Time Series Econometrics (TSE)

Henri Nyberg

Table of contents

Preface

Foreword

This material is compiled from various sources. It is largely based on the Aikasarja-analyysi (Time Series Analysis) course taught at the University of Turku from 2016 to 2025. That course material was, in turn, heavily based on the numerous time series analysis courses taught by Professor Emeritus Pentti Saikkonen at the University of Helsinki. My/our sincere thanks to him for providing access to his course materials, particularly for the course "Stationaariset aikasarjat" (2015). Thanks also to Professor Mika Meitz, whose lecture notes, "Lecture Notes on Time Series Econometrics" were valuable reference for the English translation and terminology.

Special thanks to Professor Markku Lanne for sharing his lecture notes on linear regressions with time series variables, originally part of the "Econometrics 2" course at the University of Helsinki. I also wish to thank MSocSc and doctoral researcher Roope Rihtamo for setting up the Quarto and GitHub documents of this material.

All remaining errors are solely the responsibility of the author.

Other sources and potential references include the following:

- Brockwell, P.J. & R.A. Davis (1996 or 2002). Introduction to Time Series and Forecasting. Springer (2002, 2nd ed.).

- Franses, P.H. & D. van Dijk (2000). Non-linear time series models in empirical finance. Cambridge University Press.

- Hamilton, J. (1994). Time Series Analysis. Princeton University Press.

- Lütkepohl, H. & Krätzig, M. (2004). Applied Time Series Econometrics. Cambridge University Press.

- Tsay, R.S. (2010). Analysis of Financial Time Series. Wiley

- Verbeek, M. (2008). A Guide To Modern Econometrics. John Wiley & Sons. Third Edition

This lecture manuscript primarily focus on univariate time series models. However, we will also briefly cover key aspects of multiple time series analysis, particularly vector autoregressive (VAR) models and cointegration. In addition, we will introduce some fundamental concepts of machine learning techniques relevant to time series econometrics.

The empirical examples, implemented using RStudio, serve to illustrations to theoretical concepts. While most examples are drawn from (macro)economic and financial applications, the methods are broadly applicable across various fields.

Some of the empirical examples and materials are based on well-known journal articles:

- Hall, R.E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. Journal of Political Economy, 86(6), 971–987.
- Stock, J.H. & M.W. Watson (1988). Variable trends in economic time series. Journal of Economic Perspectives, 2(3), 147–174
- Stock, J.H. & M.W. Watson (2001). Vector autoregressions. Journal of Economic Perspectives, 15(4), 101–115.

More background information (journal articles and their references)

- Engle, R.F., Lilien, D.M. & Robins, R.P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. Econometrica, 55, 391–407.
- Glosten, L.R., Jagannathan, R. & Runkle, D.E. (1993). On the relation between the expected value of the volatility of the nominal excess return on stocks. Journal of Finance, 48 , 1779–1801.
- Lanne, M. & Saikkonen, P. (2006). Why is it so difficult to uncover the risk-return tradeoff in stock returns? Economics Letters, 92, 118–125.
- Newey, W.K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703–708.
- Nyberg, H. (2012). Risk-return tradeoff in U.S. stock returns over the business cycle. Journal of Financial and Quantitative Analysis, 47, 137–158.

- Scruggs, J. (1998). Resolving the puzzling intertemporal relation between the market risk premium and conditional market variance: A two-factor approach. Journal of Finance, 53, 575–603.

Datasets and data sources used in this lecture manuscript:

- “Shiller data” refers to Robert Shiller’s updated data on Irrational Exuberance: http://www.econ.yale.edu/~shiller/data.htm.
    - See Shiller, R. (2005). Irrational Exuberance. 2nd Edition. Currency Books.
- FRED data (Federal Reserve Bank of St. Louis): https://fred.stlouisfed.org/.
    - Hereafter “FRED” (“FRED data”) refers to the Federal Reserve Economic Data (FRED) database, maintained by the Federal Reserve Bank of St. Louis.
- quantmod R package: https://cran.r-project.org/web/packages/quantmod/index.html.
    - Data retrieved via the quantmod R package is seemingly often from Yahoo Finance (https://finance.yahoo.com)
    - This package provides tools for quantitative financial modeling and data retrieval. Note that when using financial data in more “serious analyses”, like in your thesis work, ensure always that you have the appropriate license and permission to access and use the data sources, such as possibly integrated via quantmod.

Notation

On this course, we follow, for example, the following key notational selections:

- We will use the same notation for random quantities and their realized values. The context should make the difference clear.
- In vector and matrix calculations, vectors are interpreted as column vectors and we might also denote, for example, $y = [y_1 \cdots y_T]'$, where the superscripted comma denotes transpose of a vector (here $T \times 1$ vector), or elsewhere also of a matrix. Vectors (column vectors) are also at times denoted as $y = (y_1, \dots, y_T)$, indicating the same vector as above.

Concerning different (stochastic/time series) processes, we use mainly notations y_t, x_t and z_t.

- z_t is typically a zero mean process and denoting often a demeaned process. That is $z_t = y_t - \mathsf{E}(y_t)$, where y_t is the “original” time series.

- x_t, or x_t in the case of multiple time series, is used for a general time series or a vector of time series for different notational purposes, but most often they are in the role of explanatory (predictive) variables.

Chapter 1

Introduction

1.1 Time series data

A **time series** is a collection of data where observations correspond to consecutive time periods. A common notation for time series data is y_t, $t = 1, \dots, T$, where t denotes the time period and T the total number of observations. The interval between observations y_{t-1} and y_t, or y_t and y_{t+1}, represents a unit of time, for example, an hour, a month, a quarter or one year. In many cases, it is natural to use more concrete labels for the time periods, such as $t = 2000, \dots, 2024$ for yearly data, or $t = 2000 : 1, 2000 : 2, \dots, 2024 : 12$ for monthly observations. Therefore, occasionally T may also refer to the final time label, such as 2024:12 (December 2024).

- Time series data are analyzed in various fields, including economics and finance, as well as social, natural, engineering, and medical sciences.
- In (macro)economics and finance, data sets are typically time series, which has led to the specialized subfields of **macroeconometrics** and **financial econometrics**.

In this course, we assume that observations recorded an equally spaced in time intervals (or can be reasonably treated as such). Irregularly spaced data require different methods and are not covered here. If the time series observations y_t, $t = 1, \dots, T$, are scalar valued, we refer to them as a **univariate time series**. Conversely, if $\mathbf{y}_t = (y_{1t}, \dots, y_{Kt})$ is vector-valued (indicated by the bold symbol), we refer to it as a **multiple time series** or **vector-valued time series**. In these latter cases, relationships between two or more time series are of interest.

We restrict ourselves to the common case where y_t is a **real-valued** time series, or is **at least treated as such**, meaning that its realizations are real numbers.

In discrete and limited dependent time series, y_t takes only discrete values, such as binary values ($y_t = 1$ or $y_t = 0$) or values where the range is limited (such as the case where y_t is nonnegative). These types of time series require specific nonlinear models, which we will not cover in this course, except models used in volatility modelling where positive-valued variables are relevant).

When analyzing time series data, a useful and natural first step is to plot the data and visually inspect its main characteristics visually. Below, we plot the monthly Consumer Price Index (CPI) for the United States over the time period 1990:1-2025:6 (January 1990-June 2025). In this figure, as is almost always the case when plotting time series data, the line connects observation points to highlight the sequential nature of the data and the dependency between observations.

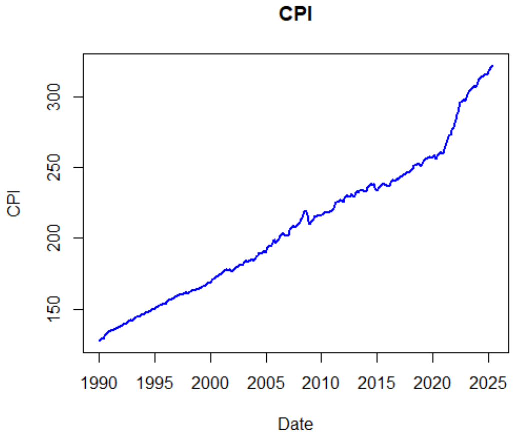


Figure: Monthly U.S. CPI index (levels) (1990:1–2025:6).

The aim of time series analysis is often to build a **statistical model** that represent the observed time series and its fluctuation over time. This typically involves examining the dependence structure among consecutive observations. **Autocorrelation** (serial correlation) between observations is a key property of time series and must be accounted for in statistical and econometric modelling.

- This violates the no-autocorrelation assumption typically imposed in linear regression models for cross-sectional datasets, as introduced in basic econometrics and statistics courses.
- With cross-sectional data, random sampling generally ensures that the error terms of the linear regression model are mutually independent (i.e., not autocorrelated), so autocorrelation is usually not an issue.

When analyzing time series data, **a useful and natural first step is to plot the series and visually inspect its main characteristics**, alongside any underlying theory or background knowledge of the phenomenon the time series represents. This provides at least a preliminary understanding of the dependence structures behind the data-generating process. Some important questions to consider when examining a time series graph include:

- Is there a **trend** (that is, do the values of the time series tend to increase or decrease over time)?
- Are there **seasonal effects** (a regularly repeating pattern of highs and lows related to calendar time, such as seasons, quarters, months or days of the week)?
- Are there major **outliers** (observations far away from other data points)?
- Is the **variability** (variance) relatively **constant over time**?
- Are there any **abrupt changes**, such as **breaks** or **regime switches**, in either the level or variance of the series?

As an example, in the U.S. CPI time series shown above, a clear feature is the persistent trend over time. However, rather than focusing on the level of the CPI, we are primarily interested in its changes as a measure of inflation, as will be discussed below.

An upward trend is also evident in the monthly S&P 500 stock market index. Stock prices have risen substantially between 1990 and 2025, with a pronounced surge is clearly visible. At the same time, several significant market corrections—commonly referred to as “bear markets”—occurred during this period. It is important to note that this trending behavior differs markedly from that observed in the CPI series. The characteristics underlying these types of stochastic trends, particularly those seen in the S&P 500 index, will be explored later in this course when we discuss nonstationary time series.

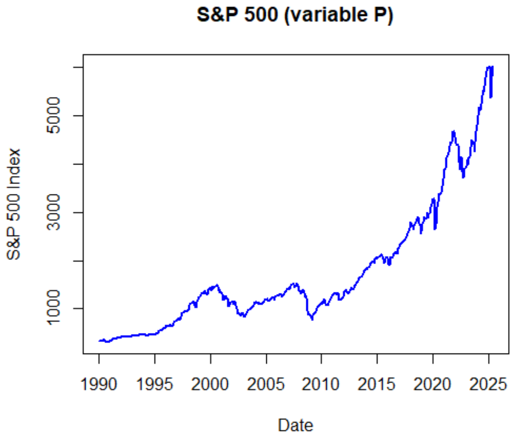


Figure: S&P 500 stock market index between January 1990 and June 2025 (source: Shiller data).

The next figure depicts the monthly number of fatalities in traffic accidents in the U.S. between 1973:1–1978:12. In this series, there is no apparent upward or downward trend. Instead, the series exhibits a clear **seasonal variation**, with the yearly maximum consistently occurring in July and the minimum in February.

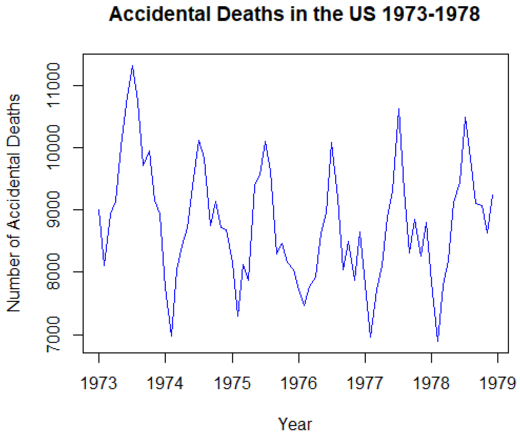


Figure: Monthly traffic accidents between January 1973 and December 1978. (source: Brockwell and Davis, 1996).

The following figure shows the yearly number of sunspots for the period 1770–1870. This series has no trend or seasonal variation, although the **cyclical pattern** could easily be mistaken for seasonality. However, the pattern is not seasonal because the nature of the cycle changes over time. In particular, the distance between consecutive peaks is not constant as it would be in the case with seasonal variation. Another apparent feature is the positive correlation between consecutive observations: a large observation is typically followed by another large one, and a small observation by another small one.

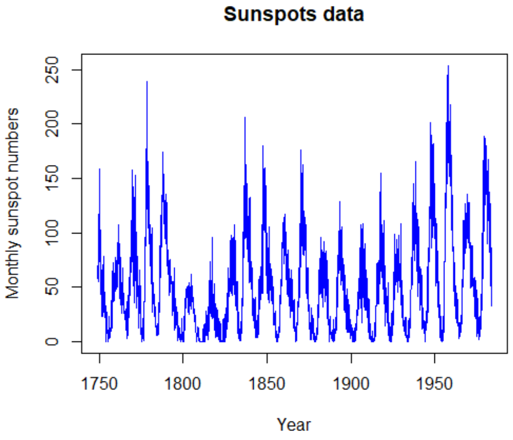


Figure: The number of sunspots between the years 1770–1870 (source: Brockwell and Davis, 1996).

Overall, these visual findings may already suggest which type of **time series model** could be appropriate. Once a model has been selected, as we will consider in more detail in the following sections, one can then:

- Estimate the unknown parameters of the model,
- Assess whether the estimated model is compatible with the observed data,
- Test hypotheses concerning the model parameters, and
- Use the model for its intended end purpose.

The ultimate purpose of the statistical model depends on the context and the field of application.

- One common use is to provide a **summarized description** of the observed dataset, which can help in understanding the underlying data-generating process.
- Another central use of a time series model is to **forecast** future values of one or more series. In the multivariate case, a series of interest can be predicted using its own past values but also information contained in other series as explanatory or predictive variables.
- A further objective is to explore potential **temporal and contemporaneous dependencies** between time series, such as the relationship between inflation and stock market returns (after applying relevant transformations, as discussed next). For example, one may investigate whether higher inflation tends to occur simultaneously with rising stock prices.

1.2 Transformations

As described above, in time series modelling a good first step is always to plot the time series under investigation. This provides an initial view of its main characteristics, such as trends, seasonal patterns, and any sudden breaks or outliers. In the following sections, we assume that such patterns are absent. If the observed series does exhibit these features, appropriate transformations should be applied to remove or mitigate them.

In time series analysis, several transformations and adjustments/"manipulations" are commonly used. These are natural and often essential components of empirical analysis in many applications.

- **The first lag** of the time series y_t is denoted by y_{t-1}. Generally k**th lag** is y_{t-k}.
- Perhaps the most commonly used transformation for time series data is **differencing**. The **first difference** of the series, Δy_t, is the **change** between y_t and y_{t-1}. In other words,
$$\Delta y_t = y_t - y_{t-1}.$$
- Often differencing is combined with taking natural **logarithms** (when assuming $y_t > 0, \forall t$). Taking logarithms is common especially when the series exhibits exponential growth or when the variation in the series increases as the level increases. The **logarithmic change** is
$$\Delta \log(y_t) = \log(y_t) - \log(y_{t-1}).$$

Importantly, the logarithmic changes can be interpreted as relative changes when the changes are small:

$$\Delta\log(y_t) \approx (y_t - y_{t-1})/y_{t-1}.$$

- The **percentage of change** between periods t and $t-1$ can be approximated as $100\Delta\log(y_t)$. For quarterly data y_t, **annualized growth rates** are obtained by multiplying 400, that is $400\Delta\log(y_t)$. Similarly, for monthly data, annualized growth rates are computed as $1200\Delta\log(y_t)$.

As an example, consider the monthly U.S. CPI time series. Taking log-differences (first applying the logarithm and then differencing) and multiplying the resulting series with 1200, we obtain a (proxy) for the annualized inflation rate. Because the CPI typically exhibits an upward trend, log-differencing removes this trend, resulting in a series without a clear deterministic trend. This detrending implies that, instead of an increasing level, the transformed series has a positive mean, which reflects the average inflation rate over the sample period.

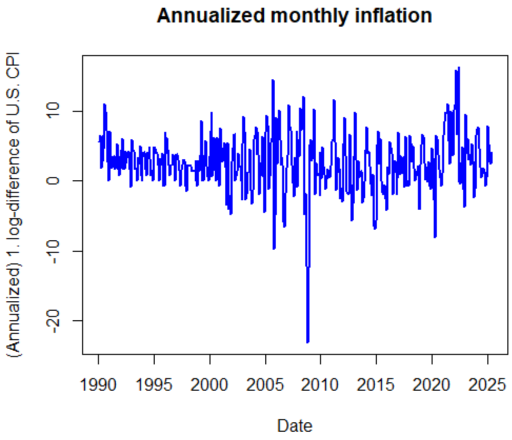


Figure: Annualized U.S. inflation (CPI inflation) rate between February 1990 to June 2025 (1990:2–2025:6).

More on differencing. Differencing may also happen with a longer lag than the first lag: $y_t - y_{t-s}$, $(s \geq 1)$.

- As discussed above, the above case $s = 1$ is the most common (representing change from one period to another).

- In the case of seasonal variation, s is often selected to match the length of the seasonal pattern. That is, for example, $s = 4$ for the quarterly data and $s = 12$ for the monthly series.

If denoting temporarily y_t as the log CPI, taking seasonal differencing $100(y_t - y_{t-12})$ yields another proxy for the U.S. inflation (cf. the time series above)

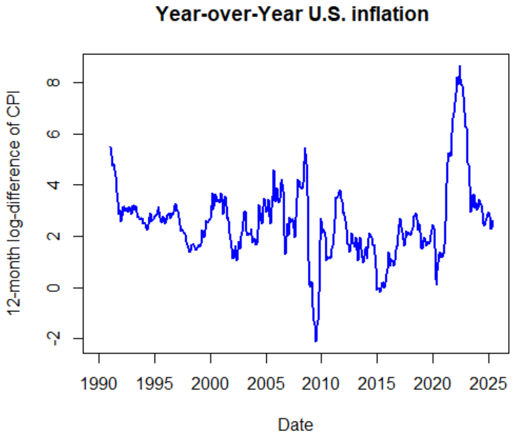


Figure: Annualized U.S. inflation (CPI inflation) obtained as year-to-year changes in the CPI (1990:2–2025:6).

As another use of differencing, especially in financial econometrics and empirical finance, is to obtain returns that different assets yield for investors. Let P_t the price of an asset at time t. Holding the asset for one period from $t-1$ to t yields the simple return

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

In addition, the so-called continuously compounded returns (or log returns) are obtained as

$$r_t = \log(1 + R_t) = \log\Big(\frac{P_t}{P_{t-1}}\Big) = p_t - p_{t-1},$$

where $p_t = \log(P_t)$. Log returns have several advantages over simple returns, such as time-additivity and better behavior under compounding. In both cases, percentage returns are obtained by multiplying r_t or R_t by 100.

- Importantly, in many financial applications, dividend (and similar) payments must be considered when computing returns. If an asset pays dividends periodically, the definition of return needs to be adjusted accordingly. We will not explore these issues in detail in this course; they are left for more advanced studies in finance.

Below, we depict the annualized log-differences of the monthly S&P 500 index, multiplied by 100 to express them as percentage returns. These log-differences approximate monthly percantage returns for a broad U.S. market index. The shape of the return time series is clearly very different from the S&P 500 index level data shown above.

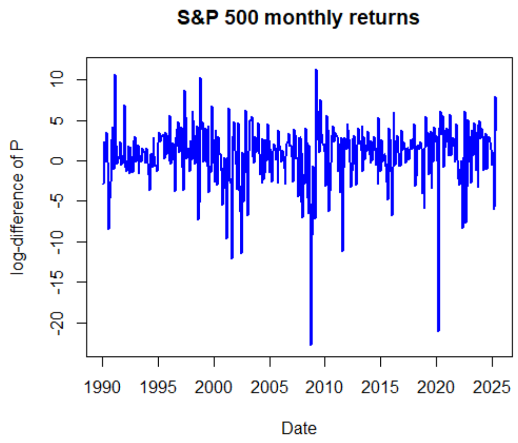


Figure: (Annualized) monthly percentage changes in the S&P 500 index.

Detrending. The trend and/or seasonal patterns in the original time series are often removed through transformations (if necessary), which may require estimating certain parameters. For example, in the case of a linear trend, one can fit a trend line using least squares regression and then consider the resulting residual series using time series methods. In that case, the observed time series y_t would be replaced by the (residual) time series

$$\hat{x}_t = y_t - \hat{\alpha} - \hat{\beta}t,\ \ t = 1, \dots, T,$$

where the estimates $\hat{\alpha}$ and $\hat{\beta}$ are obtained by regressing y_t on a constant and t. More generally, instead of the linear function of time, one can use higher-order polynomials or other functional forms to capture more complex trends.

- To model seasonal variation, one could use trigonometric functions of time.
- An often used alternative to remove seasonal variation (if not using already existing seasonally adjusted data) is the so-called seasonal indicators. As an example of this, for quarterly data, one would obtain seasonally-adjusted series by

$$\hat{x}_t = y_t - \hat{\beta}_1 d_{1t} - \cdots - \hat{\beta}_4 d_{4t},\ \ t = 1, \dots, T,$$

where the ith seasonal indicator d_{it} takes value one when the observation in question corresponds to quarter i and zero otherwise ($i = 1, \dots, 4$).

As for additional (extra) information, there is a substantial body of literature on advanced detrending techniques beyond simple differencing and seasonal dummy variables in econometrics.

1.3 Assumptions and limitations of this course

In summary, we consider time series models/processes under the assumption that the observed time series do not exhibit clear seasonal patterns. If necessary, such patterns have been removed through an appropriate transformation or a **seasonal adjustment method**, so that the modelling is applied to the transformed series.

- A common approach, especially with macroeconomic data, is to use **seasonally adjusted datasets** provided by official statistical agencies, such as Statistics Finland. Therefore, we can reasonably assume that seasonal effects have already been accounted for to a satisfactory degree.

Concerning trends, in the first part of this course (Sections 2–11) we focus on **stationary time series processes** and their realizations (i.e. observed time series). These models assume that the time series does not exhibit clear trends or very persistent/trending behaviour.

- That is, the time series considered here resemble the U.S. inflation series and stock market returns (as depicted above), rather than the CPI level series or the S&P 500 price series.

- For now, it is sufficient to understand that non-stationary time series can usually (though not always) be transformed through differencing so that the stationarity assumption holds reasonably well. This allows us to model the resulting series as a stationary process such as ARMA(p, q) process, which will be introduced later.

- An exception to this focus on stationary time series occurs in the latter part of the material, where trending behaviour is examined in more detail. The basics of non-stationary processes will be discussed in Sections 12–13, where modelling the deterministic, and, especially, stochastic trends becomes important in several applications.

Finally, we begin with univariate time series, meaning we analyze a single series at a time in Sections 2–9. Later, in Sections 10–11 and 13, we will introduce the basics of multivariate time series analysis, where multiple time series are modeled jointly.

Chapter 2

A primer on time series models

2.1 Starting point

Relation to linear regression for cross-sectional datasets. In standard econometrics or statistics courses, linear regression models are typically introduced using cross-sectional data. A key assumption in these models is the absence of autocorrelation in the error term. With time series data, however, autocorrelation is a fundamental characteristic and is almost always present.

- While linear regression can be applied to time series variables, the notation changes to reflect the temporal ordering. Variables are indexed by time, such as y_t and x_t (instead of y_i and x_i for cross-sectional observations).

We will return to multivariate regression models with time series variables later. First, we will explore **a pure time series approach**, as characterized by Verbeek (2008).

- In this approach, the current value of a univariate time series, y_t, is modeled as a function of its past values, either directly or indirectly.
- The emphasis is entirely on exploiting the information contained in y_t and its own history for modeling and forecasting purposes.

2.2 Notes on deterministic components

In time series analysis, we often decompose a process y_t into deterministic and stochastic components. This separation allows us to distinguish deterministic elements, such as the mean or the trend, from random fluctuations, which are typically the primary focus in stochastic models like ARMA or VAR models to be introduced later in this course.

The general decomposition and notation used throughout this material is

$$y_t = \mu_t + z_t,$$

where

- μ_t is the **deterministic component**, represeting features such as mean, trend or seasonality
- z_t is the **stochastic component**, capturing random fluctuations around the deterministic part. We assume $\mathsf{E}(z_t) = 0$, so z_t represents the demeaned and detrended process. Importantly, z_t typically exhibits its own autocorrelation structure and is therefore not merely pure random noise.

By construction, the expected value of y_t is then simply its deterministic component:

$$\mathsf{E}(y_t) = \mathsf{E}(\mu_t + z_t) = \mu_t + \mathsf{E}(z_t) = \mu_t$$

Common forms of the deterministic component. The functional form of μ_t depends on the underlying characteristics of the data and the application. Two of the most common specifications are:

- **Constant mean**: If the time series fluctuates around a fixed, time-invariant level, the deterministic component is simply a constant

  $$\mu_t = \mu.$$

  Here, μ (sometimes denoted μ_0) represents the unconditional mean of y_t, that is $\mathsf{E}(y_t)$. In practice, almost all real-world time series have a nonzero mean.

- **Constant and linear time trend**: If the series exhibits a persistent upward, or downward movement over time, the deterministic component includes both a constant and a linear trend

  $$\mu_t = \mu_0 + \mu_1 t$$

  where μ_0 is the constant (intercept) (the value of μ_t at $t = 0$), and μ_1 is the trend coefficient, representing the average change in y_t from one period to the next.

In practice, we may first estimate the parameters of the deterministic component μ_t, for example, using Ordinary Least Squares (OLS), and then work with z_t as the residual $y_t - \mu_t$. This approach corresponds to the simple detrending technique described in Section 1. Alternatively, and more commonly, the deterministic and stochastic components are modelled simultaneously.

Chapter 3

Stationary processes

3.1 Basic concepts

The aim is to build a statistical model for the observed time series $y_1, \dots, y_T$. At this stage, and in Sections 3–9 and 12, we assume that we have one, univariate, time series. Multiple time series and multivariate models will be briefly considered later in Sections 10–11 and 13.

On a rather general level, the observed time series can be interpreted as the observed value of the random vector $y = (y_1, \dots, y_T)$ or, in other words, the **realization** of this random vector.

- Recall the notation: Here the random vector interpretation emphasizes the fact that y_t will be interpreted as a random variable. Notice that We do not make a distinction between random variables and their realizations; the difference should be clear from the context.

When building a statistical model, we would attempt to specify the probability distribution of the T–dimensional random vector $y = (y_1, \dots, y_T)$ that produced the observed time series.

- Note and recall that in a time series, the observations are (typically) not independent as in the conventional linear regression, and therefore it is not enough to specify the marginal distributions of the components y_t $(t = 1, \dots, T)$.
- Instead, we indeed need to specify the T–dimensional joint distribution.

A concrete and often used way to specify the joint distribution of y in time series analysis is to:

- Specify a model equation that characterises the dependence structure of consecutive observations, and
- combine this with a “suitable” distributional assumption.

Before we proceed to such concrete model equations, it is useful to first briefly study some basic concepts of probability theory that form the foundation of time series analysis.

3.2 Stochastic processes

In statistical and econometric modelling of time series, one concentrates on the observed time series. That is the realization of a random vector $y = (y_1, \dots, y_T)$.

From a mathematical point of view, however, it is more convenient to consider an “extension” of this vector to all time indices when defining time series models. This leads us to the concept of (a discrete time) **stochastic process** (or just simply **process**), which is a collection of random variables $\{y_t; t = 0, \pm 1, \pm 2, \dots\}$.

- Occasionally, the cases $t = 0, 1, 2, \dots$ or $t = 1, 2, \dots$ are also considered.
- Because the set of values the time index takes is usually known, it does not need to be emphasized, and often we can simply denote the stochastic process by y_t, or $\{y_t\}$ if one wants to emphasize the distinction to a single component of the process.
- The terms “**process**” and “**model**” are often used interchangeably in time series analysis. In this material, however, “process” refers specifically to the underlying theoretical mechanism—namely, the stochastic process—that we aim to represent (learn) when specifying an empirical time series “model” for the observed data.

From a practical point of view, it is **important to understand that we have only a single realization available**, the observations $(y_1, \dots, y_T)$, of the process $\{y_t\}$.

- In other words, **we have only one realization from the process and that is only partial**, that is observations are available only for the time periods $t = 1, \dots, T$.

These above-mentioned remarks imply that in order to investigate the properties of y_t by making use of the observed data $y_1, \dots, y_T$ only, one must make some restrictive assumptions about y_t for this to be possible.

3.3 Weak stationarity

To illustrate the final remarks in the previous section, consider the **expected value function** of the process y_t (assume here $\mathsf{E}\left(y_t^2\right) < \infty$ for all t)

$$\mu_t = \mathsf{E}\left(y_t\right), \quad t = 0, \pm 1, \pm 2, ...$$

as well as the **covariance function** of y_t

$$\gamma_{s,t} = \mathsf{Cov}\left(y_s, y_t\right) = \mathsf{E}\left[\left(y_s - \mu_s\right)\left(y_t - \mu_t\right)\right], \quad t = 0, \pm 1, \pm 2, ...$$

Estimating these functions by making use of one realization $y_1, ..., y_T$ only is of course impossible, unless we somehow restrict/limit their dependence on time. **The situation simplifies if it is assumed that these quantities do not depend on time**. Indeed, it is common to assume that such a dependence on time does not (in a particular sense) exist.

A process y_t is called **weakly stationary**, or **covariance stationary**, if

$$\mathsf{E}\left(y_t\right) = \mu < \infty \quad \text{for all} \quad t = 0, \pm 1, \pm 2, ...$$

and

$$\mathsf{Cov}\left(y_t, y_{t+h}\right) = \gamma_{t,t+h} = \gamma_{0,h} < \infty \text{ for all } \ h, \ t = 0, \pm 1, \pm 2, ... \ .$$

In other words, a form of **time invariance** holds: The expected value function μ_t is constant over time, and the covariance function $\gamma_{s,t}$ does not depend on the time indices s and t but only on their distance $t - s$.

- For brevity, denote $\gamma_{0,h} \equiv \gamma_h$.
- When h is fixed, γ_h is called the **autocovariance (coefficient)** of y_t at lag h. As a function of h, the γ_h is called the **autocovariance function** of y_t.
- The μ is called the **expected value** or the **mean**.

To summarize, a characteristic feature of weakly stationary process is that the first and second moments are finite and independent of time. Unless otherwise mentioned, in what follows in this and coming few sections, **we assume that weak stationarity holds**.

A useful **practical aspect** of weak stationarity is that the plausibility of this assumption can be easily investigated by looking at a graph of the observed time series: If the observations vary with constant variance around a fixed level, then weak stationarity appears a plausible assumption.

Empirical example. Let us consider an example related to the U.S. real GDP growth during the period 1985:Q1-–2007:Q2, commonly referred to as the **Great Moderation** time period.

- More on The Great Moderation, see https://www.federalreservehistory.org/essays/great-moderation
- Real Gross Domestic Product (GDP) is the total value of a country's goods and services produced in a specific period, adjusted for inflation.
- This is the dataset GDPC1-qdata.xlsx (source: FRED)

Figure below illustrates that the time series fluctuates around a relatively stable mean and exhibits approximately constant variance over time. As discussed, this behavior is typical for a weakly stationary process.

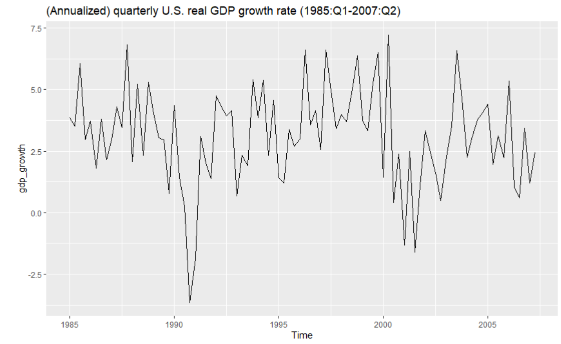


Figure: (Annualized) quarterly U.S. real GDP growth rate (1985:Q1-2007:Q2) (ggplot2 figure).

It is important to note that **weak stationarity does not imply the absence of variability**. Even in a weakly stationary process, individual observations can vary substantially due to random shocks. What matters is that the statistical properties — the mean, variance, and autocovariance structure — remain constant over time. So, occasional high or low values in the series are entirely consistent with weak stationarity, as long as they are the result of random variation and not due to major structural changes in the process.

- We will soon see more examples of weakly stationary time series through simulations.

- In Section 1, the time series depicting the U.S. inflation rate can also be interpreted as realizations from weakly stationary processes.

Because $\gamma_0 = \mathsf{Var}\,(y_t)$, it holds that $\gamma_0 \geq 0$. Furthermore, the familiar properties of the covariance function imply that

- $|\gamma_h| \leq \gamma_0$.
- In this univariate (single time series) case, we have (under weak stationary)

$$\mathsf{Cov}\,(y_t, y_{t+h}) = \mathsf{Cov}\,(y_{t-h}, y_t) = \mathsf{Cov}\,(y_t, y_{t-h})\,,$$

  which also implies that $\gamma_h = \gamma_{-h}$. This expression shows where the term "lag" ("lag length") h is coming from.

Putting the above properties together, the autocovariance function has the properties

$$\gamma_0 \geq 0,\ |\gamma_h| \leq \gamma_0 \quad \text{and} \quad \gamma_h = \gamma_{-h}.$$

The case $\gamma_0 = 0$ is obviously not of interest, and in what follows we assume that $\gamma_0 > 0$.

3.4 Autocorrelation function

In practice, a more convenient concept than the autocovariance function is the **autocorrelation function**

$$\rho_h = \mathsf{Cor}\,(y_t, y_{t+h}) = \mathsf{Cor}\,(y_t, y_{t-h}) = \gamma_h/\gamma_0.$$

The **autocorrelation coefficients** ρ_h obviously then have the properties

$$\rho_0 = 1,\ |\rho_h| \leq 1, \text{ and } \rho_h = \rho_{-h}.$$

Therefore, it suffices to consider the autocovariance and autocorrelation functions only for the lag lengths $h \geq 0$

- The latter just for the lags $h > 0$

Due to the time invariance guaranteed by weak stationarity, it now seems evident that $\mathsf{E}\,(y_t) = \mu$ and $\mathsf{Cov}\,(y_t, y_{t+h}) = \gamma_h$ can be estimated. One more problem remains: The autocorrelation function, in general, has an infinite number of quantities to be estimated. In practice, all of these cannot be estimated (without further restrictions). However, typically we can assume that

$$\gamma_h \to 0, \text{ when } h \to \infty.$$

In this case, the random variables y_t and y_{t+h} become **nearly uncorrelated when the distance h is "large"** – that is, when y_t and y_{t+h} are "far" from each other in time.

- Therefore, in practice it suffices to estimate the autocorrelation function ρ_h, when $h = 1, \dots, H$, and H is so large that $\rho_h \approx 0$ for $h > H$.

3.5 Sample autocorrelation function

Assume that the process y_t is weakly stationary. Then $\mathsf{E}(y_1) = \cdots = \mathsf{E}(y_T) = \mu$. Therefore, a natural estimator of the (population) mean is the **sample mean**

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t.$$

Because (under weak stationarity) $\mathsf{Cov}(y_1, y_{1+h}) = \cdots = \mathsf{Cov}(y_{T-h}, y_T) = \gamma_h$ $(h \geq 0)$, it also seems natural to estimate the population autocovariance coefficient γ_h using the sample covariance of the observations $(y_1, y_{1+h}), \dots, (y_{T-h}, y_T)$. Because $\mathsf{E}(y_t) = \mathsf{E}(y_{t+h})$ (again under weak stationary), a commonly used estimator is

$$\mathsf{c}_h = \frac{1}{T-h}\sum_{t=1}^{T-h}(y_t - \bar{y})(y_{t+h} - \bar{y}), \quad 0 \leq h < T,$$

which is called the (hth) **sample autocovariance coefficient**. Sometimes, the denominator $T - h$ is replaced by T.

- When c_h is interpreted as a function of h (i.e. $c_1, c_2, \dots$), it is called the **sample autocovariance function**.

A natural estimator of the autocorrelation coefficient ρ_h is hence

$$\mathsf{r}_h = \mathsf{c}_h / \mathsf{c}_0, \qquad 0 \leq h < T,$$

which is called the **sample autocorrelation coefficient** and $\mathsf{r}_1, \mathsf{r}_2, \dots$ define the **sample autocorrelation function**.

