Notes on Economic Time Series

Riccardo 'Jack' Lucchetti

4th August 2026

Contents

List of Figures

Preface

This book was originally conceived as a set lecture notes for my Econometrics course. As such, I was never particularly concerned with either rigour nor completeness. On the contrary, my main goal was to describe various concepts in an intuitive way and to try to motivate definitions and main results as explicitly as possible.

Over time, things evolved and these lecture notes grew: I no longer use them in a basic Econometrics course, but rather for more advanced courses. However, the core philosophy has remained the same: a text that can be "read" as well as "studied". Consequently, with a few exceptions, I will generally refer "to the literature" for explanations, proofs, and further insights, without citing specific sources. This is because, given my purpose, I felt it would be more useful to group the bibliographical references into a final chapter, which would also serve to guide the reader through the *mare magnum* of time series econometrics.

Over the years, I have received a great deal of *feedback* from many people, whom I thank for helping to improve the content. Among my peers, I would like to mention in particular (without holding them responsible for any remaining errors) Gianni Amisano, Marco Avarucci, Emanuele Bacchiocchi, Nunzio Cappuccio, Francesca Di Iorio, Luca Fanelli, Massimo Franchi, Carlo Favero, Roberto Golinelli, Diego Lubian, Giulio Palomba, Matteo Pelagatti, Eduardo Rossi, Maurizio Serva, Stefano Siviero, and Gennaro Zezza. Carlo Giannini deserves a special mention because, without him, I would probably have done something completely different in life, and these lecture notes would never have existed; I would certainly have been a worse person.

A grateful thought also goes to all those who were forced to endure these lecture notes as a textbook, and who prompted me to be more thorough and clear (or less incomplete and obscure, depending on one's point of view) when they pointed out to me, either in words or simply through their facial expressions, that it made no sense at all. I would prefer not to name names because there are too many, but I must make an exception for Gloria Maceratesi, whom I cannot fail to mention because her efficiency as a proofreader was nothing short of superhuman.

The fact that these lecture notes are freely available on the Internet has also led many people to download them, and some have even written me an email with advice and suggestions. In this case too, this is deeply appreciated, if only for the ego massaging.

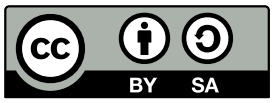
A massive thank you goes to Allin Cottrell, the unstoppable force behind

the `gretl` project: for those who do not know, `gretl` is a free[1] econometric package with which all the examples in these lecture notes were created. To learn more, and perhaps download it, go to http://gretl.sourceforge.net.

Finally, I must express my gratitude to Giulio Palomba (again), for having undertaken the task of translating this book from Italian into English, something I've shirked from for waaaay to long. Thank you, bro. I owe you.

As for the prerequisites, I assume the reader already has a certain degree of familiarity with the main probabilistic concepts (random variables, expected values, conditioning, various notions of convergence), with OLS, and with some basic concepts of estimation theory, such as identification and the properties of estimators. Therefore, those who have not already studied them might as well close this text right here and go hit the books. The rest of you, make yourselves comfortable, and let's get started.

Some passages are written in a smaller font, in two columns, like this one. They are not essential, and can be skipped without affecting the understanding of the rest. But then, if I wrote them, they must serve some purpose. It's up to you.



[1] Which *also* means open-source/gratuitous. The expression *free software*, however, means that the source code is available. In any case, it is unforgivable to confuse *free software* with *freeware*, which is simply software that can be used legally without paying.

Chapter 1

Introduction

1.1 What is a stochastic process? What is is good for?

The data we use in econometrics generally fall into two categories: *cross-sectional*, when the available observations are relative to different individuals, or *time series*, when what we have are observations, on one or more variables, collected over time[1].

In the first case, thinking of a set of N observed data as one of the possible realisations of N independent and identical random variables is not too unsustainable a hypothesis: when recording the weight and height of N individuals, there is no reason to think that

1. the physical characteristics of the i-th individual are in some way connected to those of the other individuals (*independence*);

2. the relationship between weight and height that holds for the i-th individual is different from that which holds for all the others (*identity*).

In these cases, we use the concept of the realisation of a random variable as a metaphor for the i-th observation, and the appropriate inferential apparatus is no different from the elementary one, in which independence and identity allow us to write

$$f(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} f(x_i),$$

that is, that the density function of our sample is simply the product of the density functions of the individual observations (which are all the same). When we use linear regression, this framework substantially leads us to the so-called "classical assumptions", extensively analysed at the beginning of any Econometrics course. Note that this type of reasoning is perfectly appropriate in the most cases, where the observed data come from a controlled experiment, of the type of those used by physicians or biologists.

[1] In fact, an intermediate case is given by the so-called *panel* data, but we do not deal with it here.

The time series context, however, is characterised by a basic conceptual difference that requires an extension of the probabilistic concepts to be used as a metaphor for the data: time has a direction, and history exists. Indeed, the natural tendency of many phenomena to evolve smoothly over time suggests that the data observed at a given time t is more similar to that recorded at time $t-1$ rather than in distant periods; in practice, it can be said that the time series we observe possesses some degree of "memory of itself". This characteristic is generally known as **persistence**, and it sets time series data apart from cross-sectional ones, because in the former the order of the data is of fundamental importance, while in the latter it is completely irrelevant.

The tool we use as a probabilistic metaphor for the observed time series is the **stochastic process**. A non-rigorous definition of a stochastic process, which is intuitive and substantially correct for our needs, can be the following: *a stochastic process is an infinitely long sequence of random variables* or, equivalently, a random vector of infinite dimension. A sample of T consecutive observations over time is therefore not conceived as a realisation of T distinct random variables, but rather as part of *a single* realisation of a stochastic process, whose memory is given by the degree of connection between the random variables that it contains.

