{\mathrm{ML}})'. \tag{7.51}$$

The finite sample properties of the quasi maximum likelihood estimates for GARCH(1,1) models are considered in Engle and González-Rivera (1991) and Bollerslev and Wooldridge (1992). It appears that as long as the distribution of z_t is symmetric, QML is reasonably accurate and close to the estimates obtained from exact ML methods, while for skewed distributions this is no longer the case. Lumsdaine (1995) investigates the finite sample properties of the ML method in case the series follow an IGARCH model and she concludes that this method is quite accurate.

Given that QML estimation of the parameters in the GARCH model requires an iterative optimization procedure, the results may depend on the specific implementation. McCullough and Renfro (1999) and Brooks *et al.* (2001) examine several commercially available software packages for estimating GARCH models, and report substantial differences in the QML results across packages. There are a few aspects of the estimation problem which, when given proper attention, may avoid erroneous results.

First, the iterative optimization procedures that can be used to estimate the parameters typically require the first- and second order derivatives of the log-likelihood with respect to θ, that is, the score $s_t(\theta)$ and Hessian matrix $H_t(\theta)$ defined above. For

example, the iterations in the well-known Newton-Raphson method take the form

$$\hat{\theta}^{(m)} = \hat{\theta}^{(m-1)} - \lambda \left(\sum_{t=1}^{T} H_t(\hat{\theta}^{(m-1)}) \right)^{-1} \sum_{t=1}^{T} s_t(\hat{\theta}^{(m-1)}), \tag{7.52}$$

where $\hat{\theta}^{(m)}$ is the estimate of the parameter vector obtained in the m-th iteration and the scalar λ denotes a step size. In the algorithm of Berndt *et al.* (1974) [BHHH], which is by far the most popular method to estimate GARCH models, the Hessian $H_t(\hat{\theta}^{(m-1)})$ in (7.52) is replaced by the outer product of the gradient matrix $B_T(\hat{\theta}^{(m-1)})$ obtained from (7.51). It is common to use numerical approximations to these quantities, as the analytical derivatives are fairly complex and contain recursions which are thought to be too cumbersome to compute. However, Fiorentini *et al.* (1996) show that this is not the case and suggest that it might be advantageous to use analytic derivatives. In general, convergence of the optimization algorithm requires much less iterations, whereas the standard errors of the parameter estimates are far more accurate.

Second, the numerical optimization procedures require starting values $\hat{\theta}^{(0)}$ to initialize the iterations. It is worthwhile to specify starting values that already are close to the ML estimates, as this will lead to faster convergence of the numerical algorithm and a smaller probability of ending up in a local maximum of the likelihood function. The proposal of Kristensen and Linton (2006) is very useful in this respect. Kristensen and Linton (2006) adapt the non-iterative estimation method for parameters in ARMA models discussed in Section 3.3 to obtain a closed-form estimator for the GARCH(1,1) model. Recall that the GARCH(1,1) model can be written as an ARMA(1,1) model for ε_t^2, that is,

$$\varepsilon_t^2 = \omega + \phi_1 \varepsilon_{t-1}^2 + \nu_t + \theta_1 \nu_{t-1}, \tag{7.53}$$

where $\phi_1 = \alpha_1 + \beta_1$, $\theta_1 = -\beta_1$, and $\nu_t = \varepsilon_t^2 - h_t$. Estimates of ϕ_1 and θ_1 can be obtained using (3.92)–(3.94), using the first- and second-order empirical autocorrelations of $\hat{\varepsilon}_t^2$, where the residuals $\hat{\varepsilon}_t$ are obtained by estimating the model for the conditional mean of the observed time series y_t assuming conditional homoskedasticity. Estimates of α_1 and β_1 are then given by

$$\hat{\beta}_1 = -\hat{\theta}_1 \quad \text{and} \quad \hat{\alpha}_1 = \hat{\phi}_1 + \hat{\theta}_1.$$

Finally, an estimate of the intercept ω in the GARCH(1,1) model can be obtained by using the fact that the unconditional variance of ε_t^2 is equal to $\sigma^2 = \omega/(1 - \alpha_1 - \beta_1)$. Hence, we have $\hat{\omega} = \hat{\sigma}^2(1 - \hat{\phi}_1)$, where $\hat{\sigma}^2$ denotes the sample variance of $\hat{\varepsilon}_t$.

Diagnostic checking of GARCH models

Just as it is good practice to check the adequacy of an ARMA model for the conditional mean of a time series by computing a number of misspecification tests, such diagnostic checking should also be routinely done for models for the conditional variance. Several tests might be used for this purpose.

One of the assumptions made in GARCH models is that the innovations $z_t = \varepsilon_t h_t^{-1/2}$ are independent and identically distributed. Hence, if the model is correctly specified, the standardized residuals $\hat{z}_t = \hat{\varepsilon}_t \hat{h}_t^{-1/2}$ should possess the classical properties of well-behaved regression errors, such as constant variance, lack of autocorrelation, normality, and so on. Standard test statistics as discussed in Section 3.3 can be used to determine whether this is the case or not.

Of particular interest is to check whether the estimated GARCH model fully captures the conditional heteroskedasticity in the time series y_t. If this is the case, the standardized residuals $\hat{z}_t$ have constant conditional variance and, consequently, $\hat{z}_t^2$ does not have significant autocorrelations. Li and Mak (1994) and Lundbergh and Teräsvirta (2002) develop statistics that can be used to test for remaining ARCH effects in the standardized residuals. For example, the LM test for remaining ARCH(m) in $\hat{z}_t$ proposed by Lundbergh and Teräsvirta (2002) can be computed as TR^2, where R^2 is obtained from the auxiliary regression

$$\hat{z}_t^2 = \phi_0 + \phi_1 \hat{z}_{t-1}^2 + \cdots + \phi_m \hat{z}_{t-m}^2 + \lambda' \hat{x}_t + u_t, \tag{7.54}$$

where the vector $\hat{x}_t$ consists of the partial derivatives of the conditional variance h_t with respect to the parameters in the original GARCH model, evaluated under the null hypothesis, that is, $\hat{x}_t \equiv \hat{h}_t^{-1} \partial \hat{h}_t / \partial \theta$. For example, in case of a GARCH(1,1) model

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}, \tag{7.55}$$

it follows that

$$\frac{\partial h_t}{\partial \theta'} = \left(1, \varepsilon_{t-1}^2, h_{t-1}\right) + \beta_1 \frac{\partial h_{t-1}}{\partial \theta'}. \tag{7.56}$$

As the pre-sample conditional variance h_0 is usually computed as the sample average of the squared residuals, $h_t = 1/T \sum_{t=1}^{T} \varepsilon_t^2$, h_0 does not depend on θ, and $\partial h_0 / \partial \theta = 0$. This allows (7.56) to be computed recursively. Alternatively, the partial derivatives can be obtained by recursive substitution as

$$\hat{x}_t' = \left(\frac{\sum_{i=1}^{t-1} \hat{\beta}^{i-1}}{\hat{h}_t}, \frac{\sum_{i=1}^{t-1} \hat{\beta}^{i-1} \hat{\varepsilon}_{t-i}^2}{\hat{h}_t}, \frac{\sum_{i=1}^{t-1} \hat{\beta}^{i-1} \hat{h}_{t-i}}{\hat{h}_t} \right). \tag{7.57}$$

The test statistic based on (7.54), which tests the null hypothesis $H_0 : \phi_1 = \cdots = \phi_m = 0$ is asymptotically χ^2 distributed with m degrees of freedom.

The statistic discussed above tests for correlation in the squared standardized residuals. This is closely related to the LM statistics discussed by Bollerslev (1986), which can be used to test a GARCH(p,q) specification against either a GARCH($p+r$,q) or GARCH(p,$q+s$) alternative. The test statistics are given by T times the R^2 from the auxiliary regression (7.54), with the lagged squared standardized residuals $\hat{z}^2_{t-i}$, $i = 1, \ldots, m$ replaced by $\hat{\varepsilon}^2_{t-q-1}, \ldots, \varepsilon^2_{t-q-r}$ or $\hat{h}_{t-p-1}, \ldots, \hat{h}_{t-p-s}$, respectively.

With respect to the specification of nonlinear GARCH models, such as the EGARCH and TGARCH models discussed in the previous section, it seems reasonable to start with specifying and estimating a linear GARCH model. We may then move on to a nonlinear GARCH model only if certain misspecification tests suggest that symmetry of the conditional variance function is an untenable assumption.

One possible method to test a linear GARCH specification against nonlinear alternatives is by means of the so-called Sign Bias, Negative Size Bias and Positive Size Bias tests of Engle and Ng (1993). Let S^-_{t-1} denote a dummy variable which takes the value 1 when $\hat{z}_{t-1}$ is negative and zero otherwise. The tests examine whether the squared standardized residual $\hat{z}^2_t$ can be predicted by S^-_{t-1}, $S^-_{t-1}\hat{\varepsilon}_{t-1}$, and/or $S^+_{t-1}\hat{\varepsilon}_{t-1}$, where $S^+_{t-1} \equiv 1 - S^-_{t-1}$. The test statistics are computed as the t-ratio of the parameter δ_1 in the regression

$$\hat{z}^2_t = \delta_0 + \delta_1 \hat{w}_{t-1} + \lambda' \hat{x}_t + \xi_t, \tag{7.58}$$

where $\hat{w}_{t-1}$ is one of the three measures of asymmetry defined above, $\hat{x}_t$ is the vector of partial derivatives $\hat{x}_t \equiv \hat{h}_t^{-1} \partial \hat{h}_t / \partial \theta$, and ξ_t the residual.

In case $\hat{w}_t = S^-_{t-1}$ in (7.58), the test is called the Sign Bias [SB] test, as it simply tests whether the magnitude of the square of z_t depends on the sign of the lagged shock ε_{t-1}. In case $\hat{w}_t = S^-_{t-1}\hat{\varepsilon}_{t-1}$ or $\hat{w}_t = S^+_{t-1}\hat{\varepsilon}_{t-1}$, the tests are called the Negative Size Bias [NSB] and Positive Size Bias [PSB] tests, respectively. These tests examine whether the effect of negative or positive shocks on the conditional variance also depends on their size. As the SB, NSB and PSB statistics are t-ratios, they follow a standard normal distribution asymptotically.

The tests can also be conducted jointly, by estimating the regression

$$\hat{z}^2_t = \delta_0 + \delta_1 S^-_{t-1} + \delta_2 S^-_{t-1}\hat{\varepsilon}_{t-1} + \delta_3 S^+_{t-1}\hat{\varepsilon}_{t-1} + \lambda' \hat{x}_t + \xi_t. \tag{7.59}$$

The null hypothesis $H_0 : \delta_1 = \delta_2 = \delta_3 = 0$ can be evaluated by computing n times the R^2 from this regression. The resultant test statistic has an asymptotic χ^2 distribution with 3 degrees of freedom.

The out-of-sample forecasting ability of various GARCH models is an alternative approach to judge the adequacy of different models, and an alternative approach to model selection. Obviously, if a volatility model is to be of any use to practitioners in financial markets, it should be capable of generating accurate predictions of uture volatility. Volatility forecasting is discussed in detail in the next section.

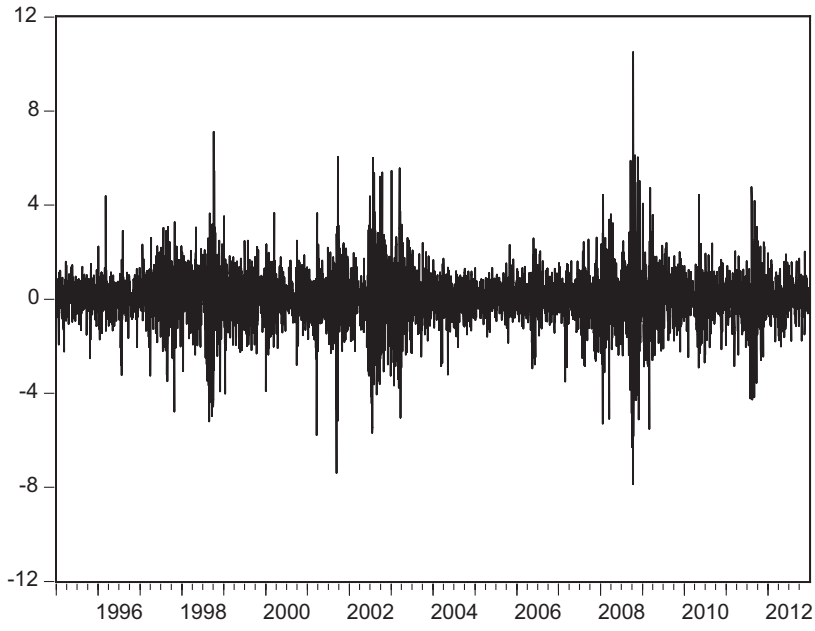


Figure 7.6: Daily MSCI Switzerland (SUI) returns from January 1, 1995–December 31, 2012.

Illustration

Consider the daily MSCI returns from Switzerland, as depicted in Figure 7.6, from January 1, 1995–December 31, 2012. We focus on the period 1995–2001 and leave the remaining years for out-of-sample forecasting. A QQ-plot would lead to the same conclusion as in case of the Dow Jones returns, indicating that the distribution is not normally distributed. Figure 7.6 shows clear signs of volatility clustering, hence a GARCH(1,1) model could be appropriate for this time series. Estimating this model with ML assuming a normal distribution of the standardized shocks gives the following results:

$$\underset{}{\hat{h}_t} = \underset{(0.008)}{0.041} + \underset{(0.012)}{0.116}\,\varepsilon_{t-1}^2 + \underset{(0.017)}{0.851}\,\hat{h}_{t-1}, \tag{7.60}$$

which resemble the typical parameter estimates we find for daily returns series, that is, a small value for α_1 and a value of $\alpha_1 + \beta_1$ close to 1. Recall that such parameter values are necessary for capturing the stylized facts of the EACF of the squared returns series. Figure 7.7 shows the Empirical AFC and the ACF implied by the GARCH(1,1) model. It seems that a GARCH(1,1) model is well suited to match the autocorrelations

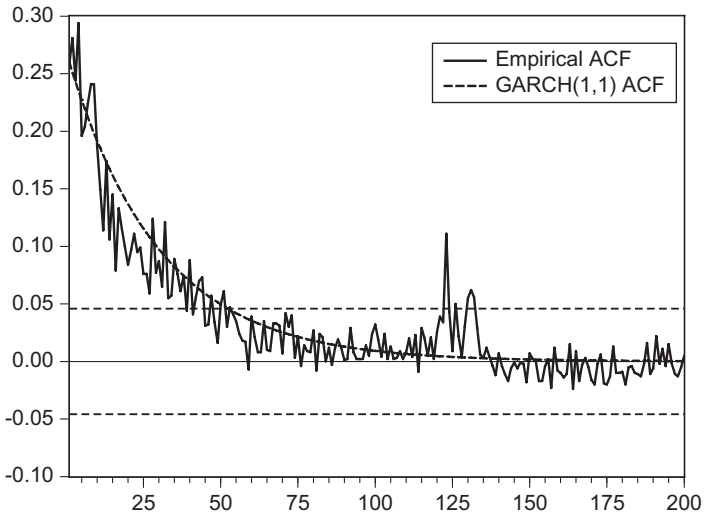


Figure 7.7: Empirical ACF and ACF implied by the GARCH(1,1) model of squared daily MSCI Switzerland returns from, January 1, 1995–December 31, 2001.

of the squared returns. Given the estimated parameters, we reconstruct the conditional standard deviation $\sqrt{\hat{h}_t}$ and plot these in Figure 7.8. It shows the effect of the Asia crisis in 1998 and 9/11 by large spikes. To provide some insights in the diagnostics of the model, we compute the autocorrelations of the (squared) standardized returns $\hat{z}_t$. As explained earlier, there should not be any autocorrelation in (the square of) this series. Figure 7.9 shows that no remaining autocorrelation exists in the standardized returns. Note that the ARCH(1) fails to capture the autocorrelation adequately. Although the above results are positive, the JB test on the normality of $\hat{z}_t$ results in a value of 317; hence normality is clearly rejected. This result is mainly due to the negative skewness of -0.33, combined with a kurtosis of 4.93. To allow for excess kurtosis in z_t, a Student-t distribution could be used instead of the Normal distribution.

Let us continue to the estimation of the TGARCH(1,1) model, allowing for different impact of returns on the volatility. The estimation results for the TGARCH(1,1) model are

$$\begin{aligned} \hat{h}_t = 0.048 &+ (0.027\ \varepsilon_{t-1}^2 + (0.162\ \varepsilon_{t-1}^2 I[\varepsilon_{t-1} < 0] + (0.851\ \hat{h}_{t-1}, \\ &(0.007)\quad (0.017)\qquad (0.015)\qquad\qquad (0.017)\qquad (7.61) \end{aligned}$$

The estimate of γ_1 is positive, suggesting that negative shocks have a six times larger effect than positive shocks of the same magnitude. Note that $\alpha_1 + \gamma_1/2$, representing

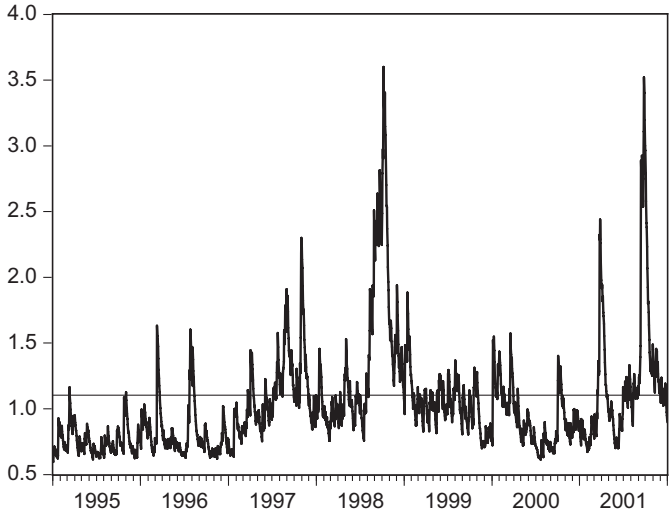


Figure 7.8: Conditional standard deviation from GARCH(1,1) model for daily returns on the MSCI Switzerland index, January 1, 1995–December 31, 2001. The black line corresponds with the unconditional standard deviation of the return series.

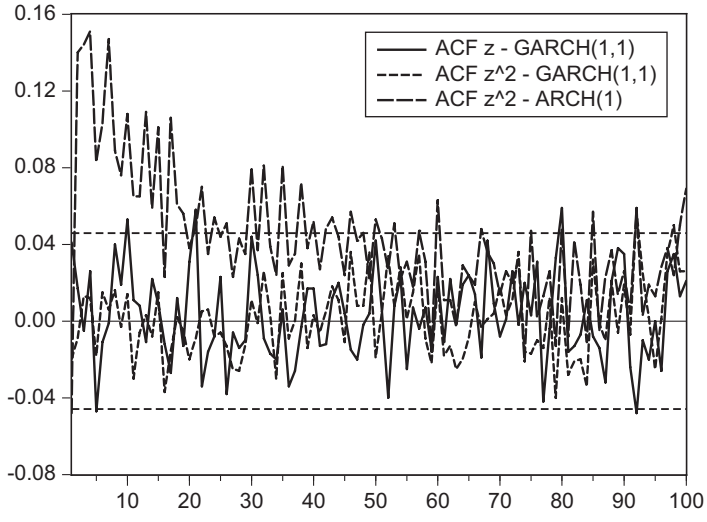


Figure 7.9: Empirical ACF of (squared) $\hat{z}_t$, for a ARCH(1) and GARCH(1,1) model, estimated on daily MSCI Switzerland returns from, January 1, 1995–December 31, 2001.

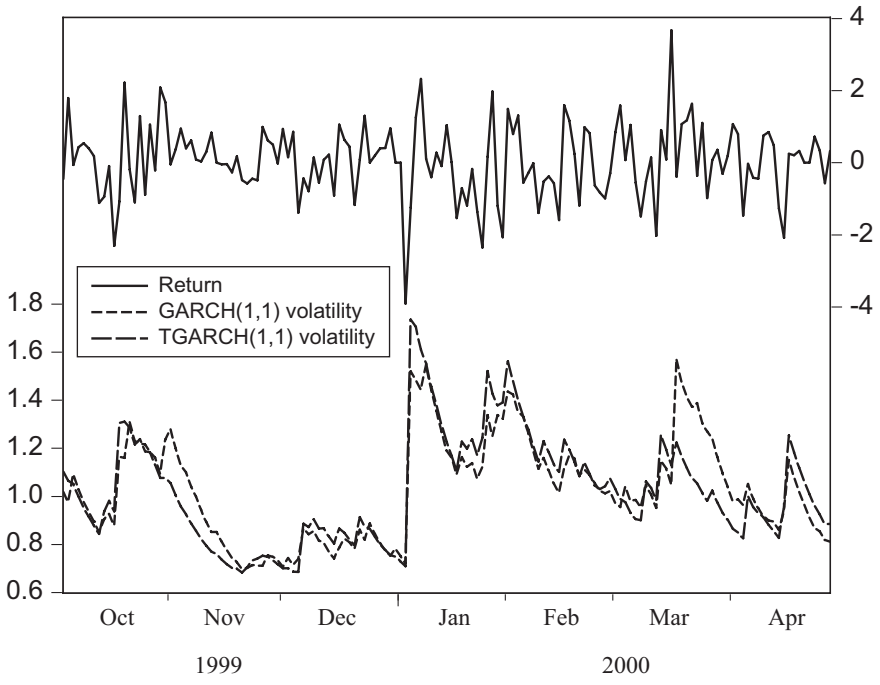


Figure 7.10: Conditional standard deviation from GARCH(1,1) and TGARCH(1,1) models for daily returns on the MSCI Switzerland index, October 1, 1999–April 30, 2000).

the average impact of positive and negative shocks, is roughly equal to the estimated value of α_1 in the symmetric GARCH(1,1) model in (7.60). To see the differences between the GARCH(1,1) and the TGARCH(1,1) model, consider Figure 7.10, showing the reconstructed series of the conditional standard deviation $\sqrt{\hat{h}_t}$ for these models. While the series of both models show similar patterns in the conditional volatility, some differences also are apparent. In particular, the spikes in the conditional volatility due to negative shocks are higher in the TGARCH(1,1) model, while some increases in volatility in the GARCH(1,1) model, due to large positive shocks, do not appear in the TGARCH volatility. For example the upward jump in h_t due to the negative shock on January 4, 2000 is considerably higher in the asymmetric TGARCH(1,1) model than in the symmetric GARCH(1,1). The positive returns on March 15 and 16 creased volatility according to the GARCH(1,1) model, but it did not have a visible effect in the TGARCH(1,1) model.

7.4 Forecasting

The presence of time-varying volatility has two important consequences for out-of-sample forecasting. First, the optimal h-step ahead point forecasts $\hat{y}_{T+h|T}$ in ARMA models are the same regardless of whether the shocks ε_t are conditionally heteroskedastic or not. Recall that the optimal h-step ahead point forecasts y_{t+h} is given by the conditional mean $\mathsf{E}[y_{T+h}|\mathcal{Y}_T]$. In deriving the analytic expressions for $\hat{y}_{T+h|T}$ in Section 3.5, no other property of the conditional distribution of y_{t+h} besides the conditional mean is used. In particular, the conditional variance $\mathsf{V}[y_{T+h}|\mathcal{Y}_t]$ does not play any role. Second, the conditional variance of the associated h-step ahead forecast error $e_{T+h|T} = y_{t+h} - \hat{y}_{T+h|T}$ becomes time-varying, which in fact was one of the main motivations for proposing the ARCH model, see Engle (1982). This makes sense as $e_{T+h|T}$ is a linear combination of the shocks that occur between the forecast origin and the forecast horizon, $\varepsilon_{t+1}, \ldots, \varepsilon_{t+h}$, see (3.119). As the conditional variance of these shocks is time-varying, the conditional variance of any function of these shocks is time-varying as well. This implies that the width of interval forecasts for y_{t+h}, which depend crucially on $\mathsf{V}[e_{T+h|T}]$, see (3.123), becomes time-varying.

Below we discuss these results in detail for the case where the observed time series y_t follows an AR(1) model,

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \tag{7.62}$$

and the conditional variance of the shocks ε_t is described by a GARCH(1,1) model

$$h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}. \tag{7.63}$$

The general case of ARMA(k,l)-GARCH(p,q) models is discussed in Baillie and Bollerslev (1992).

Forecasting the conditional mean

Baillie and Bollerslev (1992) show that the forecast that minimizes the expected value of the quadratic loss function (3.109) is the same irrespective of whether the shocks ε_t in (7.62) are conditionally homoskedastic or conditionally heteroskedastic. Thus, the optimal h-step ahead point forecast of y_{T+h} is its conditional expectation at time t, that is,

$$\hat{y}_{T+h|T} = \mathsf{E}[y_{T+h}|\mathcal{Y}_T]. \tag{7.64}$$

For the AR(1) model (7.62) this means that the optimal 1-step ahead forecast at time T is given by $\hat{y}_{T+1|T} = \phi_1 y_T$. The forecasts for $h > 1$ steps ahead can be obtained from

the recursive relationship $\hat{y}_{T+h|T} = \phi_1 \hat{y}_{T+h-1|T}$, or can be computed directly as

$$\hat{y}_{T+h|T} = \phi_1^h y_T. \tag{7.65}$$

The h-step ahead prediction error $e_{T+h|t} = y_{t+h} - \hat{y}_{T+h|T} = \phi_1 y_{T+h-1} + \varepsilon_{T+h} - \phi_1^h y_T$, and by recursive substitution for y_{T+h-j}, $j = 1, 2, \ldots, h-1$, it follows that

$$e_{T+h|t} = \sum_{i=1}^{h} \phi_1^{h-i} \varepsilon_{T+i}. \tag{7.66}$$

The variance of $e_{T+h|T}$ is given by

$$\begin{aligned}
\mathsf{V}[e_{T+h|T}] &= \mathsf{E}[\left[\left(\sum_{i=1}^{h} \phi_1^{h-i} \varepsilon_{T+i} \right)^2 |\mathcal{Y}_T \right] \\
&= \sum_{i=1}^{h} \phi_1^{2(h-i)} \mathsf{E}[\varepsilon_{T+i}^2 | \mathcal{Y}_T] \\
&= \sum_{i=1}^{h} \phi_1^{2(h-i)} \mathsf{E}[\mathsf{E}[\varepsilon_{T+i}^2 | \mathcal{Y}_{T+i-1}] | \mathcal{Y}_T] \\
&= \sum_{i=1}^{h} \phi_1^{2(h-i)} \mathsf{E}[h_{T+i} | \mathcal{Y}_T].
\end{aligned} \tag{7.67}$$

Hence, in the presence of conditional heteroskedasticity, the forecast error variance is time-varying. Before discussing the implications of this fact in more detail, we first discuss how to obtain the conditional expectations of the future conditional variances h_{T+i} at time T, or forecasts of future volatility, which are needed for computing $\mathsf{V}[e_{T+h|T}]$.

Forecasting the conditional variance

Computing the conditional expectation of h_{T+s} or, put differently, the optimal s-step ahead forecast of the conditional variance from a GARCH model is relatively straightforward. In case of the GARCH(1,1) model, the 1-step ahead forecast of h_{T+1} at time T is given by

$$\begin{aligned}
\hat{h}_{T+1|T} &= \mathsf{E}[h_{T+1} | \mathcal{Y}_T] \\
&= \mathsf{E}[\omega + \alpha_1 \varepsilon_T^2 + \beta_1 h_T | \mathcal{Y}_T] \\
&= \omega + \alpha_1 \varepsilon_T^2 + \beta_1 h_T \\
&= h_{T+1},
\end{aligned} \tag{7.68}$$

assuming that the parameters in the GARCH(1,1) model are known, such that the shocks ε_t and the conditional variance h_t can be reconstructed perfectly. The result in (7.68) illustrates the fact that GARCH models sometimes are referred to as "deterministic" volatility models, as there is no uncertainty regarding the value of h_{T+1} conditional upon $\mathcal{Y}_T$. In practice, the forecast of h_{T+1} will not be completely correct, as the parameters in the model have to be estimated, as well as the unobserved ε_T and h_T. That is, the feasible one-step ahead forecast is computed as

$$\hat{h}_{T+1|T} = \hat{\omega} + \hat{\alpha}_1 \hat{\varepsilon}_T^2 + \hat{\beta}_1 \hat{h}_T \neq h_{T+1}.$$

For the 2-step ahead forecast at time T, we obtain

$$\begin{aligned}
\hat{h}_{T+2|T} &= \mathsf{E}[h_{T+2}|\mathcal{Y}_T] \\
&= \mathsf{E}[\omega + \alpha_1 \varepsilon_{T+1}^2 + \beta_1 h_{T+1}|\mathcal{Y}_T] \\
&= \omega + \alpha_1 \mathsf{E}[\varepsilon_{T+1}^2|\mathcal{Y}_T] + \beta_1 h_{T+1} \\
&= \omega + (\alpha_1 + \beta_1) h_{T+1}, \qquad (7.69)
\end{aligned}$$

as $\mathsf{E}[\varepsilon_{T+1}^2|\mathcal{Y}_T] = h_{T+1}$. In general, the s-step ahead forecast for $s \geq 2$ can be computed recursively from

$$\begin{aligned}
\hat{h}_{T+s|T} &= \mathsf{E}[\omega + \alpha_1 \varepsilon_{T+s-1}^2 + \beta_1 h_{T+s-1}|\mathcal{Y}_T] \\
&= \omega + \alpha_1 \mathsf{E}[\varepsilon_{T+s-1}^2|\mathcal{Y}_T] + \beta_1 \mathsf{E}[h_{T+s-1}|\mathcal{Y}_T] \\
&= \omega + \alpha_1 \mathsf{E}[\mathsf{E}[\varepsilon_{T+s-1}^2|\mathcal{Y}_{T+s-2}]|\mathcal{Y}_T] + \beta_1 \hat{h}_{T+s-1|T}, \\
&= \omega + (\alpha_1 + \beta_1) \hat{h}_{T+s-1|T}. \qquad (7.70)
\end{aligned}$$

Alternatively, by recursive substitution for $\hat{h}_{T+s-1|T}$ in (7.70) it follows that the s-step ahead forecast can be computed directly as

$$\hat{h}_{T+s|T} = \omega \sum_{i=0}^{s-2} (\alpha_1 + \beta_1)^i + (\alpha_1 + \beta_1)^{s-1} h_{T+1}. \qquad (7.71)$$

If $\alpha_1 + \beta_1 < 1$, such that GARCH model is covariance stationary, (7.70) can be written as

$$\hat{h}_{T+s|T} = \sigma^2 + (\alpha_1 + \beta_1)^{s-1}(h_{T+1} - \sigma^2),$$

where $\sigma^2 = \omega/(1 - \alpha_1 - \beta_1)$ is the unconditional variance of ε_t. Hence, it follows that the conditional volatility forecasts converge to the unconditional variance, that is, $\hat{h}_{T+s|T} \to \sigma^2$ as $s \to \infty$. Also note that for the IGARCH model with $\alpha_1 + \beta_1 = 1$, (7.71) simplifies to

$$\hat{h}_{T+s|T} = \omega(s-1) + h_{T+1}, \qquad (7.72)$$

which shows that the forecasts for the conditional variance increase linearly as the forecast horizon s increases, provided $\omega > 0$.

Exercise 7.6

For the nonlinear EGARCH and TGARCH models discussed in Section 7.2, out-of-sample forecasts for the conditional variance can also be computed in a straightforward manner. As an example, consider the TGARCH model given in (7.38). The two-step ahead forecast of h_{T+2} is given by

$$\begin{aligned}\hat{h}_{T+2|T} &= \mathsf{E}[\omega + \alpha_1 \varepsilon_{T+1}^2 + \gamma_1 \varepsilon_{T+1}^2 I[\varepsilon_{T+1} > 0] + \beta_1 h_{T+1} | \mathcal{Y}_T] \\ &= \omega + (\alpha_1 + \gamma_1/2 + \beta_1) h_{T+1}, \end{aligned} \tag{7.73}$$

which follows from observing that ε_{T+1}^2 and the indicator function $I[\varepsilon_{T+1} > 0]$ are uncorrelated and $\mathsf{E}[I[\varepsilon_{T+1} > 0]] = P(\varepsilon_{T+1} > 0) = 0.5$ assuming that the median of z_t is equal to 0, and again using $\mathsf{E}[\varepsilon_{t+1}^2 | \mathcal{Y}_t] = h_{t+1}$. In general, s-step ahead forecasts can be computed either recursively as

$$\hat{h}_{T+s|T} = \omega + (\alpha_1 + \gamma_1/2 + \beta_1) \hat{h}_{T+s-1|T}, \tag{7.74}$$

or directly from

$$\hat{h}_{T+s|T} = \omega \sum_{i=0}^{s-2} (\alpha_1 + \gamma_1/2 + \beta_1)^i + (\alpha_1 + \gamma_1/2 + \beta_1)^{s-1} h_{T+1}, \tag{7.75}$$

compare the analogous expressions for the GARCH(1,1) model, as given in (7.70) and (7.71).

Exercise 7.7

Interval forecasts

We now return to the variance of the h-step ahead forecast error, which is varying over time in case of conditional heteroskedasticity, as shown by (7.67). This may be used for constructing a $100\alpha\%$ interval forecast as

$$(\hat{y}_{T+h|T} - z_{(1+\alpha)/2} \sqrt{\mathsf{V}[e_{T+h|T}]}, \hat{y}_{T+h|T} + z_{(1+\alpha)/2} \sqrt{\mathsf{V}[e_{T+h|T}]}), \tag{7.76}$$

where $z_{(1+\alpha)/2}$ is the $(1+\alpha)/2$ quantile of the distribution of z_t, see also (3.123). The fact that $\mathsf{V}[e_{T+h|T}]$ is time-varying implies that the width of these interval forecasts also varies over time. Intuitively, in periods of large uncertainty, characterized by large values of the (expected) conditional variance, the interval forecasts become wider.

In case of homoskedastic errors, $\mathsf{V}[e_{T+h|T}]$ is constant, as $\mathsf{E}[h_{t+i}|\mathcal{Y}_t]$ is constant and equal to the unconditional variance of ε_t, σ^2. To see the relation between the two cases, rewrite (7.67) as

$$\mathsf{E}[e^2_{T+h|T}|\mathcal{Y}_T] = \sum_{i=1}^{h} \phi_1^{2(h-i)}\sigma^2 + \sum_{i=1}^{h} \phi_1^{2(h-i)}(\mathsf{E}[h_{T+i}|\mathcal{Y}_T] - \sigma^2). \qquad (7.77)$$

The first term on the right hand-side of (7.77) is the forecast error variance in case of homoskedastic errors. Notice that the second term on the right-hand side can be both positive and negative, depending on the conditional expectation of future volatility. Hence, the forecast error variance in case heteroskedastic errors can be both larger and smaller than in case of homoskedastic errors.

Recall that in the homoskedastic case, the forecast error variance converges to the unconditional variance of the time series y_t as the forecast horizon increases, that is,

$$\lim_{h\to\infty} \mathsf{E}[e^2_{T+h|T}|\mathcal{Y}_T] = \lim_{h\to\infty} \sum_{i=1}^{h} \phi_1^{2(h-i)}\sigma^2 = \frac{\sigma^2}{1-\phi^2} \equiv \mathsf{V}[y_t]. \qquad (7.78)$$

Moreover, the convergence is monotonic, in the sense that the h-step ahead forecast error variance is always smaller than the unconditional variance $\mathsf{V}[y_t]$, while $\mathsf{V}[e_{T+h|T}] \geq \mathsf{V}[e_{T+h-1|T}]$ for all finite horizons h. The convergence of the forecast error variance to the unconditional variance of the time series also holds in the present case of heteroskedastic errors. This follows from the fact that the forecasts of the onditional variance $\mathsf{E}[h_{T+i}|\mathcal{Y}_T]$ converge to the unconditional variance σ^2 as shown before. However, the convergence need no longer be monotonic, in the sense that $\mathsf{V}[e_{T+h|T}]$ may be smaller than $\mathsf{V}[e_{T+h-1|T}]$. In fact, the forecast error variance may be larger than the unconditional variance of the time series for certain forecast horizons. Intuitively, large values of the conditional variance h_t suggest that it is difficult to forecast the conditional mean of the series y_t accurately. In such cases, the forecast uncertainty may be larger at shorter forecast horizons compared to longer horizons.