- Notice that by definition $\mathsf{r}_0 = 1$. Some statistical program packages report this uninteresting case.

As an example, consider the quarterly U.S. real GDP growth rate (1985:Q1–2007:Q2). Its sample autocorrelation function is plotted below for the lags $h = 0, \ldots, 20$. The two horizontal lines represent certain “confidence bands”.

- If $\rho_h = 0$ for all $h > 0$, then any individual estimator r_h has an approximately 95% probability of lying between the confidence bands $(-1.96/\sqrt{T}, 1.96/\sqrt{T})$ where T is the number of observations. These confidence bands are depicted below and typically in corresponding figures.

- If the individual sample autocorrelation coefficient r_h does not fall between these bands, there is statistically significant autocorrelation at the 5% significance level at the lag h.

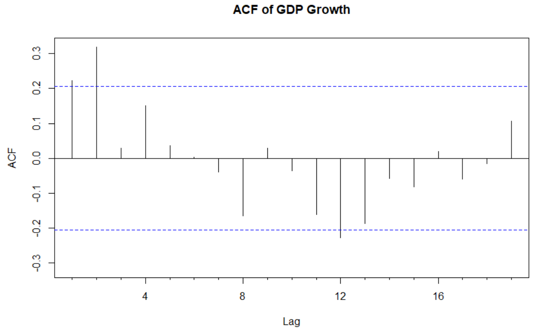


Figure: Sample autocorrelation function (ACF) of the quarterly U.S. real GDP growth rate ($T = 90$, and hence $1.96/\sqrt{T} \approx 0.207$.)

This time series exhibit clear autocorrelation, which generally dampens (slowly) as the lag length h increases.

- This dampening effect depends on the time series (process) considered

- Notice that above the first depicted lag (in this R function) is r_1. At times in different functions and packages $\mathsf{r}_0 = 1$ is depicted, but that is of course not of interest.

- If there is some (remaining) seasonal variation, say in monthly time series after seasonal adjustment (if applicable), it causes rather strong autocorrelations at lags $h = 12$, 24 and 36. Similar behaviour can also happen if the time series contain some other cyclical pattern.

3.6 White noise process

Weak stationarity is not a strong enough assumption to allow statistical analysis of time series. It alone does not suffice for the construction of a likelihood function, nor is it enough to derive distributions of estimators or test statistics. For this reason, we need to define another concept of stationarity that is stronger than weak stationarity, **strict stationarity**. We will consider that in the next section.

Before that, let us consider first a **white noise process**, which is the simplest possible example of a weakly stationary process. Let u_t be a process for which it holds that

- $\mathsf{E}\,(u_t) = \mu < \infty$, often $\mathsf{E}(u_t) = 0$,
- $\mathsf{Var}\,(u_t) = \sigma^2 < \infty$, and
- $\mathsf{Cov}\,(u_s, u_t) = 0$, when $s \neq t$.

This process is often called a **(weak) white noise** process and denoted by $u_t \sim \mathsf{wn}\,(\mu, \sigma^2)$.

- By definition, it is clear that a white noise process is weakly stationary
- Despite their simplicity, as we will see later on, white noise processes play a central role in time series analysis because they can be used to construct more complicated processes and eventually models.

3.7 Strictly stationary process

Let us now define the above-mentioned more stringent form of stationarity. A process $\{y_t; t = 0, \pm 1, \pm 2, ...\}$ is called **strictly stationary** (sometimes strongly stationary) if the collections of random variables $y_{t_1}, \dots, y_{t_m}$ and $y_{t_1+h}, \dots, y_{t_m+h}$ have the same $m-$dimensional joint distribution for all integers $t_1, \dots, t_m$, h and m $(m > 0)$.

- That is, **the entire probability structure of the process y_t is time-invariant**, not just the first and second-order moments.

Properties. A strictly stationary stochastic process has the following properties (that follow straightforwardly from the definition):

- **SS1**: The random variables y_t have the same distribution for all t.
- **SS2**: The random vectors (y_t, y_{t+h}) and (y_s, y_{s+h}) have the same distribution for all t and s and all (fixed) h.
- **SS3**: y_t is weakly stationary if $\mathsf{E}\left(y_t^2\right) < \infty$ holds.

It is clear that weak stationarity does not, in general, imply strict stationarity.

- For instance, weak stationarity says nothing about moments of order 3 and above.

Extra: Gaussian processes

In the case of so-called Gaussian processes, weak stationarity implies strict stationarity. A process y_t (not necessarily stationary in any sense) is called *normal* or *Gaussian*, if the m–dimensional probability distribution of the random variables $y_{t_1}, \dots, y_{t_m}$ is multivariate normal (Gaussian) for all integers $t_1, \dots, t_m$ and m $(m > 0)$ (for simplicity, occasionally we will simply say normal when meaning multivariate normal distribution). Recall that a multivariate normal distribution is fully determined by its first and second moments. Therefore, weak stationarity and normality together do imply strict stationarity.

A very useful property of strict (but not weak) stationarity is that it is conserved in transformations. In other words,

- **SS4**: If x_t is a strictly stationary process, then so is also the process $y_t = g\left(x_{t+m}, \dots, x_{t-n}\right), m, n \geq 0$, for any "well-behaving" (e.g. continuous or, more generally, measurable) function g. Moreover, m, n, or both can be replaced with ∞.

"Stationarity" in this material. Unless otherwise stated, stationarity will be interpreted as strict stationarity from now on.

- From practical time series analysis perspective, as introduced above, visual inspection of the time series is related to the theoretical implications of weakly stationary processes (i.e. mean, variance and autocovariances of the series are not dependent on time) is of interest.

3.8 IID and NID processes

Let u_t be a sequence of **independent and identically distributed** (IID) random variables.

- The strict stationarity of this process is easy to check.

Extra: Some further results

The cumulative distribution function of the random vector $\left(u_{t_1}, \dots, u_{t_m}\right)$ evaluated at $(x_1, \dots, x_m)$ is $F(x_1) \cdots F(x_m)$. Clearly, this is also the cumulative distribution function of the random vector $(u_{t_1+h}, \dots, u_{t_m+h})$ evaluated at $(x_1, \dots, x_m)$.

If $\mathsf{E}\left(u_t^2\right) < \infty$, then u_t is also weakly stationary, and in this case it has the same moment properties as the (weak) white noise process above.

This IID process is also called white noise, or sometimes **strong white noise**. We will use the shorthand notation

$$u_t \sim \mathsf{iid}\left(\mu, \sigma^2\right) \quad \text{or} \quad u_t \sim \mathsf{IID}\left(\mu, \sigma^2\right)$$

where iid refers to "independently and identically distributed", $\mu = \mathsf{E}(u_t)$ and $\sigma^2 = \mathsf{Var}(u_t)$ (in what follows, most often $\mu = 0$).

If it is additionally assumed that $u_t \sim \mathsf{N}\left(\mu, \sigma^2\right)$, one denotes

$$u_t \sim \mathsf{nid}\left(\mu, \sigma^2\right) \quad \text{or} \quad u_t \sim \mathsf{NID}\left(\mu, \sigma^2\right)$$

where nid refers to **normally and independently distributed**.

- In the following, the above standard notation for the iid, nid and wn processes are used in these lecture notes from now on.

Extra: Some (somewhat peculiar) example cases

The first example demonstrates that the mathematical definition of weak stationarity allows for processes that may feel awkward and surprising. Define the process

$$y_t = A\cos(\lambda t) + B\sin(\lambda t),$$

where $\lambda \in [0, \pi)$ is a constant and the random variables A and B satisfy $\mathsf{E}(A) = \mathsf{E}(B) = 0$, $\mathsf{Var}(A) = \mathsf{Var}(B) = \sigma^2$ and $\mathsf{Cov}(A, B) = 0$. Obviously, $\mathsf{E}(y_t) = 0$

holds. Using properties of trigonometric functions, straightforward calculation can be used to establish that

$$\mathsf{Cov}\left(y_t, y_{t+h}\right) = \sigma^2 \cos\left(\lambda h\right).$$

Therefore, y_t is weakly stationary (details are left as an exercise). This process is peculiar in that its realizations are quite regular functions of time (linear combinations of $\cos\left(\lambda t\right)$ and $\sin\left(\lambda t\right)$). Note also that for this process, condition @ref(eq:Acfdecay) does not hold. □

The second example: Define the process $y_t = u_t\sqrt{\omega + \alpha u_{t-1}^2}$, where $\omega > 0$, $\alpha \geq 0$ and u_t is (weak) white noise process and satisfies $\mathsf{E}\left(u_t\right) = 0$ and $\mathsf{E}\left(u_t^2\right) < \infty$. It is easy to see that y_t is weak white noise (justification as an exercise). Because $y_{t-1} = u_{t-1}\sqrt{\omega + \alpha u_{t-2}^2}$, it is clear that y_t is not strong white noise (unless $\alpha = 0$). □

As an illustration of NID process, below we depict a simulated realization from the nid$(0, 1)$ process. The length of the time series is 300 observations ($T = 300$). In other words, we can state that $z_t \sim \mathsf{nid}(0, 1)$, $t = 1, \ldots, 300$. The simulated time series exhibits completely random variation. As we can see, there are occasionally higher and lower values due to sampling variation.

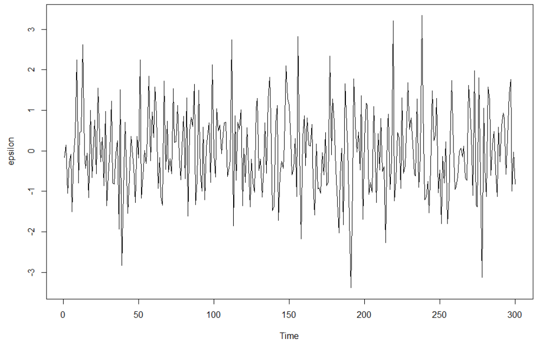


Figure: A simulated realization of $\mathsf{nid}(0, 1)$ process ($T = 300$).

Chapter 4

Linear process

4.1 MA(1) process

As was mentioned, white noise processes can be used to define more complicated processes. A simple example and special case is the **moving average process of order one**, or **MA(1) process**,

$$y_t = \mu + u_t + \theta_1 u_{t-1}, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right).$$

The values of the process are hence assumed to be generated as a weighted average of two independent and unobserved random shocks.

- In this process (model equation), we also include in the mean of the process $\mathsf{E}(y_t) = \mu$. Alternatively, we can write the MA(1) process as

$$y_t - \mu \equiv z_t = u_t + \theta_1 u_{t-1}, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right),$$

- Notice that possible other deterministic components than a nonzero mean can also be readily included in if necessary.

MA(1) processes are strictly stationary (see above and the property SS4) and also weakly stationary. Simple calculations show that

$$\mathsf{E}\left(y_t\right) = \mathsf{E}(\mu) + \mathsf{E}\left(u_t\right) + \theta_1 \mathsf{E}\left(u_{t-1}\right) = \mu,$$

$$\begin{aligned}\mathsf{Var}\left(y_t\right) &= \mathsf{E}\Big(y_t - \mathsf{E}(y_t)\Big)^2 \\ &= \mathsf{Var}\left(z_t\right) \\ &= \mathsf{Var}\left(u_t\right) + \theta_1^2 \mathsf{Var}\left(u_{t-1}\right) \\ &= \sigma^2\left(1 + \theta_1^2\right),\end{aligned}$$

and, for $h > 0$,

$$\begin{aligned}
\mathsf{Cov}\left(y_t, y_{t+h}\right) &= \mathsf{Cov}\left(z_t, z_{t+h}\right) \\
&= \mathsf{E}[\left(u_t + \theta_1 u_{t-1}\right)\left(u_{t+h} + \theta_1 u_{t+h-1}\right)] \\
&= \mathsf{E}\left(u_t u_{t+h}\right) + \theta_1 \mathsf{E}\left(u_t u_{t+h-1}\right) + \theta_1 \mathsf{E}\left(u_{t-1} u_{t+h}\right) + \theta_1^2 \mathsf{E}\left(u_{t-1} u_{t+h-1}\right) \\
&= \begin{cases} \theta_1 \sigma^2, \ h = 1, \\ 0, \ h > 1. \end{cases}
\end{aligned}$$

The latter two calculations make use of the independence of process u_t. These results show the weak stationarity of the MA(1) process.

- The same moment results are also obtained if the assumption $u_t \sim \mathsf{iid}\left(0, \sigma^2\right)$ is replaced with the milder assumption $u_t \sim \mathsf{wn}\left(0, \sigma^2\right)$, but in this case one cannot deduce the strict stationarity of y_t. The same comment holds also to the more general results to be presented in the next section.

Based on these calculations, the autocorrelation function of an MA(1) process takes the form

$$\rho_h = \begin{cases} 1, \ h = 0, \\ \theta_1 \Big/ \left(1 + \theta_1^2\right), \ h = 1, \\ 0, \ h > 1. \end{cases}$$

Therefore, a typical feature of an MA(1) process is that **the autocorrelation function drops to zero after lag one**.

- Thus, observations more than one period apart are uncorrelated and, when assumption $u_t \sim \mathsf{iid}\left(0, \sigma^2\right)$ holds, even independent.

- The same conclusion could, of course, be immediately made from the MA(1) model equation above.

The following figure presents two simulated realizations of length 150 of an MA(1) process with $u_t \sim \mathsf{nid}\left(0, 1\right)$. Sample autocorrelation function based on the observations as well as the theoretical autocorrelation function are also shown.

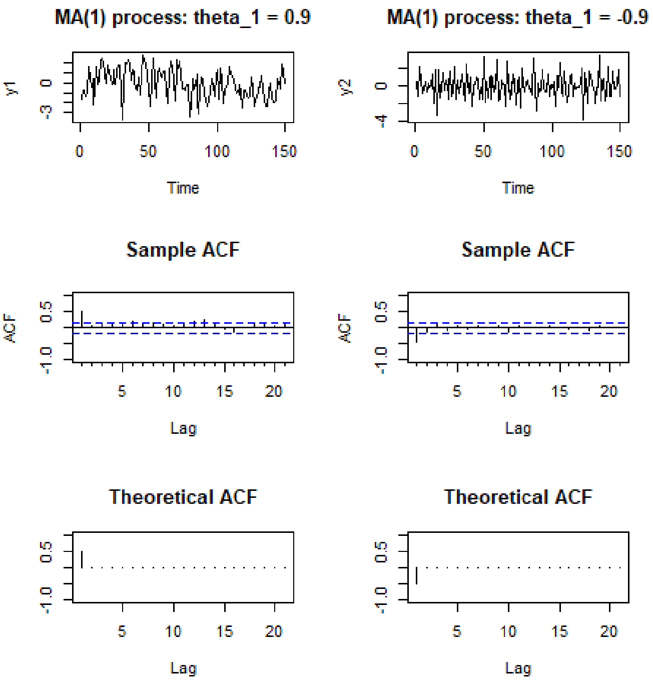


Figure: Two simulated realizations of the MA(1) process and their (sample) autocorrelations, starting from the first autocorrelations (r_1 and $\ _1$).

In Figure, on the left, the MA(1) coefficient θ_1 is 0.9, on the right $\theta_1 = -0.9$. In addition to simulated time series, the figure plots sample autocorrelations ("sample ACF") and theoretical autocorrelations ("theoretical ACF") of these two processes. In these figures $\mathsf{r}_0 = 1$ are depicted.

The figure above shows that in both simulated series, the observations vary around their mean (zero) according to their theoretical standard deviation ($\approx$ 1.345).

- In the left panel, the series has positive autocorrelation, which manifests itself as positive observations, typically followed by another positive observation.
- In the right panel, due to negative autocorrelation, positive and negative observations typically alternate.

The estimated sample autocorrelation functions resemble rather closely their theoretical counterparts. In particular, in both cases, the estimated r_1 is well outside the approximate 95% confidence bands implied by the assumption $\rho_h = 0$ $(\forall h > 0)$. The remaining estimated sample autocorrelations fall mostly within these bands.

- Due to random variation, some of the remaining 39 estimates may naturally occasionally fall outside these bands.
- Furthermore, as we will discuss later, in the case of an MA(1) process, the confidence bands used above are actually too narrow.

An obvious generalization of the MA(1) is obtained by adding a linear combination of the variables $u_{t-2}, \ldots, u_{t-q}$ $(q < \infty)$ to the right hand side of the MA(1). This leads to the so called MA(q) process and to be considered more detail in Section 5.

4.2 Causal linear process

The MA(1) process introduced in the previous section, and its generalization the MA(q) process ($q < \infty$) to be introduced later on, are special cases of the **linear process**

$$y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j u_{t-j}, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right).$$

It is clear that the infinite sum on the right hand side of this equation requires further care and can not be well defined without suitable further restrictions on the coefficients ψ_j.

- We will assume that
$$\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty.$$
Under this assumption the infinite sum on the right hand side of the linear process is well defined.

The mean and autocovariance function of y_t (general linear process) can be calculated in a similar fashion as for the MA(1) process regardless of the infinite sum:

$$\mathsf{E}\left(y_t\right) = \mu + \sum_{j=-\infty}^{\infty} \mathsf{E}\left(\psi_j u_{t-j}\right) = \mu + \sum_{j=-\infty}^{\infty} \psi_j \mathsf{E}\left(u_{t-j}\right) = \mu,$$

and denoting (cf. Section 2 on deterministic components) $z_t \equiv y_t - \mu$, we get

$$\mathsf{Var}\left(y_t\right) = \mathsf{Var}\left(z_t\right) = \sum_{j=-\infty}^{\infty} \mathsf{Var}\left(\psi_j u_{t-j}\right) = \sum_{j=-\infty}^{\infty} \psi_j^2 \mathsf{Var}\left(u_{t-j}\right) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j^2.$$

Moreover, because $\mathsf{Cov}\left(y_t, y_{t+h}\right) = \mathsf{Cov}\left(z_t, z_{t+h}\right)$, for $h > 0$,

$$\begin{aligned}
\mathsf{Cov}\left(y_t, y_{t+h}\right) &= \mathsf{E}\left(\sum_{j=-\infty}^{\infty} \psi_j u_{t-j} \sum_{i=-\infty}^{\infty} \psi_i u_{t+h-i}\right) && (4.1)\\
&= \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \mathsf{E}\left(\psi_j \psi_i u_{t-j} u_{t+h-i}\right) && (4.2)\\
&= \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \psi_j \psi_i \mathsf{E}\left(u_{t-j} u_{t+h-i}\right) && (4.3)\\
&= \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h}, && (4.4)
\end{aligned}$$

where the calculations also make use of the properties of the process u_t (compare the results to the MA(1) process). These calculations show that y_t is weakly stationary.

- Notice that the strict stationarity follows from the strict stationarity of u_t and the property SS4.

Causal linear (MA(∞)) process. Because the linear process defined above contains an infinite number of unknown parameters (the ψ_j's), it cannot be used in practice to obtain a useful statistical model unless we place some further restrictions on ψ_j.

- Despite this, the general linear model is a useful theoretical device because many processes used in practice are special cases of it.

Like in the case of an MA(1) process, it typically holds that $\psi_j = 0$ for $j < 0$, in which case the general linear process reduces to

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j u_{t-j}, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right).$$

Because y_t no longer depends on future values of the u_t variables, one often speaks of a **causal linear process** or a **causal MA(∞) process**.

- The values of the process y_t are assumed to be generated as a weighted sum of (possibly) infinitely many independent and unobserved random shocks.
- In the **noncausal case** the future shocks u_{t+j} $(j > 0)$ affect the present value of the process y_t. In this course, we do not consider noncausal models more detail.

4.3 AR(1) process

A simple special case of a causal linear process containing only one unknown parameter (and constant term) is achieved by assuming that $\psi_j = \phi_1^j$. A process defined like this leads to the autoregressive process of order one, that is an **AR(1) process**

$$y_t = \nu + \phi_1 y_{t-1} + u_t, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right).$$

Here one interprets the present value of the process to linearly depend on the previous value of the process as well as on an unobseved random shock (or error term) similarly as in the linear process. Furthermore, ν denotes the constant term of the process, whose connection to the mean $\mathsf{E}(y_t) = \mu$ will be examined below.

- Referring to the linear causal process, for the condition $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ to be satisfied, it needs to assumed that $|\phi_1| < 1$.

Taking the AR(1) model equation as a starting point, the necessity of condition $|\phi_1| < 1$ can be demonstrated by making use of repetitive substitutions.

- First, substitute $y_{t-1} = \nu + \phi_1 y_{t-2} + u_{t-1}$ on the right hand side of the AR(1) equation.
- Then substitute $y_{t-2} = \nu + \phi_1 y_{t-3} + u_{t-2}$ to the resulting expression, and so on.

Continuing in this fashion, we obtain the equation

$$y_t = \phi_1^k y_{t-k} + \nu \sum_{j=0}^{k-1} \phi_1^j + \sum_{j=0}^{k-1} \phi_1^j u_{t-j}.$$

When $|\phi_1| < 1$, this leads us to the limiting solution

$$y_t = \nu \sum_{j=0}^{\infty} \phi_1^j + \sum_{j=0}^{\infty} \phi_1^j u_{t-j}.$$

Because the AR(1) process (with $|\phi_1| < 1$) is clearly a special case of the linear process, it is strictly and weakly stationary when $|\phi_1| < 1$.

The first and second moments can be deduced from the general formulae derived in the previous section. The expected value, variance and autocovariance functions take the form

$$\mathsf{E}\left(y_t\right) \equiv \mu = \nu \sum_{j=0}^{\infty} \phi_1^j + \mathsf{E}\Big(\sum_{j=0}^{\infty} \phi_1^j u_{t-j}\Big) = \nu \sum_{j=0}^{\infty} \phi_1^j = \nu/(1-\phi_1),$$

$$\mathsf{Var}\left(y_t\right) = \sigma^2 \sum_{j=0}^{\infty} \phi_1^{2j} = \sigma^2 / \left(1-\phi_1^2\right),$$

and, for $h > 0$,

$$\begin{aligned} \gamma_h &= \mathsf{Cov}\left(y_t, y_{t+h}\right) \\ &= \sigma^2 \sum_{j=0}^{\infty} \phi_1^j \phi_1^{j+h} \\ &= \phi_1 \gamma_{h-1} \\ &= \sigma^2 \phi_1^h / \left(1-\phi_1^2\right). \end{aligned}$$

The autocorrelation function of the AR(1) process hence becomes

$$\rho_h = \begin{cases} 1,\ h = 0, \\ \phi_1^h,\ h > 0. \end{cases}$$

Unlike in the case of an MA(1) process, the autocorrelation function differs from zero for all lags (unless $\phi_1 = 0$). Note, however, that the condition $\gamma_h \to 0,$ when $h \to \infty$, is satisfied.

- Later, we will study a generalization of the AR(1) process, called an AR(p) process, which is obtained by adding a linear combination of the variables $y_{t-2}, \cdots, y_{t-p}, (p < \infty)$ on the right hand side of the AR(1) process.

In the AR(1) process, the autoregressive parameter ϕ_1 clearly measures the **persistence** of a random shock to the time series.

- If ϕ_1 is close to unity in absolute value, autocorrelation is high and the series (process) is strongly "persistent".
- If ϕ_1 is close to zero, there is no persistence and the effect of the shock is temporary.

The sign of ϕ_1 determines whether the time series is positively or negatively autocorrelated.

- If ϕ_1 is positive, then the positive values of y_t are tending to follow positive values, and similarly with negative values.
- If ϕ_1 is negative, then positive values tend to follow negative values, and vice versa.

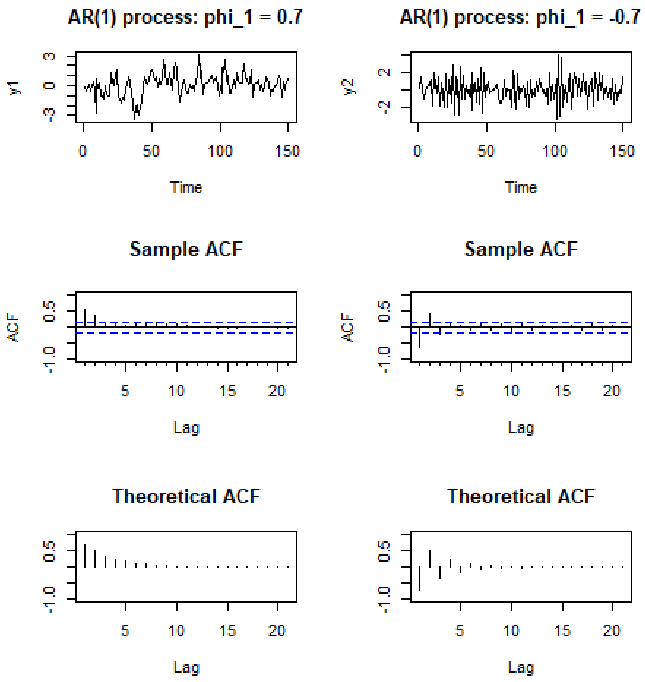


Figure: Two simulated realizations of the AR(1) process and their autocorrelations, starting from the first autocorrelations (r$_1$ and $_1$).

Figure presents two simulated realizations of length 150 of an AR(1) process with $u_t \sim \mathsf{nid}\,(0,1)$. On the left, the AR coefficient ϕ_1 is 0.7, on the right -0.7. For simplicity, the constant term (and hence the mean) is set to 0. In addition to simulated time series, the figure plots sample autocorrelations and theoretical autocorrelations of these two processes.

- As they should be stationary series, the time series vary around their mean (zero) according to their theoretical standard deviation (≈ 1.4).
- When $\phi_1 = 0.7$ (left), the observations have relatively strong positive autocorrelation and, as a consequence, several consecutive observations occur above the mean, as well as below the mean. This gives the time series a "smooth" flavour.

- When $\phi_1 = -0.7$ (right), the sign of autocorrelation between consecutive observations changes depending on the distance between them. These changes from positive to negative of the autocorrelation coefficients give the observed time series a jagged/zigzag pattern. Moreover, clusters of consecutive observations with small absolute values are also observed (the same for large absolute values).

- In both cases, the estimated sample autocorrelation functions are rather close to their theoretical counterparts. Moreover, several estimated autocorrelation coefficients are outside of the 95% confidence bands based on the assumption of $\rho_h = 0$ $(\forall h > 0)$.

Extra: Noncausal AR process

The definition of an AR(1) process is usually based on its model equation as described above. In connection with this, the condition $|\phi_1| < 1$ is often called the stationarity condition of an AR(1) process. This terminology is somewhat misleading because AR(1) equation has a stationary solution also in the case $|\phi_1| > 1$, although this solution cannot be represented in the form of linear process. When $|\phi_1| > 1$, AR(1) equation can be rewritten as

$$y_t = \phi_1^{-1} y_{t+1} - \phi_1^{-1} u_{t+1}$$

by increasing the time index t by one step. Using repetitive substitutions forward in time in a fashion similar as above, this leads to

$$y_t = \phi_1^{-k-1} y_{t+1+k} - \sum_{j=0}^{k} \phi_1^{-j-1} u_{t+1+j}.$$

This leads to the limiting solution

$$y_t = -\sum_{j=1}^{\infty} \phi_1^{-j} u_{t+j}.$$

Note that we can arrive to the (more or less) same (except for an unimportant minus sign) solution by starting from the linear process and by assuming $\psi_j = \phi_1^j$, when $j < 0$, and $\psi_j = 0$, when $j \geq 0$ $(|\phi_1| > 1)$. This solution of the AR(1) equation is called noncausal. This kind of noncausal AR processes have received some attention in the recent literature. However, in this course we restrict our attention to causal AR processes (as do most textbooks).

4.4 Random walk

For the AR(1) model, and for any initial value y_0, we obtain a representation (after recursive substitutions)

$$y_t = \phi_1^t y_0 + \nu \sum_{j=0}^{t-1} \phi_1^j + \sum_{j=0}^{t-1} \phi_1^j u_{t-j}, \quad t = 1, 2, \dots \ .$$

If we assume the initial value y_0 to be independent of the variables u_t, $t \geq 1$, one can use the assumption $u_t \sim \mathsf{iid}\left(0, \sigma^2\right)$ to deduce

$$\mathsf{E}\left(y_t\right) = \phi_1^t \mathsf{E}\left(y_0\right) + \nu \sum_{j=0}^{t-1} \phi_1^j$$

and

$$\begin{aligned} \mathsf{Var}\left(y_t\right) &= \mathsf{E}\Big(y_t - \mathsf{E}(y_t)\Big)^2 \\ &= \mathsf{Var}\left(\phi_1^t y_0\right) + \mathsf{Var}\left(\sum_{j=0}^{t-1} \phi_1^j u_{t-j}\right) \\ &= \phi_1^{2t} \mathsf{Var}\left(y_0\right) + \sigma^2 \sum_{j=0}^{t-1} \phi_1^{2j}. \end{aligned}$$

When $|\phi_1| = 1$, the expected value of y_t, or at least its variance, clearly depends on t regardless of how y_0 (or its distribution) is chosen.

- In this case, the AR(1) process therefore has no stationary solution.

When $\phi_1 = 1$ and $t \geq 1$, the AR(1) process reduces to

$$y_t = \nu + y_{t-1} + u_t, \ u_t \sim \mathsf{iid}\left(0, \sigma^2\right).$$

This is called a **random walk**. This name is due to the “wandering” nature of the realizations of the process.

- The left panel of figure below illustrates this. For simplicity, we assume here $\nu = 0$ (that is the random walk without drift) and $u_t \sim \mathsf{nid}(0, 1)$.
- The right panel illustrates the obvious fact that the differences $y_t - y_{t-1} = u_t$ of a random walk $y_t = y_0 + \sum_{j=0}^{t-1} u_{t-j}$ are stationary.

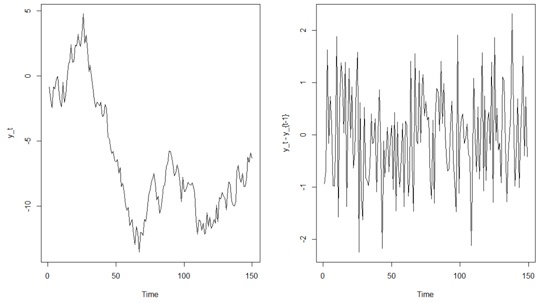


Figure: A simulated realization of the random walk process (assuming $u_t \sim \mathsf{nid}(0,1)$) (left) and its first-difference (right).

The random walk and its generalizations play a central role in the analysis of **nonstationary time series**. We will come back to this in Sections 12–13.

4.5 ARMA(1,1) process

A concept easing the algebraic manipulations of time series processes is the so-called **backshift operator** or **lag operator**. For any process (or simply a sequence of numbers) x_t, define the operation with the equation $Bx_t = x_{t-1}$. More generally, $B^2 x_t = B\left(Bx_t\right) = Bx_{t-1} = x_{t-2}$, and inductively define

$$B^k x_t = B\left(B^{k-1} x_t\right) = x_{t-k},\ B^0 x_t = x_t.$$

- Here k can also be negative, and when $k < 0$, the operator becomes a "forward shift" operator, for example $B^{-1} x_t = x_{t+1}$, $B^{-2} x_t = x_{t+2}$ etc.
- At times backshift or lag operator is denoted by L^k instead of B^k.

Using the lag operator, one can also define polynomials.

- For example, $\theta\left(B\right) = 1 + \theta_1 B$ and, e.g., $\psi\left(B\right) = \sum_{j=-\infty}^{\infty} \psi_j B^j$.

One can algebraically operate with the lag operator exactly as if B were a real or a complex number.

- For instance, the MA(1) process can be written as

$$y_t = \mu + u_t + \theta_1 u_{t-1} = \mu + \theta\left(B\right) u_t.$$

- When differencing (some) process y_t twice, we obtain

$$\begin{aligned}(1-B)^2 y_t &= (1-B)\left[(1-B)\, y_t\right] \\ &= (1-B)\,(y_t - y_{t-1}) \\ &= y_t - y_{t-1} - y_{t-1} + y_{t-2} \\ &= (1-2B+B^2)\, y_t.\end{aligned}$$

Using the lag operator, the general linear process can be defined by the equation

$$y_t - \mu = \psi(B)\, u_t, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right),$$

where $\psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j$. The operator $\psi(B)$ is sometimes thought as a **linear filter**, which transforms the white noise sequence $\{u_t\}$ to the process $\{y_t\}$.

In what follows, we will often consider the special case of the (causal) linear process in which the filter $\psi(B)$ is rational, that is,

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \theta(B)\, \phi(B)^{-1},$$

where $\phi(B)$ and $\theta(B)$ are polynomials of finite order. In the simplest case, these polynomials are of order one so that

$$\phi(B) = 1 - \phi_1 B \quad \text{and} \quad \theta(B) = 1 + \theta_1 B.$$

- It is clear that to obtain stationarity, some restrictions have to be placed on the coefficients of the polynomial $\phi(B)$.

Based on our discussion on the AR(1) process above, it is clear that in the first-order case a sufficient stationary condition is that $|\phi_1| < 1$. Then the condition $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ attached to the linear process is satisfied and the process

$$y_t - \mu = \left[\theta(B)\, \phi(B)^{-1}\right] u_t$$

is well defined. Multiplying both sides of this equation with the polynomial $\phi(B)$, we obtain the representation

$$\phi(B)\,(y_t - \mu) = \theta(B)\, u_t$$

or

$$y_t = \nu + \phi_1 y_{t-1} + u_t + \theta_1 u_{t-1}, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right),$$

where $\nu = \phi(B)\,\mu \equiv \phi(1)\,\mu = (1-\phi_1)\mu$ (see the properties of the backshift operator). A process defined like this is called the **autoregressive moving average process of order one**, or the **ARMA(1,1) process**.

- This combines the AR(1) and MA(1) processes introduced earlier, and these processes can still be obtained as special cases.
- Later we will study a generalization of this process called the ARMA(p, q) process, which can be obtained by generalizing ARMA(1,1) in a similar fashion as discussed around AR(1) and MA(1) processes.
- The coefficients ψ_j of the filter $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ can be solved fairly straightforwardly from equation $\psi(B) = \theta(B)\phi(B)^{-1}$.

The above shows that the autocorrelation function of an ARMA(1,1) process can then be derived by applying the general formulas obtained for the general linear process. Details of these calculations are left as exercises.

4.6 Wold decomposition

The following famous result (named after Herman Wold) shows that every weakly stationary non-deterministic process can be expressed as a sum of a deterministic process and a causal MA(∞) process (the proof of this result is beyond the scope of this course and omitted). In other words, every weakly stationary non-deterministic process y_t $(t = 0, \pm 1, \pm 2, ...)$ has a representation

$$y_t = \sum_{j=0}^{\infty} \psi_j u_{t-j} + \upsilon_t,$$

where

- (i) $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$,
- (ii) $u_t \sim \mathsf{wn}(0, \sigma^2)$,
- (iii) υ_t is deterministic, and
- (iv) $\mathsf{Cov}(u_t, \upsilon_s) = 0$ for all t and s. □

Part (iii) means that the process υ_t can be predicted linearly using the variables $y_{t-1}, y_{t-2}, ...$, with no error. This and part (ii) together imply that u_t can be interpreted as a forecast error when forecasting y_t linearly using the lags of y_t. Modelling υ_t can thus be thought as modelling a trend as discussed in Sections 1–2.

- In the case $\upsilon_t = 0$ the process y_t is called “purely non-deterministic”.

- A simple example of a process υ_t is that it is constant. Such a realization of υ_t can in practice be interpreted as the expected value of process y_t, $\mathsf{E}(y_t) = \mu$, or at least be included in it.

The task of modelling a weakly stationary process reduces to the task of modelling the linear filter $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$.

As a remark, we note that the significance of the Wold decomposition should be evaluated keeping in mind that it concerns weakly stationary processes and linear forecasting. Although every weakly stationary process can be represented by the Wold decomposition, this does not mean that the decomposition is the best way to describe the process. There exist (strictly) stationary processes for which linear prediction is not optimal (in the sense of minimising mean-square forecast error).