1.2 Characteristics of Stochastic Processes

The definition above (where I cunningly avoided a series of technical complications) implies several properties of stochastic processes, that are rather important for our work: given a stochastic process whose t-th element[2] is denoted by x_t,

- conceptually, it is possible to define a density function for the process $f(\ldots, x_{t-1}, x_t, x_{t+1}, \ldots)$;
- it is possible to marginalise such a density function for any subset of its components; from this it follows that the marginal density functions are defined for each of the x_t, but also for any pair of elements (x_t, x_{t+1}), any triple and so on; furthermore, the fact that the observations may not be independent of one another implies that the sample density can no longer be written as a simple product of the marginals;
- if the marginal density functions have moments, it is possible to say, for example, that $E(x_t) = \mu_t$, $V(x_t) = \sigma_t^2$, $Cov(x_t, x_{t-k}) = \gamma_{k,t}$ and so forth;
- in the same way, it is possible to define conditional distributions, with objects like $f(x_t|x_{t-1})$, and possibly the relative moments.

The properties I just described refer to stochastic processes as probabilistic structures. However, when we want to use these structures in inferential procedures, two main problems arise:

[2] To be pedantic, we should use two different notations for the whole *stochastic process*, and each generic element in it. If the latter is denoted by x_t, the process to which it belongs should be written $\{x_t\}_{-\infty}^{+\infty}$. I consider this refinement superfluous, and I will use the same notation both for a process and for its t-th element; no confusion should arise.

1. if what I observe is just *one segment* of a *single* realisation of the many possible ones, the logical possibility of making inference on the process cannot be taken for granted: there is no way to tell which features of the observed time series are specific to *that* realisation, and which instead would recur even when observing other, equally possible, ones;

2. even if it were possible, somehow, to use a single realisation to perform statistical inference on the characteristics of the process, it is necessary for it to be stable over time. In other words, the probabilistic features of such a process must remain unchanged, at least within my observation interval.

These two issues lead to the definition of two properties that stochastic processes may have or may not.

Stationarity We speak of a stationary stochastic process in two senses: **strong** (or **strict**) and **weak** stationarity.

To define strong stationarity, let us examine any possible subset of the random variables in the process; these do not necessarily have to be consecutive, but to help intuition, let us pretend that they are. Let us consider, therefore, a 'window' of length k on the process, that is a subset defined as $W_t^k = (x_t, \ldots, x_{t+k-1})$. Clearly, this is a k-dimensional random variable, with its own density function which, in general, may depend on t. If, however, it does not, then the distribution of W_t^k is equal to that of W_{t+1}^k, W_{t+2}^k and so on. We are in the presence of strong stationarity when this invariance holds for any k. In other words, when a process is strong stationary, the distributional characteristics of all the marginals remain constant over time.

Weak stationarity, on the other hand, concerns only windows of length 2, $W_t^2 = (x_t, x_{t+1})$, and implies the further requisite that the first two moments exist. In this case, we don't require all the pairs $W_t^2 = (x_t, x_{t+1})$ to have the same distribution: we just want their means and covariances to be the same.[3]

Despite their names, one definition does not imply the other; for example, a process can be strong stationary but not possess moments[4]; conversely, the constancy of the moments over time does not imply that the various marginals have the same distribution. In one case, however, the two definitions coincide: this case — which is of particularly importance in practice — is the one in which the process is **Gaussian**, that is, when the joint distribution of any subset of elements of the process is a multivariate normal. If a process is Gaussian, establishing that it is weak stationary is equivalent to establishing strong stationarity. Since Gaussianity is very often assumed in empirical applications, from an

[3] For this reason, weak stationarity is also defined as **covariance stationarity**.

[4] For example: think of a sequence of random variables $y_t = 1/x_t$, where the x_t are independent standard normal variables. The sequence of y_t is a sequence of identical and independent random variables without moments. It's easy to see that the requirements for strong stationarity are met, but the ones for weak stationarity are not.

operational point of view the definition of weak stationarity is generally adopted. When speaking of stationarity without adjectives, we mean "weak".

Limited memory Generally speaking, we will want to work with processes where elements "far apart" can be thought of as "essentially" independent random variables. That is, the information that x_{t-k} carries on the distribution of x_t is practically nil, if k is large.

The most general concept is ergodicity: in a non-ergodic process, persistence is so strong that one realisation of the process, no matter how long, is insufficient to describe the process distribution. In an ergodic process, on the contrary, the memory of the process is weak over long horizons, and as the sample size increases, the information we get also increases significantly.

The formal conditions under which a stationary stochastic process is ergodic are too complex for a book like this one; I'll just say that, by observing the process for a "sufficiently" long period of time, it is possible to observe "almost all" the subsequences that the process is capable of generating. In other words, it can be said that, in an ergodic system, if something *can* happen, then sooner or later it *must* happen. The fact that events far apart in time can be considered independent for all practical purposes is then often summarised in the following property of ergodic processes (which is sometimes used as a *definition* of an ergodic process):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathrm{Cov}(x_t, x_{t-k}) = 0.$$

Consequently, if a process is ergodic, it is possible (at least in principle) to use the information contained in its observations over time to infer its characteristics. There is a theorem (appropriately called the 'ergodic theorem') which states that, if a process is ergodic, for inferential purposes the observation of a "long enough" realisation is equivalent to the observation of a large number of shorter, separate realisations. If, for example, an ergodic process x_t has an expected value of μ, then its average *over time* is a consistent estimator of μ (in formulas, $T^{-1}\sum_{t=1}^{T} x_t \xrightarrow{\mathrm{p}} \mu$), and therefore μ can be estimated consistently as if many realisations of the process are available instead of only one.