As an illustration, we use the TGARCH(1,1) model to obtain one-step ahead forecasts of the conditional volatility of the Dow Jones index returns for the period January 1, 2002–December 31, 2012. The volatility forecasts are shown in Figure 7.11. Figure 7.12 shows 95% interval forecasts for the returns. Clearly these vary in width according to the conditional volatility forecast.

Evaluating forecasts of conditional volatility

Whereas forecasting the future conditional variance from (nonlinear) GARCH models is fairly straightforward, evaluating the forecasts is a more challenging task, due to the fact that volatility is unobserved. In the following we assume that a GARCH

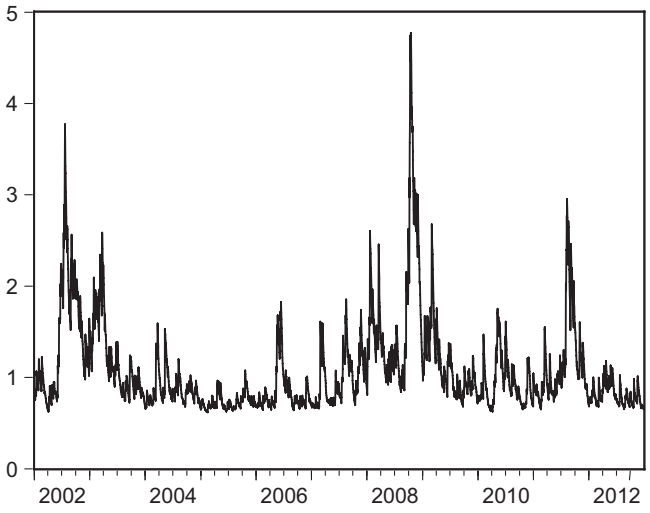


Figure 7.11: One-step ahead forecasts of conditional standard deviation from TGARCH(1,1) models for daily returns on MSCI Switzerland, January 1, 2002–December 31, 2012.

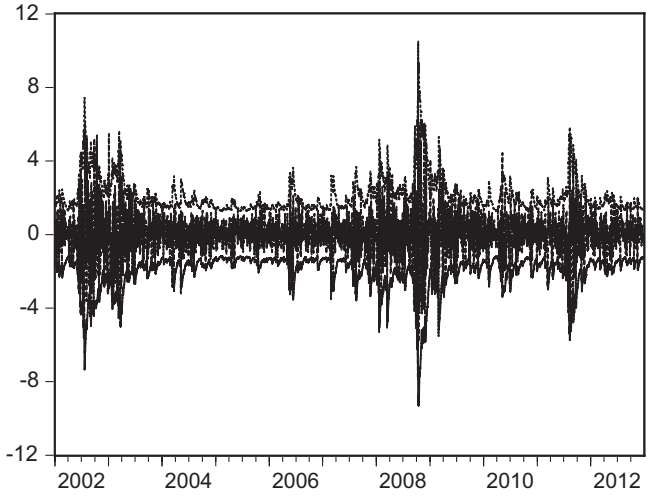


Figure 7.12: One-step ahead 95% interval forecasts from TGARCH(1,1) models for daily returns on MSCI Switzerland, January 1, 2002–December 31, 2012.

model has been estimated using a sample of T observations, whereas observations at $t = T + 1, \ldots T + P + s - 1$ are held back for evaluation of s-step ahead forecasts for the conditional variance.

Most frequently statistical criteria are used to assess the accuracy of volatility forecasts. The most popular of these criteria is the mean squared prediction error [MSPE], which for a set of P s-step ahead forecasts is computed as

$$\text{MSPE} = \frac{1}{m} \sum_{j=0}^{P-1} (\hat{h}_{T+s+j|T+j} - h_{T+s+j})^2, \tag{7.79}$$

see West and Cho (1995) and Franses and van Dijk (1996), among many others. Alternatively, the regression of true volatility on a constant and the volatility forecast, that is,

$$h_{T+s+j} = a + b\hat{h}_{T+s+j|T+j} + e_{T+s+j}, \qquad j = 0, \ldots, P-1, \tag{7.80}$$

often is considered, see Pagan and Schwert (1990), Lamoureux and Lastrapes (1993), and Jorion (1995), among others. In case $\hat{h}_{T+s+j|T+j}$ is an unbiased forecast of h_{T+s+j}, $a = 0, b = 1$ and $E(e_{T+s+j}) = 0$ in (7.80). Furthermore, we would expect a high value of the R^2 of this regression in case the volatility forecasts are accurate.

The MSPE as defined in (7.79) cannot be computed as the true volatility h_{T+s+j} is unobserved, whereas for the same reason the parameters in (7.80) cannot be estimated. To make these forecast evaluation criteria operational, an estimate of h_{T+s+j} is required. Usually the squared shock ε^2_{T+s+j} is used for this purpose. As $\mathsf{E}[\varepsilon^2_{T+s+j}|\mathcal{Y}_{T+s+j-1}] = h_{T+s+j}$, ε^2_{T+s+j} is an unbiased estimate of h_{T+s+j}. However, at the same time, $\varepsilon^2_{T+s+j} = z^2_{T+s+j} h_{T+s+j}$ is an extremely noisy estimate of h_{T+s+j} due to the multiplication with the square of z_{T+s+j}. For many forecast evaluation criteria, such as the MSPE and the regression in (7.80), this leads to the spurious conclusion that quite accurate volatility forecasts are no good, as discussed below.

A common finding is that GARCH models provide seemingly poor volatility forecasts, in the sense that the MSPE (or any other measure of forecast accuracy) is very large while the R^2 from the regression (7.80) is very small, typically below 0.10. In addition, the forecasts from GARCH appear to be biased, as it commonly found that $\hat{a} \neq 0$ in (7.80). Andersen and Bollerslev (1998) demonstrate that this poor forecasting performance need not be due to the fact that GARCH models are such bad models, but may also be caused by the approximation of the unobserved true volatility h_{T+s+j} with the noisy squared shock ε^2_{T+s+j}. As shown by Andersen and Bollerslev (1998), for a GARCH(1,1) model with a finite unconditional fourth moment the theoretical R^2 from the regression (7.80) for $s = 1$ and h_{T+s+j} replaced by ε^2_{T+s+j} is

equal to

$$R^2 = \frac{\alpha_1^2}{1 - \beta_1^2 - 2\alpha_1\beta_1}. \tag{7.81}$$

As the condition for a finite unconditional fourth moment in the GARCH(1,1) model is given by $\kappa\alpha_1^2 + \beta_1^2 + 2\alpha_1\beta_1 < 1$, it follows that the R^2 is bounded from above by $1/\kappa$. In case z_t is normally distributed, the R^2 cannot be larger than 1/3, while the upper bound is even smaller if, for example, z_t is assumed to be Student t distributed.

Thus, given that volatility is unobserved and that the squared time series is a noisy volatility measure, evaluating the accuracy of volatility forecasts appears to be rather difficult. Two possible solutions to this problem are available. First, we may attempt to obtain a more accurate measure of volatility. This is possible by using data which is sampled more frequently than the time series of interest. For example, suppose that we are interested in evaluating daily volatility forecasts for the Dow Jones index. If the index closes at the same level on day t as on day $t - 1$, the squared daily return gives a measure of volatility for day t equal to 0. This would be correct if the index level did not change at all during the trading day. However, it grossly underestimates true volatility if the index moved erratically during the day, but for some reason happened to end the day at the same level as it started. In that case, a much better assessment of the volatility in the stock market can be made if the index were observed, for example, every 15 minutes. An accurate measure of daily volatility can then be obtained by summing the squared 15-minute returns during the day. This measure is usually referred to as realized volatility. The use of such high-frequency data is not only useful for evaluating volatility forecasts from GARCH models. In addition, realized volatility may be modeled directly using, for example, an ARMA model, see Andersen *et al.* (2006) for a review.

Second, volatility forecasts from GARCH models may be evaluated "indirectly", for example, by considering the profitability of trading or investment strategies, the accuracy of option prices, Value-at-Risk estimates, or utility-based measures, which make use of conditional volatility forecasts, see Engle, Hong, Kane and Noh (1993), West *et al.* (1993), and Lopez (2001), among others. This indirect evaluation can be motivated by the fact that volatility forecasts often are not a goal in themselves, but are rather used as inputs in financial decision problems such as portfolio construction or risk management.

EXERCISES

7.1 Show that the GARCH(1,1) model of (7.13) can be written as an ARMA(1,1) model given in (7.16).

7.2 Consider the GARCH(1,1) model for the time series of daily stock returns y_t:

$$\begin{aligned} y_t &= \varepsilon_t = z_t\sqrt{h_t} \\ h_t &= \omega + \alpha y_{t-1}^2 + \beta h_{t-1} \\ z_t &\sim N(0,1) \end{aligned}$$

with $\omega, \alpha, \beta > 0$ and $\alpha + \beta < 1$.

a. Derive an expression for the probability that the conditional variance at time t is larger than the conditional variance at time $t-1$, conditional on the conditional variance at time $t-1$, that is derive $P[h_t > h_{t-1}|h_{t-1}]$.

b. Show that $P[h_t > h_{t-1}|h_{t-1}] < 0.5$ in case $h_{t-1} = \sigma^2$, where σ^2 is the unconditional variance of y_t

7.3 Consider the TGARCH(1,1) model, (a.k.a. GJR-GARCH(1,1) model) for daily stock returns y_t:

$$\begin{aligned} y_t &= \varepsilon_t = z_t\sqrt{h_t} \\ h_t &= \omega + \alpha\varepsilon_{t-1}^2 + \gamma_1\varepsilon_{t-1}^2 I[\varepsilon_{t-1} < 0] + \beta h_{t-1} \\ z_t &\sim N(0,1) \end{aligned} \tag{7.82}$$

where $I[\cdot]$ is an indicator function and the model is covariance stationary.

a. Show that the above model can also written as

$$h_t = \omega + \theta_1 I[\varepsilon_{t-1} \leq 0]\varepsilon_{t-1}^2 + \theta_2 I[\varepsilon_{t-1} > 0]\varepsilon_{t-1}^2 + \beta h_{t-1}. \tag{7.83}$$

What is the relation between α_1 and γ_1 in equation (7.82) and θ_1 and θ_2 in equation (7.83)?

b. Show that $y_t = \varepsilon_t = z_t\sqrt{h_t}$ has a symmetric marginal distribution.

c. Explain why the multi-period returns $y_{k,t} = y_t + \cdots + y_{t-k+1}$ will have a skewed marginal distribution if $\theta_1 > \theta_2$ (and $k > 1$).

7.4 Consider the Switching-GARCH(1,1) [S-GARCH(1,1)] model for the time series variable y_t:

$$\begin{aligned} y_t &= z_t\sqrt{h_t} \\ h_t &= \omega_t + \alpha\varepsilon_{t-1}^2 + \alpha_1 y_{t-1}^2 + \beta h_{t-1} \\ \omega_t &= \omega_1 I[s_t = 1] + \omega_2 I[s_t = 2] \\ z_t &\sim N(0,1) \end{aligned}$$

with $\omega_1 > \omega_2 > 0, \alpha > 0, \beta > 0$ and $\alpha + \beta < 1$ and where $I[A] = 1$ if A occurs, and 0 otherwise. The variable s_t is unobserved and takes the values 1 or 2

according to the following probabilities:

$$P(s_t = 1|s_{t-1} = P(s_t = 1) = p$$
$$P(s_t = 2|s_{t-1} = P(s_t = 2) = 1 - p \qquad \text{for some} \qquad 0 < p < 1$$

Discuss the properties of the S-GARCH(1,1) model. In particular, interpret the model for the conditional variance h_t.

7.5 Suppose you want to estimate the Treshold-GARCH(1,1) model for the time series variable y_t:

$$y_t = \varepsilon_t = z_t\sqrt{h_t}$$
$$h_t = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 I[\varepsilon_{t-1} < 0] + \beta h_{t-1}$$
$$z_t \sim t(0, 1, v)$$

with v the degrees of freedom parameter. Suppose that you have to write a program that maximizes the log likelihood function. Assume that there is an build-in maximization function `fmaxcon(loglikfunct,A,b,data)`, where the input of this function consists of the log likelihood function $\mathcal{L}(\theta)$, a matrix A and a columnvector b such that $Ax \leq b$ (linear restrictions), where x denotes the parameter vector. Moreover, A denotes a $N \times k$ matrix, (N restrictions, k the dimension of θ) and the vector b represents a $N \times 1$ vector. Finally, the term *data* represents the daily returns y_t that should be included in order to compute the likelihood function.

a. Given the data y_t, write down the log likelihood function $\mathcal{L}(\theta)$. Explain why one often optimizes the *conditional* log likelihood function.
b. Give expressions for A and b.

7.6 Consider the standard GARCH(1,1) model, as described in exercise 7.2 with $\alpha + \beta < 1$. A different way to model volatility is to use an exponentially weighted moving average. That is, one estimates/forecasts volatility by means of weighted average of squared returns:

$$h_t = (1 - \lambda)\sum_{j=1}^{\infty} \lambda^{j-1}(y_{t-j} - \bar{y})^2$$

where $0 < \lambda < 1$. For daily data $\lambda = 0.94$ gives JP Morgan's RiskMetrics model.

a. Show that (7.84) can also be written as

$$h_t = \lambda h_{t-1} + (1 - \lambda)(y_{t-1} - \bar{y})^2$$

What is the intuition behind this representation?

b. Derive the optimal 1-step, 2-step and 3-step ahead point forecasts of h_t in the GARCH(1,1) model (that is, derive expressions for $\hat{h}_{t+k|t} = \mathsf{E}[h_{t+k}|\mathcal{Y}_t]$ for

$k = 3, 4$ and 5, where $\mathcal{Y}_t$ denotes the information set available at t.) What is the optimal 100-step ahead point forecast?

c. Derive the optimal 1-step, 2-step and 3-step ahead point forecasts of h_t in the RiskMetrics model. What is the crucial difference between forecasts of the RiskMetrics model and the GARCH(1,1) model?

7.7 Let r_t denote the stock return on day t and assume that this is generated according to the Threshold GARCH(1,1) process

$$
\begin{aligned}
r_t &= \mu + \varepsilon_t \\
\varepsilon_t &= z_t \sqrt{h_t} \quad \text{with} \quad z_t \sim N(0, 1) \\
h_t &= \omega + \alpha \varepsilon_{t-1}^2 I[\varepsilon_{t-1} \leq 0] + \gamma \varepsilon_{t-1}^2 I[\varepsilon_{t-1} > 0] + \beta h_{t-1}
\end{aligned}
$$

where I[A] =1 if A occurs, and 0 otherwise. Assume that the parameters α and γ are such that $\alpha > 0$ and $\alpha = 3\gamma$. Suppose we estimate a symmetric GARCH(1,1) model for this return series (where $h_t = \omega^* + \alpha^* \varepsilon_{t-1}^2 + \beta^* h_{t-1}$) and use this to compute daily 99% Value-at-Risk (VaR) forecasts. Do we expect to find more or less than 1% violations of this VaR?

8 Non-linearity

The final typical feature of business and economic time series that is treated in this book is non-linearity. Similar to the features of trend and seasonality discussed in Chapters 4 and 5, respectively, it is difficult to define non-linearity otherwise than in the context of a model. Loosely speaking, however, a time series may be said to be linear when the impact of a shock is (i) proportional to its size, (ii) independent of its sign (in an absolute sense), and (iii) independent of the current value of the time series. Whenever one of these three properties does not hold, a time series may be said to be non-linear. As an example, consider the quarterly US unemployment rate in Figure 2.19. For this series we observe that the average increase during recessions is larger in an absolute sense than the average decrease during expansions. This suggests that the US unemployment rate is non-linear, as positive shocks (leading to more unemployment) appear to have a larger impact than negative shocks. Alternatively, the observed asymmetric behavior of the unemployment rate suggests that the effects of a given shock depend on the prevailing state of business cycle, as they seem to differ across recessions and expansions.

The three properties of shocks mentioned above are implied characteristics of ARMA models, as discussed in Chapter 3. Consequently, by definition time series exhibiting non-linear characteristics cannot be adequately described by such models. Once the linear ARMA models are dismissed in favor of some form of non-linear alternative, the problem we face is the enormous, if not unlimited, number of possible non-linear model structures that can be used to describe and forecast economic time series. In this chapter, we cannot possibly survey all non-linear time series models that are available, and therefore we focus on a few specific so-called regime-switching models, which have become popular in empirical economic applications during the recent past. These models have a clear interpretation and are plausible from an economic perspective. For more general surveys on non-linear time series models, the interested reader is referred to Tong (1990) and Terasvirta *et al.* (2010).

A natural approach to allow for non-linearity in the context of economic time series seems to define different *states of the world* or *regimes*, and to allow for the possibility

that the dynamic behavior of the variable of interest depends on the regime that occurs at any given point in time. By state-dependent dynamic behavior it is meant that certain properties of the time series such as its mean, variance and/or autocorrelations are different in different regimes. An example of such *state-dependent* or *regime-switching* behavior was in fact encountered already in Chapter 5, where it was shown that the means of economic time series may vary across the seasons. Hence, we can say that each season of the year constitutes a different regime. The interpretation of such seasonal effects as regime-switching behavior may seem somewhat odd, for in this case the regime process is *deterministic*, in the sense that the regime that occurs at any given in point in time is known with certainty in advance. By contrast, in this chapter we focus on situations where the regime process is *stochastic*. For the US unemployment rate, for example, the relevant regimes appear to be business cycle recessions and expansions. These are stochastic, in the sense that currently we do not know with complete certainty whether the economy will be in an expansion or in a recession state during the next quarter.

In recent years several time series models have been proposed which formalize the idea of regime-switching. In this chapter we restrict attention to models that assume that in each of the regimes the dynamic behavior of the time series can be adequately described by a linear AR model. In other words, the time series is modeled with an AR model, where the autoregressive parameters are allowed to depend on the regime or state. Generalizations of the MA model to a regime-switching context have been considered as well, see Wecker (1981) and de Gooijer (1998), but we abstain from discussing these models here.

In the following sections, we discuss representations and interpretation of several regime-switching models, parameter estimation, testing for the presence of regime-switching effects, model evaluation or diagnostic checking, out-of-sample forecasting. We emphasize how these different elements can be used in an empirical specification strategy.

8.1 Regime-switching models

In this section we introduce the regime-switching models and discuss their basic properties. To simplify the exposition, we focus attention on models that involve only two regimes. Assuming in addition that an AR(1) model is sufficient to characterize the time series in each of the regimes, a two-regime model is given by

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1} y_{t-1} + \varepsilon_t & \text{if regime 1 occurs,} \\ \phi_{2,0} + \phi_{2,1} y_{t-1} + \varepsilon_t & \text{if regime 2 occurs,} \end{cases} \tag{8.1}$$

where the ε_t are a white noise sequence with mean zero and variance σ^2.

The model in (8.1) needs to be completed by defining the regimes more precisely, and in particular, by specifying the way in which the regime evolves over time. Roughly speaking, two possibilities are available for that purpose, leading to two different types of regime-switching models. On the one hand, we may assume that the regime at time t can be characterized (or determined) by a variable q_t, which is observable at $t-1$ or earlier. Specifically, in the Threshold Autoregressive [TAR] model, introduced by Tong (1978) and Tong and Lim (1980), it is assumed that the regime is identified by the value of q_t relative to a threshold value, which we denote as c. The two-regime TAR model thus is given by

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1} y_{t-1} + \varepsilon_t & \text{if } q_t \leq c, \\ \phi_{2,0} + \phi_{2,1} y_{t-1} + \varepsilon_t & \text{if } q_t > c. \end{cases} \tag{8.2}$$

An alternative way to write this model is

$$y_t = (\phi_{1,0} + \phi_{1,1} y_{t-1})(1 - I[q_t > c]) + (\phi_{2,0} + \phi_{2,1} y_{t-1}) I[q_t > c] + \varepsilon_t, \tag{8.3}$$

where $I[A]$ is an indicator function with $I[A] = 1$ if the event A occurs and $I[A] = 0$ otherwise.

The so-called threshold variable q_t may be an exogenous variable z_{t-1}, or a (function of) lagged value(s) of the time series itself, for example, $q_t = y_{t-d}$ or $q_t = \Delta y_{t-d}$ for a certain integer $d > 0$. In the last two cases, the resulting model is called a self-exciting TAR (SETAR) model. The choice of q_t is guided by the relevant regimes for the time series under investigation and should be such that its value (relative to the threshold c) splits the observations in the regimes as desired. For example, for the US unemployment rate the relevant regimes appear to be business cycle recessions and expansions. A suitable exogenous transition variable in this case is output growth, which by definition is positive during expansions and negative during recessions. Alternatively, we may use the lagged change in the unemployment rate itself, which has similar characteristics (except that of course the change in unemployment is negative during expansions and positive during recessions).

The TAR model in (8.3) assumes that the border between the two regimes is given by a specific value of the threshold variable q_t, or in other words, that the switch of regimes is abrupt. A more gradual transition between the different regimes can be obtained by replacing the indicator function $I[q_t > c]$ in (8.3) by a continuous function $G(q_t; \gamma, c)$, which changes smoothly from 0 to 1 as q_t increases. The resulting model is called a Smooth Transition AR [STAR] model and is given by

$$\begin{aligned} y_t = {} & (\phi_{1,0} + \phi_{1,1} y_{t-1})\,(1 - G(q_t; \gamma, c)) \\ & + (\phi_{2,0} + \phi_{2,1} y_{t-1})\, G(q_t; \gamma, c) + \varepsilon_t, \end{aligned} \tag{8.4}$$

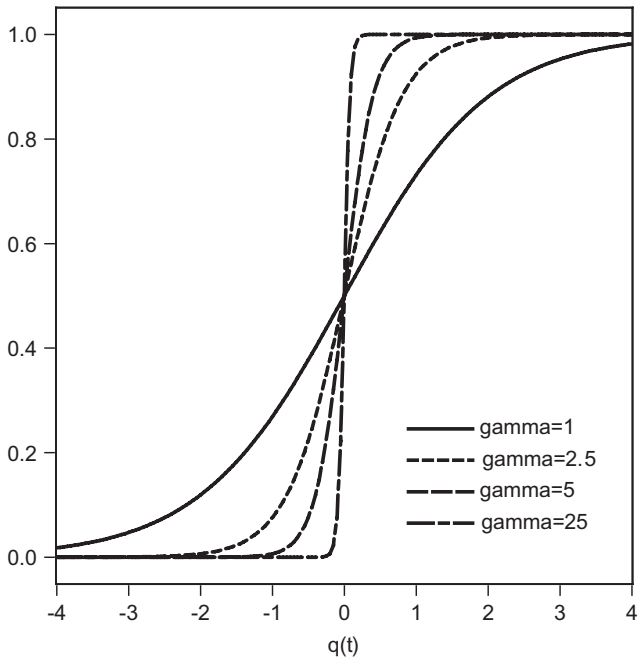


Figure 8.1: Examples of the logistic function $G(q_t; \gamma, c)$ as given in (8.5) for various values of the smoothness parameter γ and threshold $c = 0$.

introduced by Chan and Tong (1986) and Teräsvirta (1994). A popular choice for the so-called transition function $G(q_t; \gamma, c)$ is the logistic function

$$G(q_t; \gamma, c) = \frac{1}{1 + \exp(-\gamma[q_t - c])}, \tag{8.5}$$

leading to a Logistic STAR [LSTAR] model. The parameter c in (8.5) can be interpreted as the threshold between the two regimes corresponding to $G(q_t; \gamma, c) = 0$ and $G(q_t; \gamma, c) = 1$, in the sense that the logistic function changes monotonically from 0 to 1 as q_t increases, while $G(c; \gamma, c) = .5$. The parameter γ determines the smoothness of the change in the value of the logistic function, and thus the transition from one regime to the other.

Figure 8.1 shows some examples of the logistic function for various different values of the smoothness parameter γ. From this graph it is seen that as γ becomes very large, the change of $G(q_t; \gamma, c)$ from 0 to 1 becomes almost instantaneous at $q_t = c$ and, consequently, the logistic function $G(q_t; \gamma, c)$ approaches the indicator function $I[q_t > c]$. Hence the TAR model (8.3) can be approximated arbitrarily well by the

LSTAR model (8.4) with (8.5). When $\gamma \to 0$, the logistic function becomes equal to a constant (equal to 0.5) and when $\gamma = 0$, the STAR model reduces to a linear model.

An attractive feature of the TAR and STAR models is that the regimes that have occurred in the past and present are known with certainty (although they have to be found by statistical techniques, of course) as the variable q_t is observable. This also facilitates parameter estimation and testing for nonlinearity, as will be discussed in the next two sections. However, sometimes it may be difficult to find an observable variable q_t that accurately identifies the relevant regimes and splits the observations accordingly. An alternative approach in that case it to assume that the regime can be described as an underlying unobservable stochastic process s_t. In case of only two regimes, s_t can simply be assumed to take on the values 1 and 2, such that the model with an AR(1) model in both regimes is given by

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1} y_{t-1} + \varepsilon_t & \text{if } s_t = 1, \\ \phi_{2,0} + \phi_{2,1} y_{t-1} + \varepsilon_t & \text{if } s_t = 2, \end{cases} \tag{8.6}$$

or, using an obvious shorthand notation,

$$y_t = \phi_{0,s_t} + \phi_{1,s_t} y_{t-1} + \varepsilon_t. \tag{8.7}$$

To complete the model, the properties of the process s_t need to be specified. The most popular model of this type, which was advocated by Hamilton (1989), is the Markov-Switching [MSW] model, in which the process s_t is assumed to be a first-order Markov process. This implies that the current regime s_t only depends on the regime one period ago, s_{t-1}. Hence, the model is completed by defining the transition probabilities of moving from one state to the other,

$$\begin{aligned} P(s_t = 1 | s_{t-1} = 1) &= p_{11}, \\ P(s_t = 2 | s_{t-1} = 1) &= p_{12}, \\ P(s_t = 1 | s_{t-1} = 2) &= p_{21}, \\ P(s_t = 2 | s_{t-1} = 2) &= p_{22}. \end{aligned}$$

Thus, p_{ij} is equal to the probability that the Markov chain moves from state i at time $t-1$ to state j at time t or, put differently, the probability that regime i at time $t-1$ is followed by regime j at time t. Obviously, for the p_{ij}'s to define proper probabilities, they should be nonnegative, while it should also hold that $p_{11} + p_{12} = 1$ and $p_{21} + p_{22} = 1$. Also of interest in the MSW model are the *unconditional* probabilities that the process is in each of the regimes, that is, $P(s_t = i)$ for $i = 1, 2$. These unconditional

probabilities are given by

$$P(s_t = 1) = \frac{1 - p_{22}}{2 - p_{11} - p_{22}}, \tag{8.8}$$

$$P(s_t = 2) = \frac{1 - p_{11}}{2 - p_{11} - p_{22}}, \tag{8.9}$$

see Hamilton (1994, pp. 681–683) for an explicit derivation of this result.

In addition to the regimes being observable or not, a second important difference between the (S)TAR and MSW models is that in the latter the probability of switching regimes is constant and equal to p_{ij}, while it is time-varying in the former model, depending on the properties of q_t.

Higher order models

Although the TAR and STAR models with an AR(1) model in both regimes can already generate a large variety of dynamic patterns, in practice one may want to allow for higher-order AR models in the different regimes. For example, in the two-regime case, the AR orders might be set to p_1 and p_2 in the lower and upper regimes, respectively. In this case the TAR model becomes

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1} y_{t-1} + \cdots + \phi_{1,p_1} y_{t-p_1} + \varepsilon_t & \text{if } q_t \leq c, \\ \phi_{2,0} + \phi_{2,1} y_{t-1} + \cdots + \phi_{2,p_2} y_{t-p_2} + \varepsilon_t & \text{if } q_t > c, \end{cases} \tag{8.10}$$

whereas the equivalent STAR model is given by

$$\begin{aligned} y_t = {} & \left(\phi_{1,0} + \phi_{1,1} y_{t-1} + \cdots + \phi_{1,p_1} y_{t-p_1}\right)\left(1 - G(q_t; \gamma, c)\right) \\ & + \left(\phi_{2,0} + \phi_{2,1} y_{t-1} + \cdots + \phi_{2,p_2} y_{t-p_2}\right) G(q_t; \gamma, c) + \varepsilon_t. \end{aligned} \tag{8.11}$$

Stationarity

Little is known about the conditions under which TAR and STAR models generate time series that are stationary. Such conditions have only been established for the first-order model (8.2) with $q_t = y_{t-1}$. As shown by Chan and Tong (1985), a sufficient condition for stationarity is $\max(|\phi_{1,1}|, |\phi_{2,1}|) < 1$, which is equivalent to the requirement that the AR(1) models in the two regimes are stationary. Chan *et al.* (1985) show that stationarity of the first order model actually holds under less restrictive conditions. In particular, y_t may be stationary even if one of the AR(1) models contains a unit root. Testing for unit roots in TAR models is discussed in Caner and Hansen (2001) and Enders and Granger (1998). A rough-and-ready check for stationarity of nonlinear time series models in general is to determine whether or not the skeleton is stable. For example, in case of a STAR model of

(8.11), the skeleton is given by $F(x_t;\theta) = \phi_1' x_t(1 - G(q_t;\gamma,c)) + \phi_2' x_t G(q_t;\gamma,c)$, with $\phi_i = (\phi_{i,0}, \phi_{i,1}, \ldots, \phi_{i,p})'$, $i = 1, 2$, and $x_t = (1, y_{t-1}, \ldots, y_{t-p})'$. Intuitively, if the skeleton is such that the series tends to explode for certain starting values, the series is nonstationary. This can be established by what might be called *deterministic simulation*. That is, given starting values $y_0, \ldots, y_{1-p}$, with $p = \max(p_1, p_2)$, one computes the values taken by $y_1, y_2, \ldots$, while setting all ε_t, $t = 1, 2, \ldots$ equal to zero. Doing this for many different starting values gives an impression about the characteristics of the (skeleton of the) model, see Teräsvirta and Anderson (1992) and Peel and Speight (1996) for applications of this procedure. In general, one also has to resort to numerical procedures to evaluate the stationary distribution of y_t. Some of the methods that can be applied are discussed in Moeanaddin and Tong (1990) and Tong (1990, Sec. 4.2). Finally, stationarity conditions for the 2-regime MSW model are discussed in Holst *et al.* (1994).

Empirical specification procedure

We conclude this section by outlining a specification procedure for regime-switching models, which may be used in empirical applications to structure the modeling process. Granger (1993) strongly recommends to employ a specific-to-general approach when considering the use of nonlinear time series models to describe the features of a particular variable. An empirical specification procedure for TAR, STAR and MSW models that follows this approach consists of the following steps.

(i) Specify an appropriate linear AR model of order p [AR(p)] for the time series under investigation;
(ii) Test the null hypothesis of linearity against the alternative of TAR-, STAR-, and/or MSW-type nonlinearity. For the TAR and STAR models, this step also consists of selecting the appropriate variable q_t that identifies the regime at time t.
(iii) Estimate the parameters in the selected model;
(iv) Evaluate the model using diagnostic tests;
(v) Modify the model if necessary;
(vi) Use the model for descriptive or forecasting purposes.

Steps (ii)–(vi) in this specification procedure are discussed in detail in the following sections. It turns out that tests against TAR- and MSW-type nonlinearity, which are to be used in step (ii), require the input of estimates of the parameters in these models. Hence, in the next section we first discuss parameter estimation, and turn to testing for nonlinearity in Section 8.3.

Exercise 8.1–8.4

8.2 Estimation

The discussion of estimating the parameters in the different regime-switching models in this section is necessarily rather brief and only describes the general ideas of the estimation methods. For more elaborate discussions we refer to Hansen (1997,1998) for the TAR model, to Teräsvirta (1994,1998) for the STAR model, and to Hamilton (1990,1993,1994) for the MSW model. For notational convenience, we discuss the estimation problem for 2-regime models with equal AR orders in the two regimes, that is, $p_1 = p_2 = p$.

8.2.1 Estimation of TAR models

The parameters of interest in the 2-regime TAR model (8.10), that is, $\phi_{i,j}$, $i = 1, 2$ $j = 0, \ldots, p$, c and σ^2, can conveniently be estimated by sequential conditional least squares. Under the additional assumption that the ε_t's are normally distributed, the resulting estimates are equivalent to maximum likelihood estimates.

To see why least squares is the appropriate estimation method, rewrite (8.10) (with $p_1 = p_2 = p$) as

$$\begin{aligned} y_t = &\left(\phi_{1,0} + \phi_{1,1} y_{t-1} + \cdots + \phi_{1,p} y_{t-p}\right) I[q_t \leq c] \\ &+ \left(\phi_{2,0} + \phi_{2,1} y_{t-1} + \cdots + \phi_{2,p} y_{t-p}\right) I[q_t > c] + \varepsilon_t, \end{aligned} \tag{8.12}$$

or more compactly as

$$y_t = \phi_1' x_t I[q_t \leq c] + \phi_2' x_t I[q_t > c] + \varepsilon_t, \tag{8.13}$$

where $\phi_i = (\phi_{i,0}, \phi_{i,1}, \ldots, \phi_{i,p})'$, $i = 1, 2$, and $x_t = (1, y_{t-1}, \ldots, y_{t-p})'$. Note that in case the threshold c is fixed, the model is linear in the remaining parameters and estimates of $\phi = (\phi_1', \phi_2')'$ then are easily obtained by OLS as

$$\hat{\phi}(c) = \left(\sum_{t=1}^{T} x_t(c) x_t(c)'\right)^{-1} \left(\sum_{t=1}^{T} x_t(c) y_t\right), \tag{8.14}$$

where $x_t(c) = (x_t' I[q_t \leq c], x_t' I[q_t > c])'$ and the notation $\hat{\phi}(c)$ is used to indicate that the estimate of ϕ is conditional upon c. The corresponding residuals are denoted $\hat{\varepsilon}_t(c) = y_t - \hat{\phi}(c)' x_t(c)$ with variance $\hat{\sigma}^2(c) = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t(c)^2$. The least squares estimate of c

can be obtained by minimizing this residual variance, that is

$$\hat{c} = \underset{c \in C}{\operatorname{argmin}} \ \hat{\sigma}^2(c), \tag{8.15}$$

where C denotes the set of all allowable threshold values. The final estimates of the autoregressive parameters are given by $\hat{\phi} = \hat{\phi}(\hat{c})$, while the residual variance is estimated as $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{c})$.

The set of allowable threshold values C in (8.15) should be such that each regime contains enough observations for the estimator defined above to produce reliable estimates of the autoregressive parameters. A popular choice for C is to require that each regime contains at least a (pre-specified) fraction π_0 of the observations, that is,

$$C = \{c \mid q_{([\pi_0 T])} \leq c \leq q_{([(1-\pi_0)T])}\}, \tag{8.16}$$

where $q_{(1)}, \ldots, q_{(T)}$ denote the order statistics of the threshold variable q_t, $q_{(1)} \leq \cdots \leq q_{(T)}$, and $[\cdot]$ denotes integer part. A safe choice for π_0 appears to be 0.15.