Extra: Deterministic and non-deterministic processes/parts

Consider the weakly stationary process

$$y_t = A\cos(\lambda t) + B\sin(\lambda t), \qquad t = 0, \pm 1, \pm 2, \dots,$$

where $\lambda \in [0, \pi)$ is a constant and the random variables A and B satisfy the conditions $\mathsf{E}(A) = \mathsf{E}(B) = 0$, $\mathsf{Var}(A) = \mathsf{Var}(B) = \sigma^2$ and $\mathsf{Cov}(A, B) = 0$. Because

$$y_t + y_{t-2} = A\left[\cos(\lambda t) + \cos(\lambda(t-2))\right] + B\left[\sin(\lambda t) + \sin(\lambda(t-2))\right],$$

we can use the trigonometric identities

$$\sin(x_1) + \sin(x_2) = \sin((x_1 + x_2)2)\cos((x_1 - x_2)2)$$

and

$$\cos(x_1) + \cos(x_2) = 2\cos((x_1 + x_2)2)\cos((x_1 - x_2)2)$$

to derive the result $y_t + y_{t-2} = 2\cos(\lambda)\, y_{t-1}$ or, in other words,

$$y_t = 2\cos(\lambda)\, y_{t-1} - y_{t-2}, \quad t = 0, \pm 1, \pm 2, \dots.$$

The process y_t is somewhat peculiar in that when y_{t-1} and y_{t-2} (and the value of the constant λ) are known, the value for the current period y_t can be predicted using a simple linear formula with perfect precision without any forecast error. A process with such a property is called a *deterministic* process. More generally, the forecast of y_t is allowed to be any linear function of the past values of the process $y_{t-1}, y_{t-2}, \dots$ so that the forecast is a linear combination of the variables $y_{t-1}, \dots, y_{t-n}$ or a mean square limit of such linear combinations as $n \to \infty$. If a process is not deterministic, then it is called *non-deterministic*.

4.7 Properties of sample mean and autocorrelations

Because the sample mean and the sample autocorrelation function are central tools in the analysis of time series, we next briefly discuss some of their statistical properties.

We can show that (see below) that the **sample mean**

$$\bar{y} = \frac{1}{T}\left(y_1 + \cdots + y_T\right)$$

is

- **unbiased** and **consistent estimator** of the population mean $\mu = \mathsf{E}\left(y_t\right)$, and
- **asymptotically normally distributed**

Extra: Proof of unbiasedness and consistency of the sample mean

In the following discussion, we assume both weak and strict stationarity of the process.

For the sample mean $\bar{y} = T^{-1}\left(y_1 + \cdots + y_T\right)$, it holds that

$$\mathsf{E}\left(\bar{y}\right) = \frac{1}{T}\left(\mathsf{E}\left(y_1\right) + \cdots + \mathsf{E}\left(y_T\right)\right) = \mu$$

so that it is an *unbiased estimator* of the population mean $\mu = \mathsf{E}\left(y_t\right)$ (recall that in general, an estimator $\widehat{\theta}$ of a population parameter θ is called unbiased if $\mathsf{E}(\widehat{\theta}) = \theta$).

For the variance of $\bar{y}$, we can construct

$$\begin{aligned}
\mathsf{Var}\left(\bar{y}\right) &= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathsf{Cov}\left(y_t, y_s\right) \\
&= \frac{1}{T^2}\sum_{t-s=-T}^{T}\left(T - |t-s|\right)\gamma_{t-s} \\
&= \frac{1}{T}\sum_{h=-T}^{T}\left(1 - \frac{|h|}{T}\right)\gamma_h.
\end{aligned}$$

The second equality above can be justified by noting that the preceding double sum is the sum of the elements of the matrix $\left[\gamma_{t-s}\right]_{t,s=1,..,T}$. Now assume that

$$\sum_{h=-\infty}^{\infty} |\gamma_h| < \infty,$$

a requirement more stringent than condition $\gamma_h \to 0$, when $h \to \infty$. This condition is satisfied by many processes used in practice (for instance, by the AR(1) and MA(1) processes). Making use of the triangle inequality, we now obtain the result

$$\mathsf{Var}\left(\bar{y}\right) = \mathsf{E}\left(\bar{y}-\mu\right)^2 \leq \frac{1}{T}\sum_{h=-T}^{T}\left(1-\frac{|h|}{T}\right)|\gamma_h| \to 0, \quad \text{as} \quad T \to \infty.$$

In other words, the sample mean is a consistent estimator for the expected value.

If we strengthen the assumptions made above and additionally assume that the process y_t is Gaussian, the sample mean is also normally distributed with the mean and variance as indicated above. Furthermore, the following asymptotic result can be established (we omit the details)

$$\sqrt{T}\left(\bar{y}-\mu\right) \overset{d}{\to} \mathsf{N}\left(0, \textstyle\sum_{h=-\infty}^{\infty}\gamma_h\right) \quad \text{or} \quad \bar{y} \underset{as}{\sim} \mathsf{N}\left(\mu, \frac{1}{T}\textstyle\sum_{h=-\infty}^{\infty}\gamma_h\right).$$

This result can be derived without assuming y_t to be Gaussian.

The purpose here is to show the law of large numbers and the central limit theorem to apply to stationary processes with “reasonable assumptions”. To use this result to construct tests and confidence intervals for μ, we also need to estimate the infinite sum $\sum_{h=-\infty}^{\infty}\gamma_h$. A suitable estimator is (compare the expression of $\mathsf{Var}\left(\bar{y}\right)$ above) $\sum_{h=-K}^{K}\left(1-|h|/T\right)\mathsf{c}_h$, where c_h is the sample autocovariance coefficient defined earlier, and K is a “suitably” chosen number smaller than T (for example, $K \approx \sqrt{T}$).For large values of h, the estimator c_h becomes unprecise, and for this reason K should not be too large compared to T.

The statistical properties of the sample autocorrelation coefficients $\mathsf{r}_h = \mathsf{c}_h/\mathsf{c}_0$ are more complicated to derive than those of the sample mean, so we will not attempt to provide any detailed justifications. Under “reasonably general assumptions”, consistency and asymptotic normality of them can be established.

Important special case: For the important special case of $y_t \sim \mathsf{iid}\left(\mu,\sigma^2\right)$, it holds that

$$\left(\mathsf{r}_1, \ldots, \mathsf{r}_H\right) \underset{as}{\sim} \mathsf{N}\left(0, T^{-1}I_H\right),$$

with a H-dimensional zero mean vector and I_H $(H \times H)$ denotes an identity matrix. This result can be used to test whether it is realistic to consider an observed time series as an uncorrelated time series process. Under the hypothesis to be tested (uncorrelatedness), the estimators $\mathsf{r}_1, \ldots, \mathsf{r}_H$ are approximately independent with distribution $\mathsf{N}\left(0, T^{-1}\right)$. Based on this, we get

$$\mathsf{P}(|\mathsf{r}_h| \geq 1.96/\sqrt{T}) \approx 0.05,$$

a result that can be used to evaluate the significance of individual sample autocorrelation coefficients.

To obtain a joint test for several autocorrelation coefficients, that is to test the null hypothesis of no autocorrelation $H_0 = \rho_1 = \cdots = \rho_H = 0$, one can use the test statistic

$$Q = T \sum_{h=1}^{H} \mathsf{r}_h^2 \underset{as}{\sim} \chi_H^2,$$

whose large values would lead to rejection. Note that the asymptotic χ_H^2–distribution follows from the distribution result above for $(\mathsf{r}_1, \ldots, \mathsf{r}_H)$ and the definition of the chi-squared distribution. In practice, an alternative and slightly different test statistic

$$Q_{LB} = T\,(T+2) \sum_{h=1}^{H} \mathsf{r}_h^2 / \left(T - h\right) \underset{as}{\sim} \chi_H^2,$$

called the **Ljung-Box test statistic** is preferred because in small samples its distribution has been found to be closer to the χ_H^2–distribution than that of the test statistic Q. It should be clear that both tests need H not to be too large compared to T to work well.

The autocorrelation function can be used to reveal linear dependences between the observations, but not nonlinear ones (with the exception of Gaussian processes). To investigate the presence of potential nonlinear dependence over time, one (somewhat limited) approach is to test for autocorrelation in the squared observations.

- Assuming $y_t \sim \mathsf{iid}\left(\mu, \sigma^2\right)$ and $\mathsf{E}\left(y_t^4\right) < \infty$, what was said above about the sample autocorrelations also holds for sample autocorrelations computed from the squared observations y_t^2. In particular, the asymptotic result(s) remain valid when the autocorrelations are computed from squared observations, and also the Ljung-Box test has its indicated asymptotic distribution. In this context, however, the test is usually called the **McLeod-Li test**.
- Investigating the autocorrelations between squared observations is of great interest, especially in the case of financial time series, which themselves often are uncorrelated. We will return to this point later.

Empirical example (continued). As an empirical illustration, let us consider the quarterly U.S. real GDP growth rate. It turns out that we obtain the following results from the Ljung-Box test for different lag lengths H:

```
| Lag (H) |  Q_{LB}  | p-value  |
|---------|----------|----------|
|    4    |  16.487  |  0.002   |
|    8    |  19.494  |  0.012   |
|    12   |  27.942  |  0.006   |
|    16   |  32.897  |  0.008   |
|    20   |  34.679  |  0.022   |
```

These results clearly point out statistically significant autocorrelation (at the 5 % significance level) in the real GDP growth rate, which was apparent already based on the estimated autocorrelation coefficients. Furthermore, for the McLeod-Li tests, the resulting p-values are high (around 0.5 or higher) for all the lag length selections.

Chapter 5

ARMA processes

5.1 AR(p) process

Earlier, we considered the AR(1) process. This can be generalized to the **AR(p) process**

$$y_t = \nu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right),$$

which can be rewritten

$$\phi\left(B\right)\left(y_t - \mu\right) = u_t,$$

with the lag-polynomial $\phi\left(B\right) = 1 - \phi_1 B - \cdots - \phi_p B^p$. The current value of the process hence depends linearly on p past values of the process and the constant term, as well as on an unobserved random shock (or error term or "innovation", depending on the perspective and context).

- Notice again that $\phi\left(B\right)\mu \equiv \phi\left(1\right)\mu \equiv \nu$.

A sufficient condition for the stationarity of an AR(p) process. A sufficient condition for the stationarity of the AR(p) process is that the roots of the polynomial $\phi\left(z\right)$ $\left(z \in \mathbb{C}\right)$ lie outside the unit circle/disk in the complex plane, or equivalently that

$$\phi\left(z\right) \neq 0 \text{ for } |z| \leq 1.$$

- As an example, in the stationary AR(1) case ($|\phi_1| < 1$), the characteristic equation yields $\phi(z) = 0 \Leftrightarrow 1 - \phi_1 z = 0$ when $|z| > 1$. This shows why in the case $p = 1$, the condition $|\phi_1| < 1$ was found sufficient to ensure the existence of a (causal) stationary MA(∞) representation.

- Notice that the resulting root (roots), when $p > 1$, can also be complex numbers. Related to the above stationarity condition, the absolute value or norm of a complex number $z = x + iy$ $\left(i = \sqrt{-1}\right)$ is defined as $|z| = \sqrt{x^2 + y^2}$. The latter form of the stationarity condition can be further expressed as $\phi(z) = 0 \Rightarrow |z| > 1$.
- Still about possible complex roots: Because the coefficients of the polynomial $\phi(z)$ are real numbers, the potential complex roots always appear as conjugate roots, that is, if $\zeta = x + iy$ is a root, then also $\bar{\zeta} = x - iy$ is a root.

Extra: More details on stationarity condition

Still more about the possibility of getting complex roots. The absolute value or norm of a complex number $z = x + iy$ $\left(i = \sqrt{-1}\right)$ can be identified with the norm of the vector (x, y).

One way to illustrate this stationarity condition makes use of a well-known result in mathematics called the fundamental theorem of algebra. As a consequence of this result, the polynomial $\phi(z)$ can be written as (assuming that $\phi_p \neq 0$)

$$\phi(z) = \left(1 - \zeta_1^{-1} z\right) \cdots \left(1 - \zeta_p^{-1} z\right),$$

where for the roots ζ_i it therefore holds that $\phi(\zeta_i) = 0$ and $|\zeta_i| > 1$ $(i = 1, \dots, p)$. Therefore, an AR(p) process can be expressed as $\left(1 - \zeta_1^{-1} B\right) \cdots \left(1 - \zeta_p^{-1} B\right) y_t = u_t$. If all the roots of $\phi(z)$ are real, this equation can be divided with the polynomials $\left(1 - \zeta_i^{-1} B\right)$, $i = 1, \dots, p$, one at a time, and in this way it can be seen (similarly as in the case $p = 1$) that the resulting expression is a well-defined linear process. This procedure can also be generalized to the case of complex roots.

Based on the above-mentioned, a stationary AR(p) process has an MA(∞) representation

$$(y_t - \mu) = \phi(B)^{-1} u_t = \psi(B) u_t = \sum_{j=0}^{\infty} \psi_j u_{t-j},$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \phi(B)^{-1}$. The coefficients ψ_j can be solved as a function of the parameters $\phi_1, \dots, \phi_p$ from equation

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)\left(\psi_0 + \psi_1 B + \psi_2 B^2 + \cdots\right) = 1$$

by interpreting the right hand side as a power series in B, and setting the coefficients of B^j equal to each other on both sides of the equation.

- Recall that two polynomials are the same if their coefficients are the same.
- For instance, the coefficient of B^0 is 1, so that $\psi_0 = 1$. Next, the coefficient of B^1 is zero, so that $\psi_1 - \phi_1\psi_0 = 0$ from which $\psi_1 = \phi_1$ follows. The general solution is left as an exercise.

Autocorrelation function. The autocorrelation function of an AR(p) process could be derived by making use of its MA(∞) representation. Instead, we present an often used and rather practical alternative approach based on the AR(p) model equation.

Multiplying both sides of the demeaned AR(p) process presentation with $y_{t-h}-\mu$ ($h \geq 0$) and taking expectations, we obtain

$$\mathsf{E}\Big((y_t-\mu)(y_{t-h}-\mu)\Big) = \phi_1\mathsf{E}\Big((y_{t-1}-\mu)(y_{t-h}-\mu)\Big)+\cdots+\phi_p\mathsf{E}\Big((y_{t-p}-\mu)(y_{t-h}-\mu)\Big)+\mathsf{E}\Big(u_t(y_{t-h}-\mu)\Big).$$

Because y_{t-h} (and likewise $y_{t-h}-\mu$) is a linear function of the innovation terms u_{t-h}, $u_{t-h-1}, \ldots$, the variables y_{t-h} and u_t are independent when $h > 0$. Therefore, we get

- $\mathsf{E}\Big(u_t(y_{t-h} - \mu)\Big) = 0, \quad h > 0,$
- For $h = 0$, it can be seen that $\mathsf{E}\Big(u_t(y_t - \mu)\Big) = \mathsf{E}\left(u_t^2\right) = \sigma^2$.

As $\gamma_h = \gamma_{-h}$, we get

$$\gamma_h = \begin{cases} \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \sigma^2, & h = 0 \\ \phi_1\gamma_{h-1} + \cdots + \phi_p\gamma_{h-p}, & h > 0. \end{cases}$$

Dividing this (in the case $h > 0$) with the variance γ_0 leads to the autocorrelation function of an AR(p) process

$$\rho_h = \phi_1\rho_{h-1} + \cdots + \phi_p\rho_{h-p}, \quad h > 0,$$

so that it satisfies a difference equation similar to the one the AR(p) process itself satisfies.

- When the roots of $\phi(z)$ lie outside the unit disk on the complex plane, the solution ρ_h to this difference equation decays exponentially to zero as the lag length h increases.
- The solution ϕ_1^h for the AR(1) is an example of this (this solution can also be easily obtained by solving the difference equation above in the case $p = 1$ using the initial value $\rho_0 = 1$).

Partial autocorrelation function. We next define the partial autocorrelation function, which is, among other things, a useful tool for model selection. In general, α_h equals the conventional partial correlation coefficient that measures the correlation between the random variables y_t and y_{t-h} when the linear effect of the random variables $y_{t-1}, \dots, y_{t-h+1}$ has been first eliminated.

- Therefore, a particular consequence is that $|\alpha_h| \leq 1$.

In the case of an AR(p) process, the **partial autocorrelation function α_h has a special useful feature**. For $m > p$, an AR(p) process can be interpreted as an AR(m) process with $\phi_{p+1} = \cdots = \phi_m = 0$, which makes it clear that the partial autocorrelation function of an AR(p) process satisfies (also $\alpha_0 = 1$):

$$y_t \sim \mathrm{AR}(p) \Rightarrow \alpha_p = \phi_p \quad \text{and} \quad \alpha_h = 0 \text{ for } h > p.$$

In other words, **the partial autocorrelation function of an AR(p) process drops to zero after lag p** (assuming $\phi_p \neq 0$).

- The sample counterpart of the (population) partial autocorrelation function is obtained by using sample autocovariances c_h.
- The (sample) partial autocorrelation function is typically estimated using recursive algorithms or regression techniques, which are closely related to the Yule-Walker framework (see details in the Extra section below).

An important detail for model selection is that if an observed time series has been generated by an AR(p) process, its estimated autocorrelation function should decay to zero as the lag length increases **with no apparent breaks**, whereas the estimated partial autocorrelation function should have **a visible break after lag** p.

To identify the location of the break point in the estimated partial autocorrelation function, one can make use of the following result:

- In the case of an AR(p) process, the estimators $\hat{\alpha}_h$, $h > p$, are approximately independent with $\mathsf{N}\left(0, T^{-1}\right)$–distribution. Therefore,

  $$\mathsf{P}(|\hat{\alpha}_h| \geq 1.96/\sqrt{T}) \approx 0.05$$

  when $h > p$.
- On the other hand, because the estimators $\mathsf{c}_1, \dots, \mathsf{c}_p$ are consistent, the estimator $\hat{\alpha}_p$ converges in probability to the value of the theoretical partial autocorrelation coefficient α_p which for an AR(p) process is nonzero. Therefore, for $h = p$, $\mathsf{P}(|\hat{\alpha}_h| \geq 1.96/\sqrt{T})$ approaches one as the sample size increases.

- These remarks justify depicting the sample partial autocorrelation coefficients similar to that of the usual autocorrelation coefficients, and adding to it horizontal critical/confidence bands to make its interpretation easier.

Extra: Yule-Walker equations and autocorrelation and partial autocorrelation coefficients

Choosing $h = 1, \dots, p$ in the AR(p) autocorrelation function leads to the equations ($\rho_0 = 1$ and $\rho_h = \rho_{-h}$)

$$\begin{aligned}\rho_1 &= \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1}\\ \rho_2 &= \phi_1\rho_1 + \phi_2 + \cdots + \phi_p\rho_{p-2}\\ &\vdots\\ \rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p,\end{aligned}$$

which collectively are called the **Yule-Walker equations**. In matrix notation, these can be expressed as

$$\rho = P\phi,$$

where $\rho = \begin{bmatrix}\rho_1 & \cdots & \rho_p\end{bmatrix}'$, $\phi = \begin{bmatrix}\phi_1 & \cdots & \phi_p\end{bmatrix}'$ and $P = \begin{bmatrix}\rho_{i-j}\end{bmatrix}_{i,j=1,\dots,p}$ is a $p \times p$ matrix, whose row i and column j element equals ρ_{i-j}. This makes it clear that the parameter vector ϕ can be expressed as a function of the autocovariance or autocorrelation coefficients. The result is Yule-Walker equation

$$\phi = P^{-1}\rho = \Gamma^{-1}\gamma,$$

where $\gamma = \begin{bmatrix}\gamma_1 & \cdots & \gamma_p\end{bmatrix}' = \gamma_0\rho$ and $\Gamma = \begin{bmatrix}\gamma_{i-j}\end{bmatrix}_{i,j=1,\dots,p} = \gamma_0 P$. From the equations derived above it follows that also the parameter σ^2 can be expressed as a function of the autocovariance coefficients and the parameters $\phi_1, \dots, \phi_p$ as

$$\sigma^2 = \gamma_0 - \phi_1\gamma_1 - \cdots - \phi_p\gamma_p.$$

A note on P^{-1}: It is rather clear that this inverse matrix exists here. If not, there would need to be an exact linear relationship between the variables $y_{t-1}, \dots, y_{t-p}$, which (except for the case $\sigma^2 = 0$) is not possible. The sufficient condition for stationarity of an AR(p) process guarantees that P is invertible, although seeing this is somewhat technical.

For further preciseness, let us denote $\gamma = \gamma_p$ and $\Gamma = \Gamma_p$ in Yule-Walker equation. Now, the partial autocorrelation function with lag h is defined as

$$\alpha_h = \begin{cases} 1, \text{ for } h = 0\\ \text{the last component of the vector } \Gamma_h^{-1}\gamma_h, \text{ when } h \geq 1.\end{cases}$$

Because the second and third expressions of Yule-Walker equation can be defined for any weakly stationary (non-deterministic) process, so can also the partial autocorrelation function.

The sample counterpart of the (population) partial autocorrelation function is straightforward to define, simply estimating the vector γ_h and matrix Γ_h using the obvious estimators $c_h = \left[\mathsf{c}_1 \ \cdots \ \mathsf{c}_h\right]'$ and $C_h = \left[\mathsf{c}_{i-j}\right]_{i,j=1,\dots,h}$. Therefore, the sample partial autocorrelation coefficient $\hat{\alpha}_h$ equals 1 for $h = 1$, and the last component of the Yule-Walker estimate $\hat{\phi}_{YW} = C_h^{-1} c_h$ of the parameter ϕ for $h \geq 1$. As a remark, we note that there exist recursive formulas that can be used to compute both theoretical and sample partial autocorrelation coefficients that avoid computing the inverse of (the potentially large) matrix C_h, but we omit the details of this.

Empirical example. As an example, below we present the estimated sample partial autocorrelation functions (with lag lengths $h = 0, \dots, 20$) of the quarterly U.S. real GDP growth series, together with the above-mentioned critical/confidence bands.

- The corresponding sample autocorrelation coefficients are shown already above.
- The first two sample partial autocorrelation coefficients $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are clearly outside the confidence bands, whereas the rest lie within these bands (except the 11th lag), pointing towards an AR(2) process. The sample autocorrelation function partly supports this conclusion, while also MA(2) or some ARMA model might clearly be potential candidate for the GDP growth.
- We are coming back to empirical model selection later in this material.

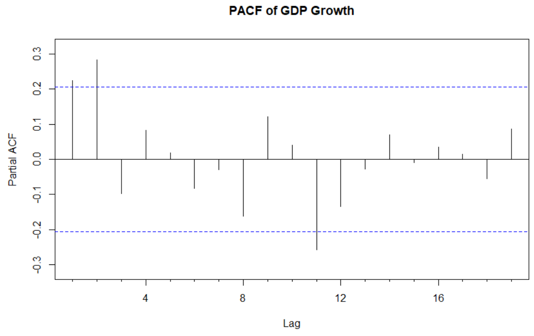


Figure: Sample partial autocorrelation function (PACF) of the quarterly U.S. real GDP growth rate.

5.2 MA(q) process and invertibility

The MA(q) process is defined as

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}, \quad u_t \sim \mathsf{iid}\left(0, \sigma^2\right),$$

or alternatively using the lag operator as

$$\left(y_t - \mu\right) = \theta\left(B\right) u_t, \; u_t \sim \mathsf{iid}\left(0, \sigma^2\right)$$

with the lag-polynomial $\theta\left(B\right) = 1 + \theta_1 B + \cdots + \theta_q B^q$. Therefore, the current value of the process is assumed to depend linearly on the present and last q unobservable random shocks. Because an MA(q) process is a special case of the (causal) linear process, it is always both weakly and strictly stationary. Moreover, we get

- $\mathsf{E}\left(y_t\right) = \mu$,
- $\mathsf{Var}\left(y_t\right) = \gamma_0 = \left(1 + \theta_1^2 + \cdots + \theta_q^2\right)\sigma^2$.

Autocorrelation function. By making use of the general results of linear process, the MA(q) process has the autocovariance function given by

$$\gamma_h = \begin{cases} \sigma^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & \text{for } \; 0 \le h \le q, \\ 0, & \text{for } \; h > q, \end{cases}$$

where $\theta_0 = 1$. The autocorrelation function is then obtained via the formula $\rho_h = \gamma_h / \gamma_0$.

- Therefore, the autocorrelation function of an MA(q) process drops to zero after lag q (assuming that $\theta_q \neq 0$).
- If one observes a similar feature in a sample autocorrelation function, one can consider an MA(q) process as a good candidate to model the time series.

When considering the suitability of an MA(q) process, one can make use of the following result: In the case of an MA(q) process, the estimators r_h, $h > q$, are approximately normally distributed with mean zero and variance $\left(1 + 2\rho_1^2 + \cdots + 2\rho_q^2\right)/T$. Therefore, if one wants to test whether an individual autocorrelation coefficient is statistically different from zero at the 5% significance level, the estimates r_h, $h > q$, should be compared with the critical bounds

$$\pm 1.96\sqrt{\hat{\mathsf{w}}_q/T}, \quad \hat{\mathsf{w}}_q = \left(1 + 2\mathsf{r}_1^2 + \cdots + 2\mathsf{r}_q^2\right).$$

In this case, $\mathsf{P}(|\mathsf{r}_h| \geq 1.96\sqrt{\hat{\mathsf{w}}_q/T}) \approx 0.05$. Note also that the bounds above are wider than in the case $y_t \sim \mathsf{iid}\left(0, \sigma^2\right)$ when $\hat{\mathsf{w}}_q$ is replaced by 1.

Invertibility. It can be shown that in a (causal) AR(p) process there exists a one to one correspondence between the parameters $\phi = (\phi_1, \ldots, \phi_p)$ and σ^2, and the autocovariance function. For an MA(q) process, a similar result does not hold.

- To see this, consider as an example the MA(1) process for which it holds that (assume that $\theta_1 \neq 0$)

  $$\gamma_0 = \sigma^2\left(1 + \theta_1^2\right) = \sigma^2\theta_1^2\left(1 + 1/\theta_1^2\right) \text{ and } \gamma_1 = \sigma^2\theta_1 = \theta_1^2\sigma^2\left(1/\theta_1\right).$$

  Define $\theta_1^* = 1/\theta_1$ and $\sigma_*^2 = \theta_1^2\sigma^2$, so that $u_t^* = \theta_1 u_t \sim \mathsf{iid}\left(0, \sigma_*^2\right)$. Now, it can be seen that the two MA(1) processes $y_t = u_t + \theta_1 u_{t-1}$ and $y_t^* = u_t^* + \theta_1^* u_{t-1}^*$ have exactly the same autocovariance function, so that they cannot be distinguished from each other based on autocovariances.

- If in addition $u_t \sim \mathsf{nid}\left(0, \sigma^2\right)$, the entire probability structures of these two processes are indistinguishable.

- A consequence is that estimating the parameters based on an observed time series, unless one "tells" the method, which of the parameter combinations $\left(\theta_1, \sigma^2\right)$ and $\left(\theta_1^*, \sigma_*^2\right)$ (that fit the data equally well) should be chosen.

The typical solution to the above problem is to set the condition $|\theta_1| < 1$. When one assumes $|\theta_1| < 1$ in an MA(1) process, one can use the equation $u_t = y_t - \mu - \theta_1 u_{t-1}$ and repetitive substitutions to first obtain $u_t = (y_t - \mu) - \theta_1(y_{t-1} - \mu) + \theta_1^2 u_{t-2}$, and eventually

$$(y_t - \mu) = -\sum_{j=1}^{k}\left(-\theta_1\right)^j\left(y_{t-j} - \mu\right) - \left(-\theta_1\right)^{k+1} u_{t-k-1} + u_t.$$

Similarly as when considering the stationarity of an AR(1) process, it can be concluded, given the condition $|\theta_1| < 1$,

$$y_t - \mu = -\sum_{j=1}^{\infty}\left(-\theta_1\right)^j\left(y_{t-j} - \mu\right) + u_t.$$

That is, when $|\theta_1| < 1$ holds, an MA(1) process can be "inverted" to an AR(∞) process, which explains why in the case $|\theta_1| < 1$ an MA(1) process is called **invertible**.

Invertibility condition of an MA(q) process . An MA(q) process is invertible, if the roots of the polynomial $\theta(z)$ ($z \in \mathbb{C}$) lie outside the unit circle/disk on the complex plane or, equivalently, if

$$\theta(z) \neq 0 \text{ for } |z| \leq 1.$$

Similarly as in the case of the stationarity condition of an AR(p) process, invertibility ensures that there exists an AR(∞) representation also for MA(q) processes. In other words, when the invertibility condition holds, one can formally solve u_t from equation $(y_t - \mu) = \theta(B) u_t$ by dividing both sides by the polynomial $\theta(B)^{-1}$. The solution is

$$\pi(B)(y_t - \mu) = u_t \text{ or } \sum_{j=0}^{\infty} \pi_j (y_{t-j} - \mu) = u_t,$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j = \theta(B)^{-1}$ and the coefficients π_j can be solved as a function of the parameters $\theta_1, \dots, \theta_q$ from the equation

$$\left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\left(\pi_0 + \pi_1 B + \pi_2 B^2 + \cdots\right) = 1$$

by interpreting the right hand side as a power series in B, setting the coefficients of B^j equal on both sides of the equation (and $\pi_0 = 1$).

As in the MA(1) case, invertibility ensures that there exists a one to one correspondence between the autocovariance function of an MA(q) process and the parameters $\theta = \left(\theta_1, \dots, \theta_q\right)$ and σ^2.

- Except for the first-order case, this correspondence is rather complicated, and we omit the details.
- In what follows (unless otherwise mentioned), MA processes are always assumed to be invertible.

Partial autocorrelation function. The general definition of a partial autocorrelation function presented in the previous subsection can also be applied in the case of an MA(q) process, but the calculations involved would become rather complicated. However, because an MA(q) process has (due to invertibility and assuming here $\theta_q \neq 0$) an AR(∞) representation, and based on what was said above about the partial autocorrelation function of an AR(p) process, it is intuitively clear that the partial autocorrelation function of an MA(q) process does not drop to zero at any point, but rather smoothly decays towards zero.

- It can be shown that the the partial autocorrelation function of an MA(1) process is
$$\alpha_h = -\left(-\theta_1\right)^h / \left(1 + \theta_1^2 + \cdots + \theta_1^{2h}\right).$$
Because $|\theta_1| < 1$, the partial autocorrelation function of an MA(1) process decays to zero at an exponential rate as the lag length h increases.

- For a general MA(q) process, it can be shown that the partial autocorrelation function decays to zero at an exponential rate, potentially following the shape of a dampening sine wave.

5.3 ARMA(p, q) process

An ARMA(p,q) process can be characterized as a combination of AR(p) and MA(q) processes. It is defined by the equation

$$y_t = \nu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q},$$

where $u_t \sim \mathsf{iid}\left(0, \sigma^2\right)$. Similarly as in the case of an AR(p) process, the current value y_t is assumed to depend linearly on p past values of the process. Unlike the AR(p) process, the error term is now (generally) not an independent random shock, but instead an autocorrelated MA(q) process.

Using the lag operator and the lag-polynomials $\phi\left(B\right) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta\left(B\right) = 1 + \theta_1 B + \cdots + \theta_q B^q$, the ARMA($p, q$) process can be expressed as

$$\phi\left(B\right)\left(y_t - \mu\right) = \theta\left(B\right) u_t.$$

- If $\theta\left(B\right) = 1$ (that is, $q = 0$), one obtains the AR(p) process as a special case.

- If $\phi\left(B\right) = 1$ (that is, $p = 0$), one obtains the MA(q) process.

Stationarity and invertibility. Based on what was said for linear processes, it is clear that an ARMA(p,q) process has a well-defined (causal) MA(∞) representation if the polynomial $\phi\left(z\right)$ satisfies the sufficient stationarity condition of an AR(p) process (what was said in the case of an AR(p) process holds, but now with MA(q) error terms).

- A sufficient condition for the stationarity of the ARMA(p,q) process is that the roots of the polynomial $\phi(z)$ ($z \in \mathbb{C}$) lie outside the unit circle/disk on the complex plane, or equivalently that

$$\phi(z) \neq 0 \text{ for } |z| \leq 1.$$

- An ARMA(p,q) process is invertible, if the roots of the polynomial $\theta(z)$ ($z \in \mathbb{C}$) lie outside the unit circle/disk on the complex plane, or equivalently if

$$\theta(z) \neq 0 \text{ for } |z| \leq 1.$$

  The general discussion concerning invertibility in the previous subsection in connection with the MA(1) process also holds here and generalizes to the case of an ARMA(p,q) process.

As can be deduced from what was discussed above, when the stationarity condition holds, an ARMA(p,q) process has an MA(∞) representation (cf. the AR(p) case)

$$(y_t - \mu) = \frac{\theta(B)}{\phi(B)} u_t = \psi(B) u_t,$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \theta(B) \phi(B)^{-1}$ and the coefficients ψ_j can be solved as a function of the parameters $\phi_1, \dots, \phi_p$ and $\theta_1, \dots, \theta_q$ by equating the coefficients of B^j on both sides of the equation

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)\left(\psi_0 + \psi_1 B + \psi_2 B^2 + \cdots\right) = 1 + \theta_1 B + \cdots + \theta_p B^q.$$

- From this equation, it can be solved that $\psi_0 = 1$, $\theta_1 = \psi_1 - \psi_0 \phi_1$, and in general,

$$\psi_j = \sum_{i=1}^{p} \phi_i \psi_{j-i} + \theta_j, \; j = 0, 1, 2, \dots,$$

  where $\theta_0 = 1$, $\theta_j = 0$, $j > q$, and $\psi_j = 0$, $j < 0$.

On the other hand, when the invertibility condition holds, an ARMA(p,q) process has an AR(∞) representation

$$\frac{\phi(B)}{\theta(B)} (y_t - \mu) = \pi(B) (y_t - \mu) = u_t,$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j = \phi(B)\,\theta(B)^{-1}$. The coefficients π_j can be solved similarly as the ψ_j's above (just exchange the roles of the polynomials $\phi(B)$ and $\theta(B)$). The end result can be expressed as

$$\pi_j = -\sum_{i=1}^{p} \theta_i \pi_{j-i} - \phi_j, \ j = 0, 1, 2, \dots,$$

where $\phi_0 = -1$, $\phi_j = 0$, $j > p$, and $\pi_j = 0$, $j < 0$. In particular, it holds that $\pi_0 = 1$ and, as in the case of the coefficients ψ_j, $\pi_j \to 0$ at an exponential rate as $j \to \infty$.

Identification condition. Earlier in connection to MA processes, the invertibility condition was motivated by the fact that it guarantees the existence of a one-to-one correspondence between the autocovariance function of a process and its model parameters. For general ARMA(p,q) processes, invertibility alone is not sufficient to guarantee the existence of such a correspondence, which is also required for maximum likelihood estimation of the parameters.

- Consider as an example the simple case of an ARMA(1,1) process with the linear representation

  $$(y_t - \mu) = \frac{1 + \theta_1 B}{1 - \phi_1 B} u_t.$$

  In the special case $\phi_1 = -\theta_1$, it is clear that the polynomials on the right hand side can be cancelled out, resulting in the equation $y_t - \mu = u_t$ so that the process is not a "real" ARMA(1,1) process, but instead just white noise. This implies that the parameters ϕ_1 and θ_1 are not identified in the sense that maximum likelihood estimation will not work because this method cannot distinguish different pairs of parameter values that satisfy the restriction $\phi_1 = -\theta_1$ (and $|\phi_1| < 1$, $|\theta_1| < 1$).

Identification (or uniqueness) condition of an ARMA(p,q) process. The polynomials $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ of a stationary and invertible ARMA(p,q) process are assumed to have no common roots and $\phi_p \neq 0$ or $\theta_q \neq 0$.

- Quite often, the requirement that $\phi_p \neq 0$ or $\theta_q \neq 0$ is not explicitly mentioned but is understood to be contained in the condition ruling out common roots.

- Unless otherwise mentioned, in what follows, we assume that this identification condition holds.

Autocorrelation function of an ARMA(p,q) process can be derived in a manner similar to the AR(p) case.

- In what follows, we only present the general principle of the solution.

Multiplying both sides of the ARMA(p, q) model equation with y_{t-h} $(h \geq 0)$ and taking expectations yields

$$\begin{aligned}\mathsf{E}\Big((y_t-\mu)(y_{t-h}-\mu)\Big) &= \phi_1\mathsf{E}\Big(y_{t-1}(y_{t-h}-\mu)\Big)+\cdots+\phi_p\mathsf{E}\Big((y_{t-p}-\mu)(y_{t-h}-\mu)\Big)+\mathsf{E}\Big(u_t(y_{t-h}-\mu)\Big)\\ &\quad+\theta_1\mathsf{E}\Big(u_{t-1}(y_{t-h}-\mu)\Big)+\cdots+\theta_q\mathsf{E}\Big(u_{t-q}(y_{t-h}-\mu)\Big).\end{aligned}$$

Making use of the MA(∞) representation and the definition of autocovariance this further leads to

$$\gamma_h = \begin{cases}\phi_1\gamma_{h-1}+\cdots+\phi_p\gamma_{h-p}+\sigma^2\sum_{j=0}^{\infty}\theta_{h+j}\psi_j, & 0 \leq h < \max\{p, q+1\},\\ \phi_1\gamma_{h-1}+\cdots+\phi_p\gamma_{h-p}, & h \geq \max\{p, q+1\}\end{cases}$$

where $\theta_0 = 1$ and $\theta_j = 0$ for $j \notin \{0, \dots, q\}$. The autocorrelation function is then obtained using the formula $\rho_h = \gamma_h/\gamma_0$. Moreover, the coefficients ψ_j can be expressed as function of the parameters $\phi_1, \dots, \phi_p$ and $\theta_1, \dots, \theta_q$.