From a practical point of view, however, is much more common to rely on an alternative (and a bit more restrictive) concept, known as "mixing". The nice thing about it is that mixing processes can be shown to be ergodic, but mixing is much easier to understand and handle in practice.

There are several ways to define mixing processes. All of them rely on choosing two events A and B as follows: A is a statement pertaining on something that happened at time $t-m$ or before; B is a statement relative to something that happens on or after time t. Therefore, $P(A)$ is something you can compute using $f(\ldots, x_{t-m-1}, x_{t-m})$, while for $P(B)$ you need $f(x_t, x_{t+1}, \ldots)$. In other words, A is something that may have

happened in the past (a very distant past, if m is large). B is an event that may happen from now on.

Of course, the two events are independent if $P(A \cap B) = P(A)P(B)$. Therefore, consider the quantity

$$\alpha_m = |P(A \cap B) - P(A)P(B)|$$

If $\lim_{m\to\infty} \alpha_k = 0$ for all possible definitions of A and B, then whatever happens up to some point becomes irrelevant after k periods (this particular definition is known as **strong mixing** — there are several more).

In general, it can be said that inference is possible only if the stochastic process we study is stationary and ergodic. It must be said, moreover, that while there are methods to test the hypothesis of non-stationarity (at least in certain contexts, which we will examine later on), the hypothesis of ergodicity is not testable if all we observe is one realisation of the process is available, no matter how long.

1.3 Moments

By definition, for weakly stationary processes every element of the process x_t has a finite and constant expectation $E(x_t) = \mu$ and a finite and constant variance $V(x_t) = \sigma^2$. Furthermore, all covariances between different elements exist:

$$\gamma_k = E\left[(x_t - \mu)(x_{t-k} - \mu)\right]. \tag{1.1}$$

These are known as **autocovariances**. Stationarity guarantees that these quantities are not functions of t; however, they are functions of k, so we cab define the "autocovariance function", that is a function of k such that $\gamma(k) = \gamma_k$. Clearly, the autocovariance of order 0 is nothing but the variance. Furthermore, the definition is such that $\gamma_k = \gamma_{-k}$, since

$$E\left[(x_t - \mu)(x_{t-k} - \mu)\right] = E\left[(x_t - \mu)(x_{t+k} - \mu)\right].$$

From here, it's trivial to define the **autocorrelations**, which are given by

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma^2} \tag{1.2}$$

Obviously, $\rho_0 = 1$. These quantities, if different from 0, describe the memory of the process, and are precisely the element that makes stochastic processes the appropriate tool to represent time series exhibiting persistence: if $\gamma_1 \neq 0$, then we have that

$$f(x_t|x_{t-1}) \neq f(x_t)$$

and consequently

$$E(x_t|x_{t-1}) \neq E(x_t), \tag{1.3}$$

which can be translated as: if x_{t-1} is known, the expected value of x_t is not the same as what we would expect if x_{t-1} were unknown. We could also extend the set of random variables on which we condition to x_{t-2}, x_{t-3}, and so on. This set of random variables is sometimes called the **information set** at time $t-1$, and is denoted by $\Im_{t-1}$.

Strictly speaking, the exact definition of an information set is a bit more complex: we would need to deal with σ-algebras and be more rigorous about what is meant by conditioning. The topic is truly fascinating, but is way beyond the scope of this book. It is not difficult, however, to think of the information set as a collection of random variables with respect to which conditioning is possible. In a time series setting, it is reasonable to assume that the past is known; consequently, it makes sense to condition a random variable at time t on its own past values, because if x_t is known, then so are x_{t-1}, x_{t-2}, and so on. Naturally, the information set at time t may also include variables other than the one we are conditioning: for example, in the theory of rational expectations, the idea of an information set at time t is used as a synonym for *all* available information about the world at time t.

After observing a realisation of size T of a stochastic process x_t, we can define the sample counterparts of the theoretical moments:

$$\begin{aligned} \text{sample mean} \quad & \hat{\mu} = T^{-1} \textstyle\sum_{t=1}^{T} x_t \\ \text{sample variance} \quad & \hat{\sigma}^2 = T^{-1} \textstyle\sum_{t=1}^{T} (x_t - \hat{\mu})^2 \\ \text{sample autocovariance} \quad & \hat{\gamma}_k = T^{-1} \textstyle\sum_{t=k}^{T} (x_t - \hat{\mu})(x_{t-k} - \hat{\mu}) \end{aligned}$$

If the process is stationary and ergodic, it can be shown that these quantities are consistent estimators of the moments of the process[5] However, these statistics are also useful as descriptive statistics. Apart from the sample mean, whose interpretation I take for granted, analysing the autocorrelation function often allows us to make interesting qualitative assessments about the persistence characteristics of the series under consideration. In the next subsection, I will give a few examples.