The minimization problem (8.15) can be solved by means of direct search. It suffices to compute the residual variance $\hat{\sigma}^2(c)$ only for threshold values equal to the order statistics of q_t, that is, for $c = q_{(i)}$ for each i such that $q_{(i)} \in C$. This follows from the observation that the value of $\hat{\sigma}^2(c)$ does not change as c is varied between two consecutive order statistics, as no observations move from one regime to the other in this case.

Chan (1993) demonstrates that the LS estimator of the threshold $\hat{c}$ is consistent at rate T and asymptotically independent of the other parameter estimates. Chan (1993) also shows that the asymptotic distribution of $\hat{c}$ depends upon many nuisance parameters, for instance the true regression parameters ϕ. Using an alternative approach, Hansen (1997) derives a limiting distribution for $\hat{c}$ that is free of nuisance parameters apart from a scale parameter. The estimates of the autoregressive parameters are consistent at the usual rate of $\sqrt{T}$ and asymptotically normal.

Confidence intervals

The asymptotic distribution of the threshold is available in closed-form, so in principle it could be used to construct confidence intervals for c. However, this requires estimation of the scale parameter in the distribution, which appears to be quite cumbersome. Therefore, Hansen (1997) recommends an alternative approach, which is based on inverting the likelihood ratio test statistic to test the hypothesis that the threshold is equal to some specific value c_0, given by

$$LR(c_0) = T\left(\frac{\hat{\sigma}^2(c_0) - \hat{\sigma}^2(\hat{c})}{\hat{\sigma}^2(\hat{c})}\right). \tag{8.17}$$

Notice that $LR(\hat{c}) = 0$. The $100 \cdot \alpha\%$ confidence interval for the threshold is given by the set $\widehat{C}_\alpha$ consisting of those values of c for which the null hypothesis is not rejected at significance level α. That is,

$$\widehat{C}_\alpha = \{c : LR(c) \leq z(\alpha)\}, \tag{8.18}$$

where $z(\alpha)$ is the $100 \cdot \alpha$ percentile of the asymptotic distribution of the LR statistic. These percentiles are given in Hansen (1997, Table 1) for various values of α. The set $\widehat{C}_\alpha$ provides a valid confidence region as the probability that the true threshold value is contained in $\widehat{C}_\alpha$ approaches α as the sample size n becomes large. An easy graphical method to obtain the region $\widehat{C}_\alpha$ is to plot the LR statistic (8.17) against c and draw a horizontal line at $z(\alpha)$. All points for which the value of the statistic is below the line are included in $\widehat{C}_\alpha$.

The estimates of the autoregressive parameters ϕ_1 and ϕ_2 are asymptotically normal distributed. Hence, one might proceed as usual and construct an asymptotic 95% confidence interval for $\phi_{2,1}$, for example, as $(\hat{\phi}_{2,1} - 1.96\hat{\sigma}_{\phi_{2,1}}, \hat{\phi}_{2,1} + 1.96\hat{\sigma}_{\phi_{2,1}})$, where $\hat{\sigma}_{\phi_{2,1}}$ is the estimated standard error of $\phi_{2,1}$. Hansen (1997) shows that the confidence intervals that are obtained in this way do not yield good finite sample approximations. He therefore recommends an alternative procedure, in which a 95% confidence interval for ϕ_1 and ϕ_2 is computed for each value of c in the set $\widehat{C}_\alpha$, and the union of these intervals is taken as the confidence interval for ϕ_1 and ϕ_2. Some simulation evidence suggests that $\alpha = 0.8$ is a reasonable confidence level for the set $\widehat{C}_\alpha$ in this case.

Choosing the threshold variable

So far, we have implicitly assumed that the threshold variable q_t, which defines the regime that occurs at any given point in time, is known. In practice, the appropriate threshold variable is of course unknown and an important question is how it can be determined. In the context of univariate time series models we might restrict attention to lagged endogenous variables y_{t-d} for positive integers d as candidate threshold variables. It turns out that in this case d can be estimated along with the other parameters in the model, by performing the above calculations for various choices of d, say $d \in \{1, \ldots, d^*\}$ for some upper bound d^*, and estimate d as the value that minimizes the residual variance.

An alternative way to interpret this procedure is that effectively the grid search in (8.15) is augmented with a search over d, that is, the minimization problem becomes

$$(\hat{c}, \hat{d}) = \underset{c \in C, d \in D}{\operatorname{argmin}} \ \hat{\sigma}^2(c, d), \tag{8.19}$$

where $D = \{1, \ldots, d^*\}$ and the notation $\hat{\sigma}^2(c, d)$ is used to indicate that the estimate of the residual variance now depends on d as well as on c. As the parameter space for d is

discrete, the least squares estimate $\hat{d}$ is super-consistent and d can be treated as known when computing confidence intervals for the remaining parameters, for example.

If one wants to allow for an exogenous threshold variable q_t, a similar procedure can be followed. In that case, the TAR model is estimated with different candidate threshold variables, and the variable that renders the best fit is selected as the most appropriate one. See Chen (1995) for alternative methods of selecting the threshold variable.

8.2.2 Estimation of STAR models

Estimation of the parameters in the STAR model (8.11) is a relatively straightforward application of nonlinear least squares [NLS], that is, the parameters $\theta = (\phi_1', \phi_2', \gamma, c)'$ can be estimated as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \; Q_T(\theta) = \underset{\theta}{\operatorname{argmin}} \; \sum_{t=1}^{T} [y_t - F(x_t; \theta)]^2 , \tag{8.20}$$

where $F(x_t; \theta)$ is the skeleton of the model, that is,

$$F(x_t; \theta) \equiv \phi_1' x_t (1 - G(q_t; \gamma, c)) + \phi_2' x_t G(q_t; \gamma, c).$$

Under the additional assumption that the errors ε_t are normally distributed, NLS is equivalent to maximum likelihood. Otherwise, the NLS estimates can be interpreted as quasi maximum likelihood estimates. Under certain regularity conditions, which are discussed in White and Domowitz (1984), Gallant (1987) and Pötscher and Prucha (1997), among others, the NLS estimates are consistent and asymptotically normal, that is,

$$\sqrt{T}(\hat{\theta} - \theta_0) \to N(0, C), \tag{8.21}$$

where θ_0 denotes the true parameter values. The asymptotic covariance-matrix C of $\hat{\theta}$ can be estimated consistently as $\widehat{A}_T^{-1} \widehat{B}_T \widehat{A}_T^{-1}$, where $\widehat{A}_T$ is the Hessian evaluated at $\hat{\theta}$

$$\widehat{A}_T = -\frac{1}{T} \sum_{t=1}^{T} \nabla^2 q_t(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \left(\nabla F(x_t; \hat{\theta}) \nabla F(x_t; \hat{\theta})' - \nabla^2 F(x_t; \hat{\theta}) \hat{\varepsilon}_t \right), \tag{8.22}$$

with $q_t(\hat{\theta}) = [y_t - F(x_t; \hat{\theta})]^2$, $\nabla F(x_t; \hat{\theta}) = \partial F(x_t; \hat{\theta}) / \partial \theta$, and $\widehat{B}_T$ is the outer product of the gradient

$$\widehat{B}_T = \frac{1}{T} \sum_{t=1}^{T} \nabla q_t(\hat{\theta}) \nabla q_t(\hat{\theta})' = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \nabla F(x_t; \hat{\theta}) \nabla F(x_t; \hat{\theta})'. \tag{8.23}$$

The estimation can be performed using any conventional nonlinear optimization procedure, see Quandt (1983), Hamilton (1994, Sec. 5.7) and Hendry (1995, Appendix A5) for surveys. Issues that deserve particular attention are the choice of starting values for the optimization algorithm, concentrating the sum of squares function and the estimate of the smoothness parameter γ in the transition function.

Starting values

Obviously, the burden put on the optimization algorithm can be alleviated by using good starting values. Note that for fixed values of the parameters in the transition function, γ and c, the STAR model is linear in the autoregressive parameters ϕ_1 and ϕ_2, similar to the TAR model. Thus, conditional upon γ and c, estimates of $\phi = (\phi_1', \phi_2')'$ can be obtained by OLS as

$$\hat{\phi}(\gamma, c) = \left(\sum_{t=1}^{T} x_t(\gamma, c) x_t(\gamma, c)' \right)^{-1} \left(\sum_{t=1}^{T} x_t(\gamma, c) y_t \right), \tag{8.24}$$

where $x_t(\gamma, c) = (x_t'(1 - G(q_t; \gamma, c)), x_t' G(q_t; \gamma, c))'$ and the notation $\phi(\gamma, c)$ is used to indicate that the estimate of ϕ is conditional upon γ and c. The corresponding residuals can be computed as $\hat{\varepsilon}_t = y_t - \hat{\phi}(\gamma, c)' x_t(\gamma, c)$ with associated variance $\hat{\sigma}^2(\gamma, c) = T^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t^2(\gamma, c)$. A convenient method to obtain sensible starting values for the nonlinear optimization algorithm then is to perform a two-dimensional grid search over γ and c and select those parameter estimates which render the smallest estimate for the residual variance $\hat{\sigma}^2(\gamma, c)$.

Concentrating the sum of squares function

As suggested by Leybourne *et al.* (1998), another way to simplify the estimation problem is to concentrate the sum of squares function. Due to the fact that the STAR model is linear in the autoregressive parameters for given values of γ and c, the sum of squares function $Q_T(\theta)$ can be concentrated with respect to ϕ_1 and ϕ_2 as

$$Q_T(\gamma, c) = \sum_{t=1}^{T} (y_t - \phi(\gamma, c)' x_t(\gamma, c))^2. \tag{8.25}$$

This reduces the dimensionality of the NLS estimation problem considerably, as the sum of squares function as given in (8.25) needs to be minimized with respect to the two parameters γ and c only.

The estimate of γ

It turns out to be notoriously difficult to obtain a precise estimate of the smoothness parameter γ. One reason for this is that for large values of γ, the shape of the logistic function (8.5) changes only little. Hence, to obtain an accurate estimate of γ one needs many observations in the immediate neighborhood of the threshold c. As this is typically not the case, the estimate of γ is rather imprecise in general and often appears to be insignificant when judged by its t-statistic. This estimation problem is discussed in a more general context in Bates and Watts (1988, p. 87). The main point to be taken is that insignificance of the estimate of γ should not be interpreted as evidence against the presence of STAR-type nonlinearity. This should be assessed by means of different diagnostics, some of which are discussed below.

8.2.3 Estimation of the Markov-Switching model

The parameters in the MSW model can be estimated using maximum likelihood techniques. However, due to the fact that the Markov-process s_t is not observed, the estimation problem is highly nonstandard. The aim of the estimation procedure in fact is not only to obtain estimates of the parameters in the autoregressive models in the different regimes and the probabilities of transition from one regime to the other, but also to obtain an estimate of the state that occurs at each point of the sample or, more precisely, the probabilities with which each state occurs at each point in time.

Consider the two-regime MSW model with an AR(p) specification in both regimes,

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1} y_{t-1} + \cdots + \phi_{1,p} y_{t-p} + \varepsilon_t & \text{if } s_t = 1, \\ \phi_{2,0} + \phi_{2,1} y_{t-1} + \cdots + \phi_{2,p} y_{t-p} + \varepsilon_t & \text{if } s_t = 2, \end{cases} \tag{8.26}$$

or in shorthand notation,

$$y_t = \phi_{s_t,0} + \phi_{s_t,1} y_{t-1} + \cdots + \phi_{s_t,p} y_{t-p} + \varepsilon_t. \tag{8.27}$$

Under the additional assumption that the ε_t in (8.26) are normally distributed (conditional upon the history $\mathcal{Y}_{t-1}$), the density of y_t conditional on the regime s_t and the history $\mathcal{Y}_{t-1}$ is a normal distribution with mean $\phi_{s_t,0} + \phi_{s_t,1} y_{t-1} + \cdots + \phi_{s_t,p} y_{t-p}$ and variance σ^2,

$$f(y_t | s_t = j, \mathcal{Y}_{t-1}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(y_t - \phi_j' x_t)^2}{2\sigma^2} \right\}, \tag{8.28}$$

where again $x_t = (1, y_{t-1}, \ldots, y_{t-p})'$, $\phi_j = (\phi_{j,0}, \phi_{j,1}, \ldots, \phi_{j,p})'$ for $j = 1, 2$, and θ is a vector that contains all parameters in the model, $\theta = (\phi_1', \phi_2', p_{11}, p_{22}, \sigma^2)'$. Notice that the parameters p_{11} and p_{22} completely define all transition probabilities because, for example, $p_{12} = 1 - p_{11}$. Given that the state s_t is unobserved, the conditional

log likelihood for the t-th observation $l_t(\theta)$ is given by the log of the density of y_t conditional only upon the history $\mathcal{Y}_{t-1}$, that is, $l_t(\theta) = \ln f(y_t|\mathcal{Y}_{t-1};\theta)$. The density $f(y_t|\mathcal{Y}_{t-1};\theta)$ can be obtained from the joint density of y_t and s_t as follows,

$$\begin{aligned} f(y_t|\mathcal{Y}_{t-1};\theta) &= f(y_t, s_t = 1|\mathcal{Y}_{t-1};\theta) + f(y_t, s_t = 2|\mathcal{Y}_{t-1};\theta) \\ &= \sum_{j=1}^{2} f(y_t|s_t = j, \mathcal{Y}_{t-1};\theta) \cdot P(s_t = j|\mathcal{Y}_{t-1};\theta), \end{aligned} \tag{8.29}$$

where the second equality follows directly from the basic law of conditional probability, which states that the joint probability of two events A and B, $P(A \text{ and } B)$, is equal to $P(A|B)P(B)$.

In order to be able to compute the density (8.29), we obviously need to quantify the conditional probabilities of being in either regime given the history of the process, $P(s_t = j|\mathcal{Y}_{t-1};\theta)$. In fact, it turns out in order to develop the maximum likelihood estimates of the parameters in the model, three different estimates of the probabilities of each of the regimes occurring at time t are needed: estimates of the probability that the process is in regime j at time t given all observations up to time $t-1$, given all observations up to and including time t, and given all observations in the entire sample. These estimates usually are called the *forecast*, *inference*, and *smoothed inference* of the regime probabilities.

Intuitively, if the regime that occurs at time $t-1$ were known and included in the information set $\mathcal{Y}_{t-1}$, the optimal *forecasts* of the regime probabilities simply are equal to the transition probabilities of the Markov process s_t. More formally,

$$\hat{\xi}_{t|t-1} = P \cdot \xi_{t-1}, \tag{8.30}$$

where $\hat{\xi}_{t|t-1}$ denotes the 2×1 vector containing the conditional probabilities of interest, that is, $\hat{\xi}_{t|t-1} = (P(s_t = 1|\mathcal{Y}_{t-1};\theta), P(s_t = 2|\mathcal{Y}_{t-1};\theta))'$, $\xi_{t-1} = (1, 0)'$ if $s_{t-1} = 1$ and $\xi_{t-1} = (0, 1)'$ if $s_{t-1} = 2$, and P is the matrix containing the transition probabilities,

$$P = \begin{pmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{pmatrix}. \tag{8.31}$$

In practice the regime at time $t-1$ is unknown, as it is unobservable. The best one can do is to replace ξ_{t-1} in (8.30) by an estimate of the probabilities of each regime occurring at time $t-1$ conditional upon all information up to and including the observation at $t-1$ itself. Denote the 2×1 vector containing the optimal *inference* concerning the regime probabilities as $\hat{\xi}_{t-1|t-1}$. Given a starting value $\hat{\xi}_{1|0}$ and values of the parameters contained in θ, one can compute the optimal forecast and inference

for the conditional regime probabilities by iterating on the pair of equations

$$\hat{\xi}_{t|t} = \frac{\hat{\xi}_{t|t-1} \odot \boldsymbol{f}_t}{\mathbf{1}'(\hat{\xi}_{t|t-1} \odot \boldsymbol{f}_t)} \tag{8.32}$$

$$\hat{\xi}_{t+1|t} = P \cdot \hat{\xi}_{t|t}, \tag{8.33}$$

for $t = 1, \ldots, n$, where $\boldsymbol{f}_t$ denotes the vector containing the conditional densities (8.28) for the two regimes, $\mathbf{1}$ is a 2×1 vector of ones and the symbol $\odot$ indicates element-by-element multiplication. The necessary starting values $\hat{\xi}_{1|0}$ can either be taken to be a fixed vector of constants which sum to unity, or can be included as separate parameters that need to be estimated. See Hamilton (1994, p.693) for an intuitive explanation of why this algorithm works.

Finally, let $\hat{\xi}_{t|T}$ denote the vector which contains the *smoothed inference* on the regime probabilities, that is, estimates of the probability that regime j occurs at time t given all available observations, $P(s_t = j|\mathcal{Y}_T; \theta)$. Kim (1993) develops an algorithm to obtain these regime probabilities from the conditional probabilities $\hat{\xi}_{t|t}$ and $\hat{\xi}_{t+1|t}$ given by (8.32) and (8.33). The smoothed inference on the regime probabilities at time t is computed as

$$\hat{\xi}_{t|T} = \hat{\xi}_{t|t} \odot (P'[\hat{\xi}_{t+1|T} \div \hat{\xi}_{t+1|t}]), \tag{8.34}$$

where $\div$ indicates element-by-element division. The algorithm runs backwards through the sample, that is, starting with $\hat{\xi}_{T|T}$ from (8.32) one applies (8.34) for $t = T-1, T-2, \ldots, 1$. For more details we refer to Kim (1993).

Returning to (8.32), notice that the denominator of the right-hand side expression actually is the conditional log likelihood for the observation at time t as given in (8.29), which follows directly from the definitions of $\hat{\xi}_{t|t-1}$ and $\boldsymbol{f}_t$. As shown in Hamilton (1990), the maximum likelihood estimates of the transition probabilities are given by

$$\hat{p}_{ij} = \frac{\sum_{t=2}^{T} P(s_t = j, s_{t-1} = i|\mathcal{Y}_T; \hat{\theta})}{\sum_{t=2}^{T} P(s_{t-1} = i|\mathcal{Y}_T; \hat{\theta})}, \tag{8.35}$$

where $\hat{\theta}$ denotes the maximum likelihood estimates of θ. It is also shown in Hamilton (1990) that these satisfy the first-order conditions

$$\sum_{t=1}^{T} (y_t - \hat{\phi}_j' x_t) x_t P(s_t = j|\mathcal{Y}_T; \hat{\theta}) = 0, \quad j = 1, 2, \tag{8.36}$$

and

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{2} (y_t - \hat{\phi}_j' x_t)^2 P(s_t = j|\mathcal{Y}_T; \hat{\theta}). \tag{8.37}$$

Notice that (8.36) implies that $\hat{\phi}_j$ is the estimate corresponding to a weighted least squares regression of y_t on x_t, with weights given by the square root of the smoothed probability of regime j occurring. Hence, the estimates $\hat{\phi}_j$ can be obtained as

$$\hat{\phi}_j = \left(\sum_{t=1}^{T} x_t(j) x_t(j)' \right)^{-1} \left(\sum_{t=1}^{T} x_t(j) y_t(j) \right), \tag{8.38}$$

where

$$y_t(j) = y_t \sqrt{P(s_t = j | \mathcal{Y}_T; \hat{\theta})},$$
$$x_t(j) = x_t \sqrt{P(s_t = j | \mathcal{Y}_T; \hat{\theta})}.$$

Finally, the ML estimate of the residual variance is obtained using (8.37) as the mean of the squared residuals from the two WLS regressions.

Putting all the above elements together suggests the following iterative procedure to estimate the parameters of the MSW model. Given starting values for the parameter vector $\hat{\theta}^{(0)}$, first compute the smoothed regime probabilities using (8.32), (8.33) and (8.34). Next, the smoothed regime probabilities $\hat{\xi}_{t|T}$ are combined with the initial estimates of the transition probabilities $\hat{p}_{ij}^{(0)}$ to obtain new estimates of the transition probabilities $\hat{p}_{ij}^{(1)}$ from (8.35). Finally, (8.38) and (8.37) can be used to obtain a new set of estimates of the autoregressive parameters and the residual variance. Combined with the new estimates of the transition probabilities, this gives a new set of estimates for all parameters in the model, $\hat{\theta}^{(1)}$. Iterating this procedure renders estimates $\hat{\theta}^{(2)}, \hat{\theta}^{(3)}, \ldots$ and this can be continued until convergence occurs, that is, until the estimates in subsequent iterations are the same.

8.3 Testing for nonlinearity

Perhaps the most important question that needs to be answered when considering regime-switching models is whether the additional regime (relative to the single regime in a linear AR model) adds significantly to explaining the dynamic behavior of the time series y_t. One possible method to address this question is to compare the in-sample fit of the regime-switching model with that of a linear model by means of a formal statistical test. A natural approach then is to take the linear model as the null hypothesis and the regime-switching model as the alternative. The null hypothesis can be expressed as equality of the autoregressive parameters in the two regimes, that is, $H_0 : \phi_1 = \phi_2$, which is tested against the alternative hypothesis $H_1 : \phi_{1,i} \neq \phi_{2,i}$ for at least one $i \in \{0, \ldots, p\}$.

The statistical tests which take either one of the three regime-switching models as the alternative all suffer from the problem of so-called *unidentified nuisance parameters* under the null hypothesis. Intuitively, this means to say that the nonlinear model contains certain parameters that are not restricted under the null hypothesis but which are not present in the linear model. In both the TAR and STAR models, the threshold c is such an unidentified nuisance parameter, whereas in the STAR model, the smoothness parameter γ is one as well. In the MSW model, the unidentified nuisance parameters are p_{11} and p_{22}, which define the transition probabilities between the two regimes. The main consequence of the presence of such parameters is that the conventional statistical theory cannot be applied to obtain the (asymptotic) distribution of the test statistics, see Davies (1977,1987) and Hansen (1996), among others. Instead, the test statistics tend to have a non-standard distribution, for which an analytical expression often is not available. This implies that critical values have to be determined by means of simulation methods.

8.3.1 Testing the TAR model

A solution to the above-mentioned identification problem when testing linearity against the alternative of a TAR model is to use the estimates of the nonlinear model to define a likelihood ratio or F-statistic, that is

$$F(\hat{c}) = T\left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\right), \tag{8.39}$$

where $\tilde{\sigma}^2$ is an estimate of the residual variance under the null hypothesis of linearity, $\tilde{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\tilde{\varepsilon}_t^2$ with $\tilde{\varepsilon}_t = y_t - \hat{\phi}' x_t$, and $\hat{\sigma}^2$ is defined just below (8.15). Notice that the statistic (8.39) is a monotonic transformation of $\hat{\sigma}^2$, in the sense that $F(\hat{c})$ always increases when $\hat{\sigma}^2$ decreases and vice versa. As $\hat{c}$ minimizes the residual variance over the set C of allowable threshold values, $F(\hat{c})$ is equivalent to the supremum over this set C of the pointwise test statistic $F(c)$,

$$F(\hat{c}) = \sup_{c \in C} F(c), \tag{8.40}$$

where

$$F(c) = T\left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2(c)}{\hat{\sigma}^2(c)}\right), \tag{8.41}$$

where $\hat{\sigma}^2(c)$ is defined just below (8.24).

Given that we are testing $p+1$ restrictions, the pointwise $F(c)$ statistic has an asymptotic χ^2 distribution with $p+1$ degrees of freedom. The test statistic (8.40)

therefore is the supremum of a number of dependent statistics, each of which follows an asymptotic χ^2 distribution. It then follows that the distribution of $F(\hat{c})$ itself is non-standard. Because the exact form of the dependence between the different $F(c)$'s is difficult to analyze or characterize, critical values are most easily determined by means of simulation, see Hansen (1997,1998) for more details.

Exercise 8.5

8.3.2 Testing the STAR model

Testing linearity against the STAR model offers the opportunity to illustrate the problems of unidentified nuisance parameters in a different manner, in the sense that more than one restriction can be used to make the STAR model collapse to a linear AR model. Besides equality of the AR parameters in the two regimes, $H_0 : \phi_1 = \phi_2$, the null hypothesis of linearity can alternatively be expressed as $H_0' : \gamma = 0$. If $\gamma = 0$, the logistic function (8.5) is equal to 0.5 for all values of y_{t-1} and the STAR model reduces to an AR model with parameters $(\phi_1 + \phi_2)/2$. Whichever formulation of the null hypothesis is used, the model contains unidentified parameters. In case H_0 is used to characterize the null hypothesis of linearity, the parameters γ and c in the transition function are the unidentified nuisance parameters. In case H_0' is used, the threshold c and the parameters ϕ_1 and ϕ_2 are. To see the latter, note that under H_0', ϕ_1 and ϕ_2 can take any value as long as their average remains the same.

The approach that has been used in this case to solve the identification problem is slightly different from the one discussed above for the TAR model. It turns out that in the case of testing against the alternative of a STAR model it is feasible to use a Lagrange Multiplier [LM] statistic which has an asymptotic χ^2 distribution. The main advantage of the ability to use this statistic is that it is not necessary to estimate the model under the alternative hypothesis.

Consider again the STAR model as given in (8.11), and rewrite this as

$$y_t = \frac{1}{2}(\phi_1 + \phi_2)' x_t + (\phi_2 - \phi_1)' x_t G^*(q_t; \gamma, c) + \varepsilon_t, \tag{8.42}$$

where $G^*(q_t; \gamma, c) = G(q_t; \gamma, c) - 1/2$. Notice that under the null hypothesis $\gamma = 0$, $G^*(q_t, 0, c) = 0$. Luukkonen *et al.* (1988) suggest to approximate the function $G^*(q_t, \gamma, c)$ with a first order Taylor approximation around $\gamma = 0$, that is,

$$T_1(q_t; \gamma, c) \approx G^*(q_t; 0, c) + \gamma \left. \frac{\partial G^*(q_t; \gamma, c)}{\partial \gamma} \right|_{\gamma=0} = \frac{1}{4}\gamma(q_t - c), \tag{8.43}$$

where we have used the fact that $G^*(q_t; 0, c) = 0$. After substituting $T_1(\cdot)$ for $G_1^*(\cdot)$ in (8.42) and rearranging terms this gives the auxiliary regression model

$$y_t = \beta_0' x_t + \beta_1' x_t q_t + \eta_t, \tag{8.44}$$

where $\beta_j = (\beta_{j,0}, \beta_{j,1}, \ldots, \beta_{j,p})'$, $j = 0, 1$. The relationships between the parameters in the auxiliary regression model (8.44) and the parameters in the STAR model (8.42) can be shown to be such that the restriction $\gamma = 0$ implies $\beta_{1,i} = 0$ for $i = 0, 1, \ldots, p$. Hence testing the null hypothesis $H_0' : \gamma = 0$ in (8.42) is equivalent to testing the null hypothesis $H_0'' : \beta_1 = 0$ in (8.44). This null hypothesis can be tested by a standard F test in a straightforward manner. Under the null hypothesis of linearity, the test statistic has a χ^2 distribution with $1 + p$ degrees of freedom asymptotically.

In small samples, the usual recommendation is to use F-versions of the LM test statistics as these have better size and power properties. The F-version of the test statistic based on (8.44) can be computed as follows:

1. Estimate the model under the null hypothesis of linearity by regressing y_t on x_t. Compute the residuals $\tilde{\varepsilon}_t$ and the sum of squared residuals $SSR_0 = \sum_{t=1}^{T} \tilde{\varepsilon}_t^2$.
2. Estimate the auxiliary regression of $\tilde{\varepsilon}_t$ on x_t and $x_t q_t$ and compute the sum of squared residuals from this regression SSR_1.
3. The LM test statistic can be computed as

$$LM = \frac{(SSR_0 - SSR_1)/(1 + p)}{SSR_1/(T - 2p - 2)}, \tag{8.45}$$

and is approximately F distributed with $1 + p$ and $T - 2p - 2$ degrees of freedom under the null hypothesis.

Choosing the transition variable

Teräsvirta (1994) suggests that the LM-type test (8.45) can also be used to select the appropriate transition variable in the STAR model. The statistic is computed for several candidate transition variables and the one for which the p-value of the test is smallest is selected as the true transition variable. The rationale behind this procedure is that the test should have maximum power in case the alternative model is correctly specified, that is, if the correct transition variable is used. Simulation results in Teräsvirta (1994) suggest that this approach works quite well, at least in a univariate setting.

Illustration

We illustrate the usefulness of non-linear models by estimation and testing a (SE)TAR and LSTAR model, using the quarterly (seasonally adjusted) growth rate of US

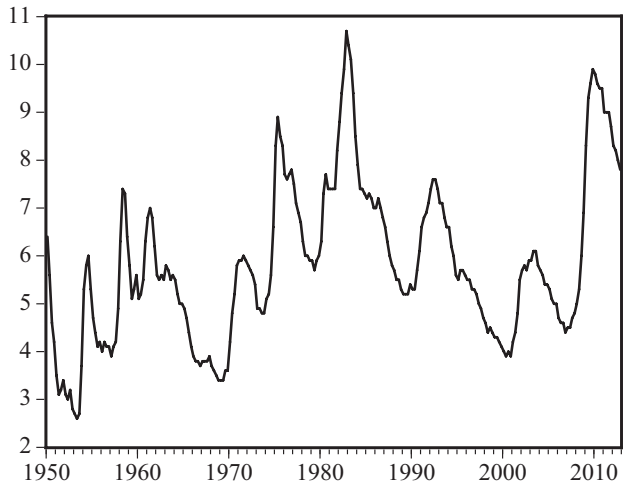


Figure 8.2: Quarterly seasonally adjusted US unemployment rates, 1950.1–2012.4.

unemployment during the period 1953.2–2012.4 (239 observations). Figure 8.2 again shows this time series. As discussed earlier in 2.5, the graph shows non-linear behavior as it increases quite rapidly, (due to economic recessions), while it decreases much more slowly in times on expansions. Let us start with a linear ARMA model. The EACF of the series suggests an AR(2) model, which results in

$$\begin{underset{(0.019)}{\Delta y_t = 0.0088} + \underset{(0.064)}{0.738\,\Delta y_{t-1}} - \underset{(0.064)}{0.171\,\Delta y_{t-2}} + \hat{\varepsilon}_t.} \tag{8.46}$$

with $\hat{\sigma}_\varepsilon = 0.295$, and AIC and BIC values of 0.411 and 0.416 respectively.

Next, we will estimate a SETAR model, where we use the lagged 12-month difference of the unemployment rate as threshold variable. That is, $q_t = \Delta_{12} y_{t-1}$. In addition, we test this model against linearity by applying the SupF test of (8.40) with a trimming fraction of $\lambda = 15\%$ and estimate the final TAR model where $\hat{c}$ corresponds with the maximum value of $F(c)$. Figure 8.3 shows the different Wald statistics for ascending values of $\Delta_{12} y_{t-1}$, combined with the asymptotical critical values of the test in this case (with $k = 3$ restrictions being tested and with $\lambda = 0.15$). The maximum Wald statistic is 24.15, which corresponds with $\hat{c} = 1.50$. Given this graph, we reject the linear AR(2) model, even at a 1% significance level. Note that $\hat{c}$ is optimal in the sense that this value minimizes the residual variance over the whole set of possible threshold

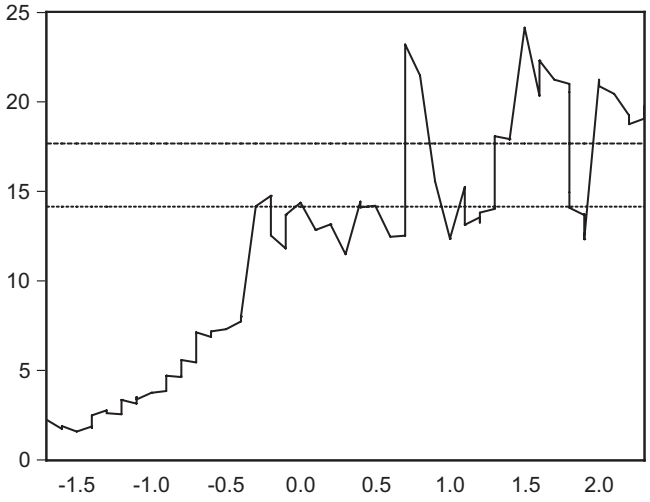


Figure 8.3: Sequence of Wald statistics for testing of threshold nonlinearity in case of an SETAR model on US unemployment rates y_t, for different values of the lagged 12-month difference $q_t = \Delta_{12} y_{t-1}$. The horizontal lines represent the critical values corresponding to the 95% and 99% significance level.

values. We end up with the following model:

$$\begin{aligned}\Delta y_t = &\underset{(0.022)}{(0.067} + \underset{(0.089)}{0.804\Delta y_{t-1}} - \underset{(0.096)}{0.017\Delta y_{t-2}}) \times I_t[\Delta_{12} y_{t-1} < 1.50] \\ &+ \underset{(0.037)}{(-0.105} + \underset{(0.086)}{0.662\Delta y_{t-1}} - \underset{(0.087)}{0.141\Delta y_{t-2}}) \times (1 - I_t[\Delta_{12} y_{t-1} < 1.50]) + \hat{\varepsilon}_t \end{aligned} \quad (8.47)$$

with $\hat{\sigma}_\varepsilon = 0.284$, and AIC and BIC values of 0.346 and 0.434. The residual standard deviation of the TAR model decreases with almost 4%. The AIC therefore decreases, however the three additional parameter causes the BIC to be higher. The estimation results above shows a different average of the change in unemployment rate in both regimes and less persistence of unemployment growth in the second regime compared to first regime. In addition, the AR(2) term, which was significant in the linear model, does not play a role anymore.

Finally, let us consider a LSTAR model, in order to see whether there is a gradual change from the one to the other regime. We minimize the concentrated sum of squared

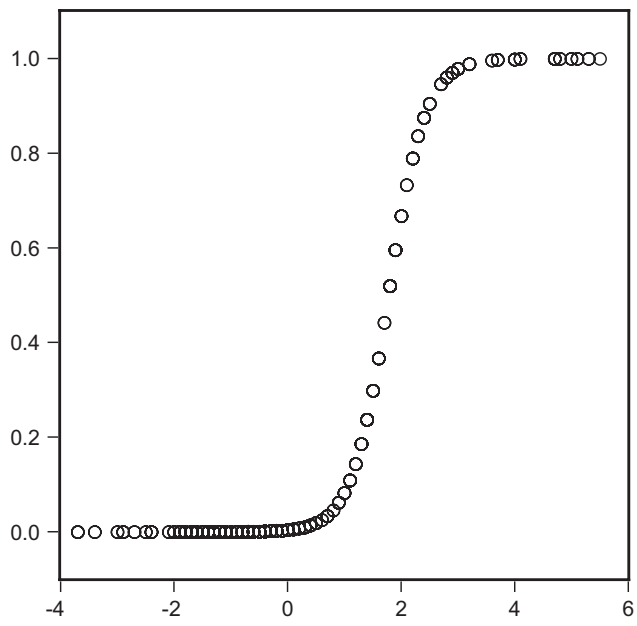


Figure 8.4: Transition function in LSTAR model for quarterly seasonally adjusted US unemployment rate against the transition variable $\Delta_{12} y_{t-1}$.

function and obtain the following estimates:

$$\begin{aligned}\Delta y_t = &(0.064 + 0.815\Delta y_{t-1} - 0.012\Delta y_{t-2}) \times G(\Delta_{12} y_{t-1}; \gamma, c) \\ &\quad (0.025) \quad (0.094) \quad (0.113) \\ &+ (-0.124 + 0.572\Delta y_{t-1} - 0.128\Delta y_{t-2}) \times (1 - G(\Delta_{12} y_{t-1}; \gamma, c) + \hat{\varepsilon}_t \\ &\quad (0.058) \quad (0.118) \quad (0.107) \end{aligned} \tag{8.48}$$

$$G(\Delta_{12} y_{t-1}; \hat{\gamma}, \hat{c}) = (1 + \exp[-3.110(\Delta_{12} y_{t-1} - 1.775)])^{-1} \tag{8.49}$$
$$(2.785 \qquad (0.0319)$$

with $\hat{\sigma}_\varepsilon = 0.284$, and AIC and BIC values of 0.358 and 0.475. The two extra parameters compared to the TAR model – where we have treated c as fixed – increases the AIC and BIC respectively. Note the similar estimation results of the constant and AR(1) term in both regimes. Figure 8.4 shows the transition function $G(\Delta_{12} y_{t-1}; \hat{\gamma}, \hat{c})$. The estimates $\hat{\gamma}$ and $\hat{c}$ are such that the change of the logistic function $G(\Delta_{12} y_{t-1}; \hat{\gamma}, \hat{c})$ from 0 to 1 takes place for values of $\Delta_{12} y_{t-1}$ between 0.3 and 3.20. Hence there is some evidence from a gradual change between the two regimes. We finally test the STAR model by

computing the LM statistic of (8.45). This gives a value of 4.90, which is significant at the 1% level according tot the F distribution with 4 and 231 degrees of freedom.