- Because this recursive solution is obtained from a difference equation similar to the one the autocovariances of an AR(p) process satisfy, it is rather clear that as the lag length h increases, the autocorrelation function of an ARMA(p, q) process decays exponentially to zero, potentially following the shape of a dampening sine wave (in the case $p = 1$ this is rather easy to verify).
- Because an ARMA(p, q) process has an AR(∞) representation when the invertibility condition holds, one can use the general definition of the partial autocorrelation function to conclude that the general shape of the partial autocorrelation function of an ARMA(p, q) process is comparable to that of the autocorrelation function.

To conclude and importantly, neither the autocorrelation nor the partial autocorrelation function of an ARMA(p, q) process ever drop to zero but instead both steadily decay towards zero.

Chapter 6

Parameter estimation

In this section, we will consider how to estimate (unknown) values of the parameters of an ARMA(p, q) model (or ARIMA(p, d, q) model when introducing it in the latter Sections). In a nutshell, we observe the estimation to be possible with different methods depending on the model at hand:

- AR(p): estimation can be done by (conditional) ordinary least squares (OLS).
- ARMA(p, q): due to the MA(q) part, estimation is more complicated. However, estimation can be carried out by the method of maximum likelihood, which is naturally also available for the AR(p) model as well.

6.1 Estimation of AR models with OLS

Estimation of the AR(p) model can be carried out with ordinary least squares (OLS), conditioning on the first p observations. The AR(p) model (including constant term) can be rewritten as

$$y_t = x_t'\beta + u_t,$$

where $x_t = [1 \quad y_{t-1} \cdots y_{t-p}]'$ and $\beta = [\nu \quad \phi_1 \cdots \phi_p]'$. The OLS estimators of β (and σ^2) can be constructed in the usual way familiar from past studies related to linear regression models. That is, the conditional OLS estimator of β is

$$\hat{\beta} = \Big(\sum_{t=1}^{T} x_t x_t'\Big)^{-1} \sum_{t=1}^{T} x_t y_t.$$

In this case (notation), it is assumed that the required p initial values $y_{-p+1}, \dots, y_0$ are available. Furthermore, the OLS estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \left(y_t - x_t' \hat{\beta} \right)^2.$$

- Notice that the summation in the OLS estimator starts from $t = 1, 2, \dots$ due to these available initial values. Depending on, mainly notational choices, the OLS estimator is alternatively at times written as

  $$\hat{\beta} = \Big(\sum_{t=p+1}^{T} x_t x_t' \Big)^{-1} \sum_{t=p+1}^{T} x_t y_t,$$

  if treating observations $y_1, \dots, y_p$ as initial values.

The usual (asymptotic) statistical properties of the resulting OLS estimator hold. In particular, it can be shown that:

- As the error term u_t is uncorrelated with the lags of y_t, we have $\mathsf{E}(x_t u_t) = \mathbf{0}$. That is, $\mathsf{E}(u_t y_{t-j}) = 0$ for $j = 1, 2, \dots, p$, and
- the OLS estimator $\hat{\beta}$ is a consistent estimator of β.
- Heteroskedasticity-autocorrelation consistent (HAC) standard errors, such as Newey-West standard errors, can, and often should, be automatically used to take the potential remaining conditional heteroskedasticity and autocorrelation in the residuals into account for detailed statistical inference (if this is the main objective in empirical analysis).

6.2 Maximum likelihood estimation

Estimation of ARMA (ARIMA) models with a moving average component is somewhat more complicated than the OLS based estimation of autoregressive models. The essential problem in estimating these models stems from the fact that u_t is not observable.

The estimation of an ARMA(p, q) model, including also the AR(p), can be based on (conditional) method of maximum likelihood by making an assumption of the distribution of the error u_t. A typical case would be to assume the error term is Gaussian $u_t \sim \mathsf{nid}(0, \sigma^2)$.

- It is worth noting that regardless of the errors being exactly Gaussian or not, the maximum likelihood estimator (MLE) can be interpreted as a **quasi-MLE (QMLE)** where the possible error in the model specification can be taken into consideration by using a robust parameter covariance matrix in formulation of the standard errors.
- We will come back to this QMLE interpretation within the context of AR-GARCH model. Similar argumentation as there can be used also in this context (without the GARCH errors).

We will next consider details on so called conditional maximum likelihood technique. The exact method (exact method of maximum likelihood) will be introduced at the end of next section in Extra material.

Conditional maximum likelihood method. The (exact) ML estimation of ARMA models requires the necessary selections for the initial values of y_t and u_t. By conditioning to the initial values, we get the conditional maximum likelihood technique. If the sample size (length of the time series) is large, the conditional and exact MLEs are close to each other (both lead to the same asymptotic distributions). Moreover, in practice, numerical methods can be used to maximize the log-likelihood function in both cases.

Let us consider a concrete example case of an ARMA(1,1) model

$$y_t = \nu + \phi_1 y_{t-1} + u_t + \theta_1 u_{t-1},$$

where $u_t \sim \mathsf{nid}(0, \sigma^2)$. That is, the normality assumption of the error term is assumed. The goal is to estimate the vector of parameters $\vartheta = (\nu, \phi_1, \theta_1, \sigma^2)$.

- We use the notation ϑ ("vartheta") to denote a general parameter vector containing all the parameters of the model.
- Even though we concentrate on the ARMA(1,1), the following argumentation generalizes readily to more general ARMA processes, which are of the form $y_t = \mathsf{E}_{t-1}(y_t) + u_t$ where $\mathsf{E}_{t-1}(y_t)$ corresponds to the employed model structure (systematic part of the model). Here $\mathsf{E}_{t-1}(\cdot)$ denotes the conditional expectation and it will be introduced more detail in Forecasting section. In this ARMA(1,1) case, we get $\mathsf{E}_{t-1}(y_t) = \nu + \phi_1 y_{t-1} + \theta_1 u_{t-1}$, whereas, e.g., in the AR(p) case we get $\mathsf{E}_{t-1}(y_t) = \nu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}$.

In the ARMA(1,1) model, taking the initial values for y_0 and u_0 as given, the sequence of $u_1, \ldots, u_T$ can be constructed from $y_1, \ldots, y_T$ by iterating

$$u_t = y_t - \nu - \phi_1 y_{t-1} - \theta_1 u_{t-1}, \quad t = 1, \ldots, T.$$

When conditioning on y_{t-1} and u_{t-1}, which are included in the information available at time $t-1$, we end up that the conditional distribution of y_t that is (conditionally) normally distributed due to the normality assumption of u_t:

$$y_t|(y_{t-1}, u_{t-1}) \sim \mathsf{N}(\nu + \phi_1 y_{t-1} + \theta_1 u_{t-1}, \sigma^2).$$

The specific form of the conditional density function of y_t, conditional on y_{t-1} and u_{t-1}, is hence

$$f_{y_t|y_{t-1},u_{t-1}}(y_t|y_{t-1}, u_{t-1}; \vartheta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_t - \nu - \phi_1 y_{t-1} - \theta_1 u_{t-1})^2}{2\sigma^2} \Big).$$

Construction of the log-likelihood function. The (conditional) log-likelihood function can based on the conditional density functions for single observations. Here we are not just restricting ourselves to the ARMA(1,1) and/or Gaussian case. That is the following steps can be generalized also to other modelling choices such as ARMA models.

Denote $\underline{\mathbf{Y}}_t = (y_1 \dots y_t)$ and, therefore, the sample of observations is denoted by $\underline{\mathbf{Y}}_T = (y_1, \dots y_T)$. The joint conditional density function (without arguments) $f_{\underline{\mathbf{Y}}_T}(\underline{\mathbf{Y}}_T)$ is obtained as the product of conditional density functions

$$f_{\underline{\mathbf{Y}}_T} = f_{y_T|\underline{\mathbf{Y}}_{T-1}} \cdot f_{\underline{\mathbf{Y}}_{T-1}} = f_{y_T|\underline{\mathbf{Y}}_{T-1}} f_{y_{T-1}|\underline{\mathbf{Y}}_{T-2}} \cdot f_{\underline{\mathbf{Y}}_{T-2}} = \cdots = \prod_{t=1}^{T} f_{y_t|\underline{\mathbf{Y}}_{t-1}} \cdot f_{\underline{\mathbf{Y}}_0}.$$

By conditioning on $\underline{\mathbf{Y}}_0$, which contains the necessary initial values (depending on, e.g., the lag lengths p and q of the ARMA(p,q) model), the conditional density function gets the form

$$\prod_{t=1}^{T} f_{y_t|\underline{\mathbf{Y}}_{t-1}},$$

which then leads to the the (conditional) log-likelihood function

$$l(\vartheta) = \sum_{t=1}^{T} \log f_{y_t|\underline{\mathbf{Y}}_{t-1}}(y_t|\underline{\mathbf{Y}}_{t-1}; \vartheta).$$

- Going back to the specific example of Gaussian ARMA(1,1) model, the conditional log-likelihood function is hence

  $$l(\vartheta) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T} u_t^2,$$

  where $u_t = y_t - \nu - \phi_1 y_{t-1} - \theta_1 u_{t-1}$ in this ARMA(1,1) case.

The maximum likelihood estimate (MLE) of ϑ, that is $\hat{\vartheta}$, is obtained by maximizing the conditional log-likelihood function.

- This requires numerical optimization methods. These numerical (iterative) algorithms require initial values and/or possibly also preliminary estimation of the parameters, but econometric and statistical program packages are carrying out these steps straightforwardly.

AR(p) case. We will next show that, in the case of the AR(p) model, the conditional maximum likelihood estimation actually leads to the (conditional) least squares estimates when the error term is Gaussian ($u_t \sim \text{nid}(0, \sigma^2)$).

Since y_t is a linear combination of $u_t, u_{t-1}, \ldots$ (given stationarity and corresponding MA(∞) representation), it follows that u_s and x_t are independent $\forall s \geq t$, where, as above in the AR(p) case, $x_t = [1 \quad y_{t-1} \cdots y_{t-p}]'$. By assuming the normality of the error term, we get

$$y_t | \underline{\mathbf{Y}}_{t-1} \sim \mathsf{N}(x_t'\beta, \sigma^2),$$

and the conditional density function of y_t is

$$f_{y_t|\underline{\mathbf{Y}}_{t-1}}(y_t|x_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(y_t - x_t'\beta)^2 \Big).$$

From this we get the (sample) log-likelihood function

$$l(\vartheta) \equiv l(\beta, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - x_t'\beta)^2.$$

where the first term can also be removed because it does not depend on the parameters ϑ. This corresponds the (conditional) log-likelihood function of a Gaussian linear regression model, but now for time series observation $(y_1, \ldots, y_T)$. Maximizing this gives us the same conditional least squares estimator $\hat{\beta}$ (and $\hat{\sigma}^2$) as obtained above.

Exact method of maximum likelihood. The exact log-likelihood function contains also the impact of initial values when constructing the log-likelihood function. Some details on this technique is provided below.

Extra: Exact maximum likelihood estimation of ARMA models

Let us take a closer look at exact maximum likelihood (ML) based estimation.

The exact likelihood function. For the purpose of estimation based on the exact method of maximum likelihood, consider the *Gaussian* ARMA(p, q) process

$$y_t = \nu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}, \;\; u_t \sim \mathsf{nid}\left(0, \sigma^2\right),$$

and assume the stationarity, invertibility and the identification condition of the ARMA process hold. Suppose $y = (y_1, \ldots, y_T)$ is a vector of observations generated by this process. As before, denote $\phi = (\phi_1, \ldots, \phi_p)$, $\theta = \left(\theta_1, \ldots, \theta_q\right)$ and $\beta = (\phi, \theta)$. Making use of the general formulas for the mean and autocovariance function of a linear process, it can be seen that $\mathsf{E}\,(y) = 0$ and $\mathsf{Cov}\,(y) = \sigma^2\Sigma$, where the (positive definite) matrix $\Sigma = \Sigma\,(\beta)$ $(T \times T)$ is a function of the parameter β (but not of the parameter σ^2).

Since linear transformations of Gaussian distributions are also Gaussian (this holds also for infinite dimensional transformations), $y \sim \mathsf{N}\left(0, \sigma^2\Sigma\right)$.

- In addition to the constant term, a trend component can be inserted in without affecting the asymptotic properties of the ML-estimators forβ and σ^2.

The density function of the random vector y is

$$f_y\,(y) = (2\pi)^{-T/2} \det\left(\sigma^2\Sigma\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} y'\Sigma^{-1} y\right\}.$$

Making use of properties of the determinant, $\det\left(\sigma^2\Sigma\right) = \sigma^{2T}\det\left(\Sigma\right)$, so that the log likelihood function becomes

$$l\left(\beta, \sigma^2\right) = -\frac{T}{2}\log\sigma^2 - \frac{1}{2}\log\det\left(\Sigma\,(\beta)\right) - \frac{1}{2\sigma^2} y'\Sigma\,(\beta)^{-1}\, y.$$

Because the matrix $\Sigma\,(\beta)$ is a rather complicated function of the parameter β, the log-likelihood function cannot be expressed in a simple form and, consequently, the above expression is difficult to be used directly to maximize the likelihood function. In practice, one has to resort to numerical methods to maximize the likelihood function. This involves computing the inverse matrix $\Sigma\,(\beta)^{-1}$ and the determinant $\det\left(\Sigma\,(\beta)\right)$ for a range of given values of β. For a given value of β, the matrix $\Sigma\,(\beta)$ can be computed for ARMA processes.

Many different algorithms have been proposed to solve the maximization problem described above.

Statistical inference. It can be shown that when the stationarity, invertibility, and identification conditions hold, the "usual" asymptototic properties of an maximum likelihood estimator (exact or conditional) hold. In particular, the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ are both consistent, asymptotically normally distributed and asymptotically independent. In other words, it holds

$$\hat{\beta} \underset{as}{\sim} \mathsf{N}\left(\beta, \mathbf{V}\,(\beta)^{-1}\right),$$

where the matrix $\mathbf{V}\,(\beta) = \mathsf{E}\left[-\partial^2 l(\beta, \sigma^2)/\partial\beta\partial\beta'\right)]$ is the so-called Fisher information matrix of the parameter β

- The asymptotic distribution result holds also without the normality assumption, although in that case the estimators are no longer (asymptotically) efficient.

- It should be noted that the stationarity, invertibility, and identification conditions are required for the above asymptotic distribution result, which can give a poor approximation also when these conditions are "nearly violated".

Hypothesis testing. Using the empirical counterpart of $\mathbf{V}(\beta)$, namely $\widehat{\mathbf{V}}(\hat{\beta},\hat{\sigma}^2) = -\partial^2 l(\hat{\beta},\hat{\sigma}^2)/\partial\beta\partial\beta'$ (the so-called empirical or observed information matrix of the parameter β), the asymptotic distribution result above can be used to construct statistical Wald tests and confidence intervals concerning the parameter β

- The partial derivatives appearing in this expression can in practice be approximated with their numerical counterparts.

In particular, the square roots of the diagonal elements of the matrix $\widehat{\mathbf{V}}(\hat{\beta},\hat{\sigma}^2)^{-1}$ can be used as approximate standard errors of the components of the estimator $\hat{\beta}$.

- For instance, an individual (ith) estimated component of $\hat{\beta}$, that is $\hat{\beta}_i$, would be interpreted as "significantly" different from zero if its absolute value is at least 2 (specifically 1.96) times the size of its approximate standard error.

Formally, let us denote $\mathbf{V}(\beta)^{-1} = [v^{ij}]$, its empirical counterpart is $\widehat{\mathbf{V}}(\hat{\beta},\hat{\sigma}^2)^{-1} = [\hat{v}^{ij}]$, and

$$\hat{\beta}_i \underset{as.}{\sim} \mathsf{N}(\beta_i, v^{ii}).$$

Thus null hypothesis $\beta_i = 0$ gets rejected at the 5% significance level if the t-ratio (cf. t-test)

$$\left|\frac{\hat{\beta}_i}{\sqrt{\hat{v}^{ii}}}\right| \geq 1.96 \Leftrightarrow |\hat{\beta}_i| \geq 1.96\sqrt{\hat{v}^{ii}}$$

and the 95% confidence intervals for β_i can be obtained by

$$\hat{\beta}_i \pm 1.96 \cdot \text{s.e.}(\hat{\beta}_i) = 1.96\sqrt{\hat{v}^{ii}}.$$

Tests constructed using likelihood ratio principles can also be used in a conventional manner.

Empirical example. In the previous section, we concluded an AR(2) model to be one possible candidate for the quarterly U.S real GDP growth (1985:Q1–2007:Q2). The conditional maximum likelihood estimation of this model gives us

$$y_t = \underset{(0.197)}{1.693} + \underset{(0.101)}{0.160} y_{t-1} + \underset{(0.101)}{0.287} y_{t-1} + \hat{u}_t, \quad \hat{\sigma}^2 = 3.536,$$

where in parentheses under the estimates are the approximate standard errors. Especially the estimate for the AR(2) coefficient is statistically significant at the 5% significance level and hence the model cannot be shrinked.

Alternatively, if reporting the estimation result for the demeaned process, we then obtain

$$\bar{y}_t = \underset{(0.101)}{0.160} \bar{y}_{t-1} + \underset{(0.102)}{0.287} \bar{y}_{t-2} + \hat{u}_t,$$

where $\bar{y}_t = y_t - 3.061$ is the demeaned GDP growth series.

Chapter 7

Model selection of ARMA model

7.1 Starting point

Consider an observed time series $y_1, \dots, y_T$, which is assumed, based on its graph or other diagnostics, to be generated by a stationary process, making an ARMA(p,q) specification reasonable.

- As mentioned earlier, achieving stationarity may require preliminary transformations, such as differencing related to ARIMA(p, d, q) models or adjustements for nonstationary processes with unit roots (to be introduced later). Common transformations include differencing (possibly combined with logarithmic transformation) and removing deterministic trends using appropriate detrending methods.
- For example, if the process follows a deterministic linear trend, the model can be written as

$$y_t = \nu + \delta t + z_t, \quad t = 1, \dots, T,$$

  where z_t is an ARMA(p, q) process (cf. the decomposition $y_t = \mu_t + z_t$). This is an example of a **trend-stationary process**. The usual approach is to estimate the parameters ν and δ using least squares and then analyze the residual series. From a parameter estimation perspective, this means modelling the deterministic part via ν and δ and the ARMA(p,q) parameters of z_t separately. An obvious alternative is to estimate all parameters jointly.

Finding a suitable ARMA(p, q) process, or, using terminology commonly employed in statistical modelling, an **ARMA(p, q) model**, traditionally involves the following interrelated stages:

1) Specifying (selecting) the model orders p and q
2) Estimating the model parameters (which may include preliminary estimations steps)
3) Evaluating the adequacy of the estimated model through diagnostic checks and, in some cases, assesing forecasting performance

If the model is found to be inadequate in Step 3, one must return to Step 1 to select new orders p and q, re-estimate the parameters, and re-evaluate the model. As noted, Steps 1–3 are interdependent and are not always performed in a strictly sequential manner.

In practice, one can never know with certainty whether the true (correct) orders of the ARMA(p, q) model have been identified.

- If the orders are chosen too small, some aspects of the time series' autocorrelation structure may remain unmodelled. This is clearly suboptimal, particularly from rigorous modelling and potentialy also from a forecasting standpoint. One might then be tempted, but errorneously, to select the overly large orders to avoid underfitting.
- However, choosing orders that are too large leads to inefficiencies in parameter estimation, which can degrade forecasting accuracy. This is especially problematic when both p and q are chosen too large, as it may result in a model where $\phi_p = 0 = \theta_q$, violating the identification condition of the ARMA process. In such cases, the model parameters are not identifiable, making meaningful estimation impossible.

For these reasons, it is generally advisable to follow the **principle of parsimony**: select a model that is as simple as possible while still being sufficiently rich to capture the essential dynamics of the data.

In what follows in this section, we assume that the error term satisfies $u_t \sim \mathsf{iid}\,(0, \sigma^2)$. In the context of maximum likelihood estimation, we adopt the stronger assumption $u_t \sim \mathsf{nid}\,(0, \sigma^2)$.

- Unless otherwise stated, we also assume that the stationarity and invertibility conditions hold.

Constant term or demeaning?. As previously discussed, the mean $\mathsf{E}(y_t) = \mu$ is typically nonzero, which means that we need to include a constant term to the model, or alternatively working with a demeaned version of the time series. A few clarifying points from empirical modelling perspective:

- The mean is typically estimated using the sample mean $\bar{y}$, after which one may consider the **centered time series** $y_t - \bar{y}$, $t = 1, \dots, T$, and proceed as if $\bar{y}$ exactly equals the unknown population mean μ. While this introduces a small approximation error, it can be shown that the error becomes negligible in large samples.
- **If the time series is not centered, you should always include the constant term** in the model equation.

7.2 Sample autocorrelations and partial autocorrelations

A good second step after visualizing the series is to investigate the autocorrelation structure of the time series using the estimated (sample) autocorrelation function (ACF) and partial autocorrelation function (PACF). In many cases, these functions may contain clues regarding the suitable lag lengths p and q concerning an adequate ARMA(p, q) model.

In particular, recall the theoretical properties of AR, MA and ARMA processes: **one should look for potential cutoffs ("sudden drops to zero") in the sample ACF and PACF**.

- A sharp cutoff in the ACF with gradual decaying PACF suggests an MA(q) process.
- A sharp cutoff in the PACF with gradual decaying ACF suggests an AR(p) process.
- Gradual decay in both ACF and PACF may indicate an ARMA$(p$,$q)$ process.

This approach is part of the **Box–Jenkins model selection procedure**. While this method does not always point to a single, definitive model, it is useful for **ruling out implausible alternatives** and narrowing down the set of candidate models. It is often used in conjunction with other model selection criteria, which will be introduced in the following sections.

7.3 Information criteria and sequential tests

As with the use of the sample ACF and PACF, the goal here is to select the orders p and q of the ARMA(p, q) model after one has first set "sufficiently large" maximum lag lengths, denoted p^* and q^*. This preliminary selection can be guided, for instance, by examining the sample autocorrelation and partial autocorrelation functions.

- Let $\tilde{\sigma}^2_{p,q}$ $(0 \leq p \leq p^*,\ 0 \leq q \leq q^*)$ be the estimator of the innovation variance σ^2 of an ARMA(p, q) process.
- Suppose also that the quantity $m > \max(p^*, q^*)$ used in this estimation method is held constant for all attempted values of p and q.

As discussed, it is preferable to avoid choosing p and q too large.

- One possible way to choose p and q would be to minimize $\tilde{\sigma}^2_{p,q}$ over the possible values $0 \leq p \leq p^*$, $0 \leq q \leq q^*$. However, this approach does not work, because due to the nature of the least squares or maximum likelihood method, this would lead one to choose orders too large.

To fix this obvious problem, one approach to select the orders p and q is to minimize the function

$$C(p, q) = \log \tilde{\sigma}^2_{p,q} + \frac{(p+q+1)\, g(T)}{T}, \quad 0 \leq p \leq p^*, 0 \leq q \leq q^*, \tag{7.1}$$

where the so-called **penalty function** $g(\cdot)$ is positive valued and satisfies $g(T)/T \to 0$ as $T \to \infty$ ("+1" due to the inclusion of the constant term ν). The idea behind the penalty function is to penalize for using an unnecessarily large model. If increasing the order p or q does not make $\tilde{\sigma}^2_{p,q}$ sufficiently much smaller, then one does not choose the larger model.

Typically used penalty functions (which have been derived based on different principles) are:

- AIC: $g(T) = 2$ (Akaike information criterion)
- HQ: $g(T) = 2\log(\log T)$ (Hannan and Quinn information criterion)
- BIC: $g(T) = \log T$ (Schwarz information criterion/Bayesian information criterion)

The first of these (AIC) penalizes the least (favors larger models) and the last (BIC) the most (favors smaller models).

- We note that of the HQ penalty function, there exist also other versions in which the constant 2 has been replaced by some other constants.

In practice, it is often advisable to use the criteria described above only as one tool in model selection, rather than relying solely on the mechanical minimization of the criterion $C(p,q)$. Ideally, the final model selection should be made after successful parameter estimation (i.e. when there are clearly no suspicious behavior in some estimates) and after conducting diagnostics checks to assess the adequacy of the model (see the next subsection).

However, from a modern and partly **machine learning–oriented perspective**, information criteria and similar tools are often applied in a more automated fashion to select the lag lengths of ARMA models. While this approach is not always ideal, it reflects a practical compromise: in the context of large and complex datasets, it may not be feasible, or even reasonable, to perform detailed diagnostic checks at every step of the model selection process.

Neighbouring models. In case of time series that appear difficult to model and select a suitable ARMA(p,q) model, it may be useful to consider as alternatives also models in which lag lengths p and q are one larger than in the model that minimizes the criterion function.

- However, it is important to keep in mind that selecting both orders too large can lead to identification problems (cf. the identification condition of the ARMA processes).

Another practical approach is to select a small set of models corresponding to the lowest values of the information criterion function and subject them to more detailed investigation. It is also common to use multiple information criteria simultaneously and examine whether they point to the same model.

Sequential tests. One way to choose the AR(p) or ARMA(p,q) model orders is based on sequential tests. In what follows, we will consider an AR(p) model for simplicity, but the following procedure can be generalized into ARMA models too.

- Start by choosing a relatively large model order $p*$
- Estimate an AR$(p*)$ model
$$y_t = \nu + \phi_1 y_{t-1} + \cdots + \phi_{p^*-1} y_{t-p^*+1} + \phi_{p^*} y_{t-p^*} + u_t.$$
- Test (evaluate) constraint $H_0 : \phi_{p*} = 0$

- If the null hypothesis cannot be rejected, estimate an AR($p * -1$) model, and test for $H_0 : \phi_{p*-1} = 0$
- This sequential testing procedure is carried out until we can reject the null hypothesis (get the first rejection).

The sequential testing above can be used mechanically (notice the obvious multiple testing problem and its impact on p-values, but this is often ignored in this context) using t-test or the Wald test statistics or using information criteria.

7.4 Evaluating the adequacy of the estimated model

In the previous sections, it is assumed that the lag lengths (orders) p and q of the ARMA(p, q) model have been chosen, in one way or another, that the model is adequate/sufficient. This means that the error term u_t satisfies at least the assumption $u_t \sim \mathsf{iid}\,(0, \sigma^2)$, but at the same time p and q are not, at least simultaneously, unnecessarily large.

- If also assuming the normality of the error term, then satisfying preferably the assumption $u_t \sim \mathsf{nid}\,(0, \sigma^2)$

As discussed above, one often follows **the principle of parsimony** when choosing the model, which can sometimes leads to a model in which at least one of the orders has been chosen too small. In such a case, the error term of that selected model would remain autocorrelated, and moving to a larger model to incorporate this remaining correlation may improve, for instance, forecasting performance of the model.

Residuals. A natural way to investigate the adequacy of an estimated model is to use residuals, whose properties should resemble those of the theoretical error terms u_t. As with linear regression models, residuals are acquired as the difference of the observed time series (y_t) and the fitted values $\widehat{y}_t$ of the selected (adequate) and estimated ARMA model

$$\widehat{u}_t = y_t - \widehat{y}_t.$$

In practice, one can also perform a further scaling and divide the residuals $\widehat{u}_t$ with their estimated standard error $\widehat{\sigma}$. The **standardized residuals** obtained in this way, $\widehat{u}_t/\widehat{\sigma}$, should in the case of a correctly specified model be approximately independent with mean zero and variance one.

Graphical and formal residual analysis. The first step in the investigation of the properties of the residuals is to plot their time series graph. If the estimated model is correct (or empirically more realistically "adequate/sufficient"), the time series of the residuals should not exhibit trends, cyclical components, systematic variation in their level over time, variation of the variance over time (that is, heteroskedasticity), too many outlying observations, etc.

- That is due to the fact that residuals $\hat{u}_t$ should resemble assumptions made on the error term u_t.

The next step is to examine whether the residuals are uncorrelated. This is done similarly to how autocorrelation is assessed in the original time series—by analyzing the sample ACF and PACF of the residuals. If significant autocorrelation remains, it suggests that the model has not fully captured the dynamics of the data, and further refinement may be necessary.

- It should be noted, though, that for the autocorrelation coefficients computed from residuals, the critical bounds $\pm 1.96/\sqrt{T}$ are not valid, not even asymptotically.
- Expressions for the asymptotically correct critical bounds do exist, but are complicated, although some computer programs plot them automatically. If such correct bounds are not (easily) available, the somewhat incorrect bounds $\pm 1.96/\sqrt{T}$ are sometimes used to give at least a rough measure of the significance of the autocorrelation coefficients computed from the residuals.

In testing remaining autocorrelation in the residuals, the **Ljung-Box test statistic** (introduced earlier) can also be used.

- However, when considering an ARMA(p, q) model with a constant term, instead of the (asymptotic) χ^2_H–distribution used earlier, one should now use the $\chi^2_{H-p-q-1}$–distribution where one has to choose H "large enough" for this (asymptotic) distribution to be valid.
- In practice, H is chosen somewhere around 10–20 depending on the sample size T. Moreover, typically a couple of different selections of H are considered to see whether the testing results are similar for different selections.

Possible (remaining) nonlinearities. Although the errors of the chosen model could be considered to have no autocorrelation, they are not necessarily independent. In line with earlier discussion on possible nonlinearities, one but clearly restricted way of investigating potential nonlinear dependencies in the errors is to use autocorrelations of the squared residuals.

It can be shown that when errors satisfy the assumption $u_t \sim \mathsf{nid}\,(0, \sigma^2)$, the sample autocorrelation coefficients of the squared residuals $\hat{u}_t^2$ $(t = 1, \ldots ., T)$ are approximately independent and $\mathsf{N}\,(0, 1/T)$–distributed. Therefore, to sample autocorrelation coefficients of squared residuals $\hat{u}_t^2$, one can apply the critical bounds $\pm 1.96/\sqrt{T}$ as well as the **McLeod-Li test** presented (that is, Ljung-Box test applied to squared residuals).

- If one notices that the errors are not autocorrelated, but that their squares are autocorrelated, one can also conclude that the errors can not be Gaussian (for Gaussian processes, being not autocorrelated is equivalent with being independent). Although the asymptotic results of the estimators presented in the previous section do hold even without the normality of the errors, it is still useful to investigate how realistic the normality assumption is by checking the histogram of the residuals or using quantile-quantile plot.

- If the errors are only serially uncorrelated, but not independent, the covariance matrix of the maximum likelihood estimator $\widehat{\beta}$ is not necessarily the one given in parameter estimation section. The consistency and asymptotic normality of $\widehat{\beta}$ can still hold under more general assumptions, but the estimated standard errors or test statistics are not necessarily valid.

Empirical example. Let us take a look at the sufficiency of the AR(2) model estimated for the U.S. real GDP growth series. Below we depict the residual time series from the fitted model, as well as their sample autocorrelations and histogram.

- It turns out that there is no (substantial) statistically significant residual autocorrelation left in the residuals. In addition to the sample autocorrelations, the Ljung-Box test statistics for different lag lengths H generally also confirm this conclusion (at least at the 5% or higher significance level).

- As an example, the Ljung-Box test statistic gets the value 17.546 from the first 16 lags. That is, to say, $H = 16$ and the corresponding p-value from the χ^2_{13}–distribution is 0.176 (degrees of freedom: $H - p - q - 1 = 13$.

The histogram of the residuals suggest that the normality assumption is reasonable (see bottom right). Residual histograms typically strictly matching the density function of the normal distribution (as depicted also in the figure below), but in this case the normality assumption seems relatively adequate.

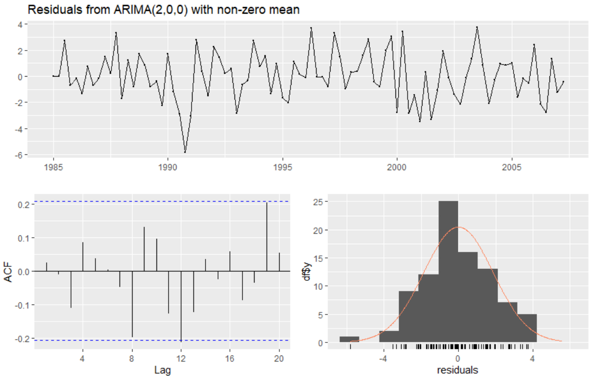


Figure: The residual time series of the AR(2) model estimated for the US real GDP growth series (upper figure), their estimated autocorrelation function ($h = 0, \ldots, 20$, bottom left) and their histogram (bottom right).

The main conclusions from residual diagnostics do not substantially change if using the AIC to select the lag length of the AR model. In the resulting AR(3) model the third lag (y_{t-3}) is not statistically significant predictor in terms of its estimated t-value. Moreover, when considering ARMA(p, q) models, the AIC selects MA(2) model, with essentially the same conclusions from residual diagnostics as in the AR(2) model.

When considering squared residuals and the McLeod-Li test statistics for different lag lengths, it turns out that basically with any lag length selection the resulting p-values are very high (typically higher than 0.50). Therefore, for the Great Moderation time period in the U.S. economy, there is no initial evidence for strong nonlinearities in the real GDP growth.

Chapter 8

Forecasting with ARMA models

8.1 Properties of the conditional expectation

Before considering forecasting with an ARMA(p,q) process, we examine a general situation where the objective is to forecast the (scalar) random variable Y using the realized value $X = x$ of a random vector X. A result from probability theory tells us that the optimal forecast, in the sense of minimized mean square forecast error, is the conditional expected value of Y given $X = x$. In other words,

$$\mathsf{E}\left[(Y - \mathsf{E}(Y\,|X = x\,))^2\right] \le \mathsf{E}\left[(Y - g(x))^2\right]$$

for any function $g(x)$ of x (assuming the expectations above are finite).

Extra: Some details on the conditional expectations

In the case of continuous distributions, the conditional expectation of Y given $X = x$ is defined by the equation

$$\mathsf{E}(Y\,|X = x\,) = \int_{-\infty}^{\infty} y f_{Y|X}(y;x)\,dy = \int_{-\infty}^{\infty} y\frac{f_{Y,X}(y,x)}{f_X(x)}dy,$$

where

- $f_{Y,X}(y,x)$ is the joint density function of the random vector (Y,X),
- $f_X(x) = \int_{-\infty}^{\infty} f_{Y,X}(y,x)\,dy$ is the marginal density function of X, and
- $f_{Y|X}(y;x) = f_{Y,X}(y,x)\,/f_X(x)$ defines the conditional density function of Y given $X = x$.

When x varies over the possible values of the random vector X, $\mathsf{E}\left(Y\,|X=x\right)$ as a function of x defines a random variable for which it is natural to use the notation $\mathsf{E}\left(Y\,|X\right)$.

The above definition can be generalized to the case where the random vector X might be infinite-dimensional. This case will be encountered in the ARMA(p,q) context assuming that all the values of the process $\{y_t,\ t=0,\pm1,...\}$ that precede the **forecast origin** (the time when the forecast is constructed) are known.

For our purposes, it suffices to know some simple properties of the conditional expected value (which also hold in the case of an infinite-dimensional conditioning variable). In this course, we will use the following properties of the conditional expected value:

- **CEV1**: $\mathsf{E}\left(aY_1+bY_2\,|X\right) = a\mathsf{E}\left(Y_1\,|X\right)+b\mathsf{E}\left(Y_2\,|X\right)$, when a are b constants.
- **CEV2**: $\mathsf{E}\left(Y\,|X\right)=\mathsf{E}\left(Y\right)$, when Y and X are independent random variables.
- **CEV3**: $\mathsf{E}\left(Y\right)=\mathsf{E}\left[\mathsf{E}\left(Y\,|X\right)\right]$ (so-called law of iterated expectations)
- **CEV4**: $\mathsf{E}\left[h\left(X\right)Y\,|X\right]=h\left(X\right)\mathsf{E}\left(Y\,|X\right)$ for any function h (assuming the expected value of $h\left(X\right)Y$ exists and is finite).