1.4 Some examples

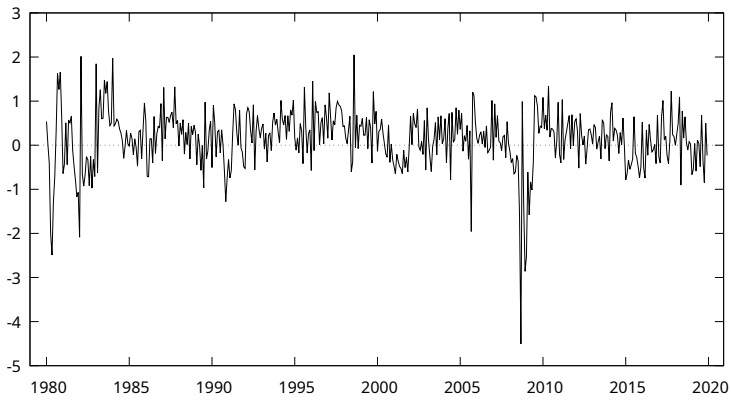
In this section, I'll use a few real-world time series to illustrate the issues discussed above. We observe a time series and assume that it can be considered a realisation of a stationary stochastic process; our aim is to find the process that represents this series "best" (whatever that may mean). More precisely, we will look for a stochastic process whose realisation are "qualitatively most similar" to our observed series.

Let us consider, for example, the time series shown in Figure 1.1, which reports the monthly percentage change in the industrial production index from the previous month for the United States, from 1980 to 2019. As can be seen, the series fluctuates fairly regularly around a central value, approximately between -5% and 3% (to be precise, its average is about 0.14%). If we were willing to assume that the process that generated these data is stationary and ergodic, we could say that this number is an estimate of the unconditional expected value of the process.

But is this process (provided that it exists) stationary? And if so, is it also ergodic? More generally: what are its persistence characteristics? It's difficult

[5]The careful reader will have noticed the absence of the 'degrees of freedom correction': in the denominator of the sample variance, for example, there is T instead of $T-1$. Asymptotically, the two formulations are obviously equivalent. What happens in finite samples is usually considered irrelevant or too complicated to be studied.

Figure 1.1: US industrial production: percentage changes



to give an answer just by looking at the graph, at least for the untrained eye. The analysis of the sample autocorrelations, reported in Figure 1.2, comes to the rescue.

Figure 1.2: US industrial production: percentage changes — correlogram

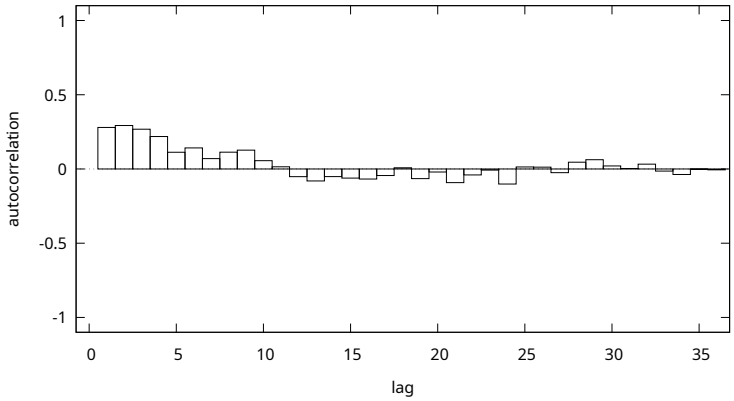


Figure 1.2 shows the **correlogram**, a bar chart in which each bar displays the value of the autocorrelation ρ_k against the order k, on on the horizontal axis. In other words, the correlogram can be interpreted as follows: if we denote by y_t the observation at time t, the correlation between y_t and y_{t-1} is about 28%, the one between y_t and y_{t-2} is 29.3%, and so on.

In an inferential setting, we would ask ourselves whether these estimate quantities (the autocorrelation of the process ρ_k) that are significantly different from 0, and proceed with some kind of hypothesis test. However, we can simply take them as descriptive statistics, with a clear meaning: nearby observations are positively correlated, and therefore they are unlikely to be independent, so there is clear evidence of some persistence.

Moreover, persistence seems to fade over time: one could say that, as the distance between observations increases, the absolute value of their correlation

(which we can, at this stage, take as an persistence indicator) tends to decrease: at a distance of 18 months, their correlation is considerably smaller (-0.53%). All in all, one could say that from a qualitative point of view, this is exactly what we expect to see in a realisation of a stationary and ergodic process: a persistence that substantially affects the series in the short run, but remains an essentially 'local' phenomenon.

Hence, one might ask whether the time series we are observing can be statistically modelled by studying its conditional mean, just as is done in a linear regression model. Indeed, if in a linear model the equation $y_t = x_t'\beta + \epsilon_t$ splits the dependent variable into a conditional mean plus an error term, nothing prevents us from making the conditional mean a function of the information set $\Im_{t-1}$, so we would use OLS on a model like:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \epsilon_t. \tag{1.4}$$

If we did so, using, for example, the values up to four months prior as the conditioning set, we would obtain the results shown in Table 1.1.

Table 1.1: OLS estimation of equation (1.4)

Coefficient	coefficient	std. error	t-statistic	p-value
β_0	0.0591	0.0293	2.0157	0.0444
β_1	0.1639	0.0457	3.5831	0.0004
β_2	0.1808	0.0459	3.9391	0.0001
β_3	0.1487	0.0459	3.2373	0.0013
β_4	0.0818	0.0458	1.7853	0.0748

Mean dependent var	0.140598	S.D. dependent var	0.664580
Sum squared resid	178.4200	S.E. of regression	0.612879
R^2	0.156640	Adjusted R^2	0.149538

The numbers in Table 1.1 can be taken as simple descriptive statistic, but also as estimators of *something*. In this case, then, their properties must necessarily be studied within an appropriate inferential framework. It is precisely for this reason that we need to study stochastic processes: to give a probabilistic meaning, if possible, to statistics like the ones we have just seen. In the following chapters, I will show how and why the estimation just performed actually makes sense, and how it should be interpreted.