8.3.3 Testing the Markov-Switching model

When assessing the relevance of the MSW model, a natural approach is to use a Likelihood Ratio [LR] statistic, which tests the null hypothesis of linearity against the alternative of a MSW model, that is, $H_0 : \phi_1 = \phi_2$ is tested by means of the test statistic

$$LR_{MSW} = \mathcal{L}_{MSW} - \mathcal{L}_{AR}, \tag{8.50}$$

where $\mathcal{L}_{MSW}$ and $\mathcal{L}_{AR}$ are the values of the log likelihood functions corresponding to the MSW and AR models, respectively. As noted in the introduction to this section, the parameters p_{11} and p_{22} defining the transition probabilities in the MSW model are unidentified nuisance parameters under the null hypothesis. As shown by Hansen (1992), the LR statistic (8.50) has a nonstandard distribution which cannot be characterized analytically. Critical values to determine the significance of the test statistic therefore have to be determined by means of simulation. The basic structure of such a simulation experiment is that one generates a large number of artificial time series y_t^* according to the model that holds under the null hypothesis. Next, one estimates both AR and MSW models for each artificial time series and computes the corresponding LR statistic LR_{MSW}^* according to (8.50). These test statistics might be used to obtain an estimate of the complete distribution of the test statistic under the null hypothesis, or simply to compute the p-value of the LR statistic for the true time series, which is given by the fraction of artificial samples for which LR_{MSW}^* exceeds the observed LR_{MSW}. Given that the estimation of the MSW model can be rather time-consuming, this procedure demands a considerable amount of computing time.

8.4 Diagnostic checking

In this section we discuss several diagnostic tests which can be used to evaluate estimated regime-switching models. First and foremost, one might subject the residuals to a battery of diagnostic tests, comparable to the usual practice in the Box-Jenkins approach in linear time series modeling, as described in Chapter 3. It turns out however, that not all test statistics that have been developed in the context of ARMA models are applicable to the residuals from nonlinear models as well. The test for normality of the residuals given in (3.106) is an example of a test which remains valid, while the Ljung-Box test statistic (3.99) is an example of a test which does not, see Eitrheim and Teräsvirta (1996). The LM approach to testing for serial correlation can still be used

however, as shown by Eitrheim and Teräsvirta (1996) and discussed in some detail below.

8.4.1 Diagnostic tests for TAR and STAR models

In this subsection we discuss three important diagnostic checks for TAR and STAR models, developed by Eitrheim and Teräsvirta (1996).

Testing for serial correlation

Consider the general nonlinear autoregressive model of order p,

$$y_t = F(x_t;\theta) + \varepsilon_t, \tag{8.51}$$

where $x_t = (1, y_{t-1}, \ldots, y_{t-p})'$ as before and the skeleton $F(x_t;\theta)$ is a general nonlinear function of the parameters θ which is at least twice continuously differentiable. An LM-test for q-th order serial dependence in ε_t can be obtained as nR^2, where R^2 is the coefficient of determination from the regression of $\hat{\varepsilon}_t$ on $\hat{z}_t \equiv \partial F(x_t;\hat{\theta})/\partial\theta$ and q lagged residuals $\hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-q}$, where hats indicate that the relevant quantities are estimates under the null hypothesis of serial independence of ε_t. The resulting test statistic is χ^2 distributed with q degrees of freedom asymptotically.

This test statistic is in fact a generalization of the LM-test for serial correlation in an AR(p) model of Breusch and Pagan (1979), which is based on the auxiliary regression (3.100). To understand why, note that for a linear AR(p) model (without an intercept) $F(x_t;\theta) = \sum_{i=1}^{p} \phi_i y_{t-i}$ and $\hat{z}_t = \partial F(x_t;\hat{\theta})/\partial\theta = (y_{t-1}, \ldots, y_{t-p})'$. In case of a STAR model, the skeleton is given by $F(x_t;\theta) = \phi_1' x_t(1 - G(q_t;\gamma,c)) + \phi_2' x_t G(q_t;\gamma,c)$. Hence, in this case $\theta = (\phi_1, \phi_2, \gamma, c)$ and the relevant partial derivatives $\hat{z}_t = \partial F(x_t;\hat{\theta})/\partial\theta$ can be obtained in a straightforward manner, see Eitrheim and Teräsvirta (1996) for details.

The nonlinear function $F(x_t;\theta)$ needs to be twice continuously differentiable for the above approach to be valid. The skeleton of the TAR model does not satisfy this requirement, as it is possibly discontinuous and in no case differentiable at the threshold value. Therefore, the LM-statistic for serial correlation cannot be applied to the residuals from an estimated TAR model. A possible way to circumvent this problem is to approximate the TAR model with a STAR model by setting γ equal to some large but finite value. Recall that in this case the logistic function (8.5) effectively becomes a step function which equals 0 for $y_{t-1} < c$ and 1 for $y_{t-1} > c$. Fixing γ at γ_0, say, the remaining parameters in the STAR model can again be estimated by NLS. When computing the test statistic for residual autocorrelation in this case, the partial derivative of the regression function with respect to γ should be omitted from the auxiliary regression as this parameter is kept fixed.

Testing for remaining nonlinearity

An important question when using nonlinear time series models is whether the proposed model adequately captures all nonlinear features of the time series under investigation. One possible way to examine this is to apply a test for remaining nonlinearity to an estimated model. For the TAR and STAR models, a natural approach is to specify the alternative hypothesis of remaining nonlinearity as the presence of an additional regime. For example, one might want to test the null hypothesis that a two-regime model is adequate against the alternative that a third regime is necessary.

It turns out that only for the STAR model an LM test is available which allows to test this hypothesis without the necessity to estimate the more complicated model. For the TAR model, testing for remaining nonlinearity necessarily involves estimating the multiple regime model. In fact, this is analogous to the situation of testing linearity against a two-regime model, compare the discussion in the introduction to Section 8.3.

For the TAR model, one can essentially apply the methodology described in Section 8.3.1 to each of the two sub-samples defined by the estimated threshold $\hat{c}$, that is, test linearity against the alternative of a two-regime TAR model on the sub-samples for which $q_t \leq \hat{c}$ and $q_t > \hat{c}$ by using the test statistic (8.40). Recall that computing the test involves estimating the two-regime model under the alternative. Hence, it appears that in case the statistics indicate the presence of an additional regime, estimates of the three-regime model are readily available by combining the original estimation results for the two-regime TAR model with those for the two-regime model on the sub-sample for which linearity is rejected. However, in case the true model is indeed a three-regime model, it can be shown that while the estimate of the second threshold $\hat{c}_2$, say, is consistent, the estimate of the first threshold $\hat{c}_1 \equiv \hat{c}$ is not. To obtain a consistent estimate of the first threshold as well, it is necessary to perform a so-called repartitioning step, in which a two-regime TAR model is estimated on the sub-sample defined by $q_t \leq \hat{c}_2$ in case $\hat{c}_1 < \hat{c}_2$ and on the sub-sample defined by $q_t > \hat{c}_2$ in case $\hat{c}_1 > \hat{c}_2$.

Eitrheim and Teräsvirta (1996) develop an LM statistic to test a two-regime STAR model against the alternative of an additive three-regime model which can be written as,

$$y_t = \phi_1' x_t + (\phi_2 - \phi_1)' x_t G_1(q_t; \gamma_1, c_1) + (\phi_3 - \phi_2)' x_t G_2(q_t; \gamma_2, c_2) + \varepsilon_t, \tag{8.52}$$

where both G_1 and G_2 are given by (8.5) and where we assume $c_1 < c_2$ without loss of generality. The null hypothesis of a two-regime model can be expressed as $H_0 : \gamma_2 = 0$. This testing problem suffers from similar identification problems as the problem of testing the null hypothesis of linearity against the alternative of a two-regime STAR model discussed in Section 8.3.2. The solution here is the same as well. The transition function $G_2(q_t; \gamma_2, c_2)$ is replaced by a Taylor approximation around the point $\gamma_2 = 0$.

In case of a third-order approximation, the resulting auxiliary model is given by

$$y_t = \beta_0' x_t + (\phi_2 - \phi_1)' x_t G_1(q_t; \gamma_1, c_1) + \beta_1' x_t q_t + \beta_2' x_t q_t^2 + \beta_3' x_t q_t^3 + \eta_t, \tag{8.53}$$

where the β_j, $j = 0, 1, 2, 3$, are functions of the parameters ϕ_1, ϕ_3, γ_2 and c_2. The null hypothesis $H_0 : \gamma_2 = 0$ in (8.52) translates into $H_0' : \beta_1 = \beta_2 = \beta_3 = 0$ in (8.53). The test statistic can be computed as TR^2 from the auxiliary regression of the residuals obtained from estimating the model under the null hypothesis $\hat{\varepsilon}_t$ on the partial derivatives of the regression function with respect to the parameters in the two-regime model, ϕ_1, ϕ_2, γ_1 and c_1, evaluated under the null hypothesis, and the auxiliary regressors $x_t q_t^j$, $j = 1, 2, 3$. The resulting test statistic has an asymptotic χ^2 distribution with $3(1 + p)$ degrees of freedom.

In the above, it has been implicitly assumed that the additional regime is determined by the same variable q_t as the original two regimes. As discussed previously, one might also consider situations where the regimes are determined by several variables, for example q_{1t} and q_{2t}. In this case, a more natural representation of the STAR model is given by

$$\begin{aligned} y_t = {} & [\phi_1' x_t (1 - G_1(q_{1t})) + \phi_2' x_t G_1(q_{1t})][1 - G_2(q_{2t})] \\ & + [\phi_3' x_t (1 - G_1(q_{1t})) + \phi_4' x_t G_1(q_{1t})] G_2(q_{2t}) + \varepsilon_t, \end{aligned} \tag{8.54}$$

The null hypothesis of a two-regime STAR model can be tested against the alternative of the four-regime model (8.54) by testing $H_0 : \gamma_2 = 0$. The LM test statistic derived by van Dijk and Franses (1999) is similar to the LM-type statistic for testing against a three-regime alternative discussed above.

Testing parameter constancy

An interesting special case of the multiple regime model (8.54) arises if the transition variable in the second transition function G_2 is taken to be time, that is $q_{2t} = t$. This gives rise to a so-called Time-Varying STAR model, which allows for both nonlinear dynamics of the STAR-type and time-varying parameters. This model is discussed in detail in Lundbergh *et al.* (2000). The point of interest here is that by testing the hypothesis $\gamma_2 = 0$ in this case, one tests for parameter constancy in the two-regime STAR model (8.11), against the alternative of smoothly changing parameters. Again this test can be adopted to test for parameter constancy in a TAR model by approximating it with a STAR model with γ_1 fixed at a large value.

8.4.2 Diagnostic tests for Markov-Switching models

Diagnostic checking of estimated Markov-Switching models has been dealt with by Hamilton (1996). He develops tests for residual autocorrelation, heteroskedasticity, misspecification of the Markov process s_t, and omitted explanatory variables. The tests are Lagrange Multiplier type tests, and thus have the attractive property that their computation only requires estimation of the model under the null hypothesis.

The tests make heavy use of the score $h_t(\theta)$, which is defined as the derivative of the log of the conditional density (or likelihood) $f(y_t|\mathcal{Y}_{t-1};\theta)$, given in (8.29), with respect to the parameter vector θ,

$$h_t(\theta) \equiv \frac{\partial \ln f(y_t|\mathcal{Y}_{t-1};\theta)}{\partial \theta}. \tag{8.55}$$

For example, for the 2-regime MSW model in (8.26) it can be shown that

$$\begin{aligned}\frac{\partial \ln f(y_t|\mathcal{Y}_{t-1};\theta)}{\partial \phi_j} = \frac{1}{\sigma^2}(y_t - \phi_j' x_t)x_t \cdot P(s_t = j|\mathcal{Y}_t) + \frac{1}{\sigma^2}\sum_{\tau=2}^{t-1}(y_\tau - \phi_j' x_\tau)x_\tau \\ \cdot(P(s_\tau = j|\mathcal{Y}_t;\theta) - P(s_\tau = j|\mathcal{Y}_{t-1};\theta)),\end{aligned} \tag{8.56}$$

for $j = 1, 2$. Hamilton (1996) describes an algorithm to compute the change in the inference concerning the state the process was in at time τ that is brought about by the addition of y_t, $P(s_\tau = j|\mathcal{Y}_t;\theta) - P(s_\tau = j|\mathcal{Y}_{t-1};\theta)$. The remaining elements of the score in (8.56) can be computed directly after estimation of the model. The same holds for the score with respect to the parameters p_{11} and p_{22}, which determine the transition probabilities of the Markov process s_t, see Hamilton (1996, Eq. (3.12)). By construction, the score evaluated at the ML estimates $\hat{\theta}$ has sample mean zero, $\sum_{t=1}^{T} h_t(\hat{\theta}) = 0$.

One of the possible uses of the conditional scores is to construct standard errors for the ML estimates of θ. To be precise, standard errors are obtained as the square roots of the diagonal elements of the inverse of the outer product of the scores,

$$\sum_{t=1}^{T} h_t(\hat{\theta})h_t(\hat{\theta})'. \tag{8.57}$$

Another use of the scores is to construct Lagrange Multiplier statistics. For example, suppose we want to test that some variables z_t have been omitted from the 2-regime MSW model, that is, we want to test (8.27) against the alternative

$$y_t = \phi_{s_t,0} + \phi_{s_t,1}y_{t-1} + \cdots + \phi_{s_t,p}y_{t-p} + \delta' z_t + \varepsilon_t. \tag{8.58}$$

The score with respect to δ, evaluated under the null hypothesis $H_0 : \delta = 0$ is equal to

$$\left.\frac{\partial \ln f(y_t|\mathcal{Y}_{t-1};\theta)}{\partial \delta}\right|_{\delta=0} = \sum_{j=1}^{2}(y_t - \hat{\phi}_j' x_t)z_t \cdot P(s_t = j|\mathcal{Y}_T;\hat{\theta}), \tag{8.59}$$

where $\hat{\theta}$ are ML estimates of the parameter vector $\theta' = (\phi_1', \phi_2, p_{11}, p_{22}, \delta)$ under the null hypothesis. The LM test statistic to test H_0 is given by

$$n\left(\frac{1}{T}\sum_{t=1}^{T} h_t(\hat{\theta})\right)'\left(\frac{1}{T}\sum_{t=1}^{T} h_t(\hat{\theta})h_t(\hat{\theta})'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} h_t(\hat{\theta})\right), \tag{8.60}$$

and has an asymptotic χ^2 distribution with degrees of freedom equal to the number of variables in z_t.

8.5 Forecasting

Nonlinear time series models may be considered for various purposes. Sometimes the main objective merely is obtaining an adequate description of the dynamic patterns that are present in a particular variable. Very often, however, an additional goal is to employ the model for forecasting future values of the time series. Furthermore, out-of-sample forecasting also can be considered as a way to evaluate estimated regime-switching models. Especially comparison of the forecasts from nonlinear models with those from a benchmark linear model might enable one to determine the added value of the nonlinear features of the model. In this section we discuss the construction of point and interval forecasts from nonlinear models.

Point forecasts

Computing point forecasts from nonlinear models is considerably more involved than computing forecasts from linear models. Consider the case where y_t is described by the general nonlinear autoregressive model of order 1,

$$y_t = F(y_{t-1};\theta) + \varepsilon_t, \tag{8.61}$$

for some nonlinear function $F(y_{t-1};\theta)$. When using a least squares criterion, the optimal point forecasts of future values of the time series are given by their conditional expectations, as discussed in Section 3.5. That is, the optimal h-step ahead forecast of y_{t+h} at time t is given by

$$\hat{y}_{t+h|t} = \mathrm{E}[y_{t+h}|\mathcal{Y}_t], \tag{8.62}$$

where $\mathcal{Y}_t$ again denotes the history of the time series up to and including the observation at time t. Using (8.61) and the fact that $\mathrm{E}[\varepsilon_{t+1}|\mathcal{Y}_t] = 0$, the optimal 1-step ahead forecast is easily obtained as

$$\hat{y}_{t+1|t} = \mathrm{E}[y_{t+1}|\mathcal{Y}_t] = F(y_t;\theta), \tag{8.63}$$

which is equivalent to the optimal 1-step ahead forecast in case the model $F(y_{t-1};\theta)$ is linear.

When the forecast horizon is longer than 1 period, things become more complicated however. For example, the optimal 2-step ahead forecast follows from (8.62) and (8.61) as

$$\hat{y}_{t+2|t} = \mathrm{E}[y_{t+2}|\mathcal{Y}_t] = \mathrm{E}[F(y_{t+1};\theta)|\mathcal{Y}_t]. \tag{8.64}$$

In general, the *linear* conditional expectation operator E cannot be interchanged with the *nonlinear* operator F, that is

$$\mathrm{E}[F(\cdot)] \neq F(\mathrm{E}[\cdot]).$$

Put differently, the expected value of a nonlinear function is not equal to the function evaluated at the expected value of its argument. Hence,

$$\mathrm{E}[F(y_{t+1};\theta)|\mathcal{Y}_t] \neq F(\mathrm{E}[y_{t+1}|\mathcal{Y}_t];\theta) = F(y_{t+1|t};\theta). \tag{8.65}$$

Rather, the relation between the 1- and 2-step ahead forecasts is given by

$$\begin{aligned}\hat{y}_{t+2|t} &= \mathrm{E}[F(F(y_t;\theta)+\varepsilon_{t+1};\theta)|\mathcal{Y}_t] \\ &= \mathrm{E}[F(y_{t+1|t}+\varepsilon_{t+1};\theta)|\mathcal{Y}_t].\end{aligned} \tag{8.66}$$

The above demonstrates that a simple recursive relationship between forecasts at different horizons, which could be used to obtain multiple-step ahead forecasts in an easy fashion analogous to the situation for linear AR(p) models, does not exist for nonlinear models in general. Of course, a 2-step ahead forecast might still be constructed as

$$\hat{y}^{(n)}_{t+2|t} = F(y_{t+1|t};\theta). \tag{8.67}$$

Brown and Mariano (1989) show that this 'naïve' approach, which lends it name from the fact that it effectively boils down to setting $\varepsilon_{t+1} = 0$ in (8.66) (or interchanging E and F in (8.64)), renders biased forecasts. Over the years, several methods have been developed to obtain more adequate multiple-step ahead forecasts, some of which are discussed below.

First, one might attempt to obtain the conditional expectation (8.66) directly by computing

$$\hat{y}^{(c)}_{t+2|t} = \int_{-\infty}^{\infty} F(y_{t+1|t}+\varepsilon;\theta) f(\varepsilon) d\varepsilon, \tag{8.68}$$

where f denotes the density of ε_t. Brown and Mariano (1989) refer to this forecast as the closed form forecast – hence the superscript (c). An alternative way to express this integral follows from (8.64) as

$$\begin{aligned}\hat{y}_{t+2|t}^{(c)} &= \int_{-\infty}^{\infty} F(y_{t+1};\theta)g(y_{t+1}|\mathcal{Y}_t)dy_{t+1} \\ &= \int_{-\infty}^{\infty} \mathrm{E}[y_{t+2}|y_{t+1}]g(y_{t+1}|\mathcal{Y}_t)dy_{t+1},\end{aligned} \tag{8.69}$$

where $g(y_{t+1}|\mathcal{Y}_t)$ is the distribution of y_{t+1} conditional upon $\mathcal{Y}_t$. This conditional distribution is in fact equal to the distribution $f(\cdot)$ of the shock ε_{t+1} with mean equal to $F(y_t;\theta)$, that is, $g(y_{t+1}|\mathcal{Y}_t) = f(y_{t+1} - F(y_t;\theta))$. As an analytic expression for the integral (8.68) (or (8.69)) is not available in general, it needs to be approximated using numerical integration techniques. An additional complication is the fact that the distribution of ε_t is never known with certainty. Usual practice is to assume normality of ε_t.

The closed form forecast becomes quite tedious to compute for forecasts more than two periods ahead. To see why, consider the Chapman-Kolgomorov relation

$$g(y_{t+h}|\mathcal{Y}_t) = \int_{-\infty}^{\infty} g(y_{t+h}|y_{t+h-1})g(y_{t+h-1}|\mathcal{Y}_t)dy_{t+h-1}. \tag{8.70}$$

where $g(y_{t+h}|y_{t+h-1})$ is the conditional distribution of y_{t+h} conditional upon y_{t+h-1}. By taking conditional expectations on both sides of (8.70) it follows that

$$\mathrm{E}[y_{t+h}|\mathcal{Y}_t] = \int_{-\infty}^{\infty} \mathrm{E}[y_{t+h}|y_{t+h-1}]g(y_{t+h-1}|\mathcal{Y}_t)dy_{t+h-1}, \tag{8.71}$$

which can be recognized as a generalization of (8.69). In order to evaluate this integral to obtain the h-step ahead exact forecast, one needs the conditional distribution $g(y_{t+h-1}|\mathcal{Y}_t)$. In principle, this distribution can be obtained recursively from (8.70), by observing that $g(y_{t+1}|y_{t+h-1})$ again is equal to the distribution of the shocks ε_{t+1} with its mean shifted to $F(y_{t+h-1};\theta)$. The recursion can be started for $h = 2$ by using the fact that $g(y_{t+1}|\mathcal{Y}_t) = f(y_{t+1} - F(y_t;\theta))$ as noted above. To obtain the conditional distribution $g(y_{t+h-1}|\mathcal{Y}_t)$ for $h > 2$ involves repeated numerical integration, which may become rather time-consuming, in particular if a large number of forecasts has to be made.

An alternative approach to computing multiple-step ahead forecasts is to use Monte Carlo or bootstrap methods to approximate the conditional expectation (8.66). The 2-step ahead Monte Carlo forecast is given by

$$\hat{y}_{t+2|t}^{(mc)} = \frac{1}{k}\sum_{i=1}^{k} F(y_{t+1|t} + \varepsilon_i;\theta), \tag{8.72}$$

where k is some large number and the ε_i are drawn from the presumed distribution of ε_{t+1}. The bootstrap forecast is very similar, the only difference being that the residuals from the estimated model, $\hat{\varepsilon}_t$, $t = 1, \ldots, n$ are used,

$$\hat{y}^{(b)}_{t+2|t} = \frac{1}{k} \sum_{i=1}^{k} F(y_{t+1|t} + \hat{\varepsilon}_i; \theta). \tag{8.73}$$

The advantage of the bootstrap over the Monte Carlo method is that no assumptions need to be made about the distribution of ε_{t+1}.

Lin and Granger (1994) and Clements and Smith (1997) compare various methods to obtain multiple-step ahead forecasts for STAR and TAR models, respectively. Their main findings are that the Monte Carlo and bootstrap methods compare favorably to the other methods.

An attractive feature of the Markov-Switching model is the relative ease with which analytic expressions for multiple-step ahead forecasts can be obtained. The essential thing to note is that the forecast of the future value of the time series, y_{t+h}, can be decomposed into a forecast of y_{t+h} conditional upon the regime that will be realized at $t + h$, s_{t+h}, and a forecast of the probabilities with which each of the regimes will occur at $t + h$. For example, the one-step ahead forecast for the two-state MSW model given in (8.6) can be written as

$$\begin{aligned} \hat{y}_{t+1|t} = {} & \mathrm{E}[y_{t+1}|s_{t+1} = 1, \mathcal{Y}_t] \cdot P(s_{t+1} = 1|\mathcal{Y}_t; \theta) \\ & \mathrm{E}[y_{t+1}|s_{t+1} = 2, \mathcal{Y}_t] \cdot P(s_{t+1} = 2|\mathcal{Y}_t; \theta). \end{aligned} \tag{8.74}$$

The forecasts of y_{t+1} conditional upon the regime at $t + 1$ follow directly from (8.6) as

$$\mathrm{E}[y_{t+1}|s_{t+1} = j, \mathcal{Y}_t] = \phi_{0,j} + \phi_{1,j} y_t,$$

whereas $P(s_{t+1} = j|\mathcal{Y}_t; \theta)$ are given by the optimal forecasts of the regime probabilities $\hat{\xi}_{t+1|t}$, which can be obtained from (8.32) and (8.33). Multiple-step ahead forecasts can be computed in a similar way, see Tjøstheim (1986) and Hamilton (1989) for details.

Interval forecasts

In addition to point forecasts one may also be interested in confidence intervals for these point forecasts. As discussed in Section 3.5, for forecasts obtained from linear models, the usual forecast confidence region is taken to be an interval symmetric around the point forecast. This is based upon the fact that the conditional distribution $g(y_{t+h}|\mathcal{Y}_t)$ of a linear time series is normal (under the assumption of normally distributed innovations) with mean $\hat{y}_{t+h|t}$.

For nonlinear models this is not the case. In fact, the conditional distribution can be asymmetric and even contain multiple modes. Whether a symmetric interval around the mean is the most appropriate forecast confidence region in this case can be questioned. This topic is discussed in detail in Hyndman (1995). He argues that there are three methods to construct a $100(1-\alpha)\%$ forecast region:

1. An interval symmetric around the mean, that is,

   $$S_\alpha = (\hat{y}_{t+h|t} - w, \hat{y}_{t+h|t} + w),$$

   where w is such that $P(y_{t+h} \in S_\alpha | \mathcal{Y}_t) = 1 - \alpha$.
2. The interval between the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the forecast distribution, denoted $q_{\alpha/2}$ and $q_{1-\alpha/2}$, respectively,

   $$Q_\alpha = (q_{\alpha/2}, q_{1-\alpha/2}).$$

3. The highest-density region [HDR], that is

   $$HDR_\alpha = \{y | g(y_{t+h}|\mathcal{Y}_t) \geq g_\alpha\}, \tag{8.75}$$

   where g_α is such that $P(y_{t+h} \in HDR_\alpha | \mathcal{Y}_t) = 1 - \alpha$.

For symmetric and unimodal distributions, these three regions are identical. For asymmetric or multimodal distributions they are not. Hyndman (1995) argues that the HDR is the most natural choice. The reasons for this claim are that first, HDR_α is the smallest of all possible $100(1-\alpha)\%$ forecast regions and, second, every point inside the HDR has conditional density $g(y_{t+h}|\mathcal{Y}_t)$ at least as large as every point outside the region. Furthermore, only the HDR will reveal features such as asymmetry or multimodality of the conditional distribution $g(y_{t+h}|\mathcal{Y}_t)$.

HDRs are straightforward to compute when the Monte Carlo or bootstrap methods described previously are used to compute the point forecast $\hat{y}_{t+h|t}$. Let $y^i_{t+h|t}$, $i = 1, \ldots, k$, denote the i-th element used in computing the Monte Carlo forecast (8.72) or bootstrap forecast (8.72) – that is, $y^i_{t+h|t} = F(y_{t+h-1|t} + \varepsilon_i; \theta)$ or $y^i_{t+h|t} = F(y_{t+h-1|t} + \hat{\varepsilon}_i; \theta)$. Note that the $y^i_{t+h|t}$ can be thought of as being realizations drawn from the conditional distribution of interest $g(y_{t+h}|\mathcal{Y}_t)$. Estimates $g_i \equiv g(y^i_{t+h|t}|\mathcal{Y}_t), i = 1, \ldots, k$, then can be obtained by using a standard kernel density estimator, that is

$$g_i = \frac{1}{k} \sum_{j=1}^{k} K([y^i_{t+h|t} - y^j_{t+h|t}]/b), \tag{8.76}$$

where $K(\cdot)$ is a kernel function such as the Gaussian density and $b > 0$ is the bandwidth. An estimate of g_α in (8.75) is given by $\hat{g}_\alpha = g_{(\lfloor \alpha k \rfloor)}$, where $g_{(i)}$ are the ordered g_i and $\lfloor \cdot \rfloor$ denotes integer part. See Hyndman (1996) for more details and some suggestions about the display of HDR's.

EXERCISES

8.1 Consider the following model, where y_t is the quarterly growth rate of US GDP:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 CDR_{t-1} + \varepsilon_t, \tag{8.77}$$

where $\varepsilon_t \sim \text{iid}N(0, \sigma^2)$, and where CDR_t is defined as follows:

$$CDR_t = Y_t - \max_{j \geq 0} Y_{t-j}, \tag{8.78}$$

where Y_t denotes the logarithm of the level of US GDP.

Estimating the parameters in the above model using observations for the period 1954Q1–2003Q4 gives estimates with the following properties: $\hat{\phi}_0 > 0$, $0 < \hat{\phi}_1 < 1$, $0 < \hat{\phi}_2 < 1$, $\hat{\phi}_1 + \hat{\phi}_2 < 1$, and $\hat{\theta}_1 < 0$. Describe in as much detail as possible which features of the US GDP growth are captured by this model. In particular, what is the function of the variable CDR_{t-1} in the model? How would you label the regimes consisting of observations for which $CDR_t = 0$ and consisting of observations for which $CDR_t < 0$?

8.2 Consider the Switching AR(1) model

$$y_t = (\mu_0 + \phi_0 y_{t-1}) + (\mu_1 + \phi_1 y_{t-1}) s_t + \varepsilon_t$$

for $t = 1, \ldots, T$ with $\varepsilon_t \sim N(0, \sigma^2)$, where s_t is a latent binary random variable with

$$\Pr[s_t = 1] = F(\delta + \gamma y_{t-1}) = \frac{\exp(\delta + \gamma y_{t-1})}{1 + \exp(\delta + \gamma y_{t-1})}.$$

a. What is/are the key difference(s) between the Switching AR(1) model and the LSTAR(1) model given by

$$y_t = (\mu_0 + \phi_0 y_{t-1}) + (\mu_1 + \phi_1 y_{t-1}) F(\delta + \gamma y_{t-1}) + \varepsilon_t$$

for $t = 1, \ldots, T$ with $\varepsilon_t \sim N(0, \sigma^2)$ or are both models the same?

b. Derive the unbiased 1-step ahead forecasts of y_{T+1} made at time T for the Switching AR(1) model and for the LSTAR(1) model.

c. Consider the Switching AR(1) model with

$$\Pr[s_t = 1] = \frac{\exp(\delta + \gamma s_{t-1})}{1 + \exp(\delta + \gamma s_{t-1})}.$$

Show that this model is the same as a Markov Switching AR(1) model where

$$\Pr[s_t = 1 | s_{t-1} = 1] = p \text{ and } \Pr[s_t = 0 | s_{t-1} = 0] = q.$$

Express p and q in terms of δ and γ.

d. Suppose that you have observed y_t for $t = 1, \ldots, T$. A recession is defined as a period where for at least 2 consecutive periods $s_t = 1$. Suppose that $S_T = 0$. Derive the probability in terms of p and q that the period $[T + 1, T + 3]$ contains a recession.

8.3 Suppose the time series y_t is generated according to the following threshold process

$$y_t = \begin{cases} -\delta + \varepsilon_t & \text{if } q_t \leq c \\ \delta + \varepsilon_t & \text{if } q_t > c \end{cases}$$

with $\delta > 0$, ε_t is a white noise series with $\mathsf{E}[\varepsilon_t] = 0$ and $\mathsf{E}[\varepsilon_t^2 = \sigma_\varepsilon^2]$ for all t, and the threshold variable q_t is i.i.d. standard normally distributed. Furthermore, ε_t and q_t are independent.

Derive an expression (in terms of the parameters δ, σ_ε^2 and c) for the following characteristics of the time series y_t:

a. the unconditional mean $\mu_y = \mathsf{E}[y_t]$,
b. the unconditional variance $\gamma_0(y) = \mathsf{E}[(y_t - \mathsf{E}[y_t])^2]$, and
c. the first-order autocorrelation $\rho_1(y) = \gamma_1(y)/\gamma_0(y)$, where $\gamma_1(y)$ is the first-order autocovariance of y_t, that is, $\gamma_1(y) = \mathsf{E}[(y_t - \mathsf{E}[y_t])(y_{t-1} - E[y_{t-1}])]$. Hint: Use the fact that the covariance between two random variables X and Z can be written as $\mathsf{E}[(X - \mathsf{E}[X])(Z - \mathsf{E}[Z])] = \mathsf{E}[XZ] - \mathsf{E}[X]\mathsf{E}[Z]$.

Interpret these expressions, and discuss how they behave as a function of the parameters.

8.4 Consider the following model, where y_t is the quarterly growth rate of US GDP:

$$y_t - \mu_t = \phi_1(y_{t-1} - \mu_{t-1}) + \varepsilon_t, \tag{8.79}$$

where $\varepsilon_t \sim \text{iid}N(0, \sigma^2)$, and μ_t is given by

$$\mu_t = \mu_0 + \mu_1 \mathtt{I}(S_t = 1) + \lambda \mathtt{I}(S_t = 0) \sum_{j=1}^{m} \mathtt{I}(S_{t-j} = 1), \tag{8.80}$$

where $\mathtt{I}(A)$ is the indicator function for the event A (that is, $\mathtt{I}(A) = 1$ if A is true, and $\mathtt{I}(A) = 0$ otherwise), and $S_t \in \{0, 1\}$ is a first-order Markov process with constant transition probabilities given by

$$P(S_t = 0|S_{t-1} = 0) = p \qquad \text{and} \qquad P(S_t = 1|S_{t-1} = 1) = q.$$

Estimating the parameters in the above model with $m = 6$ using observations for the period 1954Q1–2003Q4 gives $0 < \widehat{\phi}_1 < 1$, $\widehat{\mu}_0 > 0$, $\widehat{\mu}_0 + \widehat{\mu}_1 < 0$, $\widehat{\lambda} > 0$, $\widehat{p} = 0.94$, and $\widehat{q} = 0.78$.

Describe which features of the US GDP growth are captured by this model. In particular, how would you label the regimes $S_t = 0$ and $S_t = 1$? What is

the function of the last component of μ_t in (8.80), that is $\lambda \mathtt{I}(S_t = 0) \sum_{j=1}^{m} \mathtt{I}(S_{t-j} = 1)$?

8.5 The Eviews file gdp.wf1 contains quarterly observations for US real GDP over the period 1959Q1–2008Q2 (198 observations).

a. Estimate an AR(2) model for the quarterly real GDP growth rates and obtain the residuals. Test for a break in volatility of these residuals in 1984Q1, that is, test the null hypothesis $\mathsf{H}_0 : \delta_1 = \delta_2$ in the regression

$$\sqrt{\pi/2}|\hat{\varepsilon}_t| = \delta_1 \mathtt{I}[t \leq \tau] + \delta_2 \mathtt{I}[t > \tau] + \varepsilon_t, \quad t = 1, \ldots, n, \tag{8.81}$$

where $\hat{\varepsilon}_t$ denotes the AR(2) residual for quarter t and τ corresponds to 1984Q1.

b. Apply Quandt's Sup$F$ test to test for a break in volatility of these residuals at an unknown date. Use a trimming fraction $\lambda = 0.05$.