8.2 Forming forecasts

Forecasting an ARMA(p,q) process. As the discussion above suggests, we consider forecasting an ARMA(p,q) process assuming that the entire infinitely long history of the process is known.

In what follows, we again assume that the constant term is included in the model equation to control a nonzero mean of y_t (i.e. $\mathsf{E}(y_t)=\mu\neq0$).

- When we are interested in forecasting the levels of y_t, which is often and typically the case, it is important to consider how forecasts will be obtained for original levels if using demeaning of the time series and/or extracting the trend component. That is, all the possible transformations must be carefully taken into account in forecast computations.
- We will also assume that the ARMA(p,q) process under consideration is stationary and invertible and that the innovation term satisfies the condition $u_t\sim\mathsf{iid}\left(0,\sigma^2\right)$.

In the calculations that follow, we repeatedly make use of the above-mentioned properties of the conditional expected value (CEV1–CEV4).

AR(1) case. Consider forecasting an AR(p) process, and for simplicity, we first also assume that $p = 1$.

- In one-step forecasting, the aim is to forecast the value y_{t+1}, when the preceding history $\{y_t, y_{t-1}, ...\}$ of the process is known.
- More generally, in h-step forecasting, the aim is to forecast the value of y_{t+h}, when $\{y_t, y_{t-1}, ...\}$ is known.

For brevity, denote the conditional expectation as

$$\mathsf{E}\left(y_{t+h}\left|y_s,\ s \leq t\right.\right) \equiv \mathsf{E}_t\left(y_{t+h}\right), \quad (h \geq 1)\,.$$

Taking conditional expectations of both sides of the AR(1) equation (forwarded first by one period)

$$y_{t+1} = \nu + \phi_1 y_t + u_{t+1}$$

leads to (see CEV1)

$$\mathsf{E}_t\left(y_{t+1}\right) = \nu + \phi_1 \mathsf{E}_t\left(y_t\right) + \mathsf{E}_t\left(u_{t+1}\right).$$

- Under the stationarity condition $|\phi_1| < 1$, the variables $y_t, y_{t-1}, ...$ depend only on the variables $u_t, u_{t-1}, ...$ Recall the MA(∞) representation of y_t to see this.
- Therefore, in the conditional expectation $\mathsf{E}_t\left(u_{t+1}\right)$, the conditioning random variables $\{y_s, s \leq t\}$ are independent of u_{t+1}. Following the property CEV2, we then obtain

  $$\mathsf{E}_t\left(u_{t+1}\right) = \mathsf{E}\left(u_{t+1}\right) = 0.$$

  More generally it holds

  $$\mathsf{E}_t\left(u_{t+k}\right) = \mathsf{E}\left(u_{t+k}\right) = 0, \quad k \geq 1.$$

- Moreover, $\mathsf{E}_t\left(y_t\right) = y_t$ (see CEV4: y_t is included in the information set at time t).

Putting the above results together, we obtain

$$\mathsf{E}_t\left(y_{t+1}\right) = \nu + \phi_1 y_t.$$

When forecasting y_{t+2}, we obtain

$$\begin{aligned}\mathsf{E}_t\left(y_{t+2}\right) &= \nu + \phi_1 \mathsf{E}_t\left(y_{t+1}\right) + \mathsf{E}_t\left(u_{t+2}\right)\\ &= \nu + \phi_1 \mathsf{E}_t\left(y_{t+1}\right)\\ &= (1+\phi_1)\nu + \phi_1^2 y_t,\end{aligned}$$

and inductively (given $|\phi_1| < 1$), h-period forecast

$$\mathsf{E}_t\left(y_{t+h}\right) = (1+\phi_1+\dots+\phi_1^h)\nu + \phi_1^h y_t = \mu + \phi_1^h y_t.$$

AR(p) case. In the case of a general AR(p) process, we use a similar steps to obtain forecasts. Taking conditional expectations of both sides of

$$y_{t+1} = \nu + \phi_1 y_t + \dots + \phi_p y_{t+1-p} + u_{t+1},$$

and using similar arguments as in the case $p = 1$, including the results $\mathsf{E}_t\left(u_{t+1}\right) = 0$ and $\mathsf{E}_t\left(y_{t-j}\right) = y_{t-j}$ $(j \geq 0)$, we obtain

$$\mathsf{E}_t\left(y_{t+1}\right) = \nu + \phi_1 y_t + \dots + \phi_p y_{t+1-p}.$$

When forecasting y_{t+2}, we similarly obtain

$$\mathsf{E}_t\left(y_{t+2}\right) = \nu + \phi_1 \mathsf{E}_t(y_{t+1}) + \phi_2 y_t + \dots + \phi_p y_{t+2-p},$$

where the conditional expected value on the right hand side could be replaced with the expression obtained for it above.

- In other words, the variable y_{t+1}, unknown at time t, has been replaced with its forecast $\mathsf{E}_t(y_{t+1})$.
- The variables $y_t, \dots, y_{t+2-p}$ are known at time t. Therefore, they remain on the right hand side as they are included in the information set at the forecast origin.

Inductively, it is straightforward to see that the above generalizes to forecasting h periods ahead

$$\mathsf{E}_t\left(y_{t+h}\right) = \nu + \phi_1 \mathsf{E}_t(y_{t+h-1}) + \phi_2 \mathsf{E}_t(y_{t+h-2}) + \dots + \phi_p \mathsf{E}_t(y_{t+h-p}), \quad h \geq 1,$$

where $\mathsf{E}_t\left(y_{t+h-j}\right) = y_{t+h-j}$ for $h \leq j$. This means that forecasts can be computed recursively, starting with the one-step-ahead case $h = 1$, and proceeding one step at a time to the forecast horizons $h = 2$, $h = 3, \dots$.

ARMA(p,q) case. To obtain forecasting formulae for an ARMA(p,q) process is quite similar to forecasting with an AR(p) process.

- When the stationarity condition holds, $y_t - \mu = \sum_{j=0}^{\infty} \psi_j u_{t-j}$, from which it follows that u_{t+i}, $i \geq 1$, is independent of the variables $\{y_t, y_{t-1}, ...\}$. Therefore, $\mathsf{E}_t(u_{t+i}) = \mathsf{E}(u_{t+i}) = 0$ for all $i \geq 0$ (see CEV2).

- On the other hand, when the invertibility condition holds, $u_t = \sum_{j=0}^{\infty} \pi_j (y_{t-j} - \mu)$, and therefore $\mathsf{E}_t(u_{t-i}) = u_{t-i}$ for all $i \geq 0$ (see CEV4).

Taking conditional expectations of both sides of the equation defining an ARMA(p,q) and using the above-mentioned properties, we obtain h-period-ahead forecast

$$\mathsf{E}_t(y_{t+h}) \quad = \quad \nu + \phi_1 \mathsf{E}_t(y_{t+h-1}) + \dots + \phi_p \mathsf{E}_t(y_{t+h-p}) + \theta_1 \mathsf{E}_t(u_{t+h-1}) + \dots + \theta_q \mathsf{E}_t(u_{t+h-q}), \quad h \geq 1,$$

where $\mathsf{E}_t\left(y_{t+h-j}\right) = y_{t+h-j}$ for $h \leq j$, and

$$\mathsf{E}_t\left(u_{t+h-j}\right) = \begin{cases} u_{t+h-j}, \text{ when } h \leq j, \\ 0, \text{ when } h > j. \end{cases}$$

Using this forecast formula, forecasts can again be computed recursively, starting with the one-step-ahead case $h = 1$, and proceeding one step at a time to the horizons $h = 2, h = 3, \dots$.

The discussion above makes the unrealistic assumption that the parameters of the examined process are known.

- In practice, unknown parameters are replaced by their estimates, and in this case the above-mentioned optimality of forecasts (forecast construction) only holds approximately.

Moreover, for ARMA models, in practice the innovation terms appearing in the forecasting formulae above (not in the AR(p) case) have to be computed using the observed time series, so in the formula $u_t = \sum_{j=0}^{\infty} \pi_j (y_{t-j} - \mu)$ we have to truncate the sum at some point (e.g., $\sum_{j=0}^{t-1} \pi_j (y_{t-j} - \mu)$).

- A popular alternative is to calculate u_t using the difference equation

  $$u_t = y_t - \nu - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p} - \theta_1 u_{t-1} - \dots - \theta_q u_{t-q}, \quad t = 1, 2, \dots,$$

  where the initial values $y_0, \dots, y_{-p}$ can be assumed known (that is, observed), and the initial values $u_0, \dots, u_{-q}$ can be chosen to be $u_0 = \dots = u_{-q} = 0$ (that is, the unconditional mean of the u_t). When t is large, the effect of initial values is negligible.

8.3 Prediction intervals

An alternative way to approach forecasting an ARMA(p,q) process is via the MA(∞) representation. Taking conditional expectations of both sides of the equation

$$y_{t+h} = \mu + \sum_{j=0}^{\infty} \psi_j u_{t+h-j},$$

and using results of conditional expectations used above, we obtain

$$\mathsf{E}_t(y_{t+h}) = \mu + \sum_{j=h}^{\infty} \psi_j u_{t+h-j}.$$

This formula is not useful in practice, but it is convenient for investigating the properties of the forecast error $y_{t+h} - \mathsf{E}_t(y_{t+h})$. From the two equations above, we obtain

$$y_{t+h} - \mathsf{E}_t(y_{t+h}) = \sum_{j=0}^{h-1} \psi_j u_{t+h-j}.$$

Note, in particular, that the forecast error of the one-step forecast is u_{t+1}.

- Taking (unconditional) expectations from both sides of the last equation shows that the forecast $\mathsf{E}_t(y_{t+h})$ is unbiased in the sense that the forecast error has an expected value of zero:

$$\mathsf{E}\left[y_{t+h} - \mathsf{E}_t(y_{t+h})\right] = 0.$$

- Furthermore, we can compute the variance of the forecast error. With straightforward computation, we obtain

$$\mathsf{Var}\Big(y_{t+h} - \mathsf{E}_t(y_{t+h})\Big) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2 \equiv \sigma_h^2.$$

- Importantly, note that as h increases, the forecast converges (in mean square) to the mean of the process, whereas the variance of the forecast error σ_h^2 converges to the variance of the process being predicted (y_t), namely $\sigma^2 \sum_{j=0}^{\infty} \psi_j^2$ (cf. the properties of the ARMA process).

If one assumes that $u_t \sim \mathsf{nid}\,(0, \sigma^2)$, then

$$y_{t+h} - \mathsf{E}_t(y_{t+h}) \sim \mathsf{N}\,(0, \sigma_h^2)$$

and, therefore, y_{t+h} is contained in the interval

$$\mathsf{E}_t(y_{t+h}) \pm 1.96\sigma_h$$

with 95% probability (in repeated sampling). In practice, the unknown parameters in the forecast $\mathsf{E}_t(y_{t+h})$ and in the standard deviation σ_h are replaced by their estimates, and hence the normality of the forecast errors hold only approximately.

Empirical example (continued). In the previous two sections, we concluded that an AR(2) model to be one possible candidate for the U.S real GDP growth (1985:Q1–2007:Q2). The estimated AR(2) model

$$y_t = \underset{(0.197)}{1.693} + \underset{(0.101)}{0.160} y_{t-1} + \underset{(0.101)}{0.287} y_{t-1} + \hat{u}_t, \quad \hat{\sigma}^2 = 3.536,$$

is now used to compute 8-quarter-ahead predictions. Starting from the last observation (2007:Q2), forecasts are obtained with the forecasting formulae above applied in the case of $p = 2$ (and $q = 0$) and $h = 8$. The resulting (point) forecasts of the estimated AR(2) model and their 80 and 95% confidence intervals are presented below.

```
        Point Forecast      Lo 80    Hi 80      Lo 95    Hi 95
2007 Q3       2.427727 0.01787587 4.837578 -1.2578223 6.113277
2007 Q4       2.781453 0.34098863 5.221917 -0.9509150 6.513820
2008 Q1       2.834604 0.28036903 5.388840 -1.0717615 6.740971
2008 Q2       2.944709 0.38002888 5.509388 -0.9776305 6.867048
2008 Q3       2.977582 0.40040373 5.554759 -0.9638718 6.919035
2008 Q4       3.014464 0.43506842 5.593860 -0.9303814 6.959310
2009 Q1       3.029805 0.44884169 5.610767 -0.9174375 6.977047
2009 Q2       3.042852 0.46149492 5.624209 -0.9049929 6.990696
```

The figure below depicts the 8-step forecasts and the confidence intervals. As we can see, the forecast converges towards the mean (3.10) when the forecast horizon lengthens. The confidence intervals are quite broad.

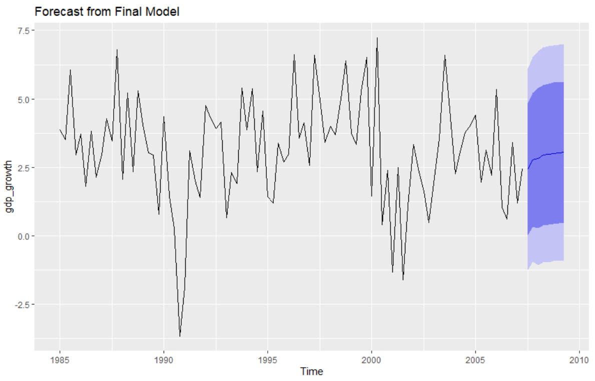


Figure: 8-step-ahead forecasts for the U.S. real GDP growth and their 80 and 95% confidence intervals from the AR(2) process.

Obviously these forecasts, constructed at 2007:Q2, for the years 2008 and 2009 turned out to be too optimistic due to the Great Financial Crisis (2008–2009) and possible endpoint of the Great Moderation! It can generally be concluded that there were no econometric model able to predict such a depression. Therefore, these presented forecasts should be seen as a scenario type of forecasts how the U.S. economy could have been evolved without the financial crisis.

8.4 Introduction to out-of-sample forecasting

In various fields, particularly in economics and finance, it is crucial to assess the reliability and accuracy of forecasts generated by estimated time series models. A popular method for this evaluation is (pseudo) **out-of-sample forecasting**. This approach aims to mimic a real-time forecasting scenario as closely as possible, using only information that would have been available at the time the forecast was made.

The fundamental idea is to evaluate the forecast performance by systematically reserving a portion of the data, unknown to the model during its estimation, to serve as a "future" against which forecasts are compared. This method is termed at times "pseudo" out-of-sample forecasting because, while we are simulating real-time forecasting, we do have full knowledge of the entire dataset beforehand in our analysis. In a "true" real-time forecasting, future observations are, by definition, unknown.

- Notice that **backtesting** in the financial industry is essentially a form of out-of-sample forecasting applied specifically to financial data and trading strategies. While the underlying principle is the same, the terminology and focus can differ slightly. While a time series analyst might use out-of-sample forecasting to evaluate how accurately they can predict a stock's price, a financial analyst would use backtesting to see if a strategy based on those predictions would have actually made money.

The out-of-sample forecasting process generally involves the following steps:

1. **Define initial estimation sample**: An initial segment of the time series, say the first T observations $(y_1, \dots, y_T)$, is used to estimate the model's

parameters. This period is often referred to as the **training sample** or **estimation sample**.

2. **Define forecasting horizon (h)**: Like above where we introduced how to construct forecasts with ARMA models, we specify how many periods ahead we want to forecast (e.g., $h = 1$ for one-period-ahead, $h = 2$ for two-periods-ahead, etc.).

3. **Define forecasting evaluation period**: A subsequent segment of the data, known as the **test sample** or **forecasting sample**, is set aside. This period spans from y_{T+1} up to $y_{T+m+h-1}$, where m is the number of distinct **forecast origins** (i.e., the number of times we re-estimate the model and generate a new h-step forecast). In other words, with this notation, we assume that we have observations $y_1, \dots, y_T, y_{T+1}, \cdots, y_{T+m+h^{\max}-1}$ where $h^{\max}$ is the maximum or the considered forecast horizon.

4. **Iterative forecasting and parameter updating**: The core of (pseudo) out-of-sample forecasting is its iterative nature. We start by estimating the model using the initial T observations, then we construct forecasts h periods ahead. We then "advance time" by one period, potentially re-estimate the model with updated information, and generate a new h-period-ahead forecast. This process is repeated m times.

Let's clarify the notation with an example. Suppose we want to evaluate two-period-ahead forecasts ($h = 2$).

- Let T denote the end of our *initial* estimation sample.
- Let j be an index for the forecast origin, running from $j = 0, 1, \dots, m - 1$. For simplicity, let $m = 4$.

The sequence of forecasts at different forecast origins and their corresponding actual values would be:

- Forecast origin T (i.e., $j = 0$):
  - Estimate the model using data $y_1, \dots, y_T$.
  - Generate a $h = 2$ period forecast, denoted $\hat{y}_{T+2|T}$.
  - This forecast is compared against the actual observation y_{T+2}.
  - The forecast error is $\hat{e}_{T+2} = y_{T+2} - \hat{y}_{T+2|T}$.
- Forecast origin $T + 1$ (i.e., $j = 1$):
  - Update the estimation sample. This is where rolling vs. expanding parameter updating window (see below) becomes relevant.
  - Generate a $h = 2$-period forecast, denoted $\hat{y}_{T+3|T+1}$.

  - This forecast is compared against the actual observation y_{T+3}.
  - The forecast error is $\hat{e}_{T+3} = y_{T+3} - \hat{y}_{T+3|T+1}$.

- Forecast origin $T + 2$ (i.e., $j = 2$):
  - Update the estimation sample.
  - Generate a $h = 2$-period forecast, denoted $\hat{y}_{T+4|T+2}$.
  - This forecast is compared against the actual observation y_{T+4}.
  - The forecast error is $\hat{e}_{T+4} = y_{T+4} - \hat{y}_{T+4|T+2}$.

- Forecast origin $T + 3$ (i.e., $j = 3$):
  - Update the estimation sample.
  - Generate a $h = 2$-period forecast, denoted $\hat{y}_{T+5|T+3}$.
  - This forecast is compared against the actual observation y_{T+5}.
  - The forecast error is $\hat{e}_{T+5} = y_{T+5} - \hat{y}_{T+5|T+3}$.

In general, for $j = 0, 1, \ldots, m - 1$, we compute the h-period forecast from the information available at $T + j$, denoted $\hat{y}_{T+h+j|T+j}$. The corresponding forecast error is:

$$\hat{e}_{T+h+j} = y_{T+h+j} - \hat{y}_{T+h+j|T+j}$$

Parameter updating strategies. In a real forecasting situation, the information set used for model estimation typically includes the most recent observations. The way this information set is updated over time defines different parameter estimation strategies:

- **Expanding window**: In this approach, the estimation sample grows by one observation in each forecast computation iteration. If the initial estimation sample is $y_1, \ldots, y_T$, the next estimation sample will be $y_1, \ldots, y_{T+1}$, then $y_1, \ldots, y_{T+2}$, and so on. This means older observations are always retained. This is suitable when one believes that all past data is equally relevant for estimating the current parameters.

- **Rolling window**: Here, the estimation sample has a fixed size (e.g., T observations) and "rolls" forward by one observation in each iteration. If the initial sample is $y_1, \ldots, y_T$, the next sample will be $y_2, \ldots, y_{T+1}$, then $y_3, \ldots, y_{T+2}$, and so on. This approach gives more weight to recent observations and is useful when model parameters are suspected to change over time (i.e., the presence of structural breaks or regime changes).

The accompanying R code (R Lab below) allows you to select either a `rolling` or `expanding` window strategy using the `window_type` parameter.

Evaluating forecast accuracy. Once the forecast errors ($\hat{e}_{T+h+j}$) for the entire test sample are computed, a typical measure for evaluating forecast accuracy is the **Mean Squared Forecast Error (MSFE)**:

$$MSFE = \frac{1}{m}\sum_{j=0}^{m-1} \hat{e}^2_{T+h+j}.$$

The MSFE penalizes larger errors more heavily due to the squaring forecast errors. Its square root, the **Root Mean Squared Forecast Error (RMSFE)**, is also commonly used as it is in the same units as the original series, making it easier to interpret.

Another popular alternative is the **Mean Absolute Forecast Error (MAFE)**:

$$MAFE = \frac{1}{m}\sum_{j=0}^{m-1} |\hat{e}_{T+h+j}|.$$

The MAFE is less sensitive to outliers than the MSFE.

The objective of this exercise is to identify the model(s) that produce the smallest MSFE and/or MAFE values, indicating superior forecasting performance.

Empirical example (continue). Let us continue an illustration of out-of-sample forecasting of the quarterly U.S. real GDP growth. Assume that we start out-of-sample forecasting in 2010:Q1 and construct one ($h = 1$) and four ($h = 4$) quarter out-of-sample forecasts for the forecasting sample (test) sample period 2010:Q1–2019:Q4. We compare the performance of AR(1), AR(2), AR(3) and AR(4) models.

- The first estimation sample is between 1985:Q1–2009:Q4.
- Obviously, based on the previous results (for the sample period excluding the Great Financial Crisis) suggest the AR(2) model.
- Parameter updating is done via expanding and rolling window approaches separately.
- MSFEs suggest that except $h = 4$ horizon and rolling window approach (there AR(4) model), AR(3) model yields the smallest MSFE in different comparisons.

```
--- MSFE Results for AR(1), AR(2), AR(3), AR(4) (h=1, expanding window) ---
$`AR(1)`
[1] 3.010018
$`AR(2)`
[1] 2.810865
```

```
$`AR(3)`
[1] 2.760154
$`AR(4)`
[1] 3.451758

--- MSFE Results for AR(1), AR(2), AR(3), AR(4) (h=1, rolling window) ---
$`AR(1)`
[1] 3.032567
$`AR(2)`
[1] 2.793521
$`AR(3)`
[1] 2.731007
$`AR(4)`
[1] 3.481733

--- MSFE Results for AR(1), AR(2), AR(3), AR(4) (h=4, expanding window) ---
$`AR(1)`
[1] 2.628715
$`AR(2)`
[1] 2.662022
$`AR(3)`
[1] 2.457973
$`AR(4)`
[1] 2.467844

--- MSFE Results for AR(1), AR(2), AR(3), AR(4) (h=4, rolling window) ---
$`AR(1)`
[1] 2.602836
$`AR(2)`
[1] 2.633853
$`AR(3)`
[1] 2.496958
$`AR(4)`
[1] 2.478106
```

Chapter 9

Volatility modelling: AR-GARCH model

9.1 Modelling conditional variance

ARMA models can be characterized as models for the conditional mean of a stationary process.

- When the invertibility condition holds, ARMA processes have the AR(∞) representation from which it can be seen that the conditional expected value of y_t, conditional on the past of the process $\{y_{t-1}, y_{t-2}, ...\}$, is $\mathsf{E}_{t-1}(y_t) = \mu - \sum_{j=1}^{\infty} \pi_j (y_{t-j} - \mu)$.

The conditional variance of an ARMA process (model) y_t is

$$\begin{aligned} \mathsf{Var}_{t-1}(y_t) &= \mathsf{E}_{t-1}(y_t - \mathsf{E}_{t-1}(y_t))^2 \\ &= \mathsf{E}_{t-1}(u_t^2) \\ &= \mathsf{E}(u_t^2) \qquad \text{(CEV2)} \\ &= \sigma^2. \end{aligned}$$

This shows that evident systematic variation in conditional variance, that several time series contain, cannot be taken into consideration with an ARMA model. Typical examples of such time series are financial time series and especially different asset return series.

Empirical example. The NASDAQ 100 (ticker symbol ^NDX) is a major stock market index that tracks the performance of 100 of the largest non-financial companies listed on the NASDAQ stock market (U.S. stock market).

- It includes 100 of the largest domestic and international companies listed on NASDAQ, weighted by a modified market capitalization method to limit the influence of the very large firms. The index excludes financial companies and is heavily concentrated in the technology sector and dominated by a few extremely large companies like (as in 2025) NVIDIA Corporation (NVDA), Microsoft Corp. (MSFT), Apple Inc. (AAPL) and Amazon.com Inc. (AMZN).
- The NASDAQ 100 is widely viewed as the benchmark for large-cap growth stocks.
- The data source: quantmod package in R, which accesses publicly available financial data—seemingly sourced from Yahoo Finance.

Let us consider excess stock returns of the NASDAQ 100 over the 3-Month Treasury Bill rate. As a simple approximation, we can think that we are interested in the percentage changes (log-differences) of the NASDAQ 100 index adjusted for risk-free rate return (see Transformations in Section 1). Below we depict the daily excess stock returns between 1.1.2003–30.9.2025.

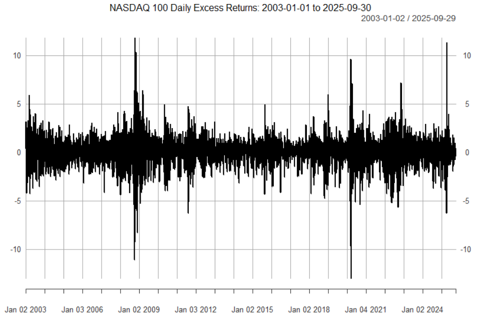


Figure: Excess stock returns of the NASDAQ 100 stock market index. (source: R package quantmod)

The return series seems quite stationary in terms of its level. Below we depict the sample autocorrelation and partial autocorrelation functions for the first 40 lags and their approximate 95% critical bounds. These suggest that there is some autocorrelation but its degree is not very high.

- Typically asset returns are almost non-autocorrelated
- For this sample period and NASDAQ returns, the resulting Ljung-Box test statistics reject the null hypothesis of no autocorrelation with all the relevant statistical significance levels.

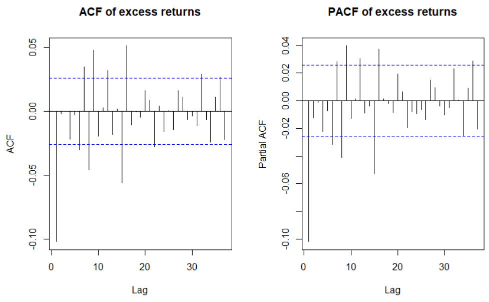


Figure: Sample autocorrelations and partial autocorrelations of the NASDAQ 100 excess returns.

The asset return series exhibits, however, some even more clear and typical variation, which is reflected by the autocorrelation function of the squared observations: All the reported autocorrelation coefficients of squared observations exhibit clear positive autocorrelation exceeding the 95%-critical bound. The approximate p-value of the McLeod Li test with $H = 10$, and also with other lag length selections, are zero with four decimal precision.

- There are periods (see above) during which the variation of the series is either larger or smaller than on average. This “volatility clustering” is certain type of heteroskedasticity.
- When the definition of variance is taken into account, it is natural to investigate this using the autocorrelation function of the squared observations. Notable are also the large difference between large and small absolute values, which suggests a more fat-tailed distribution than the normal distribution.

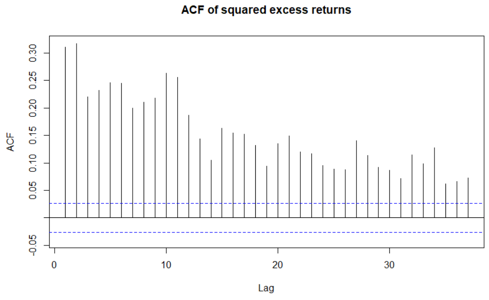


Figure: Sample autocorrelations for the squared stock return observations.

In what follows, heteroskedasticity similar to that of the stock return series above is thought to be not related with changes in unconditional variance, but rather with changes in the **conditional variance** when conditioning on the past values of the series. Next, we will first consider some general aspects of models used in modelling conditional variance, and then focus on some particular models that are most common in practice.

9.2 Model formulation

In this section, we consider an AR-GARCH model determined by the following two equations

$$y_t = \nu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + u_t,$$

$$u_t = h_t^{1/2} \varepsilon_t, \quad \varepsilon_t \sim \mathsf{iid}\,(0,1)\,,$$

where h_t is a function of the variables u_{t-j}, $j > 0$, and the error term ε_t is assumed to be independent of the variables y_{t-j}, $j > 0$, and hence also of the variables u_{t-j}, $j > 0$. In other words, we consider an **AR(p) model whose error term is conditionally heteroskedastic**.

- The coefficients $\phi_1, \dots, \phi_p$ are assumed to satisfy the (sufficient) stationarity condition of the AR(p) process: $1 - \phi_1 z - \dots - \phi_p z^p = \phi(z) \neq 0$, when $|z| \leq 1$.

Regarding the conditional variance, by its definition, the variance of a random variable is associated with the squares of the random variable. Therefore, it seems natural that the conditional variance h_t would depend on the past squared values of the process. For concreteness sake, consider a general **GARCH(r,s) model**

$$h_t = \omega + \beta_1 h_{t-1} + \dots + \beta_r h_{t-r} + \alpha_1 u_{t-1}^2 + \dots + \alpha_s u_{t-s}^2,$$

whose parameters are assumed to satisfy the required conditions for non-negativeness, identification, and strict stationarity (to be considered below).

As before, we denote the conditional expectation as $\mathsf{E}_{t-1}(\cdot) = \mathsf{E}(\cdot\,|y_s,\ s \leq t-1)$. Because u_t is a function of the variables y_t, ..., y_{t-p}, the conditional variance h_t is a function of the variables y_{t-j}, $j > 0$. Therefore, we can use the same arguments as considered, e.g., in forecast construction (see CEV4) to conclude that in the AR-GARCH model

$$\mathsf{E}_{t-1}(u_t) = h_t^{1/2} \mathsf{E}_{t-1}(\varepsilon_t) = h_t^{1/2} \mathsf{E}(\varepsilon_t) = 0.$$

Together with this result, the model equations of the AR-GARCH model can be used to justify the following two results

$$\mathsf{E}_{t-1}(y_t) = \nu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} \quad \text{and} \quad \mathsf{Var}_{t-1}(y_t) = h_t.$$

- The latter result can be justified by noticing that $y_t = \mathsf{E}_{t-1}(y_t) + u_t$ and that

$$\begin{aligned}\mathsf{Var}_{t-1}(y_t) &= \mathsf{E}_{t-1}(y_t - \mathsf{E}_{t-1}(y_t))^2 \\ &= \mathsf{E}_{t-1}(u_t^2) \\ &= h_t \mathsf{E}(\varepsilon_t^2) \\ &= h_t,\end{aligned}$$

  where the result CEV2 and the identity $u_t^2 = h_t \varepsilon_t^2$ is used (see the general model definition above).

These two results demonstrate that the conditional mean and conditional variance of the process y_t depend on the past values of the process. Based on the above points, **building a volatility model**, that is selecting the model equation for h_t, roughly consists of the following steps:

- Finding an adequate specification for the conditional mean $\mathsf{E}_{t-1}(y_t)$ (e.g., to select a suitable AR or ARMA model, and including also possible deterministic terms) is necessary to obtain a suitable specification for the conditional variance.

- Checking for the (conditional) heteroskedasticity of the error term. This can be based on residuals, as they are empirical counterparts of the error terms. That is testing “ARCH effects” using, e.g., the McLeod-Li test.

- Finding a sufficient specification for the conditional variance h_t. There are a lot of different alternative model specifications to GARCH(r, s) suggested in the econometric literature. Replacing it with some of the alternatives yield a straightforward extension of the AR-GARCH model introduced above.

- Estimation of the full model can be carried out by the method of maximum likelihood (see details below).

Extra: Relation to risk management and Value-at-Risk (VaR)

The core motivation for modeling conditional heteroskedasticity (time-varying volatility h_t) in asset returns y_t is its direct and profound impact on risk management, particularly in calculating measures like Value-at-Risk (VaR).

The question that often arises in risk management is: What is the maximum expected loss over a specific time horizon with a given confidence level? This loss is the Value-at-Risk (VaR) threshold c. The VaR threshold c is the loss level such that the probability of the actual loss y_{t+1} being less than or equal to c is a fixed, small probability π (e.g., 1% or 5%):

$$P_t(y_{t+1} \leq c) = G\left(\frac{c - \mathsf{E}_t(y_{t+1})}{h_{t+1}^{1/2}}\right) = \pi,$$

where $P_t(\cdot)$ denotes the conditional probability given information up to time t, $\mathsf{E}_t(y_{t+1})$ is the conditional mean (return forecast), h_{t+1} is the conditional variance forecast (volatility), and $G(\cdot)$ is the cumulative distribution function (CDF) of the standardized innovation $\frac{y_{t+1} - \mathsf{E}_t(y_{t+1})}{h_{t+1}^{1/2}}$.

The Value-at-Risk is the most commonly used risk measure. For the VaR model to be a useful and reliable tool in practice, the specification of the conditional variance h_t is critical. Ignoring or otherwise misspecifying the time-varying nature of volatility (for example, treating h_t as a constant) leads to severely flawed risk estimates.

9.3 GARCH(1,1) and ARCH(s) models

Instead of the general GARCH(r, s) model, the **GARCH(1,1) model**

$$h_t = \omega + \beta_1 h_{t-1} + \alpha_1 u_{t-1}^2$$

has been found adequate for most (financial) time series data.

- In this GARCH(1,1) case, the non-negativeness of h_t requires $\omega > 0$ and $\alpha_1, \beta_1 \geq 0$. Moreover, for β_1 to be identified, α_1 must be strictly positive ($\alpha_1 > 0$).
- In the GARCH(r, s) model, these non-negativeness conditions are clearly more complicated.

Another special case of GARCH models is obtained when $r = 0$. That is the "GARCH-part" is missing and the GARCH(r, s) model reduces to an **ARCH(s)** model

$$h_t = \omega + \sum_{i=1}^{s} \alpha_i u_{t-1}^2.$$

In line with the idea of capturing **volatility clustering** with ARCH and GARCH models, the periods of high (and low) conditional variance tend to persist. This is easy to see with the **ARCH(1) model** ($s = 1$)

$$h_t = \omega + \alpha_1 u_{t-i}^2,$$

where $\omega, \alpha_1 \geq 0$. A large shock (u_{t-1}^2) increases h_t, which then subsequently increases u_t^2 and h_{t+1}. This same logic holds also for more general ARCH and GARCH models. Moreover, in the ARCH(1) model, assuming normality of ε_t, the (unconditional) kurtosis of u_t is

$$\frac{\mathrm{E}(u_t^4)}{\mathrm{E}(u_t^2)^2} = \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2}.$$

Kurtosis is finite (i.e. the 4th moment exists) if $3\alpha_1^2 < 1$ and larger than 3 implied by the normal distribution. The ARCH(1) model is hence capable of capturing excess kurtosis often present in financial data.

- Large outliers (in absolute value) appear more often than implied by the nid innovations in (observed) asset returns.
- Similar result on excess kurtosis holds also for more general ARCH and GARCH models, but the formulae become more complicated.

As a summary of large past research, the GARCH(1,1) model has generally been found a successful and parsimonious alternative to the ARCH(s) model with a typically large s for asset returns. Why? By recursive substitutions, we get

$$\begin{aligned}
h_t &= \omega + \beta_1 h_{t-1} + \alpha_1 u_{t-1}^2 \\
&= \omega + \beta_1\omega + \beta_1^2 h_{t-2} + \alpha_1\beta_1 u_{t-2}^2 + \alpha_1 u_{t-1}^2 \\
&\vdots \\
&= \omega\sum_{j=0}^{k}\beta_1^j + \alpha_1\sum_{j=0}^{k}\beta_1^j u_{t-1-j}^2 + \beta_1^{k+1}h_{t-k-1}.
\end{aligned}$$

This suggests that in the case $\beta_1 < 1$,

$$h_t = \omega\sum_{j=0}^{\infty}\beta_1^j + \alpha_1\sum_{j=0}^{\infty}\beta_1^j u_{t-1-j}^2,$$

which is a particular kind of ARCH(∞) form. Therefore, we can conclude that GARCH(1,1) is able to capture volatility clustering in a parsimonious way via only three parameters.

Another perspective on the GARCH models, and specifically to the GARCH(1,1) model, is obtained by adding u_t^2 on both sides of the model equation. Therefore, the GARCH(1,1) can be rewritten

$$u_t^2 = \omega + (\alpha_1 + \beta_1)u_{t-1}^2 + \xi_t - \beta_1\xi_{t-1},$$

where $\xi_t = u_t^2 - h_t = h_t(\varepsilon_t^2 - 1)$ and $\mathrm{E}(\xi_t) = 0$. This is an ARMA(1,1) type of presentation for u_t^2.