Now let's turn to another example, where things are not so straightforward: Figure 1.3 reports the time series of inflation (defined as the annual percentage change in the consumer price index), again for the US.

Are we sure that a time series like this can be generated by a stationary process? As can be seen, periods (sometimes rather long ones) of high and low inflation alternate. Is it legitimate to think that the hypothetical process generating this series has a constant mean, as required for stationarity? Moreover: the correlogram, displayed in Figure 1.4) shows that treating persistence as a short-run phenomenon is decidedly more questionable. The autocorrelation corresponding to the 24th month is 48.1%, and the plot shows no sign of significant decay.

Highly persistent time series, like this one, are extremely common in economics and finance; to analyse them, one must find some way to represent

Figure 1.3: US inflation

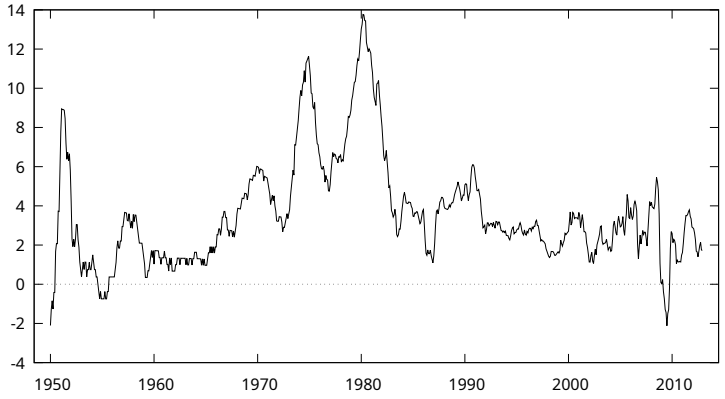
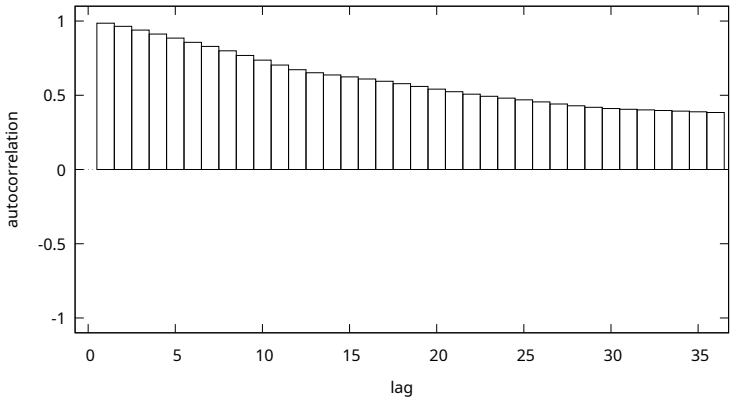
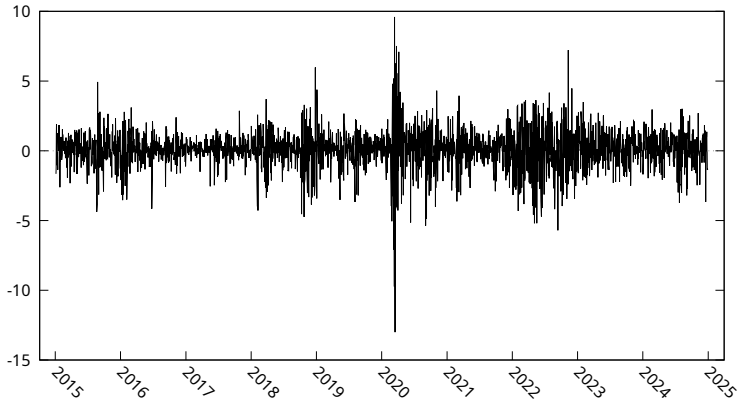


Figure 1.4: US Inflation – correlogram



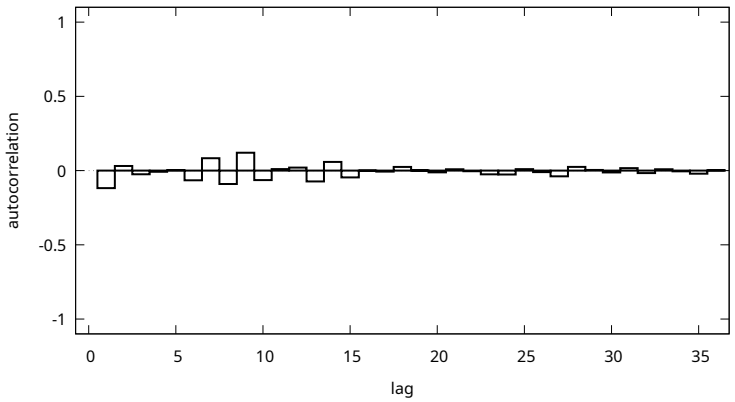
them as realisations of stationary processes, possibly after some suitable modification. In many cases, this can be done using techniques that dominated time series econometrics in the last two decades of the 20th century. These techniques, however, are somewhat complex and I'll defer their treatment to Chapter 3. Please, be patient.

Figure 1.5: Nasdaq stock market index – daily returns



I conclude this overview of examples with a quite different case: the (daily) percentage change in the Nasdaq stock index between 2015 and 2024, shown in Figure 1.5. The behaviour of this time series is — obviously — very different from the ones shown earlier: the data fluctuate around a value slightly above zero (the arithmetic mean is 0.061% — in other words, the Nasdaq index grew on average 0.061% per day over that period), and without the long fluctuations that characterised the industrial production or inflation series. This impression is confirmed by visual inspection of the correlogram in Figure 1.6.