9 Multivariate time series

Univariate time series models, as discussed in the previous chapters, can be very useful for descriptive analysis and for out-of-sample forecasting. Such univariate models are restrictive, however, in the sense that they rely purely on the history of the time series itself and do not try to exploit information present in other variables. Given that often economic variables are closely related with other variables, such information may be potentially useful and using it may lead to more realistic descriptions of the behavior of economic variables and also to more accurate out-of-sample forecasts. In fact, ignoring relationships with other variables may give rise to complications in the specification of an appropriate univariate time series model. It may be that the empirical specification of a univariate model is hampered by outliers and structural shifts, which may be attributed to one or more other variables than the y_t variable under consideration. For example, the change in the trend of US industrial production around 1979.4 (see Figure 2.3) may be caused by the large oil price increase around that time. Additionally, the same oil price variable may be responsible for the large negative growth in production in 1974/1975, which seems to correspond to an innovation outlier. In other words, it may be worthwhile to include an oil price variable in a model for US industrial production to render it more realistic and more accurate.

Including an additional regressor

When production is denoted as y_t and the price of oil as x_t, we may thus wish to consider an extended ARMA-type model such as

$$y_t = \beta x_t + (\theta_q(L)/\phi_p(L))\varepsilon_t, \tag{9.1}$$

where $\phi_p(L)$ and $\theta_q(L)$ are lag polynomials as before, and β measures the effect of x_t on y_t at time t. It may now be that the estimated residuals $\hat{\varepsilon}_t$ from (9.1) do not show typical outlier or structural break patterns, which would be present in the residuals from a pure ARMA model. Hence, including only a single extra variable can substantially

reduce the number of parameters as no additional modifications of the model to handle outliers and breaks are needed.

A model such as (9.1) introduces new complications, however. The first is that it is unknown whether the oil price x_t should be included as is done in (9.1), or that it is more appropriate to have it entering the model with a time lag, that is, whether x_{t-i} for some $i = 1, 2, \ldots, k$ should be included. Also note that in (9.1) it is assumed that the oil price has 'additive outlier'-type effects on industrial production, with x_t affecting only the contemporaneous observation y_t. An alternative and perhaps more appropriate specification could then be

$$(\phi_p(L)/\theta_q(L))y_t = \beta x_t + \varepsilon_t, \tag{9.2}$$

such that the term βx_t is similar to an innovation outlier. Which specification is more appropriate is not obvious *a priori*. In general, we can only make these decisions based on the available data, as economic theory usually provides no guidelines on lag structures in time series models.

An additional problem with models such as (9.1) is that maybe not only y_t is explained by x_t, but also in turn x_t somehow depends on current and/or past y_t. For US industrial production, we may reasonably assume that the oil price affects production but not the other way round. This may be different for other situations. Think of data as market share, distribution, and price for a particular consumer product. When distribution is low, market shares cannot be large. In turn, with increasing market shares, one may want to increase distribution. Advertising can increase market share, but it seems useless to increase advertising for a fast-moving consumer product when the product is not widely available. Finally, prices will vary because of promotional activities when a marketeer observes a decreasing market share, and lowering prices may increase sales. In sum, the relationships between these variables is complicated, with various interaction and feedback mechanisms at work with as a consequence that any of the three variables likely depends on the other two and perhaps also their past.

When we want to take all possible relations between, say, k variables into account, it seems sensible to construct a model for all these time series jointly instead of constructing models like (9.1) for all the individual series. In case we also do not know *a priori* which variable is affecting which, or in other words, when it is uncertain which variables are exogenous and which are endogenous, it seems useful to start with the construction of a general time series model for a vector time series, and it is this type of model that is addressed in this final chapter.

Spurious inference

Another motivation for starting with an unrestricted multiple time series model instead of a static regression like $y_t = \beta x_t + u_t$ is that such static univariate models may lead

to spurious inference. As this is one of the many traps in applied econometric work, we wish to spend a few lines on this issue. A simple illustration is given by the following example. Suppose that sales y_t and advertising expenditures x_t (both in terms of euros) are generated by the following set of equations:

$$y_t = \rho y_{t-1} + \varepsilon_t, \qquad \text{with } |\rho| < 1, \tag{9.3}$$

$$x_t = \delta y_{t-1} + \eta_t, \tag{9.4}$$

where ε_t and η_t are mutually independent white noise variables with variances σ_ε^2 and σ_η^2, respectively. These shocks are also independent of past realizations of x_t and y_t. On average, advertising expenditures are thus set at a proportion δ of previous sales. Clearly, but unfortunately, advertising is assumed not to generate extra sales as the equation for y_t does not include x_t or lagged x_t. However, if we were to consider the static model

$$y_t = \beta x_t + u_t, \tag{9.5}$$

we would spuriously find that advertising does affect sales, that is, we would find that β is different from zero. This can be understood by noting that the OLS estimate of β in (9.5) equals the sample covariance between y_t and x_t divided by the sample variance of x_t. From (9.3) and (9.4) it follows that the variance of x_t, denoted as $\gamma_0(x_t)$, is equal to

$$\begin{aligned}\gamma_0(x_t) &= E[(\delta y_{t-1} + \eta_t)(\delta y_{t-1} + \eta_t)] \\ &= \delta^2 \gamma_0(y_t) + \sigma_\eta^2 \\ &= \frac{\delta^2 \sigma_\varepsilon^2}{(1-\rho^2)} + \sigma_\eta^2.\end{aligned} \tag{9.6}$$

The covariance between y_t and x_t is

$$\begin{aligned}\gamma_0(y_t, x_t) &= E[(\rho y_{t-1} + \varepsilon_t)(\delta y_{t-1} + \eta_t)] \\ &= \rho\delta\gamma_0(y_t) \\ &= \frac{\rho\delta\sigma_\varepsilon^2}{(1-\rho^2)}.\end{aligned} \tag{9.7}$$

In other words, the expected value of the OLS estimate of β in (9.5) is equal to

$$\begin{aligned}\mathrm{E}[\hat{\beta}] &= \gamma_0(y_t, x_t)/\gamma_0(x_t) \\ &= \frac{\rho\delta}{\delta^2 + (1-\rho^2)\sigma_\eta^2/\sigma_\varepsilon^2}.\end{aligned} \tag{9.8}$$

As an example, when the variances of both error terms equal 1, $\delta = 0.5$, and $\rho = 0.7$, it follows that $\mathrm{E}[\hat{\beta}]$ equals about 0.46. In other words, advertising would seem to have a positive effect on sales, while in reality we know it has not. Furthermore, this spurious effect can easily be shown not to disappear when y_t is regressed on a lagged value x_{t-i} $(i = 1, 2, \ldots, k)$ instead.

Preventing spurious inference

The solution of this spurious inference trap is simple. If, given y_t and x_t, we consider the multivariate regression model

$$y_t = \phi_1 y_{t-1} + \phi_2 x_{t-1} + \varepsilon_{1,t}, \tag{9.9}$$

$$x_t = \phi_3 y_{t-1} + \phi_4 x_{t-1} + \varepsilon_{2,t}, \tag{9.10}$$

where $e_t \equiv (\varepsilon_{1,t}, \varepsilon_{2,t})'$ is a (2×1) vector white noise variable with mean zero and covariance matrix Σ, we shall find that ϕ_2 and ϕ_4 are equal to zero. The model in (9.9)–(9.10) is called a vector autoregression of order 1 [VAR(1)] as it only includes the first lag of y_t and x_t on the right-hand side.

In empirical applications, one may also wish to allow for possible contemporaneous effects of x_t on y_t and vice versa. In the example concerning sales and advertising, it may be of interest to examine whether a certain promotional activity increases sales in the same period or not. In that case, we may also consider a model like

$$y_t = \pi_1 y_{t-1} + \pi_2 x_t + \varepsilon_{1,t}, \tag{9.11}$$

$$x_t = \pi_3 y_t + \pi_4 x_{t-1} + \varepsilon_{2,t}, \tag{9.12}$$

instead of (9.9)–(9.10). This is a model with contemporaneous effects in case π_2 and π_3 are different from zero. Using matrix notation this so-called dynamic simultaneous equations model can be written as

$$\begin{pmatrix} 1 & -\pi_2 \\ -\pi_3 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \pi_1 & 0 \\ 0 & \pi_4 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}. \tag{9.13}$$

Multiplying both sides with the inverse of the matrix on the left-hand side gives

$$\begin{aligned} \begin{pmatrix} y_t \\ x_t \end{pmatrix} &= \frac{1}{1 - \pi_2 \pi_3} \begin{pmatrix} 1 & \pi_2 \\ \pi_3 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 & 0 \\ 0 & \pi_4 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \frac{1}{1 - \pi_2 \pi_3} \begin{pmatrix} 1 & \pi_2 \\ \pi_3 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \\ &= \frac{1}{1 - \pi_2 \pi_3} \begin{pmatrix} \pi_1 & \pi_2 \pi_4 \\ \pi_1 \pi_3 & \pi_4 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon^*_{1,t} \\ \varepsilon^*_{2,t} \end{pmatrix}, \end{aligned} \tag{9.14}$$

where the (2×1) vector series ε_t^* has mean zero and covariance matrix Σ^*. This expression shows that (9.13) can be written in the form of a VAR(1) model as in (9.9)–(9.10). For this example, it also holds that the parameters ϕ_i, $i = 1, 2, 3, 4$, in this VAR(1) model are uniquely related to the parameters π_i, $i = 1, 2, 3, 4$, in the simultaneous equations model in (9.11)–(9.12).

Exercise 9.1

In this last chapter we discuss linear models for multivariate time series and their representations in Section 9.1. For practical purposes the VAR model is often the most useful (particularly for analyzing stochastic trends), and hence we consider only the VAR model when discussing empirical modeling strategies, forecasting and other issues in Sections 9.2 and 9.3. For more extensive treatments of multivariate time series models, including VARMA models, the reader should consult Lütkepohl (2005).

In the last two sections of this chapter we examine the possibility that two or more time series have a common trend. This phenomenon is called cointegration, and we show that it imposes specific parameter restrictions on VAR models.

9.1 Representations

We consider m possibly related time series $y_{1,t}, y_{2,t}, \ldots, y_{m,t}$, and for the moment assume that they all have mean zero. For the shocks $\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{m,t}$ that appear in the model equations, we assume that these are individually white noise series, with mean zero, variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$, and no correlation between $\varepsilon_{i,t}$ and its own lagged values or the past of any other $\varepsilon_{j,t}$ variable ($j = 1, \ldots, m$, $j \neq i$). We do allow for contemporaneous correlation between different shocks $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$ with the covariance denoted as σ_{ij}. The variances and covariances of $\varepsilon_{i,t}$, $i = 1, \ldots, m$, are collected in the $(m \times m)$ covariance matrix Σ. To save notation we make extensive use of vector and matrix notation, that is, we stack the individual series in the $(m \times 1)$ vector series Y_t and e_t as

$$Y_t \equiv \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{m,t} \end{pmatrix} \quad \text{and} \quad e_t \equiv \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{m,t} \end{pmatrix}. \tag{9.15}$$

Vector autoregressive moving average [VARMA] model

Analogous to the ARMA model for a univariate series, we can consider a vector ARMA model of order (p, q), which is given by

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{m,t} \end{pmatrix} = \begin{pmatrix} \phi_{11,1} & \cdots & \phi_{1m,1} \\ \phi_{21,1} & \cdots & \phi_{2m,1} \\ \vdots & & \vdots \\ \phi_{m1,1} & \cdots & \phi_{mm,1} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{m,t-1} \end{pmatrix}$$
$$+ \cdots + \begin{pmatrix} \phi_{11,p} & \cdots & \phi_{1m,p} \\ \phi_{21,p} & \cdots & \phi_{2m,p} \\ \vdots & & \vdots \\ \phi_{m1,p} & \cdots & \phi_{mm,p} \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{m,t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \vdots \\ \varepsilon_{m,t} \end{pmatrix}$$
$$+ \begin{pmatrix} \theta_{11,1} & \cdots & \theta_{1m,1} \\ \theta_{21,1} & \cdots & \theta_{2m,1} \\ \vdots & & \vdots \\ \theta_{m1,1} & \cdots & \theta_{mm,1} \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \\ \vdots \\ \varepsilon_{m,t-1} \end{pmatrix} + \cdots + \begin{pmatrix} \theta_{11,q} & \cdots & \theta_{1m,q} \\ \theta_{21,q} & \cdots & \theta_{2m,q} \\ \vdots & & \vdots \\ \theta_{m1,q} & \ldots, & \theta_{mm,q} \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t-q} \\ \varepsilon_{2,t-q} \\ \vdots \\ \varepsilon_{m,t-q} \end{pmatrix}. \tag{9.16}$$

Additional to the $m(m+1)/2$ unique elements of Σ, the VARMA(p,q) model contains $m^2(p+q)$ unknown parameters. For 4 time series with $p = 1$ and $q = 1$ this already amounts to 32 parameters. Using matrix notation, the VARMA(p, q) model can be written as

$$Y_t = \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + e_t + \Theta_1 e_{t-1} + \cdots + \Theta_q e_{t-q}, \tag{9.17}$$

where Φ_i, $i = 1, \ldots, p$, and Θ_j, $j = 1, \ldots, q$, are $(m \times m)$ matrices. Even more compactly, we can write

$$\Phi_p(L) Y_t = \Theta_q(L) e_t, \tag{9.18}$$

with the matrix polynomials $\Phi_p(L)$ and $\Theta_q(L)$ defined as

$$\Phi_p(L) = I_m - \Phi_1 L - \cdots - \Phi_p L^p, \tag{9.19}$$

$$\Theta_q(L) = I_m + \Theta_1 L + \cdots + \Theta_q L^q, \tag{9.20}$$

where I_m is the m-dimensional identity matrix.

In practice, this general VARMA(p,q) model often is inconvenient. One reason is that parameters may not be identified. For example, consider the following simple

VMA(1) model for a bivariate time series $Y_t = (y_{1,t}, y_{2,t})'$,

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} + \begin{pmatrix} 0 & \pi \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{pmatrix}, \tag{9.21}$$

and compare this to the VAR(1) model

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} 0 & \pi \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}. \tag{9.22}$$

Both models imply the same relationship between $y_{1,t}$ and $y_{2,t}$, and thus are equivalent. Both specifications can be obtained from a general VARMA(1,1) model by imposing suitable restrictions on the parameters. The fact that (9.21) and (9.22) are effectively the same implies that parameter estimation in a VARMA(1,1) model for this bivariate series is cumbersome, as clearly not all model parameters are identified. Parameter estimation in a VARMA model becomes even more complicated when p and q are large, simply due to the large number of parameters in that case. Finally, the model is not particularly useful to analyze the presence of common stochastic trends and cointegration. As the latter issue, which is discussed towards the end of this chapter, is very important for modeling and forecasting economic time series, we usually stick to the VAR(p) model in many empirical applications and also below.

Stationarity of a vector autoregression

Similar to the univariate AR case, we can investigate stationarity of the VAR(p) model by checking for the presence of unit roots. For that purpose we should again consider the solutions to the characteristic equation of the autoregressive polynomial. In case of matrix polynomials in L as in the VAR model, this amounts to

$$|\Phi_p(z)| = 0, \tag{9.23}$$

where $|A|$ denotes the determinant of the matrix A. The VAR(p) model is said to be stable, and the corresponding vector time series Y_t and each of its components is said to be stationary if all solutions to (9.23) are outside the unit circle. When one or more solutions are on the unit circle, the VAR(p) model contains unit roots and one, or more or even all of the time series in Y_t are non-stationary.

In order to understand why (9.23) is the relevant condition to check for stationarity of Y_t, consider the VAR(1) model

$$Y_t = \Phi_1 Y_{t-1} + e_t, \tag{9.24}$$

where e_t is defined earlier. By recursively substituting for lagged Y_t's, we can rewrite (9.24) as

$$Y_t = \Phi_1^t Y_0 + \sum_{i=0}^{t-1} \Phi_1^i e_{t-i}, \tag{9.25}$$

which is the multivariate analogue of (3.12). For stability of the VAR(1) model, it is required that the sequence $\Phi_1^i, i = 0, 1, \ldots$ is absolutely summable, that is, $\sum_{i=0}^{t-1} \Phi_1^i = \Phi$, and that Φ_1^t converges to a zero matrix as t goes to infinity. These conditions are satisfied when all eigenvalues of the matrix Φ_1 have modulus less than one. This is equivalent to the condition that all solutions to

$$|\Phi_1(z)| = |I_m - \Phi_1 z| = 0, \tag{9.26}$$

are larger than one in modulus.

The corresponding condition for the general VAR(p) model

$$Y_t = \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + e_t \tag{9.27}$$

is most easily derived by rewriting (9.27) as a VAR(1) model for the kp-dimensional time series $Y_t^{(p)} = (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})'$, which is given by

$$Y_t^{(p)} = \Phi_1^* Y_{t-1}^{(p)} + e_t^{(p)}, \tag{9.28}$$

where

$$\Phi_1^* = \begin{pmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{p-1} & \Phi_p \\ I_k & 0 & \ldots & 0 & 0 \\ 0 & I_k & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I_k & 0 \end{pmatrix} \quad \text{and} \quad e_t^{(p)} = \begin{pmatrix} e_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{9.29}$$

The stacked time series $Y_t^{(p)}$ is stationary if all eigenvalues of Φ_1^* have modulus less than one, or all solutions to $|I_{kp} - \Phi_1^* z| = 0$ are outside the unit circle. It is not difficult to see that this corresponds with

$$|I_{kp} - \Phi_1^* z| = |I_k - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p| = 0. \tag{9.30}$$

The VAR(1) model as an example

As a specific example, consider the VAR(1) model for a bivariate time series $Y_t = (y_{1,t}, y_{2,t})'$, that is,

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}, \tag{9.31}$$

where the additional subscript 1 in (9.16) is dropped to save notation. For this VAR(1) we have

$$\Phi_1(z) = \begin{pmatrix} 1 - \phi_{11}z & -\phi_{12}z \\ -\phi_{21}z & 1 - \phi_{22}z \end{pmatrix}, \tag{9.32}$$

and hence the characteristic equation equals

$$\begin{aligned} |\Phi_1(z)| &= (1 - \phi_{11}z)(1 - \phi_{22}z) - \phi_{21}\phi_{12}z^2 \\ &= 1 - (\phi_{11} + \phi_{22})z + (\phi_{11}\phi_{22} - \phi_{21}\phi_{12})z^2 \\ &= (1 - \nu_1 z)(1 - \nu_2 z) = 0, \end{aligned} \tag{9.33}$$

where parameters ν_1 and ν_2 are defined by $\nu_j = a_j + b_j i$, $j = 1, 2$, where a_j and b_j are functions of the parameters $\phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}$, and the complex number i is defined by $i^2 = -1$. When it holds that

$$(\phi_{11} + \phi_{22}) - (\phi_{11}\phi_{22} - \phi_{21}\phi_{12}) = 1 \tag{9.34}$$

then either ν_1 or ν_2 lies on the unit circle and the VAR(1) model thus contains a single unit root. We can then write

$$|\Phi_1(z)| = (1 - z)(1 - (\phi_{11}\phi_{22} - \phi_{21}\phi_{12})z). \tag{9.35}$$

A periodic AR(1) model as an example

Consider a so-called periodic AR(1) model for a univariate quarterly time series y_t, that is

$$y_t = \phi_s y_{t-1} + \varepsilon_t, \tag{9.36}$$

where ϕ_s can take the values ϕ_1, ϕ_2, ϕ_3 and ϕ_4 in the respective quarters 1,2,3, and 4, see Franses and Paap (2004). This PAR(1) model assumes that there is a different model for different seasons as the autoregressive parameter ϕ_s varies with the quarter $s = 1, 2, 3, 4$. When the observations in season s of year t is denoted as $y_{s,t}$, the PAR(1) can be written as $y_{s,t} = \phi_s y_{s-1,t} + \varepsilon_{s,t}$ for $s = 2, 3, 4$, and $y_{1,t} = \phi_1 y_{4,t-1} + \varepsilon_{1,t}$ for $s = 1$. This (4×1) system of equations can be summarized as a dynamic simultaneous

model (or VAR(1) model) as

$$\begin{pmatrix} 1 & 0 & 0 & -\phi_1 L \\ -\phi_2 & 1 & 0 & 0 \\ 0 & -\phi_3 & 1 & 0 \\ 0 & 0 & -\phi_4 & 1 \end{pmatrix} \begin{pmatrix} y_{1,n} \\ y_{2,n} \\ y_{3,n} \\ y_{4,n} \end{pmatrix} = \begin{pmatrix} \varepsilon_{1,n} \\ \varepsilon_{2,n} \\ \varepsilon_{3,n} \\ \varepsilon_{4,n} \end{pmatrix}. \tag{9.37}$$

Note that $Ly_{4,t} = y_{4,t-1}$, that is, L operates on the years and not on the seasons. It is easy to see that the characteristic equation for this vector version of the PAR(1) model is

$$|\Phi_1(z)| = (1 - \phi_1\phi_2\phi_3\phi_4 z) = 0. \tag{9.38}$$

In other words, when $\phi_1\phi_2\phi_3\phi_4 = 1$, the VAR(1) has a unit root.

It is difficult to test hypotheses on ν_1 and ν_2 in (9.33) in order to examine the presence of unit roots. Below we will discuss methods (based on cointegration techniques) which are more useful. To obtain a preliminary and tentative impression of stationarity, we can of course, similar to the univariate AR(p) case where we consider the sum of the autoregressive parameters ϕ_i, $i = 1, 2, \ldots, p$, compute the eigenvalues of Φ_1^* as defined in (9.29) to see if these are close to unity. When they are, there may be unit roots in the VAR model.

Exercise 9.2

Implied univariate time series models

We saw that a dynamic simultaneous equation model can be written as a VAR model. Now we will show that a VAR model can be written as a set of univariate ARMA models and also as the so-called transfer function models for the individual elements of Y_t. This is important as it suggests that it is not a bad idea to perform a univariate analysis prior to eventual simultaneous-equation model building.

Consider again the VAR(p) model for a zero mean time series Y_t, that is,

$$\Phi_p(L)Y_t = e_t, \tag{9.39}$$

for which it is assumed that there are only roots on or outside the unit circle. One of the possible representations of this model follows from the equality

$$\Phi_p(z)^{-1} = |\Phi_p(z)|^{-1}\Phi_p^*(z), \tag{9.40}$$

where $\Phi_p^*(z)$ is the so-called adjoint matrix containing the cofactors. These cofactors are the determinants of $(m-1) \times (m-1)$ matrices. The cells of the adjoint matrix can contain polynomials in z of maximum order $(m-1)p$. When we pre-multiply both

sides of (9.40) with $|\Phi_p(z)|$, we get

$$|\Phi_p(z)|\Phi_p(z)^{-1} = \Phi_p^*(z). \tag{9.41}$$

Using the same idea, the VAR(p) model in (9.39) can then be written as

$$|\Phi_p(L)|Y_t = \Phi_p^*(L)e_t. \tag{9.42}$$

As $|\Phi_p(z)|$ contains terms up to z^{mp}, the resultant univariate models for $y_{1,t}$ to $y_{m,t}$ are ARMA models with maximum order $(mp, (m-1)p)$. Often, the order is much lower because common factors in the AR and MA polynomials cancel out. For a VAR(2) model for five time series the implied univariate models are ARMA of maximum order (10,8). Clearly, in practice however, one seldom finds that such highly parameterized ARMA models are required to fit the data, see also Chapter 2. When e_t in (9.39) is replaced by a VMA(q) model, the implied univariate ARMA models are of maximum order $(mp, (m-1)p+q)$. Finally, note that because each row of (9.42) is multiplied with the same determinant, the AR(mp) polynomials are the same across the m variables $y_{i,t}$.

The expression in (9.42) shows that a dynamic simultaneous equations model, which was shown to be equivalent to a VAR, implies that the univariate time series can be described by univariate ARMA models, see Zellner and Palm (1974). This means that simultaneous models, which are often claimed to be based on economic theory, can correspond quite well with univariate ARMA models, which are often based on the desire to have a simple descriptive model to forecast economic data.

As an illustration of (9.42), consider again the VAR(1) model for $(y_{1,t}, y_{2,t})'$ with $|\Phi_1(z)|$ as in (9.33) and with the adjoint matrix

$$\Phi_1^*(z) = \begin{bmatrix} 1-\phi_{22}z & \phi_{12}z \\ \phi_{21}z & 1-\phi_{11}z \end{bmatrix} \tag{9.43}$$

The implied univariate ARMA(2,1) models for $y_{1,t}$ and $y_{2,t}$ are now easily derived as being equal to

$$\begin{aligned} y_{1,t} &= (\phi_{11}+\phi_{22})y_{1,t-1} + (\phi_{21}\phi_{12} - \phi_{11}\phi_{22})y_{1,t-2} \\ &\quad + \varepsilon_{1,t} - \phi_{22}\varepsilon_{1,t-1} + \phi_{12}\varepsilon_{2,t-1}, \end{aligned} \tag{9.44}$$

$$\begin{aligned} y_{2,t} &= (\phi_{11}+\phi_{22})y_{2,t-1} + (\phi_{21}\phi_{12} - \phi_{11}\phi_{22})y_{2,t-2} \\ &\quad + \varepsilon_{2,t} - \phi_{11}\varepsilon_{2,t-1} + \phi_{21}\varepsilon_{1,t-1}. \end{aligned} \tag{9.45}$$

When the unit root restriction in (9.33) holds, the two models are ARIMA(1,1,1). In other words, *one* unit root in a bivariate VAR model leads to a unit root in *each* of the univariate models. This is a crucial result for our discussion below on cointegration as

these two univariate models apparently have the unit root in common. The phenomenon of such a common unit root is called cointegration.

Exercise 9.3

ARMAX models

A second way to represent a VARMA(p, q) model, or a VMA(q) model, is implied by the equality

$$\Theta_q(z)^{-1} = |\Theta_q(z)|^{-1}\Theta_q^*(z), \tag{9.46}$$

where $\Theta_q^*(z)$ is the adjoint matrix containing the cofactors of $\Theta_q(z)$. Using the same manipulation as before, the VARMA(p, q) model can now be rewritten as

$$\Theta_q^*(L)\Phi_p(L)Y_t = |\Theta_q(L)|e_t, \tag{9.47}$$

and this amounts to a set of so-called transfer function models or ARMAX models.

As an illustration, consider the VMA(1) model for a bivariate series, that is,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 1+\theta_{11}z & \theta_{12}z \\ \theta_{21}z & 1+\theta_{22}z \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}. \tag{9.48}$$

For the $\Theta_1(z)$ it matrix holds that

$$|\Theta_1(z)| = 1 + (\theta_{11}+\theta_{22})z + (\theta_{11}\theta_{22} - \theta_{21}\theta_{12})z^2 \tag{9.49}$$

and

$$\Theta_1^*(z) = \begin{bmatrix} 1+\theta_{22}z & -\theta_{12}z \\ -\theta_{21}z & 1+\theta_{11}z \end{bmatrix}. \tag{9.50}$$

This implies that the univariate series can now be described by

$$\begin{aligned} y_{1,t} = {} & -\theta_{22}y_{1,t-1} + \theta_{12}y_{2,t-1} + \varepsilon_{1,t} + (\theta_{11}+\theta_{22})\varepsilon_{1,t-1} \\ & + (\theta_{11}\theta_{22} - \theta_{21}\theta_{12})\varepsilon_{1,t-2} \end{aligned} \tag{9.51}$$

$$\begin{aligned} y_{2,t} = {} & -\theta_{11}y_{2,t-1} + \theta_{21}y_{1,t-1} + \varepsilon_{2,t} + (\theta_{11}+\theta_{22})\varepsilon_{2,t-1} \\ & + (\theta_{11}\theta_{22} - \theta_{21}\theta_{12})\varepsilon_{2,t-2}. \end{aligned} \tag{9.52}$$

These single equation models are called ARMAX models because they include autoregressive terms [AR], lags of explanatory variables [X] and MA terms. Notice that these ARMAX models include the error terms corresponding to each single variable, which is in contrast to the implied univariate models.

When the multivariate model would be a simultaneous model with a VMA(1) error term, then the data can be described by

$$\begin{bmatrix} 1-\gamma_1 z & -\alpha \\ -\beta & 1-\gamma_2 z \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 1+\theta_{11} z & \theta_{12} z \\ \theta_{21} z & 1+\theta_{22} z \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \tag{9.53}$$

The corresponding implied univariate ARMAX type models then also include current $y_{1,t}$ and $y_{2,t}$ variables.

In sum, even though a vector Y_t series can be described by a VARMA(p, q) model, one can also describe the component series by univariate time series models or by ARMAX models. It should be mentioned that in practice, one sometimes uses an alternative representation of an ARMAX model, which is then called a transfer function model. A simple example of such a model is

$$y_{1,t} = [\omega_w(L)/\upsilon_v(L)]y_{2,t} + [\theta_q(L)/\phi_p(L)]\varepsilon_{1,t}, \tag{9.54}$$

see for example Box and Jenkins (1970), where $\omega_w(L)$ and $\upsilon_v(L)$ are polynomials in L of orders w and v. Notice that when $y_{2,t}$ is replaced by a zero-one dummy variable, this transfer function model looks very much like some of the models for additive or innovative outliers discussed in Chapter 6.

This section has shown that there are many links between multivariate VARMA models and univariate time series models, between simultaneous equation models with lagged variables, and ARMAX models (and transfer function models). When a researcher has a set of m time series and is willing to link these series explicitly in a multivariate model, it generally seems most easy in practice to start with a VAR, see also Hendry (1995). In the sequel of this chapter, we therefore continue with an analysis of the VAR(p) model.

Exercise 9.4–9.5

9.2 Empirical model building

Similar to the creation of univariate AR models, the construction of VAR models involves several specification steps. First, one needs to specify an initial value of p. Next, one should estimate the parameters. Then, one should investigate the properties of the estimated residuals. In case one has more than one model that passes the diagnostics, one then has to select between several values of p. Finally, one may use the estimated VAR(p) model for various purposes. These purposes will be dealt with in the next section. In the current section, we review some guidelines for empirical specification.

In principle, one can derive explicit expressions for the autocorrelation function of VAR(p) processes. These would then be multivariate versions of the ACFs for the AR(p) model given in Chapter 3. For practical purposes, however, the ACFs for VAR(p) models are not very straightforward to interpret. In fact, one gets expressions for autocorrelation functions of the individual series and also for all cross-equation correlations. When m is large, this can amount to a large set of statistics, which should preferably all be investigated simultaneously. In practice, one therefore usually fits a set of VAR models with orders $1, 2, \ldots, K$, for some value of K, and evaluates whether one or more of these models fit the data well.

Estimation

The parameters in a VAR(p) model can be estimated using OLS per equation. For example, for the bivariate VAR(1) model in (9.31), one can estimate the ϕ_{ij} parameters by applying OLS to

$$y_{1,t} = \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + \varepsilon_{1,t}, \tag{9.55}$$

and to

$$y_{2,t} = \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + \varepsilon_{2,t}. \tag{9.56}$$

This gives consistent and efficient estimates even though the residual series $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ may be contemporaneously correlated. In other circumstances that would lead to the use of the seemingly unrelated regression [SUR] technique, but the fact that both regression models contain the same right-hand side variables makes that OLS per equation is the same as SUR, see Zellner (1962) and Lütkepohl (1991).

When the parameters in a VAR(p) model are estimated, some of these may seem insignificant. At this stage, however, it is not wise to set these parameters equal to zero. The main reason is that when the component series $y_{1,t}$ to $y_{m,t}$ of Y_t have stochastic trends, the t-ratios of the various parameters in a VAR(p) model will not be distributed as standard normal, see the last section of this chapter. When one is confident from the outset that the m time series are all stationary, one can of course set insignificant parameters equal to zero. It should be stressed, though, that one should use SUR in that case, as with certain parameter restrictions the m equations may not contain the same right-hand side regressors anymore.

Order selection

Suppose one estimates the parameters of several VAR models with orders from 1 to P. One may now examine the estimated residuals of each of these models. However, a more commonly applied procedure (which is slightly different from the selection

Table 9.1: Model selection results of a VAR(p) model for $p = 1, \ldots, 5$ estimated with the logarithm of gold and silver prices, 1986.1–2012.12

	lag order p				
	1	**2**	**3**	**4**	**5**
AIC	−7.067	−7.117	−7.119	−7.108	−7.092
SIC	−6.997	−7.001	−6.956	−6.898	−6.835

procedure in the univariate case), is first to choose order p, and then to examine the properties of the corresponding estimated residuals. Useful model selection criteria are the multivariate extensions of the Akaike and Schwarz information criteria, given by

$$\text{AIC} = \log|\hat{\Sigma}_p| + 2m^2 p/T, \qquad p = 1, 2, \ldots, P \tag{9.57}$$

$$\text{SIC} = \log|\hat{\Sigma}_p| + (\log n)m^2 p/T, \qquad p = 1, 2, \ldots, P \tag{9.58}$$

respectively, where $|\hat{\Sigma}_p|$ denotes the determinant of the residual covariance matrix for the VAR(p) model and T is the number of effective observations, see Lütkepohl (1991). Paulsen (1984) shows that these criteria perform well, even when the Y_t vector series contains unit roots.

As an illustration, consider the monthly gold and silver series, presented in Figure 2.21 during the period 1986.1–2012.12. A VAR(p) model for this bivariate series (all in logs) involves the estimation of $4p$ parameters. We set P equal to 5. The model selection results are summarized in Table 9.1.

The SIC obtains its smallest value for $p = 2$, while the AIC is smallest for $p = 3$. As the difference between the AIC values is small, it seems best to opt for $p = 2$. We will look later on at the estimated residuals in order to test for misspecification.

The estimation results for the VAR(2) model for $Y_t = (g_t, s_t)'$, that is,

$$Y_t = \hat{\mu} + \hat{\Phi}_1 Y_{t-1} + \hat{\Phi}_2 Y_{t-2} + \hat{\varepsilon}_t, \tag{9.59}$$

are

$$\hat{\Phi}_1 = \begin{bmatrix} 1.056^* & 0.037 \\ (-0.074) & (-0.041) \\ -0.213 & 1.224^* \\ (-0.129) & (-0.0972) \end{bmatrix} \hat{\Phi}_2 = \begin{bmatrix} -0.058 & 0.032 \\ (-0.073) & (-0.0541 \\ 0.282^* & -0.279^* \\ (-0.129) & (-0.072) \end{bmatrix} \hat{\mu} = \begin{bmatrix} -0.019 \\ (-0.024) \\ -0.072 \\ -0.042) \end{bmatrix}$$

where the estimated standard errors are given in parentheses. In case the data would be stationary, and hence the VAR(2) model would not have unit roots, the parameters

indicated with an asterisk would be considered as significant at the 5% level. These estimation results can be viewed as representative for many empirical unrestricted VAR models, that is, a substantial amount of the autoregressive parameters does not seem very relevant. Hence, it is worthwhile to see if the VAR(2) model for these gold en silver data can somehow be reduced.

As a tentative indication of the possible presence of unit roots in an empirical VAR model, one can check whether the eigenvalues of $\hat{\Phi}_1 + \hat{\Phi}_2 + \cdots + \hat{\Phi}_p$ are close to unity. For the VAR(2) model for the gold and silver data, we find that these two eigenvalues are estimated as 1.001 and 0.938. Hence the first value lies on the unit circle, while the second is close to it. A formal test for the number of unit roots will be discussed below.

Diagnostic testing

The investigation of the properties of the estimated residuals from a VAR(p) model is not a simple exercise. This is because one may wish to see whether all the individual $\hat{\varepsilon}_{i,t}$ series ($i = 1, 2, \ldots, m$) seem approximately white noise, but also whether there are no systematic patterns across current $\hat{\varepsilon}_{i,t}$ and lagged $\hat{\varepsilon}_{j,t}$, for any $i \neq j$. A practical problem with diagnostic checks for these properties is that it is often unclear in which direction one should modify the model, see also Tiao and Box (1981). In fact, one may need to add regressor variables to only a few of the m equations and not to all. Such new regressor variables may also differ across the equations, which implies that an alternative and perhaps more adequate model is not a VAR model anymore.