- As for the ARMA(1,1) process, the condition for weak and strict stationarity is $\alpha_1 + \beta_1 < 1$ (together with restrictions $0 \leq \alpha_1, \beta_1 \leq 1$ to, e.g., guarantee the non-negativity of h_t).
- Moreover, the unconditional variance (provided stationarity) is
$$\mathrm{E}(u_t^2) = \frac{\omega}{1 - \alpha_1 - \beta_1}.$$

Extra: Special case of the AR-GARCH with zero conditional mean

The AR-GARCH model presentation above contains also the special case of $\mathsf{E}_{t-1}(y_t) = 0$ often considered in various references (books etc.). This leads to the notation
$$u_t = y_t = h_t^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathrm{iid}(0,1).$$
Therefore, in this case the GARCH(1,1) model reduces to
$$h_t = \omega + \beta_1 h_{t-1} + \alpha_1 y_{t-1}^2$$
and in the ARCH(s) case, we get
$$h_t = \omega + \sum_{i=1}^{s} \alpha_i y_{t-i}^2.$$

9.4 Parameter estimation

If it is assumed that the error term is Gaussian
$$\varepsilon_t \sim \mathsf{nid}\,(0,1)\,,$$
the (conditional) likelihood function can be derived following analogous principles as in ARMA models. In other words, assuming $\varepsilon_t \sim \mathsf{nid}(0,1)$, we get
$$u_t = h_t^{1/2}\varepsilon_t, \quad \varepsilon_t \sim \mathsf{nid}(0,1),$$
where $h_t = h_t(y_{t-1}, y_{t-2}, ...)$, ε_t and vector $(y_{t-1}, y_{t-2}, ...)$ are independent, and $u_t = y_t - \mathsf{E}_{t-1}(y_t)$. Assume also that y_t is stationary. Given the above assumptions and the conditional moments (conditional mean and variance) of the AR-GARCH model, we get the conditional density function of the observation y_t as
$$y_t|\{y_{t-j},\, j \geq 1\} \sim \mathsf{N}(\nu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}, h_t).$$
Therefore, it can be concluded that the conditional distribution of y_t, conditional on $\underline{\mathbf{Y}}_{t-1} = (y_1, ... \, y_{t-1})$, $t = 1, \ldots, T$, is Gaussian with the conditional mean (see

above) $\nu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}$ and conditional variance h_t where the model equation for h_t needs to be specified.

Using the observed time series and required initial values (depending on the specific model) $\ldots, y_{-1}, y_0, y_1, \ldots, y_T$, as in the parameter estimation of ARMA models, this leads to the conditional joint density function

$$\prod_{t=1}^{T} f_{y_t|\mathbf{Y}_{t-1}} = (2\pi)^{-T/2} \cdot \prod_{t=1}^{T} h_t^{-1/2} \cdot \exp\Big(-\frac{1}{2} \sum_{t=1}^{T} \frac{u_t^2}{h_t} \Big).$$

Denote the parameter vector of the AR-GARCH model $\vartheta = (\phi, \lambda)$ where

- $\phi = (\nu, \phi_1, \ldots, \phi_p)$ contains parameters related to the conditional mean, and
- $\lambda = (\omega, \alpha_1, \ldots, \alpha_s, \beta_1, \ldots, \beta_r)$ contains the parameters of the model for the conditional variance.

The log-likelihood then becomes (omitting constant terms not dependent on the model parameters)

$$l(\vartheta) = \sum_{t=1}^{T} l_t(\vartheta) = -\frac{1}{2} \sum_{t=1}^{T} \log h_t(\phi, \lambda) - \frac{1}{2} \sum_{t=1}^{T} \frac{u_t(\phi)^2}{h_t(\phi, \lambda)},$$

with $u_t(\phi) = y_t - \nu - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p}$ and $h_t(\phi, \lambda)$ are functions of the parameters ϕ and λ.

- As an example, in the GARCH(1,1) case $r = s = 1$ and hence $h_t = \omega + \beta_1 h_{t-1} + \alpha_1 u_{t-1}^2$, for the conditional variance. This means that $\lambda = (\omega, \alpha_1, \beta_1)$.

Maximization of the likelihood function is performed using numerical methods similar to the ARMA models.

If the normality assumption of the error term $\varepsilon_t \sim \mathsf{nid}(0, 1)$ made above holds, the usual asymptotic properties of the maximum likelihood estimator (consistency and asymptotic normality) hold. If the Gaussianity assumption is found inappropriate, one could either (or both):

- Use a non-Gaussian distribution, like the t-distribution for the error term. This leads to the different form for the log-likelihood function, but the general lines to obtain it are the same as above.
- Modify the asymptotic distribution and interpret the maximum likelihood estimator (MLE) as **quasi-MLE (QMLE)**.

Concerning the QMLE approach, if the likelihood function is maximized at $\hat{\vartheta} = (\hat{\phi}, \hat{\lambda})$, it can be shown, even without the normality assumption and under general assumptions and regularity conditions, it holds

$$\hat{\vartheta} \underset{as}{\sim} \mathsf{N}\left(\vartheta, \mathbf{V}\left(\vartheta\right)^{-1}\mathbf{B}\left(\vartheta\right)\mathbf{V}\left(\vartheta\right)^{-1}\right),$$

where

$$\mathbf{V}\left(\vartheta\right) = \mathsf{E}\left[-\partial^2 l(\vartheta)/\partial\vartheta\partial\vartheta'\right]$$

and

$$\mathbf{B}\left(\vartheta\right) = \mathsf{E}\left[\sum_{t=1}^{T}\left(\frac{\partial}{\partial\vartheta}l_t(\vartheta)\right)\left(\frac{\partial}{\partial\vartheta}l_t(\vartheta)\right)'\right],$$

where $l_t(\vartheta)$ is as described above and its specific form naturally depends on the selected AR-GARCH model.

Using the empirical counterparts of the matrices $\mathbf{V}\left(\vartheta\right)$ and $\mathbf{B}\left(\vartheta\right)$, namely

$$\widehat{\mathbf{V}}(\hat{\vartheta}) = -\partial^2 l(\hat{\vartheta})/\partial\vartheta\partial\vartheta' \quad \text{and} \quad \widehat{\mathbf{B}}(\hat{\vartheta}) = \sum_{t=1}^{T}\left(\frac{\partial}{\partial\vartheta}l_t(\hat{\vartheta})\right)\left(\frac{\partial}{\partial\vartheta}l_t(\hat{\vartheta})\right)'$$

the asymptotic distribution of the estimator $\hat{\vartheta}$ presented above can be used to form approximate standard errors (when taking the square roots of the diagonal elements of the above matrix) and Wald tests about the parameter ϑ.

- Cf. the maximum likelihood estimation of AR (and ARMA) models.
- If the normality assumption (i.e. $\varepsilon_t \sim \mathsf{nid}(0,1)$) holds, then $\mathbf{V}\left(\vartheta\right) = \mathbf{B}\left(\vartheta\right)$, and the expressions above for the (QMLE) asymptotic distribution and the standard errors and Wald tests simplify to the usual MLE (maximum likelihood estimator) case.

The practical message from above is that it is often reasonable to rely on the QMLE-based asymptotic distribution result and resulting robust standard errors, such as so called Bollerslev-Wooldridge standard errors, for different parameter estimates of the AR-GARCH model.

- The rationale is that model specification, including especially the distribution assumption $\varepsilon_t \sim \mathsf{nid}(0,1)$, is that even though the model is not entirely correctly specified, allowing for the additional robustness to the possible misspecification through QMLE estimates is advisable.
- Using the QMLE specifically means that we obtain the same estimates $\hat{\vartheta}$ but we rely on the asymptotic covariance matrix and hence the resulting robust standard errors.

Empirical example (continue). Consider the estimation result of the AR-GARCH model with Gaussian error term for the excess stock returns of the NASDAQ 100 index. For simplicity, fit a relatively simple AR(1)-GARCH(1,1) model (see above with selections $p = r = s = 1$). That is, we specify

- AR(1) model for the conditional mean.
- GARCH(1,1) model for the conditional variance

That is, we estimate an AR(1) model with GARCH(1,1) errors. The estimation result of the maximum likelihood estimation, based on the normality assumption of ε_t, yields

$$\begin{aligned} y_t &= \underset{(0.013)}{0.087} - \underset{(0.013)}{0.044} y_{t-1} + \hat{u}_t, \\ \hat{h}_t &= \underset{(0.006)}{0.035} + \underset{(0.013)}{0.876} \hat{h}_{t-1} + \underset{(0.012)}{0.105} \hat{u}^2_{t-1}, \end{aligned}$$

where we report the robust standard errors under the parameter estimates (see the QMLE asymptotic distribution result). The full estimation result provided by **rugarch package** in R yields the following results:

```
*---------------------------------*
*          GARCH Model Fit        *
*---------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model : sGARCH(1,1)
Mean Model  : ARFIMA(1,0,0)
Distribution    : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.087229    0.012889   6.7679 0.000000
ar1    -0.043797    0.014149  -3.0954 0.001965
omega   0.035129    0.004868   7.2160 0.000000
alpha1  0.104804    0.008170  12.8272 0.000000
beta1   0.876000    0.008975  97.6078 0.000000

Robust Standard Errors:
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.087229    0.012567   6.9412 0.000000
```

```
ar1    -0.043797    0.012630  -3.4677 0.000525
omega   0.035129    0.006497   5.4066 0.000000
alpha1  0.104804    0.012428   8.4331 0.000000
beta1   0.876000    0.012869  68.0716 0.000000

LogLikelihood : -9030.906

Information Criteria
------------------------------------


Akaike       3.1583
Bayes        3.1641
Shibata      3.1583
Hannan-Quinn 3.1603

Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                        statistic p-value
Lag[1]                     0.1471  0.7014
Lag[2*(p+q)+(p+q)-1][2]    0.6119  0.9317
Lag[4*(p+q)+(p+q)-1][5]    2.0557  0.6952
d.o.f=1
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                        statistic p-value
Lag[1]                      1.998  0.1575
Lag[2*(p+q)+(p+q)-1][5]     4.486  0.1993
Lag[4*(p+q)+(p+q)-1][9]     6.661  0.2292
d.o.f=2

Weighted ARCH LM Tests
------------------------------------
            Statistic Shape Scale P-Value
ARCH Lag[3]   0.09937 0.500 2.000  0.7526
ARCH Lag[5]   4.68135 1.440 1.667  0.1217
ARCH Lag[7]   5.49080 2.315 1.543  0.1794
```

The estimation result shows well the point using the QMLE:

- The estimated coefficients (“AR”,“ARCH” and “GARCH” estimates) are the same for “optimal paramaters” (that is, when we assume that the normality assumption ε_t holds) and for QMLE (resulting robust estimation results), but the estimated standard errors are different. Especially for the

GARCH part this means, as often in these circumstances, that the robust standard errors are somwhat higher reflecting the additional uncertainty coming from possible model misspecification.

- All in all, all the estimated coefficients are statistically significant at the conventional significance levels. Moreover, the typical pattern of the GARCH(1,1) model is also present where the GARCH parameter (here β_1) is larger than the ARCH coefficient α_1, and their sum is relatively close to 1. Estimated volatility, which is here the conditional standard deviation $\hat{h}_t^{1/2}$) of the estimated model, and at times interpreted as a risk in financial markets is depicted below.

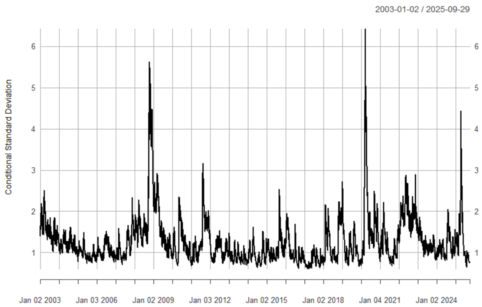


Figure: Estimated conditional standard deviation based on the AR(1)-GARCH(1,1) model for the NASDAQ 100 excess returns.

Residual diagnostics show that:

- There is no remaining autocorrelation in the residuals and squared residuals (based on the Ljung-Box and the McLeod-Li tests and residual autocorrelation coefficients).

- Without presenting details, (weighted) ARCH LM test tests the null hypothesis that there is no remaining conditional heteroskedasticity in the residuals $\hat{u}_t$. In the estimation result above, the p-values of the ARCH LM tests show that the parsimonious AR(1)-GARCH(1,1) is an adequate model.

- There is some deviation from the normality assumption in the (standardized) residuals. This suggests that the estimates should be interpreted as QMLEs, as above.

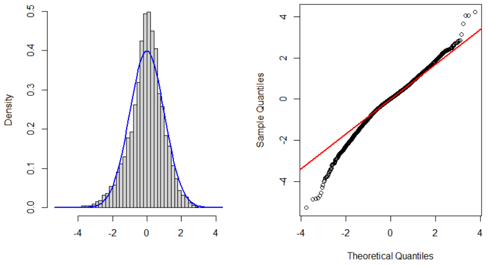


Figure: Histogram and Q-Q plot of (standardized) residuals.

9.5 Forecasting

Assume observations are available up to time t and the object of interest is to forecast the future values of y_{t+k} $(k \geq 1)$. In addition to stationarity, assume that $\mathsf{E}\left(y_t^2\right) < \infty$.

Forecasts for y_t. From the first equation of the AR(p)-GARCH model, that is the model for the conditional variance, it can be seen that the optimal (in the mean square sense) one-step-ahead forecast is

$$\mathsf{E}_t\left(y_{t+1}\right) = \nu + \phi_1 y_t + \cdots + \phi_p y_{t+1-p}.$$

When $k \geq 2$, it can be seen that (cf. forecasting formulae for AR(MA) models)

$$\mathsf{E}_t(u_{t+k}) = \mathsf{E}_t\left[\mathsf{E}_{t+k-1}(u_{t+k})\right] = 0,$$

so that

$$\mathsf{E}_t\left(y_{t+k}\right) = \nu + \phi_1 \mathsf{E}_t(y_{t+k-1}) + \cdots + \phi_p \mathsf{E}_t(y_{t+k-p}), \quad k = 1, 2, \ldots,$$

where $\mathsf{E}_t(y_{t+k-j}) = y_{t+k-j}$ for $j \geq k$.

- In conclusion, from these and the forecasting formulae for an AR(p) process with a homoskedastic error term, it can be concluded that the optimal forecasts can be formed recursively exactly in the same manner as in the (conditionally) homoskedastic case.
- In practice, when computing forecasts, unknown parameters naturally need to be replaced with their estimates, which are based on the finite sample sizes, and hence numerically here forecasts can be slightly different than obtained with an AR(p) model with conditionally homoskedastic errors.

As the above shows, conditional heteroskedasticity does not affect the forecasting formulae, which are the same as in the AR case discussed earlier. However, the presence of conditional heteroskedasticity changes how the prediction intervals are computed (see Extra below).

Extra: Prediction intervals under conditional heteroskedasticity

To see this, note that because the forecasts can be derived exactly as in the homoskedastic AR(p) model, the same also holds for the k-step forecast errors. Based on the calculations above,

$$\mathsf{E}_t\left(u_{t+k-j}\right) = \begin{cases} u_{t+k-j}, & \text{when} \quad k \leq j \\ 0, & \text{when} \quad k > j, \end{cases}$$

so that

$$y_{t+k} - \mathsf{E}_t(y_{t+k}) = \sum_{j=0}^{k-1} \psi_j u_{t+k-j} = \sum_{j=0}^{k-1} \psi_j h_{t+k-j}^{1/2} \varepsilon_{t+k-j},$$

where ψ_j are again the coefficients of the power series $\psi(z) = \phi(z)^{-1} = \sum_{j=0}^{\infty} \psi_j z^j$ $(\psi_0 = 1)$.

Let us focus our attention on the one-step-ahead forecast error, whose conditional expected value becomes

$$\mathsf{E}_t\left[y_{t+1} - \mathsf{E}_t(y_{t+1})\right] = \mathsf{E}_t(u_{t+1}) = 0$$

with conditional variance

$$\mathsf{E}_t\left[\left(y_{t+1} - \mathsf{E}_t\left(y_{t+1}\right)\right)^2\right] = \mathsf{E}_t\left(u_{t+1}^2\right) = h_{t+1}.$$

This shows that the one-step-ahead prediction error is conditionally heteroskedastic, that is, the forecast accuracy depends on what kind of values the process attained at the forecast origin (at time of forecast computation) and just before it.

- As can be seen from the expression of the forecast error, a similar result holds also for forecast horizons longer than one. Therefore, any analysis of forecast accuracy, such as prediction intervals is sensible to be based on the conditional distribution with conditioning on the history up until the date of forecasting.

If it is assumed that $\varepsilon_t \sim \mathsf{nid}(0, 1)$, then the conditional distribution of the one-step-ahead forecast error u_t (conditioned on $\{y_{t-1}, y_{t-2}, ...\}$) is Gaussian with mean zero and variance h_t.

- Using this result, one can form prediction intervals for the one-step-ahead forecast exactly as for ARMA process.

- For multiple step ahead forecasts, the situation is not that straightforward, because the conditional distributions of multiple step forecast errors are not Gaussian and have no simple expressions. Therefore, forming prediction intervals for multiple step ahead forecasts is more complicated.

Volatility forecasts. Concerning forecasting the conditional variance, we assume that observations up to and including time t (i.e. the forecast origin at time t) are available, and that forecasts for h_{t+1}, h_{t+2}, ...are desired.

As h_{t+1} is a (deterministic) function of the variables $y_t, y_{t-1}, \ldots$, and thus the first conditional variance we need to forecast is h_{t+2}. For simplicity, let us concentrate on the GARCH(1,1) case. Taking conditional expected values (conditional on $\{y_t, y_{t-1}, \ldots\}$) of both sides of

$$h_{t+2} = \omega + \beta_1 h_{t+1} + \alpha_1 u_{t+1}^2$$

yields the optimal (in the mean squared error sense) forecast of h_{t+2} as

$$\mathsf{E}_t\left(h_{t+2}\right) = \omega + \beta_1 \mathsf{E}_t\left(h_{t+1}\right) + \alpha_1 \mathsf{E}_t\left(u_{t+1}^2\right),$$

where $\mathsf{E}_t\left(h_{t+1}\right) = h_{t+1}$ (see CEV4). Moreover, we obtain that (see CEV2 and CEV4)

$$\mathsf{E}_t\left(u_{t+1}^2\right) = \mathsf{E}_t\left(h_{t+1}\varepsilon_{t+1}^2\right) = h_{t+1}\mathsf{E}_t\left(\varepsilon_{t+1}^2\right) = h_{t+1}\mathsf{E}\left(\varepsilon_{t+1}^2\right) = h_{t+1},$$

so that

$$\mathsf{E}_t\left(h_{t+2}\right) = \omega + (\alpha_1 + \beta_1)h_{t+1}.$$

When the forecast horizon is $k \geq 3$, in a similar fashion, we obtain

$$\mathsf{E}_t\left(h_{t+k}\right) = \omega + \beta_1 \mathsf{E}_t\left(h_{t+k-1}\right) + \alpha_1 \mathsf{E}_t\left(u_{t+k-1}^2\right),$$

and finally

$$\mathsf{E}_t\left(h_{t+k}\right) = \omega \sum_{j=0}^{k-2} \left(\alpha_1 + \beta_1\right)^j + \left(\alpha_1 + \beta_1\right)^{k-1} h_{t+1}, \quad k = 2, 3, \ldots,$$

where h_{t+1} is a function of variables $\{y_t, y_{t-1}, \ldots\}$ known at the time when forecasts are constructed.

Extra: Details for forecasting formulae multiple periods ahead

Consider the part

$$\mathsf{E}_t\left(h_{t+k}\right) = \omega + \beta_1 \mathsf{E}_t\left(h_{t+k-1}\right) + \alpha_1 \mathsf{E}_t\left(u_{t+k-1}^2\right),$$

Here

$$\mathsf{E}_t\left(u_{t+k-1}^2\right) = \mathsf{E}_t\left(h_{t+k-1}\varepsilon_{t+k-1}^2\right) = \mathsf{E}_t\left[\mathsf{E}_{t+k-2}\left(h_{t+k-1}\varepsilon_{t+k-1}^2\right)\right],$$

where the latter equality can be justified based on a generalization of property CEV3, that is, a generalization of the law of iterated expectations.

This generalization says that $\mathsf{E}(Y|X_2) = \mathsf{E}[\mathsf{E}(Y|X_1)|X_2]$, when the components of the (potentially infinite-dimensional) random vector X_2 are a subset of the components of X_1 (or more generally, X_2 is a function of X_1).

Because

$$\mathsf{E}_{t+k-2}\left(h_{t+k-1}\varepsilon_{t+k-1}^2\right) = h_{t+k-1}\mathsf{E}_{t+k-2}\left(\varepsilon_{t+k-1}^2\right) = h_{t+k-1},$$

we obtain

$$\mathsf{E}_t\left(u_{t+k-1}^2\right) = \mathsf{E}_t\left(h_{t+k-1}\right), \quad k = 3, 4, \ldots \ .$$

To summarize, we have shown that

$$\mathsf{E}_t\left(h_{t+k}\right) = \omega + (\alpha_1 + \beta_1)\mathsf{E}_t\left(h_{t+k-1}\right), \quad k = 2, 3, \ldots \ .$$

Because $\mathsf{E}_t\left(h_{t+1}\right) = h_{t+1}$, we inductively obtain the solution

$$\mathsf{E}_t\left(h_{t+k}\right) = \omega\sum_{j=0}^{k-2}\left(\alpha_1 + \beta_1\right)^j + \left(\alpha_1 + \beta_1\right)^{k-1} h_{t+1}, \quad k = 2, 3, \ldots,$$

where h_{t+1} is a function of variables $\{y_t, y_{t-1}, \ldots\}$ known at the time of forecasting.

In practice,

- the unknown parameters α_1, β_1, and ω have to be replaced by corresponding estimates.
- Unlike in ARMA models, the quantity being predicted is now unobserved, although it can be computed using GARCH model equation for all $t \geq 1$ as long as parameter values and required initial values for h_t and y_t are available. A common choice in practice is to use the sample variance of the observed time series as the initial value h_0. In the stationary case, the effect of the initial values diminishes as t increases.
- Forecasting with a more general GARCH(r, s) model is carried out in principle in the same way as outlined above, although the resulting forecasting formulae become more cumbersome. Deriving interval predictions is also complicated, one major reason for this being that the distribution of the conditional variance deviates heavily from a Gaussian distribution.
- Overall, **volatility forecasting** is a separate and a large area in financial econometrics that we are not considering in this course more detail.

9.6 GARCH-in-mean model

In the **GARCH-in-mean** (**GARCH-M**) **model**, the conditional variance is allowed to directly affect the conditional mean as well.

To simplify the notation, let us consider the first-order special case, i.e., the AR(1)-GARCH(1,1)-M model

$$y_t = \nu + \phi_1 y_{t-1} + \delta g(h_t) + u_t, \quad h_t = \omega + \alpha_1 u_{t-1}^2 + \beta_1 h_{t-1},$$

where, as before, $u_t = y_t - \nu - \phi_1 y_{t-1} - \delta g(h_t)$, but now the conditional variance h_t also affects the level of y_t through the function $g(h_t)$. Depending on the situation, the "in-mean effect" can be defined as

- $g(h_t) = h_t$,
- $g(h_t) = \sqrt{h_t}$, or
- $g(h_t) = \log(h_t)$.

A positive coefficient δ means that the value of y_t increases when the conditional variance increases.

- As an extension of the GARCH-M version presented above, the orders of the AR and GARCH models can naturally be greater than 1.
- In general, instead of the AR(p) model, another suitable model for the conditional mean of y_t can be used.
- Similarly, a model other than GARCH(1,1) can be chosen for the GARCH part.

The GARCH-M model is used, for example, in financial econometrics (empirical finance) to model the fundamental risk-return relation, where risk (here volatility, measured by the conditional variance or its transformation) is allowed to directly affect the expected return of a security.

Empirical example (continue). Let us consider an GARCH-in-mean extension of the AR(1)-GARCH(1,1) model (with Gaussian innovations) obtained above for the NASDAQ 100 excess stock returns. That is we consider the following model for the conditional mean

$$y_t = \nu + \phi_1 y_{t-1} + \delta \sqrt{h_t} + u_t.$$

That is we include the conditional standard deviation to the conditional mean.

- This is one possible way to examine the fundamental risk-return relationship in (excess) stock returns (here NASDAQ 100 index). The positive risk-return relation is the cornerstone of financial economics, as postulated, e.g., by the intertemporal capital asset pricing model (ICAPM). However, the existing empirical findings have been highly ambiguous: Different GARCH-M models have led to both positive and negative relationships (see, e.g., Engle, Lilien and Robins, 1987; Glosten et al., 1993; Scruggs, 1998; Lanne and Saikkonen, 2006; Nyberg, 2012).

QMLE-based estimation result obtained with the rugarch package on AR(1)-GARCH(1,1)-M model:

```
Robust Standard Errors:
        Estimate  Std. Error   t value Pr(>|t|)
mu      0.000963    0.043060  0.022373 0.982151
ar1    -0.043796    0.012596 -3.477072 0.000507
archm   0.085665    0.040331  2.124019 0.033669
omega   0.035116    0.006429  5.462160 0.000000
alpha1  0.104836    0.012296  8.526169 0.000000
beta1   0.875948    0.012702 68.958838 0.000000
```

Therefore, it appears that the estimated coefficient of δ, here “archm”, is positive and also statistically significant at the 5 % significance level based on the robust t-value. Even though the statistical significance is not very strong, the positive risk-return relation can be approved.

- Residual diagnostics of this GARCH-M model is essentially the same as obtained without the in-mean effect, and hence the model seems adequate.

Chapter 10

Multivariate time series models

So far, we have considered univariate models, such as AR and ARMA models, where only a single time series is analyzed.

- This represents the "pure time series approach" as formulated at the beginning of the material.

Modelling multiple time series simultaneously is, however, of greater interest in different applications. Multivariate time series models can provide richer dynamics and predictive information, potentially improving the accuracy of empirical analyses such as forecasting.

We will next briefly introduce three important extensions to the univariate models that have been widely used in the econometric (statistics) literature:

- Linear (predictive) regressions with time series variables
- Vector autoregressive (VAR) models
- Cointegration (in connection to nonstationary time series)

10.1 Predictive regressions: Background and starting point

Let us assume, for simplicity, that throughout this section we are working with two sets of variables: y_t reamins the dependent variable and either one

($M = 1$) or, more generally, multiple ($M > 1$) additional predictive variables ($x_{1t}, \dots, x_{Mt}$). We are mainly interested in a linear regression model, but now with time series, where the lags of predictive variables are used as regressors (predictors). For example, one possible model specification is

$$y_t = \beta_0 + x_{1,t-1}\beta_1 + \cdots + x_{M,t-1}\beta_M + u_t \equiv x_t'\beta + u_t,$$

where x_t contains predictors and u_t is iid error term or at least serially uncorrelated and uncorrelated with the regressors x_t. The parameters of this type of model, contained in β, can be estimated by the (conditional) least squares in a similar fashion as discussed, e.g., in connection to the AR(p model. That is minimizing the OLS criterion (with required initial values)

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=1}^{T} (y_t - x_t'\beta)^2,$$

resulting in the OLS estimator (and estimates)

$$\hat{\beta} = \Big(\sum_{t=1}^{T} x_t x_t' \Big)^{-1} \sum_{t=1}^{T} x_t y_t.$$

In contrast to the model specification above, at times we are also considering other linear regressions:

- Simultaneous values of the predictive variables—rather than their lags—can be used as regressors in certain circumstances and applications. As an example, and for simplicity, let us assume $M = 1$. In this case, the model can be specified as:

  $$y_t = \beta_0 + x_{1t}\beta_1 + u_t.$$

- Including also the lags of y_t on the right hand side of the model equation. These are then autoregressive models with additional predictive variables, often denoted by ARX or ADL (Autoregressive Distributed Lag) models. One example of such a model is

  $$y_t = \beta_0 + \phi_1 y_{t-1} + \beta_1 x_{1,t-1} + u_t,$$

  which can be straightforwardly extended by allowing more lags of y_t and x_{1t}.

- Including an autocorrelated error term is also possible. That is, particularly common in the earlier econometric literature, allows us to specify $u_t = \phi u_{t-1} + \varepsilon_t$, $\varepsilon_t \sim \mathsf{iid}(0, \sigma^2)$ (see, e.g., Verbeek, Chapters 4.6-4.7). Note that by including lags of y_t in the model, we can attempt to account fo the possible autocorrelation in the error term.

Subsequent (sub)sections are organized as follows:

- We consider the case where all the (dependent and predictive/explanatory) variables are stationary variables.
- At end of this section, we also briefly introduce the case where M is high-dimensional. That is, the number of predictive variables can be very large. In contrast, throughout the rest of this section, we restrict ourselves to the case where there is only one or only a few predictive variables (i.e. M is relatively small).
- When there are also nonstationary variables present (Section 13).

10.2 Stationary variables

Let us start with the case (assumption) that all variables y_t and $x_{1t}, \dots, x_{Mt}$ are stationary, which means that their respective time series plots do not contain strong stochastic or deterministic trends. Broadly speaking then, under the following two conditions (i) and (ii), the OLS estimator $\widehat{\beta}$ of the parameters of interest in a linear regression model

$$y_t = x_t' \beta + u_t,$$

where $x_t = (x_{1t}, \dots, x_{Mt})$ or, for example, $x_t = (x_{1,t-1}, \dots, x_{M,t-1})$, is consistent and asymptotically normal when

- (i) The error term u_t is serially uncorrelated and uncorrelated with the regressors included in x_t,
- (ii) All the regressors in x_t are either deterministic or stationary random variables.

If not otherwise mentioned, we assume that (i) and (ii) are valid.

After estimating the model, the same residual diagnostic methods used for ARMA models can be applied to assess the residuals and, consequently, the adequacy of the model. The resulting asymptotic distribution result for the OLS estimator is

$$\sqrt{T}(\hat{\beta} - \beta) \underset{d}{\longrightarrow} \mathsf{N}\left(0, \Big(\sum_{t=1}^{T} x_t x_t'\Big)^{-1}\right),$$

where $\underset{d}{\longrightarrow}$ denotes converge in distribution. It turns out that, in fact, even (mild) autocorrelation of the residuals can be accepted without losing many of the useful asymptotic properties of the (conditional) OLS estimator. If autocorrelation occurs after the initial formation of the model, we can proceed in essentially two different ways (see detailed discussion below):

- form a (linear) model where the error term is autocorrelated.
- (preferable alternative) adjust the covariance matrix of the estimated parameters $\widehat{\beta}$ (see below)

Extra: Remaining autocorrelation in the linear (predictive) regression

Consider a linear regression (as above), but this time we set ARMA such as an MA(1) structure for the error term. This would lead to the model

$$y_t = x_t'\beta + u_t, \quad u_t = a_t + \theta_1 a_{t-1},$$

where a_t is now an iid or white noise (wn) process. However, in macro and especially financial econometrics, it is typical to proceed using the latter option, which is discussed next.

As stated, the OLS estimator can be shown to be still consistent and asymptotically normal, even if the error term is autocorrelated and possibly also (conditionally) heteroskedastic. This is practically the message of the quasi-maximum likelihood estimator (QMLE).

- Cf., for example, the section concerning the AR-GARCH model.

In such a situation, a heteroskedasticity-autocorrelation consistent (HAC) covariance matrix estimator is used, indicating that the usual standard errors of the parameter estimates it produces can be replaced with HAC counterparts.

- Without this adjustment, the standard errors are often too small, which correspondingly increases the absolute value of the t-test statistics testing the statistical significance of individual regression coefficients (reducing the p-values).
- One formulation is based on the Newey and West (1987) estimator.

Formally, let's consider the OLS estimator and the usual covariance matrix

$$\widehat{\beta} = \Big(\sum_{t=1}^{T} x_t x_t'\Big)^{-1} \sum_{t=1}^{T} x_t y_t, \quad \mathsf{Cov}(\widehat{\beta}) = \sigma^2 \Big(\sum_{t=1}^{T} x_t x_t'\Big)^{-1},$$

where $\mathsf{Var}(u_t) = \sigma^2$. The general HAC estimator can be written as

$$\mathsf{Cov}(\widehat{\beta})_{HAC} = \Big(\sum_{t=1}^{T} x_t x_t'\Big)^{-1} \widehat{C}_{HAC} \Big(\sum_{t=1}^{T} x_t x_t'\Big)^{-1},$$

where different choices for the middle term $\widehat{C}_{HAC}$ lead to different estimators. For example, in the case of the Newey and West (1987) estimator

$$\widehat{C}_{HAC} = \sum_{t=1}^{T} \hat{u}_t^2 x_t x_t' + \sum_{j=1}^{l} w_j, \sum_{s=j+1}^{T} (x_s \hat{u}_s \hat{u}_{s-j} x_{s-j}' + x_{s-j} \hat{u}_{s-j} \hat{u}_s x_s'),$$

where l is the so-called bandwidth parameter and w_j is an appropriate weighting function.

- For example, in the case of the so-called Bartlett kernel function

$$w_j = 1 - j/l.$$

- Newey and West recommended choosing the parameter l as the integer part of the expression $4(T/100)^{2/9}$.

Furthermore, occasionally $\widehat{C}_{HAC}$ may also be specified (only) allowing for heteroskedasticity (but not autocorrelation, "HC" vs. "HAC") with the alternative (this is the so-called White estimator)

$$\widehat{C}_{HC} = \sum_{t=1}^{T} \hat{u}_t^2 x_t x_t'.$$

10.3 Forecasting with predictive variables

Let us consider forecast construction in more detail. When attempting to predict the value of y_{t+h} using external predictive variables, additional complications arise especially when the forecast horizon h lengthens. As an example, consider a simple model (the main arguments generalize to more general models) containing only two lags of one predictive variable x_t

$$y_t = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{1,t-2} + u_t.$$

As in the ARMA case, one-step forecast is the conditional expectation given the information set at time t. That is

$$\mathsf{E}_t(y_{t+1}) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1,t-1}.$$

From this, it is immediately apparent that additional challenges arise if the forecast horizon h lengthens longer than one ($h = 1$). In other words, multistep forecasts ($h > 1$) seem to require forecasts for x_{1t}.

- Vector autoregressive models provide one alternative to solve this problem when building a joint (multiple-equation) model for y_t and x_{1t}

An alternative approach is to use predictive models that are specific to each forecast horizon. In this case, when utilizing the predictive information available at time t, we can specify

$$y_{t+h} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1,t-1} + e_{t+h},$$

where e_t is a zero-mean error term and the forecast can then be constructed "directly" as

$$\mathsf{E}_t(y_{t+h}) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1,t-1}.$$

Naturally here the parameters $\beta = (\beta_0, \beta_1, \beta_2)$, and finally their estimates, are then specific to the forecast horizon h.

- The properties of the error term e_t can be complicated due to overlapping nature of the forecast horizon–from the forecast origin at time t to $t + h$.
- What is introduced above for a simple model containing only one predictive variable can be generalized straightforwardly to the case where we have M predictors such as more lags of y_t as predictors.

Overall, this same principle of h-period predictive models can be viewed as a building block for various extended predictive regressions, particularly those incorporating elements of **machine learning and statistical learning**.

10.4 Extra: Regularized predictive regressions

Let us briefly focus on predicting y_{t+h} at time t (that is the forecast origin) given the collection of M predictive variables in $x_t = (x_{1t}, \dots, x_{Mt})$, which may also contain some lagged values of y_t (as discussed above).

- This is the same predictive regression we briefly considered earlier, but now extended to the case of high-dimensional predictors.

The challenge of big dependent data. The focus here is on big dependent data, meaning we're dealing with a large collection of potential predictors where the number of variables, M, can be substantial.

- In some cases, M might even be greater than the sample length T ($M > T$) used for estimating the model parameters.

- Here for “big dependent data” we mean that the number of (predictive) time series is large. “Dependent” refers to autocorrelation (serial correlation) in the data. In this section, our dependent variable y_t is still scalar-valued (i.e. only one time series) and hence we consider single-equation models.

Traditional limitation. When M is large relative to T, traditional estimation techniques, like Ordinary Least Squares (OLS), as introduced above, fail or produce highly unstable and unreliable results.

- The classic OLS estimator is not well-defined if $M > T$ (instead $M << T$).
- Even if M is less than T but still large, the resulting model can suffer from overfitting, where it fits the noise in the historical data too closely, leading to poor out-of-sample forecasting performance.