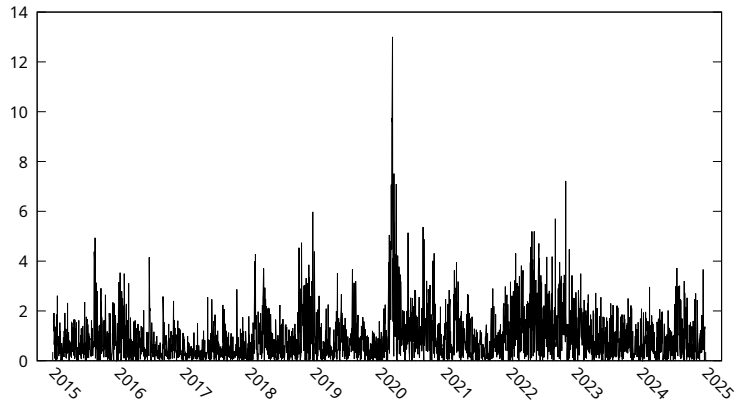
Figure 1.6: Nasdaq stock market index – Correlogram



We see very little persistence here. But then, there's a very good reason for this: no matter what technical analysis fans may say here, if there were a

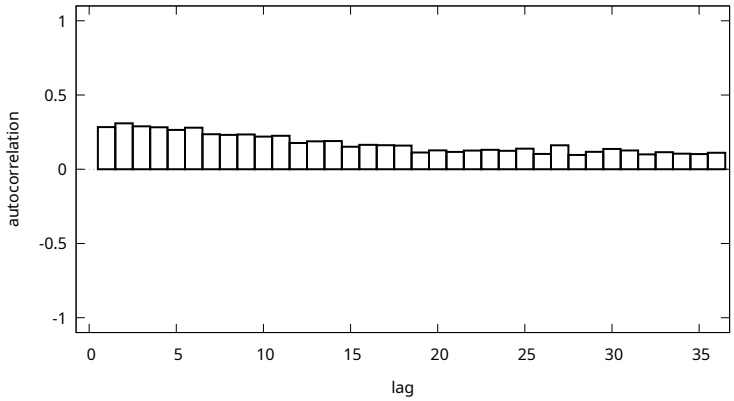
'simple' rule linking returns to their past values, any idiot could follow that rule and make unlimited profits.[6]

Figure 1.7: Nasdaq stock market index – absolute values of daily returns



And yet, we have something interesting here too: in fact, the plot in Figure 1.5 is typical of almost all high-frequency financial time series. The most interesting feature is the alternation of periods when absolute returns are larger with periods when they are smaller. Technically, these periods are referred to as high-volatility or low-volatility regimes. It should be no surprise that the high-volatility episodes coincide with events that had a major impact on financial markets: for example, the COVID pandemic and the start of the war in Ukraine. This feature is clearly visible by considering the time series of the absolute values of the returns (Figure 1.7).

Figure 1.8: Nasdaq stock market index – correlogram of absolute values



Evidently, we have quite a lot of persistence here. In this case, the aim is not so much to model the persistence of the series itself, but rather that of its

[6]Experts will notice that I am oversimplifying the so-called 'efficient market hypothesis' (EMH) with a recklessness you'd only expect in a random YouTube video. My apologies.

volatility.

The statistical concept into which the word 'volatility' translates is, clearly, the variance (provided that the second moments exist). As will be seen later, to analyse series of this type, we use stochastic processes of a specific nature, in which any persistence exhibited by the series translates into the time varying dependence of the variance, rather than the mean. In other words, the persistence of these processes is summarised by the fact that

$$V(x_t|x_{t-1}) \neq V(x_t). \tag{1.5}$$

Compare this with equation (1.3): in these processes, which are called **conditionally heteroskedastic processes**, what makes the difference between the marginal and conditional distributions with respect to the information set $\Im_{t-1}$ is precisely the structure of the second moments, rather than the first moments. Processes of this type have been widely used in advanced empirical finance for nearly four decades.

By now, our rationale should be clear enough. In the following chapter, we will introduce a class of stochastic processes which is the cornerstone of all time series econometrics: ARMA processes.

Chapter 2

ARMA processes

ARMA processes are, by far, the most widely used family of stochastic processes in time series econometrics, for theoretical as well as practical reasons, as we'll see. Before analysing the main characteristics of these processes, however, a few basic definitions are necessary.

2.1 The lag operator

Time series are nothing but sequences of numbers, with a natural ordering given by time. In the same way, stochastic processes are sequences of random variables. Therefore, it is natural to look for a general way to manipulate sequences by means of appropriate operators. The **lag operator** is generally denoted by the letter L by econometricians (statisticians use B instead — savages!); it's an operator that turns a sequence x_t into another sequence, that contains the same objects as x_t, but shifted back by one period.[1] If you apply L to a constant, the result is the same constant. In formulae,

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

Repeated application of the L operator n times is indicated by L^n, and therefore $L^n x_t = x_{t-n}$. By convention, $L^0 = 1$. The L operator is *linear*, which means that, if a and b are constant, then $L(ax_t + b) = aLx_t + b = ax_{t-1} + b$. These simple properties have the nice consequence that, in many cases, we can manipulate the L operator algebraically as if it was a number. This trick is especially useful when dealing with *polynomials* in L, that is operators of the form

$$a_0 + a_1 L + a_2 L^2 + \dots a_p L^p,$$

where the maximum exponent p is known as the "order" of the polynomial (also known as its "degree"). Polynomials in the lag operators are also known as **filters**, so we speak of "filtered" sequences.