To obtain insights into the autocorrelation properties of the estimated residuals, one can consider the multivariate extensions of the portmanteau and LM tests for serial correlation, see Hosking (1980) and Poskitt and Tremayne (1982). A practical problem with these system tests is that a large number of estimated autocorrelations and cross-equation correlations is considered, and that many of these are required to be significant in order to make the overall test statistic significant. Put differently, the power of these tests is not very large in practically relevant cases.

One may therefore also consider the estimated autocorrelations of the residual series $\hat{\varepsilon}_{i,t}$ and some cross-equation correlations, that is, one can consider matrices like

$$r_k(\hat{e}_t) = \begin{bmatrix} r(\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{1,t-k}) & \ldots & r(\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{m,t-k}) \\ \vdots & & \vdots \\ r(\hat{\varepsilon}_{m,t}, \hat{\varepsilon}_{1,t-k}) & \ldots & r(\hat{\varepsilon}_{m,t}, \hat{\varepsilon}_{m,t-k}) \end{bmatrix} \tag{9.60}$$

The theoretical standard deviations of these correlations are not equal to $T^{-\frac{1}{2}}$, as shown in Chapter 4 of Lütkepohl (1991). In fact, when one uses $2T^{-\frac{1}{2}}$ as the supposedly 95% confidence interval, the true significance level is smaller than 5%.

Note that it is the judgement of the modeler to decide whether to modify the model when some of these correlations are unequal to zero. As an illustration of this situation, consider this matrix for $k = 1$ and $k = 2$ for the VAR(2) model for the gold and silver series, that is,

$$r_1(\hat{e}_t) = \begin{bmatrix} 0.014 & 0.006 \\ 0.013 & 0.019 \end{bmatrix}, \tag{9.61}$$

and

$$r_2(\hat{e}_t) = \begin{bmatrix} -0.132^* & -0.085^* \\ -0.101^* & -0.122^* \end{bmatrix}. \tag{9.62}$$

With 228 observations and assuming normality, the 95% confidence interval would be ± 0.056. This means that four correlations are significant at the 5% level, which in turn suggests that the VAR(2) model seems not reasonably adequate. It appears that estimating a VAR(3) model implies less significant cross correlations in the residuals. Hence, although the Information criteria suggest an VAR(2) model, diagnostics checks could lead to a different choice of the lag-order p.

Additional diagnostic tests for a VAR(p) model can concern parameter stability, normality, outliers, ARCH and nonlinearity. Tests for the first two features are given in Lütkepohl (1991). Multivariate tests for ARCH and nonlinearity are also available. In practice one typically starts with a range of VAR(p) models for a range of values for p. Then one selects a model using criteria like AIC or SIC and then one checks using (9.60) whether there are not too many significant correlations.

9.3 Applying VAR models

When the appropriate order p for the VAR model is found, that is, a VAR($p-1$) is misspecified and a VAR($p+1$) contains too many redundant parameters, and the parameters are estimated, the resultant empirical model can be used for various purposes. These are out-of-sample forecasting and, what is called, structural analysis. The latter can include the investigation of exogeneity and causality properties of certain variables, of impulse response functions and of forecast error variance decompositions. Under the assumption that a VAR model has not yet been simplified by setting individual parameters equal to zero, we will briefly treat each of these issues in this section. Some of the theoretical illustrations involve the bivariate VAR(1) model as in (9.31).

Forecasting

Similar to the univariate AR model, forecasts from a VAR(p) model amount to simple extrapolation schemes. Consider the VAR(1) model

$$Y_t = \Phi_1 Y_{t-1} + e_t, \tag{9.63}$$

and assume for notational convenience that the parameters are known. The one-step-ahead forecast at time T is given by

$$\hat{Y}_{T+1|T} = \Phi_1 Y_T, \tag{9.64}$$

and hence the h-step-ahead forecast equals

$$\hat{Y}_{T+h|T} = \Phi_1^h Y_T. \tag{9.65}$$

For a VAR(p) model one has

$$\hat{Y}_{T+h|T} = \Phi_1 \hat{Y}_{T+h-1|T} + \cdots + \Phi_h Y_T + \cdots + \Phi_p Y_{T+h-p}, \tag{9.66}$$

where $\hat{Y}_{T+h-1|T}, \ldots, \hat{Y}_{T+1|T}$ are forecasted using similar schemes, and where $\hat{Y}_j = Y_j$ for $j = 1, 2, \ldots, T$. In practice, one of course needs to substitute the estimated parameter matrices for the Φ_i, $i = 1, 2, \ldots, p$.

The theoretical one-step-ahead forecast error for a VAR(1) model equals e_{n+1}. The two-step-ahead forecast from a VAR(1) model is

$$\hat{Y}_{T+2|T} = \Phi_1^2 Y_T, \tag{9.67}$$

while the true observation at time $T + 2$ is

$$Y_{T+2} = \Phi_1 Y_{T+1} + e_{T+2} = \Phi_1^2 Y_T + e_{T+2} + \Phi_1 e_{T+1}. \tag{9.68}$$

The $(m \times m)$ forecast error covariance matrix for two steps ahead thus equals

$$\text{SPE}_2 = \Sigma + \Phi_1' \Sigma \Phi_1. \tag{9.69}$$

With the normality assumption, one can derive simple expressions for forecast intervals for the individual time series by considering the diagonal values of the estimated SPE_h matrices. As the forecast errors for $y_{i,t}$ are also affected by the forecasts for the other $m - 1$ $y_{j,t}$ variables, one may consider the determinant or the trace of the SPE_h matrices in order to compare rival empirical models.

As an example of an evaluation of the forecasting performance, consider the comparison of one-step ahead forecasts for gold en silver from the VAR(2) model and from univariate AR(p) models. For this purpose, we estimate the parameters in ((9.59)) for the period 1982.1–2000.12 and generate one-step-ahead forecasts for the period 2001.1–2012.4. The univariate AR models appear to be of order 2, and 1 for g_t and s_t, respectively, when we have used the methods outlined in Chapter 2 to arrive at this

decision. The mean squared prediction errors for g_t and s_t from the VAR(2) model are 0.0627, and 0.0850, and from the univariate models these are 0.0628 and 0.0827, respectively. Hence, for silver a univariate model yields better forecasts, while for gold the multivariate VAR(2) is slightly better.

Exogeneity and causality

Sometimes it may be useful to know (for economic theory considerations, but also for forecasting) whether a variable is exogenous to key parameters in the model. Loosely speaking, this means that one can then limit attention to a smaller set of equations. In other words, imposing exogeneity can imply a reduction of the number of parameters to be estimated and also an improved precision in forecasting. For example, consider the bivariate VAR(1) model again, that is,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}, \tag{9.70}$$

with e_t distributed with mean zero and covariance matrix Σ. When $\phi_{21} = 0$ and $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$, the $y_{2,t}$ variable is said to be strictly exogenous, see Engle *et al.* (1983) for more exogeneity concepts and test procedures. Notice that in this case, (9.70) reduces to a simple recursive system with the advantage that in order to construct a forecasting model for $y_{1,t}$, one can simply consider the first equation of (9.66).

The concept of Granger-causality, put forward in Granger (1969), bears similarities with the concept of exogeneity in the sense that it allows to draw inference on the dynamic impact of one variable on another. Such inference can be given an economically meaningful interpretation. The concept of Granger-causality draws upon the concept of forecastability. For example, for a bivariate series, the variable $y_{2,t}$ is said to be Granger-non-causal for $y_{1,t}$ if

$$E(y_{1,t}|Y_{1,t-1}, Y_{2,t-1}) = E(y_{1,t}|Y_{1,t-1}), \tag{9.71}$$

that is, the past of $y_{2,t}$ does not help to forecast $y_{1,t}$. For the bivariate model (9.70), Granger non-causality of $y_{2,t}$ implies that $\phi_{12} = 0$.

Impulse response functions

To create out-of-sample forecasts as discussed before, one substitutes known and forecasted values of Y_t into the forecasting schemes. As such, it is not easy to see what are the dynamic effects of the error process e_t. To understand that, one can calculate the so-called impulse-response function. Suppose Y_t can be described by

$$Y_t = \Phi_1 Y_{t-1} + e_t, \quad e_t \sim (0, \text{diag}[\sigma_1^2, \ldots, \sigma_m^2]), \tag{9.72}$$

and suppose there is an interest in the effect of shocks corresponding to the first variable. One can then calculate

$$X_0 = e_0^* = (\sigma_1 \quad 0 \quad 0 \quad 0 \quad \dots 0)' \tag{9.73}$$

$$X_1 = \Phi_1 e_0^*$$

$$X_2 = \Phi_1^2 e_0^*$$

$$\vdots$$

$$X_h = \Phi_1^h e_0^*. \tag{9.74}$$

The resulting trajectories, that is, the m elements of the X_t vector series, $t = 0, 1, 2, \dots, h$ are called the impulse-response functions. When e_t has covariance matrix Σ with non-zero off-diagonal elements, e_0^* is replaced by Pe_0^*, where P is defined by $\Sigma = PP$, and P is a lower triangular matrix. Notice that when the m components of Y_t are arranged differently, one has a different P matrix. This implies that the m series contained in X_t can become different across different arrangements of Y_t.

The long-run impulse-response matrix for a VAR(p) model can be calculated as

$$\Psi_\infty = (I_m - \Phi_1 - \dots - \Phi_p)^{-1}. \tag{9.75}$$

For the empirical VAR(2) model for our gold and silver series, where Y_t is arranged as (g_t, s_t), this matrix is estimated as

$$\Psi_\infty = \begin{bmatrix} -234 & -21 \\ -293 & -8.5 \end{bmatrix}, \tag{9.76}$$

for which we can conclude for example that silver has a negative long-run effect on gold and vice versa. These long-run impact results are illustrated by the estimated impulse response functions for the two variables based on a VAR(2) model in Figure 9.1.

The most obvious feature of these graphs is that, even after 50 periods, some of the effects of shocks do not become zero. Because of this large persistence of the shocks, there may be unit roots in this VAR model, which we will investigate below.

Variance decomposition

A final structural application of VAR models is that one can investigate which part of the error variance is caused by the error variance of which variable. Consider again the

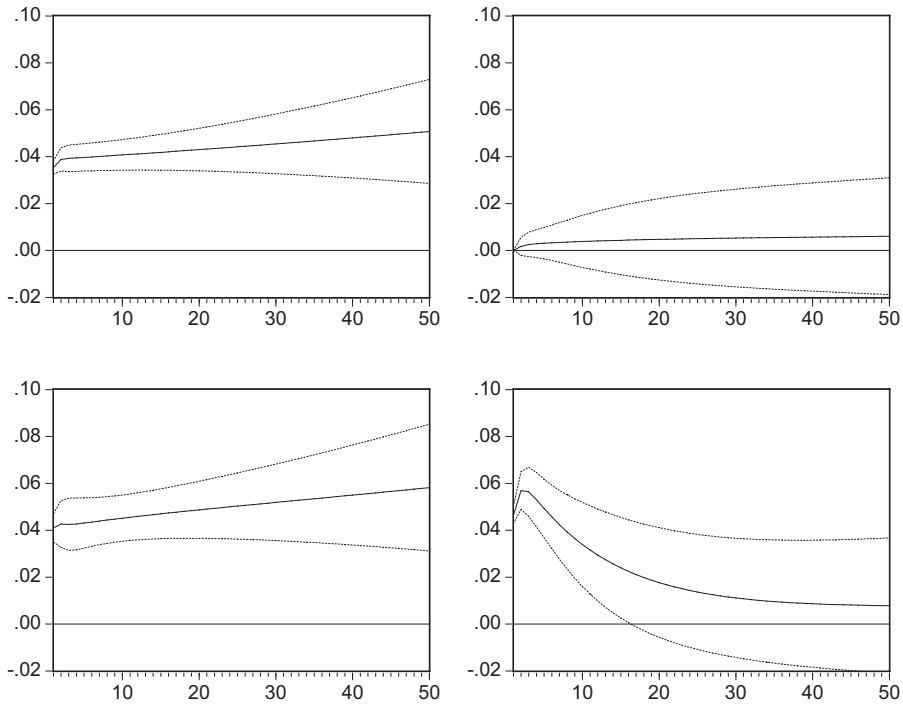


Figure 9.1: Impulse response function with 95% confidence bounds. The upper panel shows the effect of a one standard deviation shock at time $t = T$ in gold prices on gold prices(left) and silver prices (right) for period $t = T + 1, \ldots T + 50$. The lower panel shows the effect of a one standard deviation shock in silver prices on gold prices (left) and silver prices (right).

bivariate model in (9.31) with its corresponding two step ahead forecast errors

$$\begin{bmatrix} \varepsilon_{1,t+2} \\ \varepsilon_{2,t+2} \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t+1} \\ \varepsilon_{2,t+1} \end{bmatrix} \tag{9.77}$$

For the first variable, and assuming that $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$, the (forecast) error variance equals

$$\sigma_1^2(1 + \phi_{11}^2) + \phi_{12}^2\sigma_2^2 \tag{9.78}$$

The first part of (9.78) can be seen to be due to the own error variance, while the $\sigma_{12}^2\sigma_2^2$ part is due to the second variable. Expressed as percentages, one can thus examine the relative importance of the error variance of variable $y_{i,t}$ for forecasting $y_{j,t}$.

Table 9.2: Decomposed error variance in percentages of a VAR(2) model for 1 until 10 periods ahead, estimated using (the logarithm of) gold and silver prices, 1986.1–2012.12

	Decomposition of gold			Decomposition of silver		
	S.E.	% Gold	% Silver	S.E.	% Gold	% Silver
1	0.035	100	0.0	0.062	43.7	56.3
2	0.053	99.9	0.1	0.094	39.3	60.7
3	0.066	99.8	0.2	0.118	38.2	61.8
4	0.077	99.7	0.3	0.136	38.5	61.5
5	0.086	99.6	0.4	0.151	39.4	60.6
6	0.095	99.6	0.4	0.164	40.6	59.4
7	0.103	99.5	0.5	0.175	42.0	58.0
8	0.111	99.5	0.5	0.184	43.4	56.6
9	0.118	99.5	0.5	0.193	44.9	55.1
10	0.125	99.4	0.6	0.201	46.4	53.6

As an illustration, consider the decomposed error variances for the VAR(2) model for gold and silver series in Table 9.2. The first (left-hand side) panel shows that errors in gold prices are mainly due to uncertainty in gold prices itself. The second panel for silver shows that these errors are also large due to gold. This can be interpreted as that silver is much related with gold, while the gold price depends less on the silver price.

Conclusion

So far, we have reviewed various concepts in multivariate time series analysis. The key assumption underlying most of the previous discussion is that the data are stationary. For several business and economic time series, this assumption does not hold. It is therefore important to carefully analyze trend properties in multivariate systems. In the next two sections, we will focus on methods for such an analysis.

9.4 Cointegration: some preliminaries

When two or more time series each have a stochastic trend, they may have such trends in common. Incorporating common trends in a multivariate time series model reduces the number of parameters and it helps with forecasting. For example, when x_t and y_t have a trend, while $x_t - y_t$ has not, one may need only to model the trend in y_t in order to forecast the trend in x_t. A common trend can also be relevant from an economic point of view. For example, economic theory predicts that consumption and income should somehow have similar trends as in equilibrium these two quantities should be approximately equal.

Since the appearance of article of Engle and Granger (1987) (for which they were awarded the Nobel Prize in Economics in 2003), the issue of cointegration has attracted an enormous amount of attention. An important reason for this attention is that cointegration allows for a solution to the problem of spurious regressions. Consider the two time series variables that are independently generated as

$$y_{1,t} = y_{1,t-1} + \varepsilon_{1,t}$$
$$y_{2,t} = y_{2,t-1} + \varepsilon_{2,t},$$

that is, as independent random walks, and suppose one considers the static regression

$$y_{1,t} = \beta y_{2,t} + u_t,$$

the estimated β parameter will seem significant and it will have a large absolute t-ratio and also the R^2 of this regression will be close to unity, see Granger and Newbold (1974) for simulation evidence and Phillips (1986) for a formal statistical treatment. Apparently, if one does not properly take account of stochastic trends, one tends to find spurious results. Cointegration analysis is then very useful.

In this section we introduce some concepts and notation, where the analysis is restricted to a bivariate VAR(1) model. To save space, we will highlight only three often applied cointegration methods. These are the methods proposed in Engle and Granger (1987), Johansen (1991) and Boswijk (1994). We abstain from many technical details and refer the interested reader for more rigorous treatments of cointegration to Banerjee *et al.* (1993), Hendry (1995), and Johansen (1995). The Johansen method is the one that is most often used, and we discuss more details in the final section of this chapter.

Representation

Consider two time series $y_{1,t}$ and $y_{2,t}$, and assume that these can be described by the following bivariate model, that is,

$$y_{1,t} + \delta y_{2,t} = v_t, \qquad v_t = \mu_1^* + \rho_1 v_{t-1} + \varepsilon_{1,t}^*, \qquad 0 \leq \rho_1 \leq 1 \tag{9.79}$$

$$y_{1,t} + \eta y_{2,t} = w_t, \qquad w_t = \mu_2^* + \rho_2 w_{t-1} + \varepsilon_{2,t}^*, \qquad 0 \leq \rho_2 \leq 1 \tag{9.80}$$

where $\delta \neq \eta$. The latter restriction prevents δ and η from being equal to zero at the same time, although either δ or η may be equal to zero. As we will show below, this is simply a VAR(1) model, but for the moment this notation is quite convenient. The μ_1^* and μ_2^* are intercept terms, and it is assumed that $\varepsilon_{1,t}^*$ and $\varepsilon_{2,t}^*$ are standard white noise error processes, which are mutually independent at all lags. The two equations (9.79)–(9.80) reflect that two distinct linear combinations of $y_{1,t}$ and $y_{2,t}$ can be described by AR(1) models.

The interpretation of the two linear combinations depends on the values of ρ_1 and ρ_2. In this bivariate case, there are three relevant cases, each of which implies a different interpretation of (9.79) and/or (9.80). The first is that $\rho_1 = \rho_2 = 1$, which implies that any linear combination of $y_{1,t}$ and $y_{2,t}$ is a random walk variable (possibly with drift if μ_1^* or μ_2^* is unequal to zero). In turn, this implies that $y_{1,t}$ and $y_{2,t}$ are I(1) variables and hence are non-stationary themselves. In this first case, $y_{1,t}$ and $y_{2,t}$ individually have a stochastic trend as $\rho_1 = \rho_2 = 1$, but they do not have such a trend in common as no linear combination of $y_{1,t}$ and $y_{2,t}$ is stationary. These variables are now said not to be cointegrated.

The second case is that both $0 \leq \rho_i < 1$ for $i = 1, 2$. This implies that any linear combination of $y_{1,t}$ and $y_{2,t}$ is a stationary AR(1) process, and therefore $y_{1,t}$ and $y_{2,t}$ themselves are stationary variables. The individual series thus do not need to be differenced in order to obtain stationary time series.

The third and often most interesting case is when $\rho_1 = 1$ and $0 \leq \rho_2 < 1$ (or, similarly, when $\rho_2 = 1$ and $0 \leq \rho_1 < 1$). There is then one linear combination of $y_{1,t}$ and $y_{2,t}$ which is a stationary AR(1) process, while the other combination is a random walk (with drift). This implies that, although both $y_{1,t}$ and $y_{2,t}$ individually are I(1) time series, which can be seen by suitably subtracting (9.80) from (9.79), there is one linear combination which is stationary. In this case, the two time series are said to be cointegrated, see Engle and Granger (1987). When there is such a stationary relationship between $y_{1,t}$ and $y_{2,t}$, when they individually have a stochastic trend, cointegration among $y_{1,t}$ and $y_{2,t}$ implies that these series also have a common stochastic trend (as will become clear below).

To illustrate matters, Figure 9.2 provide a graph of generated series in case $\delta = 0$, $\eta = -1$, $\mu_1^* = \mu_2^* = 0.5$, $\rho_1 = 1$, $\rho_2 = 0.9$ in Figure 9.2. The graphs in this Figure

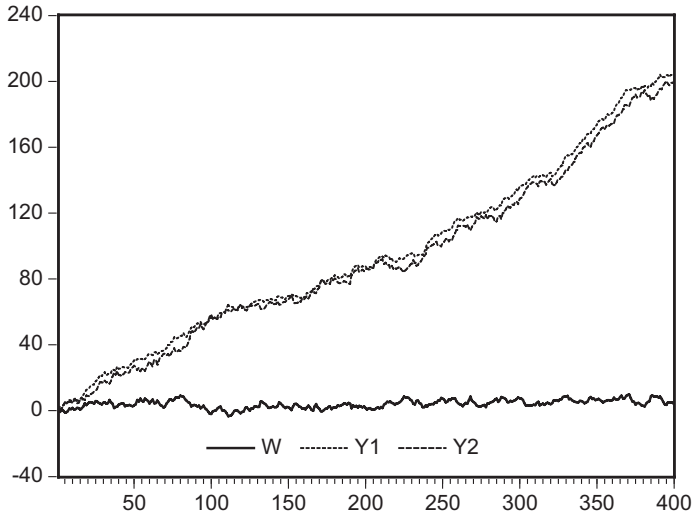


Figure 9.2: Generated series according to (9.79) and (9.80).

clearly show that both $y_{1,t}$, and $y_{2,t}$ are trending, but that the linear combination $w_t = y_{1,t} - y_{2,t}$ does not have such a trend. In other words, $y_{1,t}$ and $y_{2,t}$ appear to have a common stochastic trend.

The model framework in (9.79)–(9.80) immediately suggests a simple method to test for cointegration between two variables, which is the one initially put forward in Engle and Granger (1987). It amounts to estimating δ or η in (9.79) or (9.80), and then testing whether either ρ_1 or ρ_2 is equal to 1 using the ADF regression outlined in Chapter 4. More formally, one first performs the regression

$$y_{1,t} = \hat{\psi} + \hat{\lambda} y_{2,t} + \hat{u}_t, \tag{9.81}$$

and then considers the auxiliary test regression

$$\Delta_1 \hat{u}_t = \pi + \rho \hat{u}_{t-1} + \theta_1 \Delta_1 \hat{u}_{t-1} + \cdots + \theta_p \Delta_1 \hat{u}_{t-p} + \nu_t, \tag{9.82}$$

and evaluates the t-test for the significance of ρ. When $\rho = 0$, $\hat{u}_t$ has a unit root and $y_{1,t} - \hat{\psi} - \hat{\lambda} y_{2,t}$ is a nonstationary time series. In that case, (9.81) does not reflect a stationary cointegration relationship. When $\rho < 0$, that is, when the $t(\hat{\rho})$ value is significantly negative, $y_{1,t}$ and $y_{2,t}$ are cointegrated. Notice that (9.79) and (9.80) imply that $p = 0$ in (9.82). In practice, the value of p can be selected using diagnostic tests or the AIC or SIC. Some critical values of this t-test for ρ are given in Table 9.3 in case of m variables, with $m = 2, 3, 4$, and 5.

Table 9.3: Asymptotic critical values for the Engle and Granger (1987) method. The test regression is (9.82) where the variable of interest is u_t, with u_t is the estimated residual series from the regression of (9.81) where both regressions may include a constant and/or a trend.

Number of variables	Significance level		
m	0.01	0.05	0.10
Regressions contain a constant			
2	−3.96	−3.37	−3.07
3	−4.31	−3.77	−3.45
4	−4.73	−4.11	−3.83
5	−5.07	−4.45	−4.16
Regressions contain a constant and a trend			
2	−4.36	−3.80	−3.52
3	−4.65	−4.16	−3.84
4	−5.04	−4.49	−4.20
5	−5.58	−5.03	−4.73

Source: Phillips and Ouliaris (1990).

Asymptotic theory for the Engle-Granger method is presented in Phillips and Ouliaris (1990). Comparing the relevant critical values in Table 9.3 with those in Table 4.1 for $m = 1$, it can be seen that the ADF critical values shift to the left in case of more than one variable. For small sample sizes, one can use the critical values tabulated in MacKinnon (1991). Finally, it should be mentioned that one may also evaluate a regression of $y_{2,t}$ on $y_{1,t}$ using the similar method. When the R^2 of these regressions is very close to 1, it may not matter much which variable is regressed on the other. When this is not the case, it becomes relevant which variable is chosen to be the regressand, see Ng and Perron (1997).

As an illustration of the Engle-Granger method, consider again the gold and silver price series. The graphs of these series suggest a common pattern, that is, perhaps a linear combination is stationary. We analyze $y_{1,t} = \log(\text{gold price})$ and $y_{2,t} = \log(\text{silver price})$ for the sample 1986.01–2012.12. The first regression

result is

$$y_{1,t} = \underset{(0.072)}{0.967} + \underset{(0.013)}{0.817} y_{2,t}, \qquad R^2 = 0.928 \tag{9.83}$$

where the estimated standard errors are given in parentheses. As the individual series have a unit root, the t-ratios for these parameters cannot be compared with the standard normal tables. The ADF test result for the estimated residuals, where p can be set equal to 0, is -3.658. Comparing this with the 5 percent critical value of -3.37 in Table 9.1 shows that $y_{1,t} - 0.817 y_{2,t}$ seems a stationary variable. The reversed regression results in

$$y_{2,t} = \underset{(0.099)}{-0.692} + \underset{(0.018)}{1.135} y_{1,t}, \qquad R = 0.929 \tag{9.84}$$

with an ADF value of -3.773 also for $p = 0$. Combining (9.84))with (9.83) we can be reasonably confident that a linear combination of the logs of the gold and silver prices is stationary, while they individually have the stochastic trend feature.

The Engle-Granger method is useful when one analyses only two time series. It becomes less useful for more than two time series as then the number of possible cointegration relations increases with the number of time series, and this implies an increasing difficulty to decide on which variable is on the left-hand side. Also, with three of more series there may be two or more cointegration relations. The corresponding parameters cannot be identified from a single regression. The method still is very useful if one knows a priori that there is only a single cointegrating relation. In other situations, multivariate methods may be more useful.

Vector autoregressive representation

The expressions in (9.79)–(9.80) can be summarized as

$$\begin{bmatrix} 1 & \delta \\ 1 & \eta \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1^* \\ \mu_2^* \end{bmatrix} + \begin{bmatrix} \rho_1 & \delta\rho_1 \\ \rho_2 & \eta\rho_2 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t}^* \\ \varepsilon_{2,t}^* \end{bmatrix}. \tag{9.85}$$

Multiplying both sides of (9.85) with the inverse of the left-hand matrix and subtracting the one period lagged Y_{t-1} from both sides gives

$$\Delta_1 Y_t = \mu + \Pi Y_{t-1} + e_t \tag{9.86}$$

with $e_t = (\varepsilon_{1,t}, \varepsilon_{2,t})$, where the two $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ series are functions of $\varepsilon^*_{1,t}$ and $\varepsilon^*_{2,t}$ and η and δ, and with

$$\mu = \begin{bmatrix} (\eta\mu_1^* - \delta\mu_2^*)/(\eta - \delta) \\ (\mu_2^* - \mu_1^*)/(\eta - \delta) \end{bmatrix} \text{ and} \tag{9.87}$$

$$\Pi = \begin{bmatrix} (\eta\rho_1 - \delta\rho_2 - \eta + \delta)/(\eta - \delta) & \eta\delta(\rho_1 - \rho_2)/(\eta - \delta) \\ (\rho_2 - \rho_1)/(\eta - \delta) & (\eta\rho_2 - \delta\rho_1 - \eta_\delta)/(\eta - \delta) \end{bmatrix}. \tag{9.88}$$

When $0 \leq \rho_i < 1$ for $i = 1, 2$ (and given that η and δ are not both equal 0), the Π matrix in (9.88) has full rank 2. On the other hand, when $\rho_1 = \rho_2 = 1$, it can easily be observed that all elements of Π equal zero, and hence that the rank of Π is equal to 0.

The interesting intermediate case is again the cointegration case. For example, when $\rho_1 = 1$ and $0 \leq \rho_2 < 1$, the Π matrix can be written as

$$\Pi = \alpha\beta', \tag{9.89}$$

with

$$\alpha = \begin{bmatrix} \delta(1 - \rho_2)/\eta - \delta) \\ -(1 - \rho_2)/(\eta - \delta) \end{bmatrix} \text{ and } \beta = \begin{bmatrix} 1 \\ \eta \end{bmatrix}. \tag{9.90}$$

The (2×2) matrix Π equals the outer product of two (2×1) matrices, which defines that matrix Π has rank 1. This leads to a reduction from 4 to 3 parameters in Π. In general, cointegration reduces the number of parameters in a VAR model. Notice however that the decomposition in (9.90) is not unique, and that we can find other nonlinear parameter restrictions on Π which also correspond with cointegration. In this cointegration case, the characteristic polynomial of the VAR(1) model in (9.86), that is, $| I_m - (\Pi + I_m)z | = 0$, can be shown to yield one solution on the unit circle. The implied univariate models for $y_{1,t}$, and $y_{2,t}$ are ARIMA(1,1,1) models. Hence, both series have a unit root, while the vector series only has a single unit root.

The parameter vector β in (9.89) contains the cointegration parameters, that is, the parameters that indicate an equilibrium relation between $y_{1,t}$ and $y_{2,t}$. In this bivariate example, the equilibrium (or long-run) relation is $y_{1,t} + \eta y_{2,t}$. The (2×1) parameter matrix α in (9.89) contains the two so-called adjustment parameters. These α_1 and α_2, say, reflect the speed of adjustment towards equilibrium, which is most easily seen from writing the two equations in (9.86) with (9.89) as

$$\Delta_1 y_{1,t} = \mu_1 + \alpha_1(y_{1,t-1} + \eta y_{2,t-1}) + \varepsilon_{1,t} \tag{9.91}$$

$$\Delta_1 y_{2,t} = \mu_2 + \alpha_2(y_{1,t-1} + \eta y_{2,t-1}) + \varepsilon_{2,t}, \tag{9.92}$$

where each equation is called an error correction model [ECM]. Together, this system is called a vector error correction model [VECM].

Impact of the constant μ

With the cointegration restriction that $\rho_1 = 1$ and $0 \leq \rho_2 < 1$, the VAR(1) model in (9.86) can be written as

$$\Delta_1 Y_t = \mu + \alpha\beta' Y_{t-1} + e_t. \tag{9.93}$$

In Johansen (1995) it is shown that this model can be solved by recursive substitution as

$$Y_t = Y_0^* + C\mu t + C\sum_{i=1}^{t} e_i + X_t, \tag{9.94}$$

where X_t is a stationary bivariate time series, and where C is defined by

$$C = \beta_\perp(\alpha_\perp'\beta_\perp)^{-1}\alpha_\perp', \tag{9.95}$$

and $\alpha_\perp$ and $\beta_\perp$ are defined by

$$\alpha_\perp'\alpha = 0 \text{ and } \beta_\perp'\beta = 0. \tag{9.96}$$

When $m = 1$, $C = 1$, we see that (9.94) equals (4.12) for a univariate random walk with drift series.

The (2×1) vector series $\sum_{i=1}^{t} e_i$ in (9.94) contains the accumulated sums of the errors $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$. As cointegration relation amongst two time series implies the presence of a common trend, (9.94) suggests that $z_t = \alpha_\perp' \sum_{i=1}^{t} e_i$ is the common trend, see Johansen (1995). Note that in our bivariate cointegrated example, with one common stochastic trend, z_t is a univariate time series. Multiplying both sides of (9.94) with $\alpha_\perp'$ results in

$$\alpha_\perp' Y_t = \alpha_\perp' Y_0^* + \alpha_\perp'\mu t + \alpha_\perp' \sum_{i=1}^{t} e_i + \alpha_\perp' X_t, \tag{9.97}$$

and hence $\alpha_\perp' Y_t$ has a stochastic trend component. For our bivariate example in (9.79)–(9.80), and given (9.90), it is easy to see that

$$\alpha_\perp' Y_t = (1 \;\; \delta)\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = y_{1,t} + \delta y_{2,t} = v_t, \tag{9.98}$$

where v_t is defined in (9.79). Indeed, with $\rho = 1$, the v_t series is a random walk with drift. Equation (9.97) shows that $\alpha_\perp' Y_t$ has a deterministic trend component $\alpha_\perp'\mu t$, where in our example

$$\begin{aligned} \alpha_\perp'\mu &= (\eta - \delta)^{-1}[(\eta\mu_1^* - \delta\mu_2^*) + \delta(-\mu_1^* + \delta\mu_2^*)] \\ &= \mu_1^*. \end{aligned} \tag{9.99}$$

Hence, when $\alpha'_{\perp}\mu = 0$, that is, when $\mu_1^* = 0$, the common stochastic trend in $y_{1,t}$ and $y_{2,t}$ does not have a drift. In case $\mu_1^* = 0$ in (9.99), while $\rho_1 = 1$ and $0 \leq \rho_2 < 1$, the ECM equations in (9.91) and (9.92) become

$$\Delta_1 y_{1,t} = \alpha_1 \left[\frac{-\mu_2^*}{1-\rho_2} + y_{1,t-1} + \eta y_{2,t-1} \right] + \varepsilon_{1,t} \tag{9.100}$$

$$\Delta_1 y_{2,t} = \alpha_2 \left[\frac{-\mu_2^*}{1-\rho_2} + y_{1,t-1} + \eta y_{2,t-1} \right] + \varepsilon_{2,t}. \tag{9.101}$$

The constant terms in the equilibrium relations are restricted to be equal. These two equations show that the error correction variable $y_{1,t} + \eta y_{2,t}$ has mean $\mu_2^*/(1-\rho_2)$, and this could already be seen from (9.80). If one would generate data from (9.100) and (9.101), we would observe that $y_{1,t}$ and $y_{2,t}$ display random walk behavior, although there does not seem to be a clear trend in the data.

In sum, there are three relevant cases concerning the intercept terms for a cointegrated bivariate system based on (9.79)–(9.80). The first is that both μ_1^* and μ_2^* are equal to zero. The second is that the intercepts are both unrestricted, that is, the common stochastic trend has a deterministic trend component. The third case is the case where the intercept terms are restricted such that the underlying common stochastic trend does not display a trending pattern and hence neither do the individual time series.

When the v_t and w_t series in (9.79) and (9.80) are generated by

$$v_t = \mu_1^* + \tau_1^* t + \rho_1 v_{t-1} + \varepsilon_{1,t}^* \qquad 0 \leq \rho_1 \leq 1 \tag{9.102}$$

$$w_t = \mu_2^* + \tau_2^* t + \rho_2 w_{t-1} + \varepsilon_{2,t}^* \qquad 0 \leq \rho_2 \leq 1, \tag{9.103}$$

the common common stochastic trend displays quadratic trend behavior when the τ_1^* and τ_2^* parameters are unequal to zero. When the data do not have quadratic trend-like properties, one may want to restrict the parameters τ_1^* and τ_2^* to zero. In that case, the term $[\frac{-\tau_2^*}{1-\rho_2}]t$ should enter the expression in parentheses in equations (9.100) and (9.101). We will return to the most relevant cases concerning intercepts and trends when reviewing the cointegration test statistics based on VAR models.

9.5 Inference on cointegration

Consider again the VAR(p) model

$$Y_t = \mu + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} + e_t, \tag{9.104}$$

for an $(m \times 1)$ time series Y_t, containing $y_{1,t}$ through $y_{m,t}$. Similar to (9.86), it is convenient to write (9.104) in error correction format, that is

$$\Delta_1 Y_t = \mu + \Gamma_1 \Delta_1 Y_{t-1} + \cdots + \Gamma_{p-1} \Delta_1 Y_{t-p+1} + \Pi Y_{t-p} + e_t, \qquad (9.105)$$

where

$$\Gamma_i = (\Phi_1 + \Phi_2 + \cdots + \Phi_i) - I_m, \quad \text{for } i = 1, 2, \cdots, p-1, \qquad (9.106)$$

$$\Pi = \Phi_1 + \Phi_2 + \cdots + \Phi_p - I_m, \qquad (9.107)$$

where the $m \times m$ matrix Π contains the information on possible cointegrating relations between the m elements of Y_t. Notice that when $\sum_{i=1}^{p} = \Phi_i$ has eigenvalues close to unity, Π has eigenvalues close to 0. The latter implies that the matrix Π is close to rank deficiency, and hence there may be cointegration.