Need for structure: To overcome the curse of dimensionality - the problems that arise when the number of variables grows - we need methods that can impose structure or constraints on the predictive relationship. This is essential for selecting the most relevant predictors and stabilizing the parameter estimates in a data-rich environment.

Regularized estimation and sparsity. To enable effective predictive regressions in this high-dimensional setting, where the predictor space can be very large, a class of techniques known as regularized estimators is employed. Regularization involves adding a penalty term to the standard loss function (like the sum of squared errors) that we aim to minimize. This penalty discourages the model from assigning large values to the coefficients, effectively shrinking them towards zero. Key examples (briefly without details in this course):

- **Ridge estimator**: This technique adds a penalty proportional to the sum of the squared coefficients. It stabilizes the estimates by shrinking all coefficients but doesn’t set any exactly to zero.
- **LASSO** (Least Absolute Shrinkage and Selection Operator): This method uses a penalty based on the sum of the absolute values of the coefficients. Crucially, the LASSO has the ability to perform automatic variable selection by setting the coefficients of irrelevant predictors from x_t exactly to zero. This produces a sparse model, meaning that only a few (the most important) predictors are ultimately used.
- **Elastic Net**: This is a hybrid that combines the penalties of both ridge and LASSO. It’s often used to leverage the variable selection of LASSO while retaining the grouping effect and stability of Ridge, especially when predictors are highly correlated.

Relevance: These estimators are foundational in modern forecasting and (statistical) machine learning, particularly when integrating vast amounts of data—such as financial indicators, survey data, or text-based predictors—into an economic forecasting model. They provide a principled way to manage the trade-off between bias (from shrinkage) and variance (from the large number of predictors) to achieve superior out-of-sample forecasting accuracy.

A more detailed treatment of this topic and these estimators is reserved for the **Advanced Time Series Econometrics course**.

Chapter 11

Basics of vector autoregression

Univariate ARMA models can be generalized to the multivariate case where **multiple time series are analyzed simultaneously**. In this course, we concentrate on a brief introduction of basic ideas of a **vector autoregressive (VAR) model**. The VAR model provides a very common framework for analyzing time series dynamics, such as the effects of random shocks (e.g., changes in economic policy or technology).

- VAR model and its statistical details will be considered more detail in **Advanced Time Series Econometrics course**.

11.1 VAR(p) process

Let us start with a bivariate model. That is we have two stationary time series y_{1t} and y_{2t} with $\mathsf{E}(y_{1t}) = \mu_1$ and $\mathsf{E}(y_{2t}) = \mu_2$. **A bivariate (i.e., a two-variable) first-order VAR process** (i.e., a bivariate **VAR(1) process**) is given as

$$\begin{aligned} y_{1t} &= \nu_1 + a_{11} y_{1,t-1} + a_{12} y_{2,t-1} + u_{1t} \\ y_{2t} &= \nu_2 + a_{21} y_{1,t-1} + a_{22} y_{2,t-1} + u_{2t}, \end{aligned}$$

where u_{1t} and u_{2t} are two error terms (independent of the history of y_{1t} and y_{2t}) that may be correlated. In a matrix notation, the VAR(1) process can be rewritten

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}.$$

- Equivalently,

$$y_t = \nu + A_1 y_{t-1} + u_t,$$

where now $y_t = (y_{1t}, y_{2t})$, $\nu_t = (\nu_1, \nu_2)$ and $u_t = (u_{1t}, u_{2t})$ are two 2×1 vectors.

- In this notation, likewise throughout this material, with bolded font we emphasize vectors and matrices, to distinguish scalar-valued components and processes.

As an extension of two-variable and first-order case, **K-variable VAR(p) process** is

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \quad u_t \sim \mathsf{iid}(0, \Sigma_u)$$

where $y_t = (y_{1t} \cdots y_{Kt})$ is a $K \times 1$ vector of variables and $A_1, \ldots, A_p$ are $K \times K$ coefficient matrices. In other words, the VAR(p) process depends on the p lags of $y_{1t}, \ldots, y_{Kt}$. The error term ("shock" or "innovation") $u_t = (u_{1t}, \ldots, u_{Kt})$ is an unobserved $K \times 1$ (vector) iid process.

- Hence $\mathsf{E}(u_t) = 0$, $\mathsf{E}(u_t u_t') = \Sigma_u$, and $\mathsf{E}(u_t u_s') = 0$, $s \neq t$.
- The bivariate and more general model show the basic idea: The idea of the VAR(p) model is to regress each component of y_t on its own lags and on the lags of the other components. The model can describe lagged or dynamic dependencies among variables. For example, one may ask how y_{2t} affects the future path of y_{1t}, and other way round.

A companion form: The VAR(p) process can be rewritten

$$Y_t = \underline{\nu} + \mathbf{A} Y_{t-1} + U_t,$$

where $Y_t = [y_t' \cdots y_{t-p+1}']'$ and $\underline{\nu} = [\nu' \; 0' \cdots 0']'$ are $(Kp \times 1)$ and $U_t = [u_t' \; 0' \cdots 0']'$ are $(Kp \times 1)$ vectors, and

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_K & 0_K & \cdots & 0_K & 0_K \\ 0_K & I_K & \cdots & 0_K & 0_K \\ \vdots & & \ddots & \vdots & \vdots \\ 0_K & 0_k & \cdots & I_K & 0_K \end{bmatrix} \quad (Kp \times Kp)$$

In other words,

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \nu \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_K & 0_K & \cdots & 0_K & 0_K \\ 0_K & I_K & \cdots & 0_K & 0_K \\ \vdots & & \ddots & \vdots & \vdots \\ 0_K & 0_K & \cdots & I_K & 0_K \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

indicating that there holds equalities $y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t$, $y_{t-1} = y_{t-1}$, and so on to $y_{t-p+1} = y_{t-p+1}$.

Lag-operator and the **lag-polynomial presentation of the VAR(p)**. Define again the lag-operator as $B^k x_t = x_{t-k}$, $k = 0, \pm 1, \pm 2, \ldots$, where x_t might be a vector. Therefore, using the lag-operators, the VAR(p) process can be rewritten

$$A(B)y_t = \nu + u_t,$$

where $A(B) = (I_K - A_1 B - \cdots - A_p B^p)$.

Stationarity (stability) of the VAR(p) process. A VAR(p) process is (strictly) stationary (stable) (given $u_t \sim \text{iid}(0, \Sigma_u)$) if the roots of the determinant

$$\det\Big(\mathbf{I}_K - A_1 z - \cdots - A_p z^p\Big) = 0$$

are outside the unit circle in absolute value (i.e., $|z| > 1$).

Notice, however, that many statistical software packages, including the **vars package in R**, use an alternative but equivalent presentation.

- When using the companion form of the VAR(p), the stability of the VAR(p) process is determined by the eigenvalues of the companion matrix **A**.
- In "vars package" and "VAR-function", the part "Roots of the characteristic polynomial" reported by R are precisely these eigenvalues. The key relationship is that these eigenvalues of the companion matrix are the reciprocals of the roots of the determinant polynomial. Therefore, the two conditions are equivalent:
  - Roots of $\det\Big(\mathbf{I}_K - A_1 z - \cdots - A_p z^p\Big) = 0$ are outside the unit circle ($|z| > 1$).
  - Eigenvalues of the companion matrix **A** are inside the unit circle ($|\lambda| < 1$). That is, we the eigenvalues of the matrix **A** satisfy $\det\Big(\mathbf{I}_K \lambda^p - A_1 \lambda^{p-1} - \cdots - A_p\Big) = 0$.

- So, when using R's output, we should check if the moduli of the reported roots (i.e. the eigenvalues) are **all less than 1** to confirm that the VAR model is stable.

Linear process. As the AR(p) process, the stationary (stable) VAR(p) process can be written

$$y_t = \mu + \sum_{j=0}^{\infty} \Psi_j u_{t-j} = \Psi(B) u_t,$$

where $\Psi(B) = [A(B)]^{-1} = \Psi_0 + \Psi_1 B + \Psi_2 B^2 + \cdots$ defines matrix-valued lag-polynomial. In addition to this linear process presentation, the sequence $\Psi_0, \Psi_1, \Psi_2, ...$ have also other important role in macroeconometrics. It turns out that the matrix Ψ_s ($\Psi_0 = I_K$) measures the effect of one-unit increase in u_{jt} upon $y_{i,t+s}$, , $i, j = 1, ..., K$, or equivalently

$$\Psi_s = \frac{\partial y_{t+s}}{\partial u_t'}.$$

A function of elements Ψ_0, Ψ_1, $\Psi_2, ...$ is the **(forecast error) impulse response function** (IRF) of the VAR(p) process.

- In other words, the row i, column j element of Ψ_s gives the effect of a one-unit increase in the error term of the jth variable at time t on the ith variable at time $t + s$, holding all other errors constant.
- For instance, the first column of Ψ_1 gives the effect in period 1 of a unit increase in the error term of the first variable (first shock) in period 0 on each variable in the system.
- Especially in (structural) macroeconometrics, the VAR model is extended by economic identification assumption, which leads to **structural VAR models** and identified impulse response functions. These will be considered more detail in **Macroeconometrics** course.

As in the univariate case, here defined element-by-element, due to the typical assumption $\Psi_s \longrightarrow 0$, $s \longrightarrow \infty$, the effects decay in time. In empirical analysis, the unknown Ψ_s must also be replaced by their estimates, obtained from the estimation result of the VAR model.

VAR(p) model: Estimation and model selection. The stationary VAR(p) model can be estimated by ordinary least squares (OLS).

- More specifically, estimation can be done equation by equation by OLS (cf. the model structure).

- If the distribution of u_t is known (asssumed to be known), the model can also be estimated by the method of maximum likelihood (ML). Under normality assumption, that is $u_t \sim \mathsf{nid}(0, \Sigma_u)$, it can be shown that the conditional ML estimator is numerically equivalent to the OLS estimator.
- The OLS estimator is consistent and asymptotically normally distributed. Thus, the t-statistics of the parameter estimates can be used as usual to test for the statistical significance of individual parameter coefficients.

As for the univariate AR(p) model, there are several approaches to determine the lag length p for the VAR(p) model.

- Sequential testing procedure (cf. the sequential testing for univariate models)
- Information criteria (including intercept parameters):
  - AIC: $\log(\det(\Sigma_u)) + \frac{2}{T}K(Kp+1)$
  - BIC: $\log(\det(\Sigma_u)) + \frac{\log(T)}{T}K(Kp+1)$

For the competing models (recall that exactly the same sample period must be used!) the one with the lowest value of the selected information criterion is the selected (preferable) model.

VAR(p): Model selection and residual diagnostics

Overall, the model selection can be based on the specification cycle presented in the case of univariate AR models but modified suitable to VARs. However, due to multiple time series and hence more complicated dynamic interrelationships between the variables, autocorrelation functions cannot be directly used to select the lag length p.

- Especially information criteria, such as AIC and BIC, can be used like in the univariate case.
- At times the lag length selections correspond the frequency of the data. That is, for the quarterly data $p = 4$ and for monthly data $p = 12$.

The adequacy of the selected and estimated VAR(p) model can be evaluated by diagnostic checks (like for ARMA models). A well-specified model should capture all the systematic, linear dynamics present in the (multivariate) time series data. The primary goal of these checks is to determine if there is any significant autocorrelation left in the vector of residuals $\hat{u}_t = y_t - \hat{y}_t$, where $\hat{y}_t$ are the fitted values of the selected VAR model.

- If autocorrelation is still present in residuals, it means the model has failed to capture some of the predictable patterns in the data, making it unreliable for forecasting or structural analysis.

To formally test for remaining autocorrelation in the residuals, we use a multivariate extension of the Ljung-Box test. This is called the **Portmanteau test**. It jointly examines whether a group of residual autocovariance matrices up to a certain lag are statistically different from zero. The **null hypothesis** (H_0) of the test is that there is no serial correlation in the residuals up to a specified lag H. The Portmanteau test statistic, often denoted as Q_H, is calculated as follows:

$$Q_H = T \sum_{h=1}^{H} \operatorname{tr}(\widehat{C}_h' \widehat{C}_0^{-1} \widehat{C}_h \widehat{C}_0^{-1}),$$

where $\widehat{C}_j = \frac{1}{T} \sum_{t=j+1}^{T} \hat{u}_t \hat{u}_{t-j}'$ is the estimated residual autocovariance matrix at lag j and $\operatorname{tr}(\cdot)$ is the trace operator, which sums the diagonal elements of a matrix. Under the null hypothesis of no serial correlation, the Q_H statistic approximately follows a χ^2-distribution with $K^2(H-p)$ degrees of freedom and the associated p-value can be constructed.

- Should the model fail in the Portmanteau test or overall in diagnostic checking, the model should be respecified, such as to increase the lag length of the model (e.g., from VAR(p) to VAR($p+1$) or higher) and re-estimate it. The logic is that by including more lags, the model can better capture the complex temporal dependencies in the data. After respecification, the diagnostic checks must be performed again to ensure the new model is adequate.

- However, especially with complex multivariate data, it is often challenging to find a model that can be considered completely adequate from a diagnostic perspective. When the diagnostic checks indicate that the model is misspecified (for instance, due to some persistent residual autocorrelation), the standard assumptions for maximum likelihood estimation are violated. In such cases, the parameter estimates can be interpreted as Quasi-Maximum Likelihood Estimates (QMLE), a concept that was introduced on a general level in the AR-GARCH section and can be extended to the VAR models (with necessary changes in notation etc.).

VAR(p): Forecasting

Similarly as in the univariate AR(p) model, forecasts can be obtained as conditional expectations, conditional on the information at time t. The properties CEV1–CEV4 of the conditional expectation can be extended to the vector-valued case straigtforwardly (element-by-element treatment).

- One-period-ahead forecasts with the VAR(p):

$$\mathsf{E}_t(y_{t+1}) = \nu + A_1 y_t + \cdots + A_p y_{t-p+1}.$$

- Multiperiod h-step forecasts can be obtained iteratively following the recursion (cf. the AR(p) case):

$$\mathsf{E}_t(y_{t+h}) = \nu + A_1 \mathsf{E}_t(y_{t+h-1}) + \cdots + A_p \mathsf{E}_t(y_{t+h-p}),$$

  where $\mathsf{E}_t(y_{t+h-j}) = y_{t+h-j}$ for $h \leq j$.

Similarly as in the univariate AR models, to include nonzero mean vector to the VAR model is assumed above. Alternatively, you can construct the VAR model for the deamed time series and add the mean vector to the forecasts obtained with the VAR model with demeaned data.

Empirical example. Consider an application of a VAR model to the classic three-variable case (as described, e.g., in Stock and Watson, 2001), which is a benchmark in macroeconomic and especially monetary policy analysis.

- See J.H. Stock and M.W. Watson (2001). Vector autoregressions. Journal of Economic Perspectives, 115(4), 101–115.

The VAR model contains the following three ($K = 3$) key quarterly U.S. macroeconomic variables (source: FRED):

- Inflation: Measured as the percent change from the preceding period in the Gross Domestic Product Implicit Price Deflator, expressed at a seasonally adjusted annual rate.
- Unemployment: The seasonally adjusted quarterly average of the civilian unemployment rate.
- Interest Rate: The quarterly average of the effective Federal Funds Rate, which is not seasonally adjusted.

These three series are modeled as an interdependent system where each variable is influenced by its own past values and the past values of the other two variables. The following analysis covers the sample period from 1960:Q1 to 2000:Q4 as in Stock and Watson (2001).

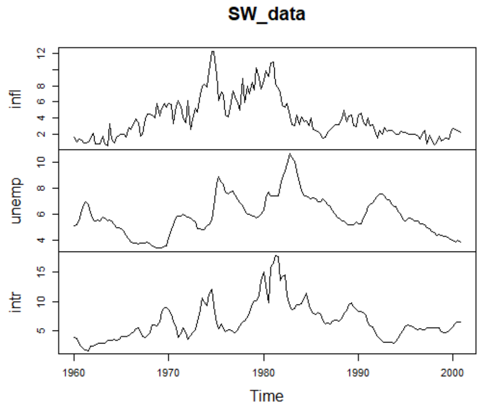


Figure: U.S. macroeconomic time series (inflation, unemployment rate and interest rate). See Stock and Watson (2001). Source of data: FRED.

Lag length selection: For this analysis and illustration, we follow Stock and Watson (2001) and specify a VAR(4) model. This choice is common in quarterly macroeconomic data as it captures potential year-over-year dynamics.

Parameter estimation: The VAR(4) model is estimated assuming the error terms are normally distributed. Because each equation in the VAR system has the same set of explanatory variables (the lagged values of all variables), Ordinary Least Squares (OLS) applied equation-by-equation is an efficient estimation method.

- A key feature of VAR models is the large number of parameters. In this three-variable VAR(4) model, each of the three equations includes an intercept and four lags of all three variables, resulting in $3\times(1+4\times3)=39$ parameters to be estimated. Due to this high dimensionality, interpreting individual coefficient estimates is often impractical. Instead, analysis typically focuses on system-wide properties through Granger causality tests, as well as impulse response functions and forecast error variance decompositions (see Macroeconometrics course)

Residual diagnostics: After estimation, diagnostic tests are crucial for assessing the model's adequacy.

- The Portmanteau test for serial correlation in the residuals indicates some remaining unmodelled dynamics. The test's statistically significant result suggests that the VAR(4) specification may not fully capture all the (linear) dependencies within the data.

- While the residuals from the different variable equations appear to have low cross-correlation, an analysis of the squared residuals reveals a different pattern. The presence of autocorrelation in the squared residuals is an indicator of conditional heteroskedasticity. This means the volatility of the shocks is not constant over time. While the VAR(4) model may adequately capture the conditional mean of the variables, it fails to account for this time-varying variance in the error terms. A more advanced model, such as a VAR-GARCH (not considered in this course), would be required to model volatility dynamics.

Details of the estimated VAR(4) model

```
VAR Estimation Results:
=========================
Endogenous variables: Inflation, Unemployment, Fedfunds
Deterministic variables: const
Sample size: 160
Log Likelihood: -396.638
Roots of the characteristic polynomial:
0.9696 0.9696 0.7928 0.7928 0.686 0.686 0.5674 0.5674 0.4609 0.4609 0.2053 0.2053
Call:
VAR(y = SW_data, p = 4, type = c("const"), exogen = NULL, lag.max = NULL)


Estimation results for equation Inflation:
==========================================
Inflation = Inflation.l1 + Unemployment.l1 + Fedfunds.l1 + Inflation.l2 + Unemployment.l2 + Fedfu

                  Estimate Std. Error t value Pr(>|t|)
Inflation.l1     0.5886863  0.0821552   7.166 3.47e-11 ***
Unemployment.l1 -0.8356913  0.4006087  -2.086  0.03870 *
Fedfunds.l1      0.2370805  0.1082335   2.190  0.03007 *
Inflation.l2     0.0902554  0.0937902   0.962  0.33747
Unemployment.l2  1.3545379  0.6841360   1.980  0.04958 *
Fedfunds.l2     -0.2168750  0.1453158  -1.492  0.13773
Inflation.l3     0.1205897  0.0944654   1.277  0.20377
Unemployment.l3 -1.0953222  0.6779667  -1.616  0.10833
Fedfunds.l3     -0.0005773  0.1455216  -0.004  0.99684
Inflation.l4     0.1866837  0.0846107   2.206  0.02891 *
Unemployment.l4  0.4122130  0.3765054   1.095  0.27538
Fedfunds.l4     -0.0231108  0.1113078  -0.208  0.83581
const            1.0641191  0.3968301   2.682  0.00817 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.053 on 147 degrees of freedom
Multiple R-Squared: 0.8509, Adjusted R-squared: 0.8387
F-statistic:  69.9 on 12 and 147 DF,  p-value: < 2.2e-16


Estimation results for equation Unemployment:
=============================================
Unemployment = Inflation.l1 + Unemployment.l1 + Fedfunds.l1 + Inflation.l2 + Unemployme

                  Estimate Std. Error t value Pr(>|t|)
Inflation.l1     0.0066525  0.0181442   0.367 0.714410
Unemployment.l1  1.4696483  0.0884753  16.611  < 2e-16 ***
Fedfunds.l1      0.0006271  0.0239036   0.026 0.979106
Inflation.l2    -0.0115173  0.0207138  -0.556 0.579040
Unemployment.l2 -0.5155262  0.1510930  -3.412 0.000833 ***
Fedfunds.l2      0.0604859  0.0320933   1.885 0.061446 .
Inflation.l3     0.0317894  0.0208629   1.524 0.129724
Unemployment.l3 -0.0057705  0.1497305  -0.039 0.969310
Fedfunds.l3     -0.0399995  0.0321388  -1.245 0.215265
Inflation.l4    -0.0200840  0.0186865  -1.075 0.284230
Unemployment.l4 -0.0050194  0.0831521  -0.060 0.951947
Fedfunds.l4      0.0114985  0.0245826   0.468 0.640657
const            0.0854091  0.0876408   0.975 0.331392
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.2326 on 147 degrees of freedom
Multiple R-Squared: 0.9786, Adjusted R-squared: 0.9768
F-statistic: 559.3 on 12 and 147 DF,  p-value: < 2.2e-16


Estimation results for equation Fedfunds:
=========================================
Fedfunds = Inflation.l1 + Unemployment.l1 + Fedfunds.l1 + Inflation.l2 + Unemployment.l

                Estimate Std. Error t value Pr(>|t|)
Inflation.l1     0.07293    0.06945   1.050  0.29536
Unemployment.l1 -1.38408    0.33865  -4.087 7.16e-05 ***
Fedfunds.l1      0.95325    0.09149  10.419  < 2e-16 ***
Inflation.l2     0.20379    0.07928   2.570  0.01115 *
Unemployment.l2  1.27436    0.57832   2.204  0.02911 *
Fedfunds.l2     -0.40302    0.12284  -3.281  0.00129 **
Inflation.l3    -0.07175    0.07985  -0.899  0.37035
Unemployment.l3 -0.49829    0.57311  -0.869  0.38601
```

```
Fedfunds.l3      0.34476    0.12301   2.803  0.00575 **
Inflation.l4    -0.04306    0.07152  -0.602  0.54804
Unemployment.l4  0.49476    0.31827   1.555  0.12221
Fedfunds.l4      0.03129    0.09409   0.333  0.73997
const            0.53914    0.33545   1.607  0.11015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.8905 on 147 degrees of freedom
Multiple R-Squared: 0.9273, Adjusted R-squared: 0.9214
F-statistic: 156.3 on 12 and 147 DF,  p-value: < 2.2e-16



Covariance matrix of residuals:
             Inflation Unemployment Fedfunds
Inflation     1.109681     0.001595   0.1486
Unemployment  0.001595     0.054126  -0.0928
Fedfunds      0.148558    -0.092796   0.7930

Correlation matrix of residuals:
             Inflation Unemployment Fedfunds
Inflation     1.000000     0.006508   0.1584
Unemployment  0.006508     1.000000  -0.4479
Fedfunds      0.158370    -0.447925   1.0000

=====
Portmanteau Test (adjusted)

serial.test(VAR_lag_p, lags.pt = 12, type = "PT.adjusted") #

Chi-squared = 132.39, df = 72, p-value = 1.904e-05

> serial.test(VAR_lag_p, lags.pt = 15, type = "PT.adjusted") #

Chi-squared = 149.51, df = 99, p-value = 0.0007889

> serial.test(VAR_lag_p, lags.pt = 6, type = "PT.adjusted") #

Chi-squared = 48.748, df = 18, p-value = 0.0001165
```

11.2 Granger causality in VARs

Granger causality is a concept used to determine whether one time series can predict another. It doesn't imply true causality in the philosophical or strict scientific sense, but rather predictive causality based on temporal dependence and informational content.

- Granger causality can be examined for single-equation models, such as the ones considered in the previous section examining whether the additional predictor Granger cause y_t.
- Granger causality is most often studied in the context of VAR models.

Bivariate VAR model ($K = 2$): Granger-causality can be easily tested within a VAR model. Consider first, for simplicity, a bivariate VAR(p) process

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} a_{11,j} & a_{12,j} \\ a_{21,j} & a_{22,j} \end{bmatrix} \begin{bmatrix} y_{1,t-j} \\ y_{2,t-j} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}.$$

If $a_{12,j} = 0$, $j = 1, 2, \dots, p$, then the lags of y_{2t} do not help to forecast y_{1t} and there is no Granger-causality from y_{2t} to y_{1t} (that is, y_{2t} is not Granger-causal for y_{1t}).

The restriction $a_{12,1} = a_{12,2} = \cdots = a_{12,p} = 0$ implies that there is not such Granger causality relationship.

- When working with estimated VAR model, this hypothesis of no Granger causality from y_{2t} to y_{1t} can be tested by the Wald or likelihood ratio (LR) test, following χ^2_p-distribution under the null hypothesis.
- At times, such as in vars package in R, also the F-test statistic (instead of the Wald test statistic) will be used where the critical values are coming from the F-distribution.

Larger VAR models ($K > 2$): When the VAR model contains more than two variables, Granger causality between two variables cannot be tested by testing zero-restrictions in a straightforward manner (as above). As an example, consider the following three-variable VAR process

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} a_{11,j} & a_{12,j} & a_{13,j} \\ a_{21,j} & a_{22,j} & a_{23,j} \\ a_{31,j} & a_{32,j} & a_{33,j} \end{bmatrix} \begin{bmatrix} y_{1,t-j} \\ y_{2,t-j} \\ y_{3,t-j} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix}.$$

If, for instance, $a_{12,1} = \cdots = a_{12,p} = 0$, the lags of y_{2t} do not help forecasting y_{1t} one period ahead. However, because the lags of y_{2t} may affect y_{3t}, which, in turn, affects y_{1t}, the lags of y_{2t} may help forecasting $y_{1,t+1}$, $y_{1,t+2}$ etc. In other words, there may be Granger causality through indirect effects.

Generally, Granger causality from one **group of variables** to another is defined similarly. Partition $y_t = [y'_{1t} \quad y'_{2t}]'$, where y_{1t} is $K_1 \times 1$ and y_{2t} is $K_2 \times 1$, $(K_1 + K_2 = K)$, and write the VAR(p) process

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} A_{11,j} & A_{12,j} \\ A_{21,j} & A_{22,j} \end{bmatrix} \begin{bmatrix} y_{1,t-j} \\ y_{2,t-j} \end{bmatrix} + \begin{bmatrix} \zeta_{1t} \\ \zeta_{2t} \end{bmatrix},$$

where $A_{\cdot\cdot,j}$ are generally matrices (but of course may reduce to scalar-valued components in certain situations)

- If, for instance, $A_{12,j} = 0$, $\forall j = 1, \dots, p$, then the variables in y_{2t} are not Granger-causal for the variables in y_{1t}.

The general restriction $A_{12,1} = A_{12,2} = \dots = A_{12,p} = 0$ can be tested by the Wald (F-test) or likelihood ratio (LR) test.

- The above restrictions and test statistic apply also when there is a constant term included in the VAR(p).

The hypothesis of interest

$$A_{12,1} = A_{12,2} = \cdots = A_{12,p} = 0$$

implies that there is $K_1 K_2 p$ zero restrictions. Wald and LR tests can be used to test the above null hypothesis.

- Under the null hypothesis, the Wald and LR tests are asympotically $\chi^2_{K_1 K_2 p}$-distributed and large values of the test statistics are critical for the null hypothesis.
- As mentioned, at times F-test statistic is used where the number of restrictions (the first degree-of-freedom parameter in the F-distribution) is the above number of restrictions.

Example: Three-variable process/model. Consider the VAR(1) process/model:

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} y_{t-1} + u_t.$$

Partitioning $y_t = [y_{1t} \quad y_{2t}']'$ so that $y_{2t} = [y_{2t} \quad y_{3t}]'$, y_{2t} does not Granger cause $y_{1t} = y_{1t}$ because the matrix $A_{12,1} = 0$. On the other hand, y_{1t} Granger causes y_{2t}.

Instantaneous causality. In the bivariate VAR, there is no instantaneous causality between y_{1t} and y_{2t} if the errors u_{1t} and u_{2t} are uncorrelated. In general (not necessarily related to the VAR), a variable y_{1t} is said to be instantaneously causal for another variable y_{2t}, if knowing the value of y_{1t} in the forecast period helps to improve forecasts of y_{2t}.

It can be proved that instantaneous causality is symmetric so that if y_{1t} causes y_{2t}, then also y_{2t} causes y_{1t}. That is the concept is symmetric: This definition does not account for the “direction” of causality which must be determined from other sources.

- For instance, economic theory may suggest that the (instantaneous) causality runs only in one direction. Then significant correlation can be interpreted in favour of a causal relation.

In systems with more than two variables, there is no instantaneous causality between two groups of variables if the covariance matrix of the errors, Σ_u, is block diagonal.

- As an example, if the covariance matrix in a VAR process for $y_t = (y_{1t}, y_{2t}, y_{3t})$ is
$$\Sigma_u = \begin{bmatrix} 2.0 & 0 & 0 \\ 0 & 1.5 & 0.4 \\ 0 & 0.4 & 1.0 \end{bmatrix},$$
then there is no instantaneous causality between y_{1t} and $y_{2t} = (y_{2t}, y_{3t})$.
- Instantaneous causality can be tested by a Wald type of test of zero restrictions on the corresponding elements of the covariance matrix of the error terms of a VAR model.

Empirical example: Three-variable monetary policy VAR. Continue with the three-variable VAR(4) model selected above for the Stock-Watson (2001) type of data. Consider Granger causality and instantaneous causality tests. The goal is to determine if past values of certain variables are useful in predicting others. The null hypothesis for each test is that no Granger causality/instantaneous causality exists.

The results, obtained using R, strongly indicate a highly interconnected dynamic system. We find statistically significant Granger causality in all tested directions, meaning the variables have substantial predictive power over each other. A summary of the test results is provided below:

```
| (Infl., Unemp.) -> Int. rate  | Granger (F-Test) | 4.86 | < 0.001 | Inst. causality (\chi^2) |
| (Unemp., Int. rate) -> Infl.  | Granger (F-Test) | 3.40 | < 0.001 | Inst. causality (\chi^2) |
| (Infl., Int. rate) ->  Unemp. | Granger (F-Test) | 4.82 | < 0.001 | Inst. causality (\chi^2) |
```

In short, the past values of each pair of variables are useful for predicting the third.

- That is we are rejecting the null hypotheses of no Granger causality at the conventional statistical significance levels.

Significant instantaneous (contemporaneous) relationships are also present in almost all cases, with the only exception being the link between inflation and the other two variables (unemployment rate and interest rate), which is not significant at the 5% level. Overall, these results confirm strong dynamic interlinkages among the interest rate, inflation, and unemployment rate.

Chapter 12

Nonstationary processes

12.1 Integrated process

In some application fields, such as economics and finance, the existence of trends in the analyzed time series is common.

- Here the term “trend” is used quite generally.

Earlier in this material, it was discussed that nonstationary time series containing a trend can be investigated using methods developed for stationary time series if the trend is first removed using a suitable transformation. The most popular of such transformations is taking differences, which amounts to modelling the differenced series $\Delta y_t = y_t - y_{t-1}$ where $\Delta = 1 - B$ is the difference operator.

- These series might seem stationary and thus be modelled with stationary models reasonably well.
- If the differences still include some sort of trend, it is natural to extend this procedure to the difference of these differences $\Delta^2 y_t = y_t - 2y_{t-1} + y_{t-2}$, and so on. This, however, is very rare in practice, and typically only first differences are considered when differencing data.

Based on the aforementioned, let us introduce the following definition (here for univariate time series processes):

Integrated process. A stochastic process y_t $\{y_t, t = 1, 2, ...\}$ is said to be **an integrated process of order** d, or an $I(d)$ **process** if the d times differenced process

$$\Delta^d y_t$$

is stationary (or asymptotically stationary), but the $(d-1)$ differenced process $\Delta^{d-1}y_t$ is not.

- In some definitions, the process is first centered by subtracting its mean, that is
$$\Delta^d(y_t - \mathsf{E}(y_t)), \quad t = 1, 2, \ldots,$$
is stationary, while $\Delta^{d-1}(y_t - \mathsf{E}(y_t))$ is not. This form accounts for the deterministic component of the process.

If y_t is an $I(d)$ **process**, we denote

$$y_t \sim \mathrm{I}(d) \;\; (d \geq 1).$$

In this course, we will consider only $d = 1$ and $d = 0$.

Random walk. The simplest example of an $I(1)$ process is **random walk**, which was briefly covered already in connection to the AR(1) process. Here the random walk is defined with a drift term ν:

$$\Delta y_t = \nu + u_t, \;\; t = 1, 2, \ldots,$$

where $u_t \sim \mathsf{iid}(0, \sigma^2)$. By recursive substitutions, we get

$$y_t = y_0 + \nu t + \sum_{j=1}^{t} u_j, \;\; t = 1, 2, \ldots$$

- Sometimes, depending on the time series and application that we are considering. it is assumed that $\nu = 0$, that is, we have a **random walk without drift**, and hence the term νt above vanishes. Furthermore, in this case, differences $y_t - y_{t-1} = u_t$ from the process $y_t = y_0 + \sum_{j=0}^{t-1} u_{t-j}$ are stationary, and this holds even if instead of $u_t \sim \mathsf{iid}(0, \sigma^2)$ we only assume u_t is stationary.

Assuming the initial (starting) value y_0 as constant, for the random walk we get (details are left as an exercise)

$$\mathsf{E}(y_t) = y_0 + \nu t,$$
$$\mathsf{Var}(y_t) = \mathsf{Var}\left(\sum_{j=1}^{t} u_j\right) = \sigma^2 t,$$
$$\mathsf{Cor}(y_t, y_{t+k}) = \frac{1}{\sqrt{1 + k/t}}, \;\; k \geq 0.$$

Here we can observe that random walk is not stationary even if the value of y_0 is altered.

- As $\Delta y_t \sim \mathsf{iid}(\nu, \sigma^2)$ is stationary, we can conclude that the random walk is an $I(1)$ process.
- From the calculations above, we can observe that for random walk $\mathsf{Var}(y_t) \to \infty$ and $\mathsf{Cor}(y_t, y_{t+k}) \to 1$ with any $k > 0$ as $t \to \infty$. Therefore, random walk processes are strongly autocorrelated, and due to their wandering nature, they exhibit trend-like features.

In other words, we can summarize:

- An $I(0)$ process is stationary: it fluctuates around a constant mean, has constant variance, and autocorrelations decay quickly. It is often described as mean-reverting and has short memory.
- An $I(1)$ process is non-stationary: it exhibits a **stochastic trend** (see below) and **wanders** over time. Each innovation has a permanent effect, and autocorrelations decay slowly, indicating long memory.

ARIMA processes and trends in time series. Random walk is the simplest example of so-called **ARIMA processes**.

- The acronym ARIMA stands for AutoRegressive Integrated Moving Average, where the letter I refers to integration — the idea that the process is formed by summing (or "integrating") past innovations $y_t = \sum_{j=0}^{t-1} u_{t-j}$. This cumulative sum introduces a **stochastic trend**, meaning the trend component is random and evolves over time due to the accumulation of shocks.
- If the process includes a drift term $\nu \neq 0$, the stochastic trend fluctuates around a **deterministic trend** $y_0 + \nu t$. A deterministic trend is a non-random, predictable function of time — typically linear — around which the series may fluctuate.

Trend-stationary vs. difference-stationary processes. A process of the form

$$y_t = \alpha + \beta t + \psi(B) u_t,$$

where $\psi(B)$ is the MA(∞) representation of an ARMA process (including the special case $\psi(B) = 1$), is called a trend-stationary process. If the deterministic trend βt (or $\alpha + \beta t$) is removed, the residual process is stationary.

- In contrast, a difference-stationary process (e.g., a random walk) requires differencing to achieve stationarity, as it contains a stochastic trend.

Types of trends. Many economic and financial time series exhibit trends, which can be either deterministic or stochastic.

- A **deterministic trend** is a fixed function of time (e.g., βt), around which the series fluctuates due to stationary noise. Removing the trend yields a stationary process and implies mean reversion.
- A **stochastic trend** results from the accumulation of random shocks, causing the series to drift persistently. It implies permanent effects of innovations and requires differencing for stationarity.

Empirical evidence suggests that stochastic trends are often more appropriate for modeling macroeconomic and financial time series.

12.2 ARIMA(p,d,q) process

If the process y_t, $t = 0, 1, ...$, is non-stationary, but the difference Δy_t is stationary and follows an invertible ARMA(p, q) process, y_t is called an **ARIMA$(p, 1, q)$ process**.