Allow me to exemplify:

[1] In certain cases, you might want to use the "lead operator", usually notated as F, which is defined as the inverse to the lag operator ($Fx_t = x_{t+1}$, or $F = L^{-1}$). I'm not using it in this book, but its usage is very common in economic models with rational expectations.

Example 2.1.1 *Call b_t the money you have at time t, and s_t the difference between the money you earn and the money you spend between $t-1$ and t (in other words, your savings). Of course,*

$$b_t = b_{t-1} + s_t.$$

Now the same thing with the lag operator::

$$b_t = Lb_t + s_t \to b_t - Lb_t = (1-L)b_t = \Delta b_t = s_t$$

The Δ operator, which I suppose not unknown to the reader, is defined *as $(1-L)$, that is a polynomial in L of order 1. The above expression simply says that the variation in the money you have is your net saving.*

Example 2.1.2 *Call q_t the GDP for the Kingdom of Verduria in quarter t. Obviously, yearly GDP is given by*

$$y_t = q_t + q_{t-1} + q_{t-2} + q_{t-3} = (1 + L + L^2 + L^3)q_t$$

Since $(1 + x + x^2 + x^3)(1 - x) = (1 - x^4)$, if you "multiply" the equation above[2] *by $(1-L)$ you get*

$$\Delta y_t = (1 - L^4)q_t = q_t - q_{t-4};$$

The variation in yearly GDP between quarters is just the difference between the quarterly figures a year apart from each other.

A polynomial $P(x)$ may be evaluated at any value, but two cases are of special interest. Obviously, if you evaluate $P(x)$ for $x = 0$ you get the "constant" coefficient of the polynomial, since $P(0) = p_0 + p_1 \cdot 0 + p_2 \cdot 0 + \cdots = p_0$; instead, if you evaluate $P(1)$ you get the sum of the polynomial coefficients:

$$P(1) = \sum_{j=0}^{n} p_j 1^j = \sum_{j=0}^{n} p_j.$$

This turns out to be quite handy when you apply a lag polynomial to a constant, since

$$P(L)\mu = \sum_{j=0}^{n} p_j \mu = \mu \sum_{j=0}^{n} p_j = P(1)\mu.$$

There are two more routine results that come in very handy: the first one has to do with inverting polynomials of order 1. It can be proven that, if $|\alpha| < 1$,

$$(1 - \alpha L)^{-1} = (1 + \alpha L + \alpha^2 L^2 + \cdots) = \sum_{i=0}^{\infty} \alpha^i L^i; \tag{2.1}$$

the other one is that a polynomial $P(x)$ is invertible if and only if all its roots are greater than one in absolute value:

$$\frac{1}{P(x)} \quad \text{exists iff} \quad P(x) = 0 \Rightarrow |x| > 1. \tag{2.2}$$

The proofs are in subsection 2.8.

[2] To be precise, we should say: 'if you apply the $(1-L)$ operator to the expression above'.

Example 2.1.3 (The Keynesian multiplier) *Let me illustrate a possible use of polynomial manipulation by a very old-school macro example: the simplest possible version of the Keynesian multiplier idea. Suppose that*

$$Y_t = C_t + I_t; \tag{2.3}$$
$$C_t = \alpha Y_{t-1}; \tag{2.4}$$

where Y_t is GDP, C_t is aggregate consumption and I_t is investment; $0 < \alpha < 1$ is the marginal propensity to consume; note that, in this version of the Keynesian model, equation (2.4) defines consumption as a fraction of past *income, as opposed to* current *income.*

By combining the two equations,

$$Y_t = \alpha Y_{t-1} + I_t \to (1 - \alpha L) Y_t = I_t.$$

therefore, by applying the first degree polynomial $A(L) = (1 - \alpha L)$ to the Y_t sequence (national income), you get the time series for investments, simply because $I_t = Y_t - C_t = Y_t - \alpha Y_{t-1}$.

Now invert the $A(L) = (1 - \alpha L)$ operator and use equation (2.1)

$$Y_t = \frac{1}{1 - \alpha L} I_i = (1 + \alpha L + \alpha^2 L^2 + \cdots) I_t = \sum_{i=0}^{\infty} \alpha^i I_{t-i} :$$

aggregate demand at time t can be seen as a weighted sum of past and present investment. Suppose that investment goes from 0 to 1 at time 0. This brings about a unit increase in GDP via equation (2.3); but then, at time 1 consumption goes up by α, by force of equation (2.4), so at time 2 it increases by α^2 and so on. Since $0 < \alpha < 1$, $a^n \to 0$ and this effect eventually dies out.

If investments were constant through time, then $I_t = \bar{I}$; therefore, $A(L) Y_t = \bar{I}$ becomes

$$Y_t = \frac{1}{A(L)} \bar{I} = \frac{1}{A(1)} \bar{I} = \frac{\bar{I}}{1 - \alpha}$$

where the second equality comes from the fact that $\bar{I}$ is constant. The rightmost expression is nothing but the familiar "Keynesian multiplier" formula.

A word of caution: in many cases, it's OK to manipulate L algebraically as if it was a number, but sometimes it's not: the reader should always keep in mind that the expression Lx_t does not mean 'L times x_t', but 'L applied to x_t'. The following example should, hopefully, convince you.