A statistical method is now needed to investigate whether the rank of Π differs from zero or from m. An elegant approach is proposed in Johansen (1988), see also Johansen (1991, 1995). The Johansen method essentially amounts to a multivariate extension of the univariate ADF method. In fact, like in the univariate case, for the VAR(p) model with autoregressive polynomial $\Phi_p(L)$, one can write

$$\Phi_p(L) = -\Pi L^p - \Gamma_{p-1}(L)(1 - L), \qquad (9.108)$$

where Π is defined in (9.107) and where $\Gamma_{p-1}(L)$ is a $(p-1)$-th order matrix polynomial.

In the VAR model in (9.105) there are three interesting cases. First, the matrix Π can be the zero matrix, which implies that the rank of Π equals 0. Second, the matrix Π can have full rank m. Third, matrix Π has rank deficiency, that is, $0 < \text{rank}\Pi < m$, and it can be decomposed as $\Pi = \alpha\beta'$, where α and β are $(m \times r)$ parameter matrices. In the bivariate example above we had $m = 2$ and $r = 1$. The matrix β contains the r cointegrating relations, while the matrix α contains the adjustment parameters. Again, these α and β matrices are not unique, so it is more accurate to say that the columns of β span the space with cointegration vectors. When there are r cointegration relations, there are $m - r$ common stochastic trends, see Engle and Granger (1987) and Johansen (1991).

The Johansen maximum likelihood cointegration testing method aims to test the rank of the matrix Π in (9.105) using the reduced rank regression technique based on canonical correlations. The idea for using this technique (for a bivariate example) is that we want to find the linear combination $y_{1,t} + \eta y_{2,t}$ which has the largest partial correlation with any linear combination of the stationary variables $\Delta_1 y_{1,t}$ and $\Delta_1 y_{2,t}$. This amounts to the following computations for the case where the intercepts in the m equations are unrestricted. First, we regress $\Delta_1 Y_t$ and Y_{t-p} on a constant and on lagged $\Delta_1 Y_{t-1}$ through $\Delta_1 Y_{t-p+1}$ variables. This results in $(m \times 1)$ vectors of residuals r_{0t}

and r_{1t} and the $(m \times m)$ residual product matrices

$$S_{ij} = \frac{1}{n} \sum_{t=1}^{T} r_{it} r'_{jt}, \qquad \text{for } i, j = 0, 1 \tag{9.109}$$

respectively. When we wish to restrict the constants, we should also regress a unity vector on the $\Delta_1 Y_{t-1}$ to $\Delta_1 Y_{t-p+1}$ variables. Adding these covariances to the rows and columns of (9.99) results in S_{ij} matrices of dimension $(m+1) \times (m+1)$. In the other two cases concerning possible quadratic trends, see (9.102)–(9.103), one should perform similar calculations with either a trend variable on the right-hand side or on the left-hand side of the auxiliary regressions.

The next step is to solve the eigenvalue problem

$$|\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0 \tag{9.110}$$

which gives the eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_m$ and the corresponding eigenvectors $\hat{\beta}_1$ through $\hat{\beta}_m$. A test for the rank of Π can now be performed by testing how many eigenvalues λ_i equal unity. The first test statistic for the resulting number of cointegration relations proposed in Johansen (1988) is the so-called Trace test statistic

$$Trace = -T \sum_{i=r+1}^{m} \log(1 - \hat{\lambda}_i). \tag{9.111}$$

The null hypothesis for this *Trace* test is that there are at most r cointegration relations.

We begin with testing whether there is no cointegration ($r = 0$) versus at most 1 such relation. If this is rejected, one tests whether there are at most two cointegration relations. When we would finally reject the null hypothesis of at most $r = m - 1$ cointegration relations, we would find that the Y_t vector series is stationary. In practice, often the null hypothesis of at most $r - 1$ cointegrating relations is rejected, while the hypothesis of at most r such relations is not, indicating that there are r cointegrating relations. Another useful test statistic is given by testing the significance of the estimated eigenvalues themselves, or

$$\lambda_{max} = -T \log(1 - \hat{\lambda}_r), \tag{9.112}$$

which can be used to test the null hypothesis of $r - 1$ against r cointegration relations. Similar to the univariate ADF test for unit roots, critical values for the *Trace* and λ_{max} tests have to be simulated. Asymptotic distributions of the test statistics are summarized in Johansen (1995). Critical values for three practically relevant cases (case I, II and IV) are given in Table 9.4.

To illustrate the practical implementation of the Johansen method, consider the white and black pepper series, see Figure 9.3. The $y_{1,t}$ and $y_{2,t}$ series concern monthly data on the logarithm of white and black pepper during the period 1973.10-1996.04. We

Table 9.4: Asymptotic critical values for the Johansen cointegration method

	Trace test statistic				**λ_{max} test statistic**			
$m-r$	0.20	0.10	0.05	0.01	0.20	0.10	0.05	0.01
I: The regression model contains no constants and no deterministic trends								
1	1.82	2.86	3.84	6.51	1.82	2.86	3.84	6.51
2	8.45	10.47	12.53	16.31	7.58	9.52	11.44	15.69
3	18.83	21.63	24.31	29.75	13.31	15.59	17.89	22.99
4	33.16	36.58	39.89	45.58	18.97	21.58	23.80	28.82
5	51.13	55.44	59.46	66.52	24.83	27.62	30.04	35.17
II: The regression model contains constants but no deterministic trends while the data display linear trending patterns (the parameters for the intercepts are unrestricted)								
1	4.82	6.50	8.18	11.65	4.82	6.50	8.18	11.65
2	13.21	15.66	17.95	23.52	10.77	12.91	14.90	19.19
3	25.39	28.71	31.52	37.22	16.51	18.90	21.07	25.75
4	41.65	45.23	48.28	55.43	22.16	24.78	27.14	32.14
5	61.75	66.49	70.60	78.87	28.09	30.84	33.32	38.78
III: The regression model contains constants but no deterministic trends, while the data do not display linear trending patterns (the parameters for the intercepts are restricted)								
1	5.91	7.52	9.24	12.97	5.91	7.52	9.24	12.97
2	15.25	17.85	19.96	24.60	11.54	13.75	15.67	20.20
3	28.75	32.00	34.91	41.07	17.40	19.77	22.00	26.81
4	45.65	49.65	53.12	60.16	22.95	25.56	28.14	33.24
5	66.91	71.86	76.07	84.45	28.76	31.66	34.40	39.79

Table 9.4: (*cont.*)

IV: The regression model contains constants and deterministic trends, while the data display quadratic trending patterns (the parameters for the trends are unrestricted)								
1	7.78	9.66	11.55	15.78	7.78	9.66	11.55	15.78
2	18.30	20.87	23.37	28.80	13.76	15.99	18.04	22.41
3	32.60	36.03	39.04	45.37	19.42	22.06	23.97	28.65
4	50.49	54.79	58.57	65.73	25.28	27.76	30.31	35.60
5	72.48	77.77	82.18	90.83	30.89	33.96	36.65	42.05
V: The regression model contains constants and deterministic trends, while the data have linear trends but no quadratic trending patterns (the parameters for the trends are restricted)								
1	8.65	10.49	12.25	16.26	8.65	10.49	12.25	16.26
2	20.19	22.76	25.32	30.45	14.70	16.85	18.96	23.65
3	35.56	39.06	42.44	48.45	20.45	23.11	25.54	30.34
4	54.80	59.14	62.99	70.05	26.30	29.12	31.46	36.65
5	77.83	83.20	87.31	96.58	31.72	34.75	37.52	42.36

Sources: Osterwald-Lenum (1992) and Johansen (1995).
Note: The statistics are given in (9.111) and (9.112). m denotes the number of time series in Y_t, r is the number of cointegration relationships.

find that a VAR(2) model describes the two level series most adequately. The Johansen method, based on a VAR(2) model without an intercept, results in $\hat{\lambda}_1 = 0.048$ and $\hat{\lambda}_2 = 0.001$. This gives *Trace* test statistics with value 13.26 and 0.16 and λ_{max} test statistics with values 13.10 and 0.16. Comparing these values with the relevant critical values (case I) in Table 9.4 shows that the values 13.26 and 13.10 are significant at the 5 percent level. Hence, there is evidence of cointegration among our bivariate pepper series. After rescaling, the cointegration relation can be written as

$$\text{cointegration} = y_{1,t} - 1.042 y_{2,t} \tag{9.113}$$

This equilibrium relation implies that in the long run white pepper and black pepper are correlated positively. Figure 9.4 shows this cointegration relation. As indicated by

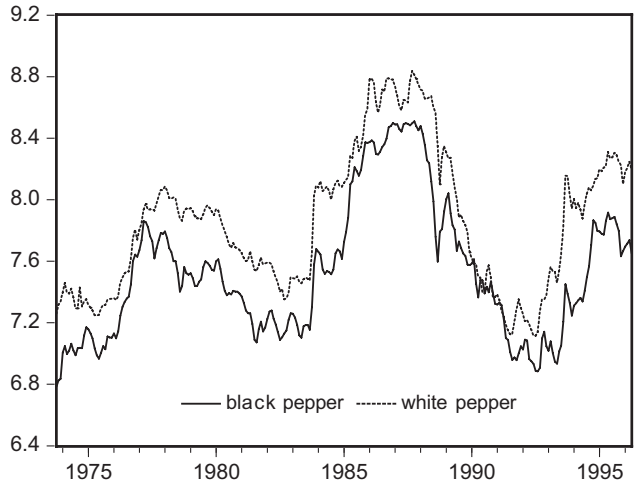


Figure 9.3: Monthly white and black pepper price series (in logarithms) during 1973.10-1996.04.

this graph, this series does not contain a unit root (the ADF test is rejected at the 1% level). Table 9.5 shows that the estimated autocorrelation function of this cointegration variable shows a decline to zero while the partial autocorrelation is only significant at the first order.

The VAR(2) model for the bivariate pepper series $y_{1,t}$, and $y_{2,t}$ can now be written in VECM format including the variables $\Delta_1 z_{t-1}$ for $z_t = y_{1,t}, y_{2,t}$ and $y_{1,t} - 1.042y_{2,t}$, which is the cointegrating variable. Since all these variables do not have a unit root, standard t-tests can now be used to delete insignificant variables in order to retain only the relevant ones in the final simplified VECM. The parameters in this final model should then be estimated using SUR. The estimation results are

$$\begin{bmatrix} 1 - 0.133L & -0.321L \\ (0.068) & (0.067) \\ 0 & 1 - 0.325L \\ & (0.069) \end{bmatrix} \begin{bmatrix} \Delta_1 y_{1,t} \\ \\ \Delta_2 y_{2,t} \\ \\ \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ \\ 0.062 \\ (0.026) \end{bmatrix} (y_{1,t-2} - 1.042y_{2,t-2}) + \begin{bmatrix} \hat{\varepsilon}_{1,t} \\ \\ \hat{\varepsilon}_{2,t} \\ \\ \end{bmatrix}, \qquad (9.114)$$

Table 9.5: Empirical (partial) autocorrelation functions for the cointegration variable $\text{cointegration} = y_{1,t} - 1.042y_{2,t}$ where $y_{1,t}$ and $y_{2,t}$ are monthly data on the logarithm of white and black pepper during the period 1973.10-1996.04.

Lag	EACF	EPACF
1	0.904*	0.904*
2	0.810*	−0.045
3	0.738*	0.075
4	0.670*	−0.024
5	0.597*	−0.049
6	0.532*	−0.004
7	0.468*	−0.036
8	0.407*	−0.022
9	0.360*	0.042
10	0.290*	−0.170*
11	0.217*	−0.041
12	0.159*	−0.003
13	0.097	−0.08

Note: An asterisk indicates significance at the 5% level. The estimated standard error for sample is 0.122.

where standard errors are given in parentheses. The unrestricted VAR(2) model with 10 parameters can thus be simplified to a VECM with 5 parameters. It is clear that error correction only occurs for the black pepper series, that is, deviations from the equilibrium relation in (9.113) have an effect on future black pepper. Additionally, as white pepper seems to be explained only by past black pepper prices, this variable is called weakly exogenous, see Johansen (1992).

Testing hypotheses

Sometimes economic theory suggests that the cointegration relations have a specific form. For example, in the case of the pepper prices the only reasonable cointegrating

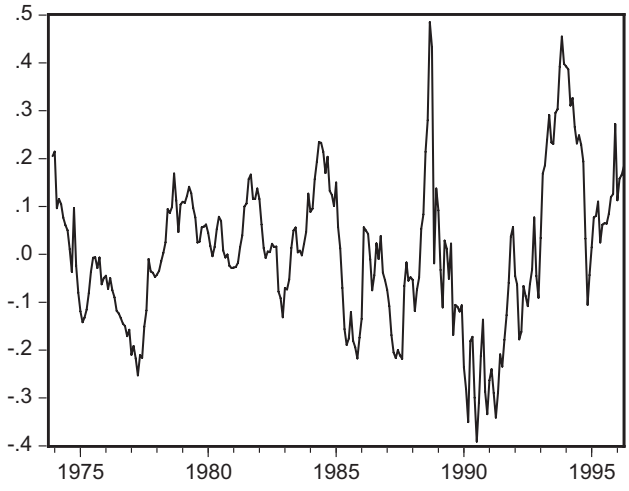


Figure 9.4: Cointegration relation between the logarithm of white and black pepper prices, 1973.10–1996.04.

relationship between these substitution goods is governed by the $(1, -1)$ restriction. Hence in case one has found cointegrating relations, one may wish to test hypotheses on these cointegrating vectors. To test for linear restrictions on the cointegrating vectors in β, it is useful to define the $(m \times r)$ matrix H, which reduces β to the $(r \times r)$ parameter matrix φ, or $\beta = H\varphi$.

To test these parameter restrictions, one compares the estimated eigenvalues $\hat{\xi}_i$, $i = 1, \ldots, r$, from

$$| \xi H' S_{11} H - H' S_{10} S_{00}^{-1} S_{01} H | = 0 \tag{9.115}$$

with the $\hat{\lambda}_i$ from (9.110), $i = 1, \ldots, r$, via the test statistic

$$Q = T \sum_{i=1}^{r} \log\{(1 - \hat{\xi}_i)/1 - \hat{\lambda}_i)\}, \tag{9.116}$$

see Johansen (1995). Under the null hypothesis, and conditional on the correct value of r, the test statistic Q asymptotically follows the $\chi^2(r(m - q))$ distribution. Notice that the matrix H can also contain unit vectors. In that case one investigates whether one or more of the univariate time series variables are stationary, and as such one can circumvent the problem of prior testing for unit roots in univariate time series. In other words, one may straightaway analyze the data using the Johansen test.

Common stochastic trends

When there are r cointegration relations among m variables, there are $m - r$ common stochastic trends in the system. Stock and Watson (1989) exploit this feature to propose an alternative method to investigate cointegration by testing for the number of stochastic trends. For some empirical purposes it is useful not only to obtain estimates of the cointegration relations, but also to obtain insight into the driving non-stationary forces. cointegrating rank $m - 1$, but that this overwhelming long-run correspondence between the freight rates cannot be exploited for forecasting since the driving common stochastic trend dominates the variation in the data.

There are several methods to estimate the stochastic trends, but we limit attention to the method proposed in Gonzalo and Granger (1995), which explicitly exploits the duality of cointegration and stochastic trends, see also Johansen (1995). The canonical correlation method, used to find those combinations of the elements of Y_t which have maximum partial correlation with the stationary variables, can also be reversed to find those combinations which have minimum correlation. Gonzalo and Granger (1995) show that the relevant eigenvalue problem then becomes

$$\mid \lambda S_{00} - S_{01} S_{11}^{-1} S_{10} \mid = 0, \tag{9.117}$$

which is the dual version of (9.110), and where the S_{ij} matrices are defined by ((9.109)). The solutions to (9.117) are the same eigenvalues $\hat{\lambda}_1$ to $\hat{\lambda}_m$ as before, but now one obtains eigenvectors $\hat{w}_1, \ldots, \hat{w}_m$, which are different from those of the cointegration approach. Gonzalo and Granger (1995) show that in case of r cointegration relations, stochastic trend variables can be constructed as $\hat{w}'_{r+1} Y_t$ to $\hat{w}'_m Y_t$.

Testing for cointegration in a conditional ECM

The third and final method to test for cointegration discussed here is the method proposed in Boswijk (1994). This method is useful in case one has a system of variables but there is a major interest in a single equation. For example, in a system relating money, income and prices, one may be interested only in the money equation. The method makes use of the fact that cointegration implies error correction. Hence, a test for the significance of error correction variables can be used to examine if there is cointegration. A second aspect of this method is the assumption that the variables $y_{2,t}$ through $y_{m,t}$ are so-called weakly exogenous for the cointegration parameters of interest, see Engle *et al.* (1983) for a discussion of various exogeneity concepts.

Consider again the bivariate ECM in (9.91)–(9.92). Johansen (1992) shows that if $\alpha_1 = 0$, the $y_{1,t}$ process is weakly exogenous for η. If this holds true, one can perform cointegration analysis in a conditional ECM [CECM] for $y_{2,t}$. In our bivariate example, the restriction $\alpha_1 = 0$ implies that $\delta = 0$, that is, that (9.79) reduces to

Table 9.6: Asymptotic critical values for the cointegration test based on a conditional error correction model.

Number of variables	Significance level			
m	0.20	0.10	0.05	0.01
Regression contains no constant and no trend				
2	4.73	6.35	7.95	11.83
3	7.37	9.40	11.37	15.87
4	9.89	12.12	14.31	18.78
5	12.28	14.79	17.13	21.51
Regression contains a constant				
2	7.52	9.54	11.41	15.22
3	9.94	12.22	14.38	18.68
4	12.38	14.93	17.18	21.43
5	14.83	17.38	19.69	24.63
Regression contains a constant and a trend				
2	10.16	12.32	14.28	18.53
3	12.49	14.91	17.20	21.62
4	14.84	17.57	19.81	24.65
5	17.08	19.96	22.47	27.52

Source: Boswijk (1994)

$y_{1,t} = \mu_1^* + y_{1,t-1} + \varepsilon_{1,t}^*$, and that the CECM for $y_{2,t}$ can be written as

$$\Delta_1 y_{2,t} = \mu_2 + \alpha_2 y_{1,t-1} + \psi y_{2,t-1} + \nu \Delta_1 y_{1,t} + \varepsilon_{2,t}, \tag{9.118}$$

where $\psi = \alpha_2 \eta$, see Boswijk (1994) for details. The parameters in this model can be estimated using OLS, and the significance of the error correction variable can be evaluated using a joint Wald test for $\alpha_2 = 0$ and $\psi = 0$. Asymptotic theory for this Wald test is provided in Boswijk (1994). Critical values are given in Table 9.6.

When the null hypothesis of no cointegration (that is, no error correction) is rejected, one can apply NLS to estimate η. A test for the weak exogeneity of $y_{1,t}$ can be performed by regressing $\Delta_1 y_{1,t}$ on the estimated error correction term obtained in the first round, and on lagged $\Delta_1 y_{1,t}$ and $\Delta_1 y_{2,t}$ variables, and testing the significance of this error correction variable. Conditional on the presence of cointegration, the relevant test statistic asymptotically follows the χ^2 distribution. The same applies to tests for the values of α_2 and η in (9.118).

As an example of this CECM method, consider again the white ($y_{1,t}$) and black ($y_{2,t}$) pepper prices. One can obtain the following estimation results for the sample ending in 1992.12:

$$\begin{aligned}\Delta_1 y_{2,t} &= \underset{(0.060)}{0.072} - \underset{(0.029)}{0.106} y_{2,t-1} + \underset{(0.027)}{0.093} y_{1,t-1} \\ &+ \underset{(0.060)}{0.186} \Delta_1 y_{2,t-1} + \underset{(0.062)}{0.486} \Delta_1 y_{1,t} + \hat{\varepsilon}_t,\end{aligned} \tag{9.119}$$

The Wald test statistic value of 13.773 is significant at the 1 percent level (as the relevant critical value is 11.41). The estimate of η obtained using NLS is 0.877 with standard deviation 0.072. As the value of 1 is included in the 5 percent confidence interval for η, one can set it equal to 1. The final CECM is then given by

$$\begin{aligned}\Delta_1 y_{2,t} &= \underset{(0.060)}{0.072} - \underset{(0.028)}{0.093} (y_{2,t-1} - y_{1,t-1}) + \underset{(0.060)}{0.180} \Delta_1 y_{2,t-1} \\ &+ \underset{(0.062)}{0.491} \Delta_1 y_{1,t} + \hat{\varepsilon}_t,\end{aligned} \tag{9.120}$$

for which the adjustment parameter obtains a t-ratio of -3.368. For the first differences of the white pepper series, one needs to include $\Delta_1 y_{2,t}$ with lags of 1 and 3 and $\Delta_1 y_{1,t}$ with lags at 1 and 2. Adding the error correction variable $y_{2,t-1} - y_{1,t-1}$ to this model yields an estimated parameter value of -0.047 with standard deviation of 0.030. This suggests that the white pepper price is weakly exogenous for η.

Some further practical issues

The empirical power and size of the above discussed tests for cointegration have been studied in many simulation studies. Generally, one finds that the tests become oversized and that their power becomes low in case one considers large systems, that is, when m gets large relative to the sample size. In practice it is therefore recommended to keep m small, say 4 or 5, and to use large enough samples covering large spans

of time. Additionally, graphs of the estimated elements of β can be informative by indicating which linear combinations possess trend or random walk behavior.

When sets of variables are cointegrated, their out-of-sample forecasts are tied together as well, see Engle and Yoo (1987). This is particularly useful when one aims to forecast many observations out of sample. Simulation results in Lee and Tse (1996) show that when the correct rank r is imposed on Π in the rewritten VAR in (9.105), the forecasts are better than when no such cointegration restrictions are imposed. Hence, for out-of-sample forecasting it is relevant to take account of cointegration. Explicit expressions of forecast intervals for cointegrated series are given in Reimers (1997), and measures to evaluate point forecasts are given in Clements and Hendry (1995). Point forecasts for cointegrated series can simply be found using the expressions earlier in this chapter, when imposing the nonlinear restrictions that correspond to cointegration. Tests for Granger causality in cointegrated systems are derived in Lütkepohl and Reimers (1992).

Cointegration with outliers, ARCH and nonlinearity

Many economic time series data do not only display trends and seasonality, but also other features as outliers, ARCH, and nonlinearity. Using simulation experiments, Lin and Tsay (1996) examine the effects of ARCH, and find that these appear to be not too substantial when the ARCH parameters are not too large. With simulations and theoretical arguments, Franses and Haldrup (1994) show that one is inclined to find spurious cointegration when there are neglected additive outliers. Therefore, Franses and Lucas (1998) propose to use outlier robust estimation techniques to reduce the effect of aberrant data points.

Nonlinearity may be directly incorporated in the cointegration framework. For example, consider the nonlinear error correction model

$$\Delta_1 y_{1,t} = \mu + (\alpha_1 F(z_{t-1}))z_{t-1} + \varepsilon_{1,t}, \tag{9.121}$$

where $z_t = y_{1,t} - y_{2,t}$ reflects a cointegration relation, and the $F(.)$ function is for example the logistic function, see Balke and Fomby (1997). Using simulations, Van Dijk and Franses (2000) show that the cointegration parameter can be estimated using the above linear methods with great precision.

Cointegration across time series with seasonality

If seasonally unadjusted data are available, it is most sensible to use these raw data when investigating common long-run non-seasonal trends. The main reason is that it can be shown that most seasonal adjustment methods have an effect on the trend behavior of individual series and hence may affect cointegration analysis. For example,

simulation and empirical results in Franses (1996) and Lee and Siklos (1997) show that seasonal adjustment can lead to less cointegration, while the empirical results in Ermini and Chang (1996) show that adjustment can also lead to spurious cointegration.

Suppose we have two quarterly observed time series $y_{1,t}$ and $y_{2,t}$, each of which has a unit root 1, and we aims to test whether $y_{1,t} + \eta y_{2,t}$ does not have this unit root. When the seasonal fluctuations in each series can be described by seasonal dummies, and the data do not have seasonal unit roots, one can simply replace the intercept vector μ in (9.104) by

$$\mu_0 + \mu_1(D_{1,t} - D_{4,t}) + \mu_2(D_{2,t} - D_{4,t}) + \mu_3(D_{3,t} - D_{4,t}), \qquad (9.122)$$

see page 84 of Johansen (1995). The asymptotic theory for the tests reviewed in the previous section can now be applied, and hence one can use the critical values given in the various tables.

When the data have seasonal unit roots, one may also examine whether their associated stochastic trends are common across m time series. The presence of seasonal cointegration in a multivariate time series Y_t can be analyzed using an extension of the Johansen approach. This approach is developed in Lee (1992), Lee and Siklos (1995) and notably in Johansen and Schaumburg (1999), see also Kunst (1993) for a useful outline of its practical implementation. It amounts to testing the ranks of matrices that correspond to variables which are transformed using the filters to remove the roots 1, -1 or $\pm i$. More precise, consider the $(m \times 1)$ vector process Y_t, and assume that it can be described by the VAR(p) process

$$Y_t = \Theta D_t + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + e_t, \qquad (9.123)$$

where D_t is the (4×1) vector process $D_t = (D_{1,t}, D_{2,t}, D_{3,t}, D_{4,t})'$ containing the seasonal dummies, and where Θ is an $(m \times 4)$ parameter matrix. Similar to the Johansen approach and conditional on the assumption that $p > 4$, model (9.123) can be rewritten as

$$\begin{aligned} \Delta_4 Y_t = {} & \Theta D_t + \Pi_1 Y_{1,t-1} + \Pi_2 Y_{2,t-1} + \Pi_3 Y_{3,t-2} + \Pi_4 Y_{3,t-1} \qquad (9.124) \\ & + \Gamma_1 \Delta_4 Y_{t-1} + \cdots + \Gamma_{p-4} \Delta_4 Y_{t-(p-4)} + e_t, \end{aligned}$$

where

$$\begin{aligned} Y_{1,t} &= (1 + L + L^2 + L^3) Y_t \\ Y_{2,t} &= (1 - L + L^2 - L^3) Y_t \\ Y_{3,t} &= (1 - L^2) Y_t. \end{aligned}$$

Obviously, (9.124) is a multivariate extension of the univariate HEGY model in Chapter 5. The ranks of the matrices Π_1, Π_2, Π_3 and Π_4 determine the number of cointegration relations at a certain frequency. Similar to the Johansen approach, one

can now construct residual processes from regressions of $\Delta_4 Y_t$ and the $Y_{1,t-1}$, $Y_{2,t-1}$, $Y_{3,t-2}$ and $Y_{3,t-1}$ on lagged $\Delta_4 Y_t$ time series and deterministics, and construct the relevant moment matrices as in (9.110). Solving four eigenvalue problems results in sets of estimated eigenvalues which can be checked for their significance using the Trace-test statistic. Critical values of the various test statistics are given in Lee (1992) and Lee and Siklos (1995). Applications of this method can be found in these studies, and in Kunst (1993), Ermini and Chang (1996) and Reimers (1997). The latter study also shows through simulations that imposing the correct cointegration rank can improve out-of-sample forecasting. Finally, Kunst and Franses (1998) discuss the impact of seasonal constants on seasonal cointegration analysis.

EXERCISES

9.1 Show that the following relations hold between the VAR(1) model of (9.9) and the SEM model of (9.11) and (9.12): $\pi_1 = \phi_1(1 - \phi_2\phi_3/\phi_1\phi_4)$, $\pi_2 = \phi_2/\phi_4$, $\pi_3 = \phi_3/\phi_1$, and $\pi_4 = \phi_4(1 - \phi_2\phi_3/\phi_1\phi_4)$.

9.2 Verify the result in (9.38).

9.3 Verify the result in (9.44) and (9.45).

9.4 Give expressions for these implied univariate ARMAX models corresponding to (9.53).

9.5 Consider a trivariate AR(2) model for $Y_t = (y_{1,t}, y_{2,t}, y_{3,t})'$, that is $Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \varepsilon_t$, where $\varepsilon_t \sim NID(0, \Sigma)$ and

$$A_1 = \begin{bmatrix} 0.6 & 1 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

a. Do the series of Y_t contain unit roots?

b. Show that $y_{1,t}$ can be described by a univariate ARMA(3,1) model, y_{2t} by an ARMA(1,0) model, and $y_{3,t}$ by an ARMA(1,1) model.

9.6 Consider the following bivariate model

$$y_t - y_{t-1} = \gamma(y_{t-1} - x_{t-1}) + \alpha(y_{t-1} - y_{t-2}) + v_t \tag{9.125}$$

$$x_t - x_{t-1} = \beta(y_{t-1} - y_{t-2}) + w_t \tag{9.126}$$

where v_t and w_t are mutually independent white noise variables with variances σ_v^2 and σ_w^2 and covariance σ_{vw}.

a. Show that the two equations can be written as a VAR model for the variables y_t and x_t, like $Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \varepsilon_t$, with $Y_t = (y_t, x_{2,t})'$, $\varepsilon_t = (v_t, w_{2,t})'$ $A_1 = \begin{bmatrix} 1+\gamma+\alpha & -\gamma \\ \beta & 1 \end{bmatrix}$ and $A_2 = \begin{bmatrix} -\alpha & 0 \\ -\beta & 0 \end{bmatrix}$

b. Show that there is cointegration between y_t and x_t.

c. Determine the implied univariate models for y_t and x_t.

9.7 Consider the bivariate model

$$\begin{pmatrix} 1 & 0 \\ -\phi_2 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0 & \phi_1 \\ 0 & \phi_3 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ u_t \end{pmatrix}, \tag{9.127}$$

where ε_t and u_t are white noise variables with variances σ_ε^2 and σ_u^2.

a. Give the univariate models for y_t and x_t, which are implied by this model.

b. Show that y_t and x_t are I(1) and that they are cointegrated when

$$\phi_1\phi_2 + \phi_3 = 1$$

9.8 Consider the three-variable system

$$\begin{pmatrix} x_t - x_{t-1} \\ y_t - y_{t-1} \\ z_t - z_{t-1} \end{pmatrix} = \alpha\beta' \begin{pmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} u_t \\ v_t \\ w_t \end{pmatrix} \tag{9.128}$$

where $\alpha' = (-0.5, 0.5, 0.5)$ and $\beta' = (1, -1, -1)$.

a. Write this model as a VAR(1) model for $Y_t = (x_t, y_t, z_t)'$.

b. Are x_t, y_t and z_t cointegrated? If yes, why?

c. Show that the VAR model for Y_t contains 2 unit roots.

d. Derive the implied models for the univariate variables x_t, y_t and z_t.

Bibliography

Abraham, B. and J. Ledolter (1983), *Statistical Methods for Forecasting*, New York: John Wiley & Sons.

Agiakloglou, C. and P. Newbold (1992), Empirical evidence on Dickey-Fuller type tests, *Journal of Time Series Analysis* **13**, 471–483.

Akaike, H. (1974), A new look at statistical model identification, *IEEE Transactions on Automatic Control* **AC-19**, 716–723.

Andersen, T. and T. Bollerslev (1998), Answering the skeptics: Yes, standard volatility models do provide accurate forecasts, *International Economic Review* **39**, 885–906.

Andersen, T.G., T. Bollerslev, F.X. Diebold and P. Labys (2003), Modeling and forecasting realized volatility, *Econometrica* **71**, 579–625.

Andersen, T.G., T. Bollerslev, P.F. Christoffersen and F.X. Diebold (2006), Volatility and correlation forecasting, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland, pp. 777–878.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, New York: John Wiley & Sons.

Bai, J. and S. Ng (2005), Tests for skewness, kurtosis, and normality for time series data, *Journal of Business and Economic Statistics* **23**, 49–61.

Baillie, R.T. (1996), Long memory processes and fractional integration in econometrics, *Journal of Econometrics* **73**, 5–59.

Baillie, R.T. and T. Bollerslev (1992), Prediction in dynamic models with time-dependent conditional variances, *Journal of Econometrics* **52**, 91–113.

Balke, N.S. and B.T. Fomby (1997), Threshold cointegration, *International Economic Review* pp. 627–645.

Banerjee, A., J.J. Dolado, J.W. Galbraith and D.F. Hendry (1993), *Co-integration, Error-correction, and the Econometric Analysis of Nonstationary Data*, Oxford: Oxford University Press.

Banerjee, A., R.L. Lumsdaine and J.H. Stock (1992), Recursive and sequential tests of the unit-root and trend-break hypotheses: theory and international evidence, *Journal of Business and Economic Statistics* **10**, 271–287.

Bass, F.M. (1969), A new product growth model for consumer durables, *Management Science* **15**, 215–227.

Bates, D.M. and D.G. Watts (1988), *Nonlinear regression and its applications*, New York: John Wiley.

Bera, A.K. and C.M. Jarque (1982), Model specification tests: A simultaneous approach, *Journal of Econometrics* **20**, 59–82.

Beran, J. (1995), Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models, *Journal of the Royal Statistical Society* **B 57**, 654–672.

Bernardo, J.M. and A.F.M. Smith (1994), *Bayesian Theory*, New York: John Wiley & Sons.

Berndt, E.R., B.H. Hall, R.E. Hall and J.A. Hausman (1974), Estimation and inference in nonlinear statistical models, *Annals of Economic and Social Measurement* **3**, 653–665.

Bhardwaj, G. and N.R. Swanson (2006), An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series, *Journal of Econometrics* **131**, 539–578.

Black, F. (1976), The pricing of commodity contracts, *Journal of Financial Economics* **3**, 167–179.

Bollerslev, T. (1986), Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics* **31**, 307–327.

Bollerslev, T. (1987), A conditionally heteroskedastic time series model for speculative prices and rates of return, *Review of Economics and Statistics* **69**, 542–547.

Bollerslev, T. (1988), On the correlation structure for the generalized autoregressive conditional heteroskedastic process, *Journal of Time Series Analysis* **9**, 121–131.

Bollerslev, T. and J.M. Wooldridge (1992), Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances, *Econometric Reviews* **11**, 143–172.

Bollerslev, T., R.F. Engle and D.B. Nelson (1994), ARCH models, in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics IV*, Amsterdam: Elsevier Science, pp. 2961–3038.

Bollerslev, T., R.Y. Chou and K.F. Kroner (1992), ARCH modeling in finance: a review of the theory and empirical evidence, *Journal of Econometrics* **52**, 5–59.

Boswijk, H.P. (1994), Testing for an unstable root in conditional and structural error correction models, *Journal of Econometrics* **63**, 37–60.

Boswijk, H.P. and P.H. Franses (2005), On the econometrics of the Bass diffusion model, *Journal of Business and Economic Statistics* **23**, 255–268.

Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis; Forecasting and Control*, San Francisco: Holden-Day.

Box, G.E.P., G.M. Jenkins and G.C. Reinsel (1994), *Time Series Analysis; Forecasting and Control*, 3rd edition, Englewood Cliffs: Prentice Hall.

Breitung, J. (1994), Some simple tests of the moving-average unit root hypothesis, *Journal of Time Series Analysis* **15**, 351–370.

Breusch, T.S. and A.R. Pagan (1979), A simple test for heteroscedasticity and random coefficient variation, *Econometrica* **47**, 1287–1294.

Brodsky, J. and C.M. Hurvich (1999), Multi-step forecasting for long-memory processes, *Journal of Forecasting* **18**, 59–75.

Brooks, C., S.P. Burke and G. Persand (2001), Benchmarks and the accuracy of GARCH model estimation, *International Journal of Forecasting* **17**, 45–56.