- The order in the middle refers to the fact, that stationarity is achieved by differencing the process once.
- If the process is non-stationary after being differenced once, but the second differences $\Delta^2 y_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$ follow a stationary and invertible ARMA(p, q) process, y_t is called an ARIMA$(p, 2, q)$ process.

In general, y_t is called an **ARIMA(p, d, q) process** if it becomes a stationary and invertible ARMA(p, q) process after differencing d times.

- That is, $\Delta^d y_t \sim$ ARMA(p, q).
- In practice, $d = 1$ is by far the most common case, $d = 2$ can be encountered quite rarely, and $d > 2$ can be safely ignored.
- With $d = 0$, we obtain the special case of an ARMA(p, q) process.

The typical properties of an ARIMA(p,1,q) process are generally the same as for random walk. The realizations exhibit a wandering nature explained by the variance growing as a function of t and strong autocorrelation even though the autocorrelation function cannot be defined like with stationary processes.

A warning on overdifferencing. While differencing is a fundamental tool for making a nonstationary time series stationary, applying it too many times, a problem known as **overdifferencing**, can be counterproductive. It is a common pitfall to assume that if one difference helps, a second one might help more. Even more often, **you should not mechanically difference a time series that is already "stationary enough"**.

- Doing so without a strong theoretical justification (e.g., an economic model suggesting a process is integrated of order two, I(2)), is highly unlikely and can even harm your statistical/econometric analysis.

Overdifferencing does not make a series "more stationary". Instead, it introduces new artificial patterns into the data that were not present in the original process. This complicates model building and can lead to poor analyses and forecasts.

- The most significant problem is that overdifferencing creates a specific, predictable correlation structure.

Illustration: Overdifferencing white noise. Let see what happens when we difference a series that is already stationary: a white noise process, $y_t = \varepsilon_t$, where $\varepsilon_t \sim \mathsf{wn}(0, \sigma^2)$. This series has by definition zero autocorrelation at all lags. If we mistakenly difference it, we create a new series, $z_t = \Delta y_t = y_t - y_{t-1} = \varepsilon_t - \varepsilon_{t-1}$. This new series, z_t, is no longer white noise. Let's examine its properties:

- Variance: $\mathsf{Var}(z_t) = \mathsf{Var}(\varepsilon_t - \varepsilon_{t-1}) = \mathsf{Var}(\varepsilon_t) + \mathsf{Var}(\varepsilon_{t-1}) = 2\sigma^2$. We have doubled the variance!
- Autocorrelation: The autocovariance at lag 1 is $\mathsf{Cov}(z_t, z_{t-1}) = \mathsf{Cov}(\varepsilon_t - \varepsilon_{t-1}, \varepsilon_{t-1} - \varepsilon_{t-2}) = -\mathsf{Var}(\varepsilon_{t-1}) = -\sigma^2$.

The autocorrelation at lag 1 is therefore:

$$\rho_1 = \frac{\mathsf{Cov}(z_t, z_{t-1})}{\sqrt{\mathsf{Var}(z_t)\mathsf{Var}(z_{t-1})}} = \frac{-\sigma^2}{\sqrt{2\sigma^2 \cdot 2\sigma^2}} = \frac{-\sigma^2}{2\sigma^2} = -0.5.$$

By differencing a series with zero autocorrelation, we created a new series with a perfectly defined negative autocorrelation of -0.5 at the lag 1.

As we will discuss later on, always check for stationarity using tools like the ACF plot and unit root tests (e.g., ADF test, see the coming section) before applying another round of differencing. If the series already appears stationary enough, stop!

12.3 Unit root process

Let's consider the (univariate) AR(p) process

$$a(B)z_t = u_t,$$

where $a(B) = 1 - a_1 B - \cdots - a_p B^p$.

- Here notation z_t, as in Section 2, will be used when the unit root process is not assumed to have a constant or other deterministic trends.

It can be shown (see the Extra tag below) that the polynomial $a(\mathsf{B})$ can be rewritten as

$$a\left(B\right) = \Delta - \phi B - \phi_1 \Delta B - \cdots - \phi_{p-1} \Delta B^{p-1}, \quad \phi = -a\left(1\right),$$

so the process z_t can be represented as

$$\Delta z_t = \phi z_{t-1} + \phi_1 \Delta z_{t-1} + \cdots + \phi_{p-1} \Delta z_{t-p+1} + u_t, \quad t = 1, 2, \dots,$$

where $u_t \sim \mathsf{iid}(0, \sigma^2)$ (or that u_t is white noise).

Let's assume that for $a(z)$ the following holds:

$$\text{If } a(z) = 0, \text{ then } |z| > 1 \text{ or } |z| = 1.$$

In other words, the roots of the polynomial $a(z)$ lie outside or at the unit circle on the complex plane.

- If all the roots lie outside the unit circle, z_t is (at least) asymptotically stationary (as meeting the stationarity condition of the AR(p) process).

Let us now assume there is exactly one **unit root**. From the known properties of polynomials, it follows that

$$a(z) = (1 - z)b(z),$$

where $b(z)$ is a polynomial of the degree (at most) $p-1$ with its roots inevitably lying outside the unit circle (see the Extra below).

- It is clear that this is equal to $\phi = 0$ above. When $\phi = 0$, the above-mentioned polynomial becomes $b(z) = 1 - \phi_1 z - \cdots - \phi_{p-1} z^{p-1} := \phi(z)$.

Thus, Δz_t is $I(0)$ and as the roots of the polynomial $\phi(z)$ lie outside the unit circle, such initial values $\Delta z_{t-1}, ..., \Delta z_{t-p+1}$ can be found that Δz_t is stationary.

- Therefore, the process Δz_t can be written as $\Delta z_t = \phi(B)^{-1}u_t$, which is a special case of the random walk we looked at earlier but with the stationary AR$(p-1)$ process $\phi(\mathsf{B})^{-1}u_t$ replacing the innovation u_t.

Extra: Polynomials and power series

Consider polynomial function of power m

$$p\left(z\right) = \sum_{k=0}^{m} a_k z^k,$$

where a_k and z are real or complex numbers and $a_m \neq 0$. In the case of complex numbers $z \in \mathbb{C}$, by the *fundamental theorem of algebra* the polynomial $p(z)$ can be given the form

$$p\left(z\right) = a_m\left(z-\zeta_1\right)\cdots\left(z-\zeta_m\right),$$

where $\zeta_1, ..., \zeta_m$ are the roots of $p\left(z\right)$ and thus $p\left(\zeta_k\right) = 0$. The case $\zeta_k = \zeta_l$ $(k \neq l)$ is possible and in the case of a real coefficient $a_k \in \mathbb{R}$ the roots occur as complex conjugates for all k; that is if $\zeta_k = x_k + iy_k$ $(i^2 = -1)$, for some $l \neq k$ it holds $\zeta_l = x_l - iy_k =: \bar{\zeta}_k$.

Let's return to a normal polynomial $p\left(z\right) = \sum_{k=0}^{m} a_k z^k$. It is easy to conclude that $p(z)$ can be written as

$$p\left(z\right) = p\left(1\right) + \left(1-z\right)q\left(z\right),$$

where

$$q\left(z\right) = \sum_{k=0}^{m-1} b_k z^k, \quad b_k = -\sum_{j=k+1}^{m} a_j \quad k = 1, ..., m-1.$$

A corresponding alternative way to present this would be

$$p\left(z\right) = p\left(1\right)z + \left(1-z\right)r\left(z\right),$$

where

$$r\left(z\right) = \sum_{k=0}^{m-1} c_k z^k, \quad c_0 = a_0 \quad \text{and} \quad c_k = -\sum_{j=k+1}^{m} a_j,\ k = 1, ..., m-1$$

These can be generalized to the power series, that is to say to the case

$$p\left(z\right) = \sum_{k=0}^{\infty} a_k z^k \quad \left(\left|z\right| \leq 1\right).$$

If we assume

$$\sum_{k=1}^{\infty} k\left|a_k\right| < \infty,$$

it can be easily concluded that the above equations hold when $|z| \leq 1$ and m are replaced with ∞. In addition $\sum_{k=1}^{\infty} |b_k| < \infty$ and $\sum_{k=1}^{\infty} |c_k| < \infty$.

As discussed above in connection to integrated and ARIMA processes, the presence of a unit root is particularly interesting question from an economic point of view. In models (and processes) with unit roots, shocks (such as policy or technological interventions and disruptions) have persistent effects that last forever, whereas with stationary cases such shocks have only a temporary effect. Therefore, it is of particular interest to test the unit root hypothesis in y_t.

12.4 Testing for a unit root: ADF test

Let us continue to work with unit root processes and examine how the unit root can be tested. Among various alternatives, we focus on the Augmented Dickey-Fuller (ADF) test, which is the most commonly used unit root test.

Additional background and motivation for unit root testing are provided below in the Extra tab before proceeding to the ADF test.

Extra: Details on the above unit root test statistics

Let us continue with the autoregressive unit root process from the previous section

$$\Delta z_t = \phi z_{t-1} + u_t, \quad t = 1, 2, ...,$$

where for the sake simplicity $z_0 = 0$ and $u_t \sim \mathsf{iid}(0, \sigma^2)$. The question of interest is to **test the unit root hypothesis**

$$H_0 : \phi = 0$$

against the stationary (or stable) alternative of $\phi < 1$.

It seems natural to base the test on the least squares estimator (or maximum likelihood estimator) of ϕ:

$$\hat{\phi} = \left(\sum_{t=1}^{T} z_{t-1}^2 \right)^{-1} \sum_{t=1}^{T} z_{t-1} \Delta z_t \overset{H_0}{=} \left(\sum_{t=1}^{T} z_{t-1}^2 \right)^{-1} \sum_{t=1}^{T} z_{t-1} u_t,$$

where T is the number of observations. The unit root test can be based directly on the estimator $\hat{\phi}$, but it is more common to use the "t-ratio" printed by different statistical programs

$$\tau := \frac{\sqrt{\sum_{t=1}^{T} z_{t-1}^2}}{\hat{\sigma}} \hat{\phi},$$

which follows, under the null (unit root) hypothesis, nonstandard asymptotic distribution.

- The limiting distribution is independent of the unknown parameters and can be tabulated (tables are often presented in books and computer programs).
- The percentage points of the limiting distributions usually require numerical methods or simulations. In fact both the estimator $\hat{\phi}$ and the "t-ratio" τ have a non-Gaussian and left skewed asymptotic distribution. For example, with the τ test statistic the 5%-percentage point is about -1.94 while with a $\mathsf{N}(0,1)$ it is -1.645, and thus the unit root hypothesis gets rejected too easily.

The above unit root test can be generalized into a case with deterministic constant and trend (if appropriate due to shape of the time series or background information on it). A common model for this is

$$y_t = \mu_t + z_t, \quad t = 1, 2, \dots,$$

where z_t is as determined above (using x_t) and the deterministic trend is

$$\mu_t = \mu \quad \text{or} \quad \mu_t = \mu_1 + \mu_2 t.$$

Indicator variables can also be used to account for seasonal variation if necessary.

- The limiting distribution of the "t-ratio" can again be tabled. It deviates from the one with no constant and is more skewed to the left (5% percentage point is now -2.86).
- The skewedness of the asymptotic distribution keeps growing when the "t-ratio"-based test is generalized for the linear trend $d_t = \mu_1 + \mu_2 t$ (the details will be skipped now).

The asymptotic distribution of the estimator $\hat{\phi}$: Under the null hypothesis, then $z_t = \sum_{j=1}^{t} u_j$ is a (Gaussian) random walk, so the usual limit theorems do not hold. Moreover, it can be shown that the "t-ratio" follows the following (non-standard) asymptotic distribution

$$\tau := \frac{\sqrt{\sum_{t=1}^{T} z_{t-1}^2}}{\hat{\sigma}} \hat{\phi} \xrightarrow{d} \left(\int_0^1 W(u)^2 \, du \right)^{-1/2} \int_0^1 W(u) \, dW(u),$$

where $\hat{\sigma}^2 = (T-1)^{-1} \sum_{t=1}^{T} (\Delta z_t - \hat{\phi} z_{t-1})^2$ is the usual regression estimator for the error variance σ^2 (T can also be used as the denominator) and $W(u)$ is standard Brownian motion.

In the latter case with the trend component, by multiplying the equation above on both sides by $\Delta - \phi \mathsf{B}$, reorganizing the terms and setting $\mu_t = \mu$ we get

$$\Delta y_t = \nu + \phi y_{t-1} + u_t,\ t = 1, 2, ...,$$

where $u_t \sim \mathsf{nid}\left(0, \sigma^2\right)$ and $\nu = -\phi\mu$. When the unit root hypothesis holds, for the "t-ratio" τ_μ of the parameter ϕ holds

$$\tau_\mu \xrightarrow{d} \left(\int_0^1 \bar{W}\left(u\right)^2 du \right)^{-1/2} \int_0^1 \bar{W}\left(u\right) dW(u),$$

where $\bar{W}$ is the centered version of the standard Brownian motion.

Augmented Dickey-Fuller (ADF) test. Based on the discussion in the previous section, assume (for simplicity) that $y_t = \mu + z_t$, where $z_t \sim \mathrm{AR}(p)$, and hence y_t also follows an AR(p) process. To test for a unit root in y_t, we use the following test regression model

$$\Delta y_t = \nu + \phi y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + u_t, \quad u_t \sim \mathsf{iid}(0, \sigma^2), \quad t = 1, 2, ...$$

The null hypothesis is $H_0 : \phi = 0$, which implies that y_t has a unit root and is non-stationary. Under the alternative hypothesis $H_1 : \phi < 0$, the process is stationary around a deterministic mean.

- Here in the ADF test the intercept term ν is related to the mean of the process via $\nu = -\phi\mu$.
- The test statistic is the **t-ratio** of the estimated coefficient $\widehat{\phi}$, often denoted $\tau_\phi = \frac{\widehat{\phi}}{\mathrm{s.e.}(\widehat{\phi})}$. Its asymptotic distribution under the null is non-standard but known, and critical values are tabulated in the literature.
- Importantly, the asymptotic properties of the test hold even without assuming normality of the errors; it is sufficient that u_t are independent and identically distributed with finite variance.

If the null hypothesis of unit root ($H_0 : \phi = 0$) cannot be rejected, according to the above equation, an AR($p - 1$) model can be built on the differences.

- In other words, this implies the need to take first differences to get stationary time series.

Determining the lag length p and deterministic terms in the ADF test. Executing the ADF test regression requires specifying the lag length p, which can be selected using information criteria (such as AIC or BIC), sequential testing (covered in the next section), or based on common fixed choices related to the data frequency.

- It is generally advisable to try a few different values of p and compare the resulting test outcomes for robustness.

The ADF test regression can also be extended to include a trend (together with the constant). In this case, the test regression includes the term $\mu_0 + \mu_1 t$, allowing for a deterministic trend in the data:

$$\Delta y_t = \mu_0 + \mu_1 t + \phi y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + u_t.$$

The null hypothesis remains $H_0 : \phi = 0$, indicating a unit root.

Deciding which deterministic terms to include is a crucial step. A constant term is typically included by default, while the inclusion of a trend term depends on whether a trending pattern is evident in the data. This decision can be guided by graphical inspection or theoretical considerations relevant to the application.

Empirical examples. Let us consider unit root testing for the few time series considered above in different parts. As discussed, the first thing in the ADF test is to determine which deterministic terms to include in the test regression. Basically you should always include a constant term, but whether the deterministic trend component should be included as well should be determined based on visual inspection and background knowledge on the time series at hand.

In the following illustrations, for simplicity, we fix the number of lagged differenced lags p in the ADF test to 6, 8 or 10 lags. The lag length could and should be varied to examine whether the testing results are intact for this selection.

(i) Consider **log** of the CPI time series (Introduction): In this case, there is clearly an upward trend present. It seems reasonable to argue that due to different (economic) reasons we can think that price level is steadily going up and a deterministic trend seems reasonable to be included in the ADF test.

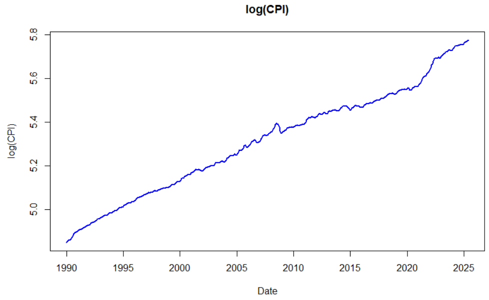


Figure: Monthly log(CPI) time series (1990:1–2025:6).

```
ADF-test (deterministic terms: constant and trend)
p=8: Test statistic: -1.113, p-value: 0.925
p=6: Test statistic: -1.517, p-value: 0.823
p=10: Test statistic: -1.451, p-value: 0.845
```

(ii) Log of the S&P 500 index. Increasing stock prices point out very well a possible stochastic trend, but it seems difficult to argue that there is deterministic upward trend in stock prices. Therefore, including a constant is surely good choice, but for robustness we could also examine the unit root hypothesis also when including deterministic trend as well (together with the constant) in the ADF regression.

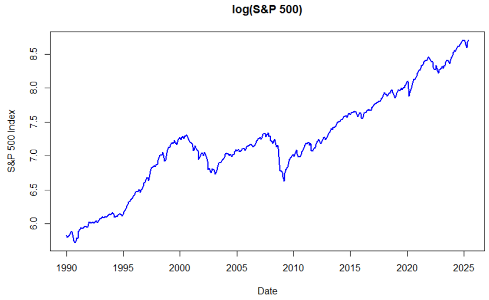


Figure: Monthly log(S&P 500) time series (1990:1-2025:6).

```
ADF-test (deterministic terms: constant)
p=8: Test statistic: -0.769, p-value: 0.826
p=6: Test statistic: -0.302, p-value: 0.922
p=10: Test statistic: -0.772, p-value: 0.825

ADF-test (deterministic terms: constant and trend)
p=8: Test statistic: -2.193, p-value: 0.492
```

(iii) Taking the log-differences of the CPI and S&P 500 indices results in the time series previously shown in the Introduction. A visual inspection alone suggests that the unit root hypothesis is no longer reasonable for these differenced series. To confirm this, let's examine the Augmented Dickey-Fuller (ADF) test results below. The test includes a constant, as a deterministic trend is clearly no longer present in the data.

Putting the testing results together, the results show that the ADF tests fail to reject the unit root hypothesis for the original level series of both the CPI and the S&P 500.

- For the CPI levels, even after including a constant and a linear trend in the test regression, we still cannot reject the null hypothesis of a unit root. This strongly suggests that the CPI can be treated as an $I(1)$ process (integrated of order one).
- A similar conclusion holds for the S&P 500 index.

These conclusions are reinforced by the test results for their log-differences, which appear to be stationary (i.e., $I(0)$ variables). This provides a consistent picture: the original series are non-stationary $I(1)$, but their first differences are stationary ($I(0)$).

```
ADF-tests (deterministic terms: constant)

logdiff(CPI):
p=8: Test Statistic: -7.744 , p-value: 1.248e-11
p=6: Test Statistic: -4.764, p-value: 7.568e-05
p=10: Test Statistic: -7.886, p-value: 5.331e-12

logdiff(S&P 500)
p=8: Test statistic: -6.095, p-value: 1.303e-07
p=6: Test statistic: -7.011, p-value: 8.741e-10
p=10: Test statistic: -5.451, p-value: 3.299e-06
```

A critical limitation: The low power of the ADF test. An important limitation of the ADF test is its low statistical power. In practical terms, this means: **The ADF test often fails to reject the null hypothesis of a unit root, even when the time series is actually stationary**.

- In other words, the test is not always very good at correctly identifying a stationary process, especially if the process is close to non-stationary (e.g., an AR(1) process where the autoregressive coefficient ϕ_1 is close to 1).
- If we are unable to reject the presence of a unit root, it does not necessarily mean that it is true and the process is necessarily $I(1)$. It could just be (but of course not always!) that there is not sufficient amount of information in the data to reject the unit root.
- The practical takeaway is that it's common for a truly stationary time series to produce a p-value from the ADF test that is too high to be considered statistically significant. You can observe this phenomenon by running the following code, which simulates AR(1) processes and computes the ADF test statistic, with a varying value of the AR(1) coefficient.

From empirical point of view, the choice whether the process is a unit root process (nonstationary) or a "near" unit root process (stationary) is interesting but at times complicated.

- Particularly ambiguous are, for example, interest rates. Interest rates are highly persistent and the unit root hypothesis cannot often be rejected, although nonstationary such as random walk interest rates do not seem to be very plausible from an economic point of view.

```
# Simulating AR(1) process with different values of phi_1
# - Change T and phi_1

# Load the libraries
library(urca)
library(tseries)

T=200         # number of observations
phi_1=0.98

epsilon=rnorm(T)  # random draws from nid(0,1)
y=epsilon
y[1]=0 # for simplicity and E(y_t) = 0
for(i in 2:T) y[i] = phi_1*y[i-1]+epsilon[i]
plot(y,type="l", main="Simulated realization")
```

```
test_ur_none <- ur.df(y, type = "none", lags = adf_lags)
stat_none <- test_ur_none@teststat[1, 1]
p_val_none <- punitroot(q = stat_none, N = length(test_ur_none@res), trend = "nc")
cat("P-value:", p_val_none, "\n")
```

Chapter 13

Linear regressions with I(1) variables

Instead of stationary variables, as considered earlier in Section 10, let us now assume that there are two variables, y_t and x_t, and either of them or both are nonstationary and specifically $I(1)$ variables.

- y_t is again the dependent variable and x_t is the predictive/explanatory variable.

Before continuing to examine linear (predictive) regressions between y_t and x_t in different cases, let us introduce an important concept of cointegration.

13.1 Basics of cointegration

An important special case of the linear regression between y_t and x_t, where y_t and x_t are nonstationary $I(1)$ variables, arises when there is a common stochastic trend in both series.

In other words, suppose that there is a linear relationship between variables so that there exists some value δ such that $y_t - \delta x_t$ is stationary ($I(0)$) although y_t and x_t are nonstationary ($I(1)$).

- In such case, it is said that y_t and x_t are **cointegrated** and they share the common trend.

In principle, an alternative way to construct a linear regression model, when y_t and x_t are $I(1)$, is based on the differences Δy_t and Δx_t which are then stationary ($I(0)$).

- This is not an optimal strategy if there is indeed a cointegration relationship between the variables.
- Cf. the discussion on overdifferencing in Section 12.

Let us examine cointegration more detail. **If y_t and x_t are $I(1)$ and cointegrated**, then

$$z_t = y_t - \delta x_t = [y_t \quad x_t][1 \quad -\delta]' \sim I(0).$$

The vector $[1 \quad -\delta]'$ is called the **cointegration vector**.

- Cointegration is often interpreted as a long-run relationship between the variables. Assume that the equilibrium of the variables y_t and x_t is defined by the relationship $y_t = \tilde{\delta} x_t$ for some fixed $\tilde{\delta}$. Then $\hat{z}_t = y_t - \hat{\delta} x_t$ is the "equilibrium error" which measures the extent y_t deviates from its "equilibrium value".
- As z_t is $I(0)$, the equilibrium error is stationary and fluctuating around zero. In other words, on average, the system is in equilibrium.

Notice that we can also use a slightly modified definition of cointegration for the above, especially when there will be more than two variables involved. We will consider this definition, and cointegration in connection to the VAR model more detail in the **Advanced Time Series Econometrics course**.

13.2 Testing cointegration between two variables

From the discussion above, it is obvious that it is important to distinguish whether there is a cointegration relationship between the variables.

Consider a "cointegration regression" between two $I(1)$ variables y_t and x_t:

$$y_t = \alpha + \delta x_t + u_t.$$

If y_t and x_t are cointegrated, then u_t is $I(0)$. If not, u_t will be $I(1)$. Therefore, after the model is estimated by OLS, the presence of a possible cointegration relationship can be evaluated by testing for a unit root in the residuals $\hat{u}_t$.

Testing for cointegration with a known coefficient. In some applications, the cointegration coefficient δ (and the intercept α) is known in advance due to economic theory or other application-specific knowledge. When the coefficient is known, the cointegration relationship can be tested as follows. For the sake of simplicity, we will assume $\alpha = 0$ in the procedure described below

- Construct the series $z_t = y_t - \delta x_t$.
- Use the ADF test for the null hypothesis of unit root, which implies $z_t \sim I(1)$. If H_0 is rejected, then there is a cointegration relationship.

Testing for cointegration with unknown coefficients. If δ (and α) are unknown, then we can proceed as follows (Engle-Granger ADF-test):

- Estimate the following model by using OLS

$$y_t = \alpha + \delta x_t + u_t.$$

- Construct the series $\hat{u}_t = y_t - \hat{\alpha} - \hat{\hat{\delta}} x_t$, which is the residual of the regression.
- Test the null hypothesis implying $\hat{u}_t \sim I(1)$.
- Critical values are now different than in the above case, and in the ADF test, as unit root testing is based on the (estimated) residuals.

Notice that if y_t and x_t are indeed cointegrated, OLS yields a consistent estimator for the cointegration coefficient δ. However, the OLS estimator of δ has a non-normal asymptotic distribution, and the inferences based on the standard t-test statistic can be misleading.

Emprirical example: Consumption and income. Let us consider an important empirical question derived from Hall's (1978) Permanent Income Hypothesis (PIH) in the context of cointegration.

- R. E. Hall (1978). Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. Journal of Political Economy, 86(6), 971–987.
- Source of data: FRED.

The testable hypothesis is that there is a stable long-run relationship exist between real consumption and real income (that is for inflation adjusted consumption and income), such that the difference between them (the cointegration residual) is stationary.

- In other words, the cointegration theory predicts that since consumption c_t is proportional to permanent income y_t so that $c_t = a\, y_t$, $0 \leq a \leq 1$ (the present value of expected future income), and both are driven by a common stochastic trend (permanent income shocks), they should be cointegrated. That is there is a long-term equilibrium between c_t and y_t, and if c_t and y_t are now $I(1)$ processes, then there should be a cointegration relationship between the variables.

So, let us test the prediction of the PIH that real per capita consumption ($\log(c_t)$) and real per capita disposable income ($\log(y_t)$) share a single cointegrating relationship.

- Hypothesis: The two variables ($\log(c_t)$ and $\log(y_t)$) are $I(1)$ and cointegrated.
- Long-run equation: $\log(c_t) = \alpha + \delta \log(y_t) + u_t$
- PIH implication: The cointegrating vector should be $(1, -\beta_1)$, where β_1 is theoretically close to 1 (or exactly 1 in the simplest PIH model) because, in the long run, consumption should grow proportionally with permanent income. The error term u_t and the resulting residuals $\hat{u}_t$ must be $I(0)$ (stationary).

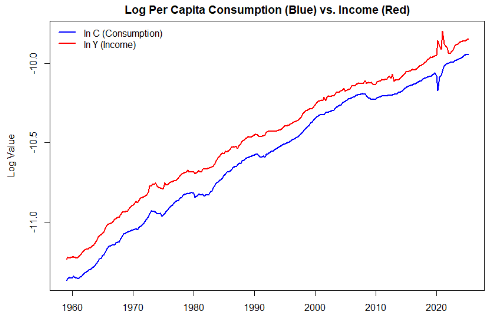


Figure: Monthly U.S. data (sample period 1959:1–2025:8) on (log) consumption and income. (Source: FRED)

Check first the unit root hypothesis for the two time series. In these ADF tests, a constant and linear trend component is included in the test regression and the number lagged differences p is 4:

```
ADF tests (deterministic terms: constant and trend)

log(c_t):
p=4: Test Statistic: -2.005, p-value: 0.595
p=6: Test Statistic: -2.160, p-value: 0.510

log(y_t)
p=4: Test Statistic: -2.583, p-value: 0.289
p=6: Test Statistic: -2.473, p-value: 0.341
```

Estimation result of the cointegration regression:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.383122   0.044448    8.62 6.29e-16 ***
ln_Y        1.048293   0.004244  246.99  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02692 on 264 degrees of freedom
Multiple R-squared:  0.9957,    Adjusted R-squared:  0.9957
F-statistic: 6.1e+04 on 1 and 264 DF,  p-value: < 2.2e-16
```

Based on the Engle-Granger two-step procedure, the null hypothesis of no cointegration can be rejected at 5% level (see testing result below). Despite one large outlier related to the Covid-10, the cointegration residual series seems to fluctuate as stationary I(0) variable.

- In the ADF test below, we are using the critical values of the ADF test even though, as discussed above, that is not completely correct.

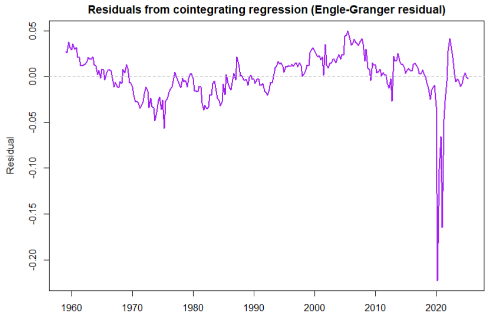


Figure: PIH residual time series (from the estimated cointegrated model).

```
Test Statistic: -4.561388
> cat("P-value:", p_val_drift, "\n")
P-value: 0.000197051
```

13.3 Linear regressions containing I(1) variables

Recall the linear regression

$$y_t = x_t'\beta + u_t,$$

and the two assumptions (i)–(ii) set in Section 10:

- (i) The error term u_t is serially uncorrelated and uncorrelated with the regressors included in x_t.
- (ii) All the regressors in x_t are either deterministic or stationary random variables.

Assume now that only (i) **holds**, that is the error term u_t is serially uncorrelated and uncorrelated with the regressors included in x_t.

- Even if the model cannot be written in this way, the OLS estimator of the coefficients of the $I(1)$ regressors x_t is consistent. However, its asymptotic distribution is, in general, nonstandard such that usual inference does not apply.
- If the model can be written such that all the parameters of interest are coefficients of mean zero stationary variables, their OLS estimator is consistent and asymptotically normal.

To examine these points more detail, let us consider again the following relatively simple regression model

$$y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + u_t,$$

where $x_t \sim I(1)$. The model can be rewritten as

$$y_t = \beta_0 + (\beta_1 + \beta_2)x_{t-1} - \beta_2(x_{t-1} - x_{t-2}) + u_t,$$

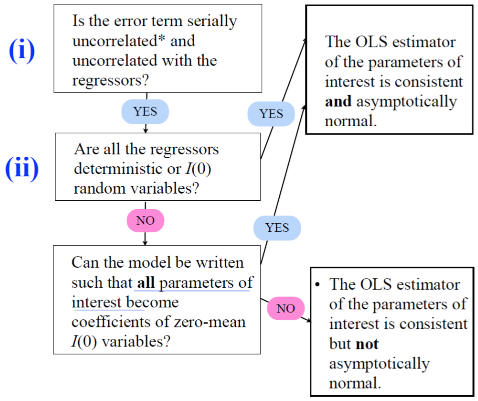
or

$$y_t = \beta_0 + \beta_1(x_{t-1} - x_{t-2}) + (\beta_1 + \beta_2)x_{t-2} + u_t.$$

As $(x_{t-1} - x_{t-2}) \sim I(0)$, and hence, standard inference on β_2 (or β_1) holds. Therefore, both β_1 and β_2 cannot simultaneously be written as coefficients of $I(0)$ variables (unless higher lags are included in the model).

- The OLS estimator of β_1 and β_2 is not, in general, jointly asymptotically normal.
- The test statistic on a hypothesis concerning both coefficients (for instance $H_0 : \beta_1 = \beta_2$), in general, does not have the usual asymptotic χ^2 distribution.

These remarks are compiled to the next figure.


(i)
Is the error term serially uncorrelated* and uncorrelated with the regressors?
YES
(ii)
Are all the regressors deterministic or I(0) random variables?
NO
Can the model be written such that all parameters of interest become coefficients of zero-mean I(0) variables?
YES
YES
The OLS estimator of the parameters of interest is consistent and asymptotically normal.
NO
• The OLS estimator of the parameters of interest is consistent but not asymptotically normal.

* Not autocorrelated after H-1 lags.

Figure: Cheat sheet on inference in OLS regression models with $I(1)$ variables (References: Stock and Watson (1988) and lecture notes by Markku Lanne).

Let us continue with linear regression where now also the assumption (i) does not hold and the dependent variable is $I(1)$.

- If the dependent variable is not cointegrated with any of the regressors, the OLS estimator of the coefficients of the $I(1)$ regressors is inconsistent.
- If the dependent variable is cointegrated with at least one of the regressors, the OLS estimator of the parameters of interest
  - that can be written as coefficients of stationary variables, is inconsistent.
  - that cannot be written as coefficients of stationary variables, is consistent, but not asymptotically normal.

For instance, consider the following regression model:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t,$$

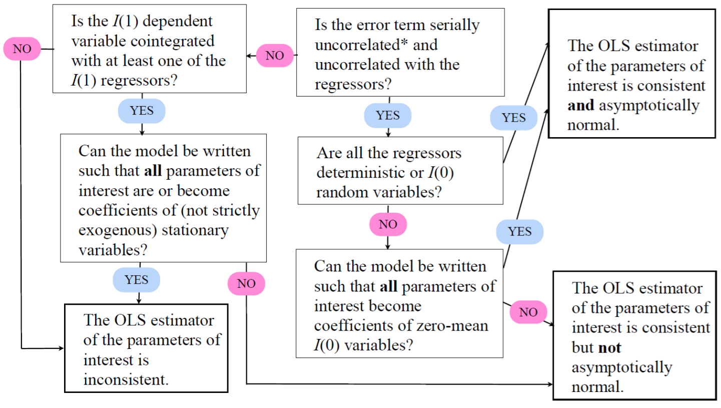
where all variables are $I(1)$, and y_t and x_{2t} are cointegrated and assumption (i) does not hold.

- If x_{1t} and x_{2t} are cointegrated such that $(x_{1t} - \gamma x_{2t}) \sim I(0)$, the model can be written as

  $$y_t = \beta_0 + \beta_1(x_{1t} - \gamma x_{2t}) + (\beta_1\gamma + \beta_2)x_{2t} + u_t,$$

  and the OLS estimator of β_1 is inconsistent.
- If x_{1t} is not cointegrated with x_{2t}, β_1 cannot be written as a coefficient of an $I(0)$ variable, and hence, its OLS estimator is consistent, but not asymptotically normal.

The complete *cheat sheet* on linear regression with $I(0)$ and $I(1)$ variables is summarized hereby:


Is the I(1) dependent variable cointegrated with at least one of the I(1) regressors?
NO
NO
Is the error term serially uncorrelated* and uncorrelated with the regressors?
YES
YES
YES
The OLS estimator of the parameters of interest is consistent and asymptotically normal.
Can the model be written such that all parameters of interest are or become coefficients of (not strictly exogenous) stationary variables?
Are all the regressors deterministic or I(0) random variables?
NO
YES
YES
NO
Can the model be written such that all parameters of interest become coefficients of zero-mean I(0) variables?
NO
The OLS estimator of the parameters of interest is consistent but not asymptotically normal.
The OLS estimator of the parameters of interest is inconsistent.

* Not autocorrelated after H-1 lags.

Figure: Complete cheat sheet on inference in OLS regression models with $I(1)$ variables.

A spurious regression emerges when a statistically significant relationship between two $I(1)$ variables y_t and x_t is found although those are completely unrelated. Assume that a researcher estimates a linear regression

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

but assuming wrongly that the variables y_t and x_t are stationary. Instead, they are generated by two independent random walks (i.e., y_t and x_t are $I(1)$ variables)

$$\begin{aligned} y_t &= y_{t-1} + u_{1t}, \quad u_{1t} \sim \text{iid}(0, \sigma_1^2) \\ x_t &= x_{t-1} + u_{2t}, \quad u_{2t} \sim \text{iid}(0, \sigma_2^2). \end{aligned}$$

This typically leads to the estimation result of the above linear regression characterized by a fairly high R^2 (the coefficient of determination), highly autocorrelated residuals $\hat{u}_t$ and a (highly) statistically significant estimate of β_1. This result is clearly spurious given that the variables are completely independent!

- Estimation results like this should not be taken seriously.
- The reason is that with y_t and x_t being $I(1)$ variables, the error term u_t will also be $I(1)$, not stationary ($I(0)$).

Solution to this problem is to include the lags of y_t and x_t as predictors in the model. In other words, we end up a model (cf. dynamic regression models in Section 10)

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \beta_3 x_{t-1} + u_t.$$

In this case the OLS estimator is consistent. Thus, in general, including lagged values in the regression is sufficient to solve many of the problems associated with possibly spurious regressions.

An example of a spurious regression situation can be replicated with the following program code in R lab.