Brown, B.Y. and R.S. Mariano (1989), Predictors in dynamic nonlinear models: large sample behaviour, *Econometric Theory* **5**, 430–452.

Burridge, P. and A.M.R. Taylor (2004), Bootstrapping the HEGY seasonal unit root tests, *Journal of Econometrics* **123**, 67–87.

Bustos, O.H. and V.J. Yohai (1986), Robust estimates for ARMA models, *Journal of the American Statistical Association* **81**, 155–168.

Campbell, J.Y. and P. Perron (1991), Pitfalls and opportunities: What macroeconomists should know about unit roots, in O.J. Blanchard and S. Fisher (eds.), *NBER Macroeconomics Annual*, Cambridge, MA: MIT Press, pp. 141–220.

Caner, M. and B.E. Hansen (2001), Threshold autoregressions with a unit root, *ectrica* **69**, 1555–1596.

Carpenter, R.E. and D. Levy (1998), Seasonal cycles, business cycles, and the comovement of inventory investment and output, *Journal of Money, Credit and Banking* **30**, 331–346.

Cecchetti, S.G., A.K. Kashyap and D.W. Wilcox (1997), Interactions between the seasonal and business cycles in production and inventories, *American Economic Review* **87**, 884–892.

Chan, K.S. (1993), Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model, *Annals of Statistics* **21**, 520–533.

Chan, K.S. and H. Tong (1985), On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations, *Advances in Applied Probability* **17**, 666–678.

Chan, K.S. and H. Tong (1986), On estimating thresholds in autoregressive models, *Journal of Time Series Analysis* **7**, 179–190.

Chan, K.S., J.D. Petrucelli, H. Tong and S.W. Woolford (1985), A multiple threshold AR(1) model, *Journal of Applied Probability* **22**, 267–279.

Chandrasekaran, D. and G.J. Tellis (2008), Global takeoff of new products: Cluture, wealth, or vanishing differences?, *Marketing Science* **27**, 844–860.

Chen, C. and L.M. Liu (1993), Forecasting time series with outliers, *Journal of Forecasting* **12**, 13–35.

Chen, R. (1995), Threshold variable selection in open-loop threshold autoregressive models, *Journal of Time Series Analysis* **16**, 461–481.

Cheung, Y-W. (1993), Long memory in foreign-exchange rates, *Journal of Business and Economic Statistics* **11**, 93–101.

Cheung, Y.-W. and K.S. Lai (1995), Lag order and critical values of the augmented Dickey-Fuller test, *Journal of Business and Economic Statistics* **13**, 277–280.

Christoffersen, P.F. (1998), Evaluating interval forecasts, *International Economic Review* **39**, 841–862.

Christoffersen, P.F. and F.X. Diebold (1996), Further results on forecasting and model selection under asymmetric loss, *Journal of Applied Econometrics* **11**, 561–571.

Christoffersen, P.F. and F.X. Diebold (1997), Optimal prediction under asymmetric loss, *Econometric Theory* **13**, 808–817.

Clements, M.P. (2005), *Evaluating Econometric Forecasts of Economic and Financial Variables*, New York: Palgrave-MacMillan.

Clements, M.P. and D.F. Hendry (1995), An empirical study of seasonal unit roots in forecasting, *Journal of Applied Econometrics* **10**, 127–146.

Clements, M.P. and D.F. Hendry (1998), *Forecasting Economic Time Series*, Cambridge: Cambridge University Press.

Clements, M.P. and D.F. Hendry (1999), *Forecasting Non-stationary Economic Time Series*, Cambridge, MA: MIT Press.

Clements, M.P. and D.F. Hendry (2002), *A Companion to Economic Forecasting*, Blackwell Publishing.

Clements, M.P. and J. Smith (1997), The performance of alternative forecasting methods for SETAR models, *International Journal of Forecasting* **13**, 463–475.

Corradi, V. and N.R. Swanson (2006), Predictive density evaluation, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland, pp. 197–284.

Crato, N. and B.K. Ray (1996), Model selection and forecasting for long-range dependent processes, *Journal of Forecasting* **15**, 107–125.

Davies, R.B. (1977), Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* **64**, 247–254.

Davies, R.B. (1987), Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* **74**, 33–43.

de Gooijer, J.G. (1998), On threshold moving-average models, *Journal of Time Series Analysis* **19**, 1–18.

De Gooijer, J.G., B. Abraham, A. Gould and L. Robinson (1985), Methods for determining the order of an autoregressive-moving average process: A survey, *International Statistical Review* **85**, 301–329.

Denby, L. and R.D. Martin (1979), Robust estimation of the first-order autoregressive parameter, *Journal of the American Statistical Association* **74**, 140–146.

Dhrymes, P.J. (1981), *Distributed Lags: Problems of Estimation and Formulation*, Amsterdam: North-Holland.

Dickey, D.A. and S.G. Pantula (1987), Determining the order of differencing in autoregressive processes, *Journal of Business and Economic Statistics* **5**, 455–461.

Dickey, D.A. and W.A. Fuller (1979), Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association* **74**, 427–431.

Dickey, D.A. and W.A. Fuller (1981), Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica* **49**, 1057–1072.

Diebold, F.X. and J.A. Lopez (1995), Modelling volatility dynamics, in K. Hoover (ed.), *Macroeconometrics – Developments, Tensions and Prospects*, Boston, Kluwer, pp. 427–472.

Diebold, F.X. and L. Kilian (2000), Unit root tests are useful for selecting forecasting models, *Journal of Business and Economic Statistics* **18**, 265–273.

Diebold, F.X. and R.S. Mariano (1995), Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**, 253–263.

Ding, Z. and C.W.J. Granger (1996), Modeling volatility persistence of speculative returns: a new approach, *Journal of Econometrics* **73**, 185–215.

Doornik, J. and M. Ooms (2003), Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models, *Computational Statistics and Data Analysis* **42**, 333–348.

Doornik, J. and M. Ooms (2004), Inference and forecasting for ARFIMA models with an application to US and UK inflation, *Studies in Nonlinear Dynamics and Econometrics* **8**, 1–25.

Durbin, J. and S.J. Koopman (2001), *Time Series Analysis by State Space Methods*, Oxford: Oxford University Press.

Eitrheim, Ø. and T. Teräsvirta (1996), Testing the adequacy of smooth transition autoregressive models, *Journal of Econometrics* **74**, 59–76.

Elliott, G. (2006), Forecasting with trending data, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland, pp. 99–134.

Elliott, G., C.W.J. Granger and A. Timmermann (2006), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland.

Enders, W. and C.W.J. Granger (1998), Unit-root tests and asymetric adjustment with an example using the term structure of interest rates, *Journal of Business and Economic Statistics* **16**, 304–311.

Engle, R.F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* **50**, 987–1007.

Engle, R.F. (1983), Estimates of the variance of US inflation based upon the ARCH model, *Journal of Money, Credit and Banking* **15**, 286–301.

Engle, R.F. and B.S. Yoo (1987), Forecasting and testing in co-integrated systems, *Journal of Econometrics* **35**, 143–159.

Engle, R.F., C.-H. Hong, A. Kane and J. Noh (1993), Arbitrage valuation of variance forecasts with simulated options, in D.M. Chance and R.R. Trippi (eds.), *Advances in Futures and Options Research*, Greenwich, CN: JAI Press.

Engle, R.F. and C.J.W. Granger (1987), Co-integration and error correction: representation, estimation, and testing, *Econometrica* **55**, 251–276.

Engle, R.F., C.W.J. Granger, S. Hylleberg and H.S. Lee (1993), Seasonal cointegration: The Japanese consumption function, *Journal of Econometrics* **55**, 275–298.

Engle, R.F., D.F. Hendry and J.F. Richard (1983), Exogeneity, *Econometrica* **51**, 277–304.

Engle, R.F., D.M. Lilien and R.P. Robins (1987), Estimating time varying risk premia in the term structure: the ARCH-M model, *Econometrica* **55**, 391–407.

Engle, R.F. and G. González-Rivera (1991), Semiparametric ARCH models, *Journal of Business and Economic Statistics* **9**, 345–360.

Engle, R.F. and T. Bollerslev (1986), Modelling the persistence of conditional variances, *Econometric Reviews* **5**, 1–50 (with discussion).

Engle, R.F. and V.K. Ng (1993), Measuring and testing the impact of news on volatility, *Journal of Finance* **48**, 1749–1778.

Ermini, L. and D. Chang (1996), Testing the joint hypothesis of rationality and neutrality under seasonal cointegration: the case of Korea, *Journal of Econometrics* **74**, 363–386.

Fan, J. and I. Gijbels (1970), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fiorentini, G., G. Calzolari and L. Panatoni (1996), Analytic derivatives and the computation of GARCH estimates, *Journal of Applied Econometrics* **11**, 399–417.

Fok, D., P.H. Franses and R. Paap (2006), Performance of seasonal adjustment procedures, in K. Patterson and T.C. Mills (eds.), *The Palgrave Handbook of Econometrics Volume 1: Econometric Theory*, Palgrave/MacMillan, pp. 1035–1055.

Franses, P.H. (1994), A multivariate approach to modeling univariate seasonal time series, *Journal of Econometrics* **63**, 133–151.

Franses, P.H. (1995), A differencing test, *Econometric Reviews* **14**, 183–193.

Franses, P.H. (1996), *Periodicity and stochastic trends in economic time series*, Oxford: Oxford University Press.

Franses, P.H. (1998), *Time series models for business and economic forecasting*, Cambridge University Press.

Franses, P.H. and A. Lucas (1998), Outlier detection in cointegration analysis, *Journal of Business and Economic Statistics* **16**, 459–468.

Franses, P.H. and B. Hobijn (1997), Critical values for unit root tests in seasonal time series, *Journal of Applied Statistics* **24**, 25–47.

Franses, P.H. and D. van Dijk (1996), Forecasting stock market volatility using nonlinear GARCH models, *Journal of Forecasting* **15**, 229–235.

Franses, P.H. and D. van Dijk (2000), *Nonlinear Time Series Models in Empirical Finance*, Cambridge: Cambridge University Press.

Franses, P.H. and D. van Dijk (2005), The forecasting performance of various models for seasonality and nonlinearity for quarterly industrial production, *International Journal of Forecasting* **21**, 87–102.

Franses, P.H. and N. Haldrup (1994), The effects of additive outliers on tests for unit roots and cointegration, *Journal of Business and Economic Statistics* **12**, 471–478.

Franses, P.H. and R. Paap (1999), Does seasonality influence the dating of business cycle turning points?, *Journal of Macroeconomics* **21**, 79–92.

Franses, P.H. and R. Paap (2002), Forecasting with periodic autoregressive time series models, in M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, Blackwell Publishing, pp. 432–452.

Franses, P.H. and R. Paap (2004), *Periodic Time Series Models*, Oxford: Oxford University Press.

Franses, P.H. and T.J. Vogelsang (1998), On seasonal cycles, unit roots, and mean shifts, *Review of Economics and Statistics* **80**, 231–240.

Fuller, W.A. (1976), *Introduction to Statistical Time Series*, New York: John Wiley & Sons.

Galbraith, J.W. and V. Zinde-Walsh (1994), A simple, noniterative estimator for moving average models, *Biometrika* **81**, 143–155.

Gallant, A.R. (1987), *Nonlinear Statistical Models*, New York: John Wiley.

Geweke, J. (2005), *Contemporary Bayesian Econometrics and Statistics*, New York: John Wiley & Sons.

Geweke, J. and C. Whiteman (2006), Bayesian forecasting, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland, pp. 3–80.

Geweke, J. and S. Porter-Hudak (1983), The estimation and application of long memory time series models, *Journal of Time Series Analysis* **4**, 221–237.

Ghysels, E. (1994), On the economics and econometrics of seasonality, in C.A. Sims (ed.), *Advances in Econometrics, Sixth World Congress of the Econometric Society*, Cambridge: Cambridge University Press, pp. 257–316.

Ghysels, E., C.W.J. Granger and P.L. Siklos (1996), Is seasonal adjustment a linear or nonlinear data filtering process?, *Journal of Business and Economic Statistics* **14**, 374–386.

Ghysels, E. and D.R. Osborn (2001), *The Econometric Analysis of Seasonal Time Series*, Cambridge: Cambridge University Press.

Ghysels, E., D.R. Osborn and P.M.M. Rodrigues (2006), Forecasting seasonal time series, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland, pp. 659–711.

Ghysels, E., H.S. Lee and J. Noh (1994), Testing for unit roots in seasonal time series, *Journal of Econometrics* **62**, 415–442.

Ghysels, E. and P. Perron (1993), The effect of seasonal adjustment filters on tests for a unit root, *Journal of Econometrics* **55**, 57–99.

Ghysels, E. and P. Perron (1996), The effect of linear filters on dynamic time series with structural change, *Journal of Econometrics* **70**, 69–97.

Glosten, L.R., R. Jagannathan and D.E. Runkle (1993), On the relation between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance* **48**, 1779–1801.

Godfrey, L.G. (1979), Testing the adequacy of a time series model, *Biometrika* **66**, 67–72.

Gonzalo, J. and C.W.J. Granger (1995), Estimation of common long-memory components in cointegrated systems, *Journal of Business and Economic Statistics* **13**, 27–35.

Gourieroux, C. (1997), *ARCH Models and Financial Applications*, Berlin: Springer-Verlag.

Granger, C.W.J. (1966), A typical spectral shape of an economic variables, *Econometrica* **34**, 150–161.

Granger, C.W.J. (1969), Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **36**, 150–161.

Granger, C.W.J. (1993), Strategies for modelling nonlinear time-series relationships, *The Economic Record* **69**, 233–238.

Granger, C.W.J., M.L. King and H. White (1995), Comments on testing economic theories and the use of model selection criteria, *Journal of Econometrics* **67**, 173–188.

Granger, C.W.J. and P. Newbold (1974), Spurious regressions in econometrics, *Journal of Econometrics* **2**, 111–120.

Granger, C.W.J. and P. Newbold (1976), Forecasting transformed time series, *Journal of the Royal Statistical Society* **B 38**, 189–203.

Granger, C.W.J. and P. Newbold (1986), *Forecasting Economic Time Series*, 2nd edition, London: Academic Press.

Granger, C.W.J. and R. Joyeux (1980), An introduction to long-memory time series models and fractional differencing, *Journal of Time Series Analysis* **1**, 15–39.

Haldrup, N. (1994), Semi-parametric tests for double unit roots, *Journal of Business and Economic Statistics* **12**, 109–122.

Hall, A.D. and M. McAleer (1989), A Monte Carlo study of some tests of model adequacy in time series analysis, *Journal of Business and Economic Statistics* **7**, 95–106.

Hall, A.R. (1994), Testing for a unit root in time series with pretest data-based model selection, *Journal of Business and Economic Statistics* **12**, 461–470.

Hamilton, J.D. (1989), A new approach to the economic analysis of nonstationary time series subject to changes in regime, *Econometrica* **57**, 357–384.

Hamilton, J.D. (1990), Analysis of time series subject to changes in regime, *Journal of Econometrics* **45**, 39–70.

Hamilton, J.D. (1993), Estimation, inference and forecasting of time series subject to changes in regime, in G.S. Maddala, C.R. Rao and H.D. Vinod (eds.), *Handbook of Statistics, vol. 11*, Elsevier, North Holland, pp. 231–260.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton: Princeton University Press.

Hamilton, J.D. (1996), Specification testing in Markov-switching time series models, *Journal of Econometrics* **70**, 127–157.

Hansen, B.E. (1992), The likelihood ratio test under nonstandard assumptions: testing the Markov Switching model of GNP, *Journal of Applied Econometrics* **7**, S61–S82; erratum (1996) **11**, 195–198.

Hansen, B.E. (1996), Inference when a nuisance parameter is not identified under the null hypothesis, *Econometrica* **64**, 413–430.

Hansen, B.E. (1997), Inference in TAR models, *Studies in Nonlinear Dynamics and Econometrics* **2**, 1–14.

Hansen, B.E. (2000), Sample splitting and threshold estimation, *Econometrica* **68**, 575–603.

Härdle, W., H. Lütkepohl and R. Chen (1997), A review of nonparametric time series analysis, *International Statistical Review* **65**, 49–72.

Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press.

Harvey, A.C. (2006), Forecasting with unobserved components time series models, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland, pp. 327–412.

Harvey, D.I. and D. van Dijk (2006), Sample size, lag order and critical values of seasonal unit root tests, *Computational Statistics and Data Analysis* **50**, 2734–2751.

Hassler, U. and J. Wolters (1995), Long memory in inflation rates: international evidence, *Journal of Business and Economic Statistics* **13**, 37–46.

He, C. and T. Teräsvirta (1999), Fourth moment structure of the GARCH(p,q) process, *Econometric Theory* **15**, 824–846.

Hendry, D.F. (1995), *Dynamic Econometrics*, Oxford: Oxford University Press.

Holst, U., G. Lindgren, J. Holst and M. Thuvesholmen (1994), Recursive estimation in switching autoregressions with a Markov regime, *Journal of Time Series Analysis* **15**, 489–506.

Hong, P.Y. (1991), The autocorrelation structure for the GARCH-M process, *Economics Letters* **37**, 129–132.

Hosking, J.R.M. (1980), The multivariate portmanteau statistic, *Journal of the American Statistical Association* **75**, 602–608.

Hosking, J.R.M. (1981), Fractional differencing, *Biometrika* **68**, 165–176.

Hylleberg, S. (1995), Tests for seasonal unit roots: General to specific or specific to general?, *Journal of Econometrics* **69**, 5–25.

Hylleberg, S. and G.E. Mizon (1989), A note on the distribution of the least squares estimator of a random walk with drift, *Economics Letters* **29**, 225–230.

Hylleberg, S., R.F. Engle, C.W.J. Granger and B.S. Yoo (1990), Seasonal integration and cointegration, *Journal of Econometrics* **44**, 215–238.

Hyndman, R.J. (1995), Highest-density forecast regions for nonlinear and nonnormal time series, *Journal of Forecasting* **14**, 431–441.

Hyndman, R.J. (1996), Computing and graphing highest-density regions, *American Statistician* **50**, 120–126.

Johansen, S. (1988), Statistical analysis of cointegration vectors, *Journal of Economic Dynamics and Control* **12**, 231–254.

Johansen, S. (1991), Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models, *Econometrica* **59**, 1551–1580.

Johansen, S. (1992), Testing weak exogeneity and the order of cointegration in UK money demand data, *Journal of Policy Modeling* **14**, 313–334.

Johansen, S. (1995), Identifying restrictions of linear equations with applications to simultaneous equations and cointegration, *Journal of Econometrics* **69**, 111–132.

Johansen, S. and E. Schaumburg (1999), Likelihood analysis of seasonal cointegration, *Journal of Econometrics* **88**, 301–339.

Jorion, P. (1995), Predicting volatility in the foreign exchange market, *Journal of Finance* **50**, 507–528.

Kim, C.-J. (1993), Unobserved-components time series models with Markov-Switching heteroskedasticity: changes in regime and the link between inflation rates and inflation uncertainty, *Journal of Business and Economic Statistics* **11**, 341–349.

Koning, A.J., P.H. Franses, M. Hibon and H.O. Stekler (2005), The M3 competition: Statistical tests of the results, *International Journal of Forecasting* **21**, 397–409.

Koop, G. (2003), *Bayesian Econometrics*, New York: John Wiley & Sons.

Krane, S. and W. Wascher (1999), The cyclical sensitivity of seasonality in US employment, *Journal of Monetary Economics* **44**, 523–553.

Kristensen, D. and O. Linton (2006), A closed-form estimator for the GARCH(1,1) model, *Econometric Theory* **22**, 323–337.

Kunst, R.M. (1993), Seasonal cointegration in macroeconomic systems: Case studies for small and large European countries, *Review of Economics and Statistics* pp. 325–330.

Kunst, R.M. and P.H. Franses (1998), The impact of seasonal constants on forecasting seasonally cointegrated time series, *Journal of Forecasting* **17**, 109–124.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt and Y. Shin (1992), Testing the null hypothesis of stationarity against the alternative of a unit root, *Journal of Econometrics* **54**, 159–178.

Lamoureux, C.G. and W.D. Lastrapes (1993), Forecasting stock return variances: towards understanding stochastic implied volatility, *Review of Financial Studies* **6**, 293–326.

Lee, H.S. (1992), Maximum likelihood inference on cointegration and seasonal cointegration, *Journal of Econometrics* **54**, 1–47.

Lee, H.S. and P.L. Siklos (1995), A note on the critical values for the maximum likelihood (seasonal) cointegration tests, *Economics Letters* **49**, 137–145.

Lee, H.S. and P.L. Siklos (1997), The role of seasonality in economic time series reinterpreting money-output causality in U.S. data, *International Journal of Forecasting* **13**, 381–391.

Lee, J.H.H. (1991), A Lagrange multiplier test for GARCH models, *Economics Letters* **37**, 265–271.

Lee, S.-W. and B.E. Hansen (1994), Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator, *Econometric Theory* **10**, 29–52.

Lee, T.H. and Y. Tse (1996), Cointegration tests with conditional heteroskedasticity, *Journal of Econometrics* **73**, 401–410.

Leone, R.P. (1995), Generalizing what is known about temporal aggregation and advertising carryover, *Marketing Science* **14**, 141–150.

Leybourne, S., P. Newbold and D. Vougas (1998), Unit roots and smooth transitions, *Journal of Time Series Analysis* **19**, 83–97.

Li, W.K. and T.K. Mak (1994), On the squared residual autocorrelations in nonlinear time series analysis with conditional heteroskedasticity, *Journal of Time Series Analysis* **15**, 627–636.

Lilien, G.L., A. Rangaswamy and C. van den Bulte (2000), Diffusion models: Managerial applications and software, in V. Mahajan, E. Muller and Y. Wind (eds.), *New-Product Diffusion Models, Ch. 12*, Boston: Kluwer, pp. 295–332.

Lin, J.L. and C.W.J. Granger (1994), Forecasting from nonlinear models in practice, *Journal of Forecasting* **13**, 1–9.

Lin, J.L. and R.S. Tsay (1996), Co-integration constraint and forecasting: An empirical examination, *Journal of Applied Econometrics* **11**, 519–538.

Ling, S. and M. McAleer (2002a), Necessary and sufficient moment conditions for the GARCH(r,s) and asymmetric power GARCH(r,s) models, *Econometric Theory* **18**, 722–729.

Ling, S. and M. McAleer (2002b), Stationarity and the existence of moments of a family of GARCH processes, *Journal of Econometrics* **106**, 109–117.

Ljung, G.M. and G.E.P. Box (1978), On a measure of lack of fit in time series models, *Biometrika* **65**, 297–303.

Lomnicki, Z.A. (1961), Tests for departure from normality in the case of linear stochastic processes, *Biometrika* **4**, 27–62.

Lopez, J.A. (2001), Evaluating the predictive accuracy of volatility models, *Journal of Forecasting* **20**, 87–109.

Lopez, J.H. (1997), The power of the ADF test, *Economics Letters* **57**, 5–10.

Lucas, A. (1995), An outlier robust unit root test with an application to the extended Nelson-Plosser data, *Journal of Econometrics* **66**, 153–173.

Lucas, A. (1996), *Outlier robust unit root analysis*, PhD thesis, Erasmus University Rotterdam.

Lumsdaine, R.L. (1995), Finite-sample properties of the maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models: a Monte Carlo investigation, *Journal of Business and Economic Statistics* **13**, 1–10.

Lumsdaine, R.L. (1996), Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models, *Econometrica* **64**, 575–596.

Lumsdaine, R.L. and S. Ng (1999), Testing for ARCH in the presence of possibly misspecified mean, *Journal of Econometrics* **93**, 257–279.

Lundbergh, S. and T. Teräsvirta (2002), Evaluating GARCH models, *Journal of Econometrics* **110**, 579–609.

Lundbergh, S., T. Teräsvirta and D. van Dijk (2000), Time-varying smooth transition autoregressive models, Working papers in Economics and Finance No. 376, Stockholm School of Economics.

Lütkepohl, H (1991), *Introduction to Multiple Time Series Analysis*, Springer, New York.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Berlin: Springer-Verlag.

Lütkepohl, H. and H.E. Reimers (1992), Granger-causality in cointegrated VAR processes The case of the term structure, *Economics Letters* **40**, 263–268.

Luukkonen, R., P. Saikkonen and T. Teräsvirta (1988), Testing linearity against smooth transition autoregressive models, *Biometrika* **75**, 491–499.

MacKinnon, J.G. (1991), Numerical distribution functions for unit root and cointegration tests, *jae* **11**, 601–618.

Mahajan, V., E. Muller and F.M. Bass (1993), New-product diffusion models, in J. Eliashberg and G.L. Lilien (eds.), *Handbooks in Operations Research and Management Science, Vol. 5 (Marketing)*, Amsterdam: North-Holland, pp. 349–408.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen and R. Winkler (1982), The accuracy of extrapolation (time series) methods: Results of a forecasting competition, *Journal of Forecasting* **1**, 111–153.

Makridakis, S. and M. Hibon (2000), The M3 competition: Results, conclusions and implications, *International Journal of Forecasting* **16**, 451–476.

McCullough, B.D. and C.G. Renfro (1999), Benchmarks and software standards: A case study of GARCH procedures, *Journal of Economic and Social Measurement* **25**, 59–71.

Meade, N. and T. Islam (1995), Prediction intervals for growth curve forecasts, *Journal of Forecasting* **14**, 413–430.

Miron, J.A. (1996), *The Economics of Seasonal Cycles*, Cambridge, MA: MIT Press.

Miron, J.A. and J.J. Beaulieu (1996), What have macroeconomists learned about business cycles from the study of seasonal cycles?, *Review of Economics and Statistics* **78**, 54–66.

Moeanaddin, R. and H. Tong (1990), Numerical evaluation of distributions in nonlinear autoregression, *Journal of Time Series Analysis* **11**, 33–48.

Nelson, D.B. (1990), Stationarity and persistence in the GARCH(1,1) model, *Econometric Theory* **6**, 318–334.

Nelson, D.B. (1991), Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**, 347–370.

Nelson, D.B and C.Q. Cao (1992), Inequality constraints in the univariate GARCH model, *Journal of Business and Economic Statistics* **10**, 229–235.

Newbold, P. and D.I. Harvey (2002), Forecast combination and encompassing, in M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, Blackwell Publishing, pp. 268–283.

Newey, W.K. and K.D. West (1987), A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* **55**, 703–708.

Newey, W.K. and K.D. West (1994), Automatic lag selection in covariance matrix estimation, *Review of Economic Studies* **61**, 631–653.

Ng, S. and P. Perron (1995), Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag, *Journal of the American Statistical Association* **90**, 268–281.

Ng, S. and P. Perron (1997), Estimation and inference in nearly unbalanced nearly cointegrated systems, *Journal of Econometrics* **79**, 53–81.

Ng, S. and P. Perron (2001), Lag length selection and the construction of unit root tests with good size and power, *Econometrica* **69**, 1519–1554.

Osborn, D.R. (2002), Unit-root versus deterministic representations of seasonality for forecasting, in M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, Blackwell Publishing, pp. 409–431.

Osborn, D.R., S. Heravi and C.R. Birchenhall (1999), Seasonal unit roots and forecasts of two-digit industrial production, *International Journal of Forecasting* **15**, 27–47.

Osterwald-Lenum, Michael (1992), A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics1, *Oxford Bulletin of Economics and Statistics* **54**, 461–472.

Paap, R., P.H. Franses and H. Hoek (1997), Mean shifts, unit roots and forecasting seasonal time series, *International Journal of Forecasting* **13**, 357–368.

Pagan, A. (1996), The econometrics of financial markets, *Journal of Empirical Finance* **3**, 15–102.

Pagan, A.R. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge: Cambridge University Press.

Pagan, A.R. and G.W. Schwert (1990), Alternative models for conditional stock volatility, *Journal of Econometrics* **45**, 267–290.

Palm, F.C. (1996), GARCH models of volatility, in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics, vol. 14*, Amsterdam: Elsevier Science, pp. 209–240.

Parzen, E. (1984), *Time Series Analysis of Irregularly Observed Data*, Lecture Notes in Statistics, Vol. 25, New York: Springer-Verlag.

Paulsen, J. (1984), Order determination of multivariate autoregressive time series with unit roots, *Journal of Time Series Analysis* **5**, 115–127.

Peel, D.A. and A.E.H. Speight (1996), Is the US business cycle asymmetric? Some further evidence, *Applied Economics* **28**, 405–415.

Perron, P. (1989), The great crash, the oil price shock, and the unit root hypothesis, *Econometrica* **57**, 1361–1401.

Perron, P. (1990), Testing for a unit root in a time series with a changing mean, *Journal of Business and Economic Statistics* **8**, 153–162.

Perron, P. (2006), Dealing with structural breaks, *Palgrave handbook of econometrics*, Palgrave Macmillan, pp. 278–352.

Perron, P. and T.J. Vogelsang (1992), Nonstationarity and level shifts with an application to purchasing power parity, *Journal of Business and Economic Statistics* **10**, 301–320.

Phillips, P.C.B. (1986), Time series regression with a unit root, *Journal of Econometrics* **33**, 311–340.

Phillips, P.C.B. (1987), Time series regression with a unit root, *Econometrica* **55**, 277–301.

Phillips, P.C.B. and P. Perron (1988), Testing for a unit root in time series regression, *Biometrika* **75**, 335–346.

Phillips, P.C.B. and S. Ouliaris (1990), Asymptotic properties of residual based tests for cointegration, *Econometrica* **58**, 165–193.

Phillips, P.C.B. and Z. Xiao (1998), A primer on unit root testing, *Journal of Economic Surveys* **12**, 423–470.

Poirier, D.J. (1995), *Intermediate Statistics and Econometrics: A Comparative Approach*, Cambridge, MA: MIT Press.

Poskitt, D.S. and A.R. Tremayne (1982), Diagnostic tests for multiple time series models, *Annals of Statistics* **10**, 114–120.

Pötscher, B.M. and I.V. Prucha (1997), *Dynamic Nonlinear Econometric Models – Asymptotic Theory*, Berlin: Springer-Verlag.

Priestley, M.B. (1981), *Spectral Analysis and Time Series*, London: Academic Press.

Quandt, R. (1983), Computational problems and methods, in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics I*, Amsterdam: Elsevier Science, pp. 699–746.

Reimers, H.E. (1997), Forecasting of seasonal cointegrated processes, *International Journal of Forecasting* **13**, 369–380.

Said, S.E. and D.A. Dickey (1984), Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika* **71**, 599–607.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

Schwert, G.W. (1989), Tests for unit roots: a Monte Carlo investigation, *Journal of Business and Economic Statistics* **7**, 147–160.

Shephard, N. (1996), Statistical aspects of ARCH and stochastic volatility, in O.E. Barndorff-Nielsen D.R. Cox and D.V. Hinkley (eds.), *Statistical Models in Econometrics, Finance and Other Fields*, London: Chapman & Hall, pp. 1–67.

Smith, R.J. and A.M.R. Taylor (1998), Additional critical values and asymptotic representations for seasonal unit root tests, *Journal of Econometrics* **85**, 269–288.

Sowell, F. (1992), Maximum likelihood estimation of stationary univariate fractionally integrated time series models, *Journal of Econometrics* **53**, 165–188.

Srinivasan, S. and C.H. Mason (1986), Nonlinear least squares estimation of new product diffusion models, *Marketing Science* **5**, 169–178.

Stock, J.H. and M.W. Watson (1989), Interpreting the evidence on money-income causality, *Journal of Econometrics* **40**, 161–181.

Stock, J.H. and M.W. Watson (1999), A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, in R.F. Engle and H. White (eds.), *Cointegration, Causality and Forecasting. A Festschrift in Honour of Clive W.J. Granger*, Oxford: Oxford University Press, pp. 1–44.

Talukdar, D., K. Sudhir and A. Ainslie (2002), Investigating new product diffusion across products and countries, *Marketing Science* **21**, 97–114.

Tellis, G.J., S. Stremersch and E. Yin (2003), The international takeoff of new products: The role of economics, culture and country innovativeness, *Marketing Science* **22**, 118–208.

Teräsvirta, T. (1994), Specification, estimation, and evaluation of smooth transition autoregressive models, *Journal of the American Statistical Association* **89**, 208–218.

Teräsvirta, T. (1998), Modelling economic relationships with smooth transition regressions, in A. Ullah and D.E.A. Giles (eds.), *Handbook of Applied Economic Statistics*, New York: Marcel Dekker, pp. 507–552.

Terasvirta, T., D. Tjostheim and C.W.J. Granger (2010), *Modelling nonlinear economic time series*, Oxford University Press.

Teräsvirta, T. and H.M. Anderson (1992), Characterizing nonlinearities in business cycles using smooth transition autoregressive models, *Journal of Applied Econometrics* **7**, S119–S136.

Tiao, G.C. and G.E.P. Box (1981), Modeling multiple time series with applications, *Journal of the American Statistical Association* **76**, 802–816.

Tjøstheim, D. (1986), Some doubly stochastic time series models, *Journal of Time Series Analysis* **7**, 51–72.

Tong, H. (1978), On a threshold model, in C.H. Chen (ed.), *Pattern Recognition and Signal Processing*, Amsterdam: Sijhoff & Noordoff, pp. 101–141.

Tong, H. (1990), *Non-Linear Time Series: A Dynamical Systems Approach*, Oxford: Oxford University Press.

Tong, H. and K.S. Lim (1980), Threshold autoregressions, limit cycles, and data, *Journal of the Royal Statistical Society* **B 42**, 245–292 (with discussion).

Tsay, R.S. (1988), Outliers, level shifts, and variance changes in time series, *Journal of Forecasting* **7**, 1–20.

Tsay, R.S. (1993), Testing for noninvertible models with applications, *Journal of Business and Economic Statistics* **11**, 225–233.

Tuan, P.-D. (1979), The estimation of parameters for autoregressive-moving average models from sample autocovariances, *Biometrika* **66**, 555–560.

Van den Bulte, C. and G.L. Lilien (1997), Bias and systematic change in the parameter estimates of macro-level diffusion models, *Marketing Science* **16**, 338–353.

van Dijk, D. and P.H. Franses (1999), Modeling multiple regimes in the business cycle, *Macroeconomic Dynamics* **3**, 311–340.

Van Dijk, D. and P.H. Franses (2000), Nonlinear error-correction models for interest rates in the Netherlands, in W. Barnett, D.F. Hendry, S. Hylleberg, T. Teräsvirta, D. Tjøstheim and A.W. Würtz (eds.), *Non-linear Econometric Modelling in Time Series Analysis*, Cambridge: Cambridge University Press, pp. 203–227.

van Dijk, D., P.H. Franses and A. Lucas (1999), Testing for ARCH in the presence of additive outliers, *Journal of Applied Econometrics* **14**, 539–562.

Wallis, K.F. (2003), Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts, *International Journal of Forecasting* **19**, 165–175.

Wecker, W.E. (1981), Asymmetric time series, *Journal of the American Statistical Association* **76**, 16–21.

West, K.D. (2006), Forecast evaluation, in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting. Vol. 1*, Amsterdam: North-Holland, pp. 99–134.

West, K.D. and D. Cho (1995), The predictive ability of several models of exchange rate volatility, *Journal of Econometrics* **69**, 367–391.

West, K.D., H.J. Edison and D. Cho (1993), A utility based comparison of some models of exchange rate volatility, *Journal of International Economics* **35**, 23–45.

White, H. and I. Domowitz (1984), Nonlinear regression with dependent observations, *Econometrica* **52**, 143–161.

Zellner, A. (1962), An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association* **57**, 348–368.

Zellner, A. (1970), *Bayesian Analysis in Econometrics and Statistics*, Amsterdam: North-Holland.

Zellner, A. and F.C. Palm (1974), *The structural econometric time series analysis approach*, Cambridge University Press.

Zivot, E. and D.W.K. Andrews (2002), Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis, *Journal of Business and Economic Statistics* **20**, 25–44.

Subject index