Example 2.1.4 *Given two sequence x_t and y_t, define the sequence z_t as $z_t = x_t \cdot y_t$. Obviously, $z_{t-1} = x_{t-1} y_{t-1}$; however, one may be tempted to argue that*

$$z_{t-1} = x_{t-1} y_{t-1} = L x_t L y_t = L^2 x_t y_t = L^2 z_t = z_{t-2}$$

which is obviously absurd.

2.2 White noise processes

The **white noise** (WN) is the simplest stochastic process one can imagine[3]: in fact, it is a process that possesses moments of (at least) up to the second order; these are constant over time (hence the process is stationary), but they give the process no memory of its own past.

The same thing can be expressed more formally as follows: a white noise process, whose t-th element will be denoted by ϵ_t, exhibits the following characteristics:

$$E(\epsilon_t) = 0 \tag{2.5}$$

$$E(\epsilon_t^2) = V(\epsilon_t) = \sigma^2 \tag{2.6}$$

$$\gamma_k = 0 \quad \text{per } |k| > 0 \tag{2.7}$$

A white noise is therefore, in essence, a sequence containing infinitely many random variables, all with zero mean and constant variance; furthermore, these random variables are all uncorrelated with one another. Strictly speaking, this does not mean that they are independent. If, however, the process is a *Gaussian* white noise, where the joint distribution of all pairs $(\epsilon_t, \epsilon_{t+k})$ is a bivariate normal, then they are. Two aspects deserve to be highlighted:

- in the case of normality, a realization of size N of a white noise can also be quite legitimately considered a realization of N independent and identically distributed random variables. In this sense, a cross-sectional sample can be viewed as a special case;
- there is no substantial difference between the conditions that define a white noise and the so-called 'classical assumptions' on the disturbance in the OLS model, except for the lack of correlation between regressors and disturbances; summarizing the classical assumptions in the OLS model with the phrase 'the disturbance term is a white noise uncorrelated with the regressors' would not be far from the truth.

A white noise process, therefore, is a stochastic process with no persistence. As such, one might think it inadequate for achieving the goal we set ourselves in the introduction, namely finding a probabilistic structure that can be used as a metaphor for persistent time series. The decisive step forward, described in the next section, lies in considering what happens when we filter a white noise via a lag polynomial.

If I had wanted to be impeccable, I should have made a distinction between different types of 'memoryless' stochastic processes. Strictly speaking, in fact, the only type of process with no trace of persistence is whose elements are independent random variables. Often, however, it is useful to consider processes that are not so restrictive: for example, the so-called *martingale difference*, which is a concept very commonly employed both in statistics (especially in asymptotic theory) and in economics (theory of rational expectations).

In a martingale difference sequence (MDS), the distribution is left unspecified; what charac-

[3]The reason why this process bears the evocative name of 'white noise' would be very interesting, but the analytical tools we use in this book are too limited to develop this point. Oh well.

terizes this type of sequence is the property $E(x_t|\Im_{t-1}) = 0$. In this context, our primary interest is the conditional mean of the process, which must not depend in any way on the past. A white noise, on the other hand, is a different concept altogether: the lack of correlation between different elements ensures only that the conditional expectation is not a *linear* function of the past. Zipped proof:

$$\begin{aligned} E(x_t|\Im_{t-1}) = bx_{t-1} &\implies \\ E(x_t x_{t-1}|\Im_{t-1}) = bx_{t-1}^2 &\implies \\ E(x_t x_{t-1}) = E\left[E(x_t x_{t-1}|\Im_{t-1})\right] &= \\ = bE(x_{t-1}^2) \neq 0 & \end{aligned}$$

(we thank the law of iterated expectations for its kind cooperation).
However, it is entirely possible that the conditional mean from being a non-zero non-linear function. In fact, one can construct examples of white noise processes that are not martingale difference sequences. Furthermore, not all martingale difference sequences are white noise processes: indeed, the definition of white noise entails very specific conditions on the second moments, which may not even exist in a MDS.
That being said, it must be noted that in practice these concepts can be overlapped fairly painlessly: a Gaussian WN, for example, is a sequence of independent random variables with a mean of 0, so it is also a MDS.

2.3 MA processes

A **moving average** process, or **MA process** for short, is a sequence of random variables that can be written in the form

$$y_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i} = C(L)\epsilon_t$$

where $C(L)$ is a q-order lag polynomial and ϵ_t is a white noise process. In most cases, we set $C(0) = \theta_0 = 1$. If $C(L)$ is a polynomial of degree q, y_t is also said to be an MA(q) process, which is read as 'MA process of order q'. Let us examine its moments: for the first moment, we have

$$E(y_t) = E\left[\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right] = \sum_{i=0}^{q} \theta_i E(\epsilon_{t-i}) = 0.$$

Therefore, an MA process has mean 0. At first glance, one might think that this characteristic is a serious shortcoming for applying MA processes to real-world situations, since in general observed time series do not necessarily fluctuate around the value 0. However, this limitation is more apparent than real: for any process x_t for which $E(x_t) = \mu_t$, one can always define a new zero-mean process $y_t = x_t - \mu_t$ [4]. If y_t is covariance-stationary, then it is sufficient to study y_t and then add back the mean to obtain x_t.

As for the variance, the fact that the $E(y_t) = 0$ makes it possible to express it as the second moment, namely

$$V(y_t) = E(y_t^2) = E\left[\left(\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right)^2\right]$$

[4] Note in passing that in this simple example the process x_t is not stationary, according to the definition we have already provided, unless μ_t is constant, but the process y_t is.

By expanding the square[5], the sum can be split into two distinct parts:

$$\left(\sum_{i=0}^{q}\theta_i\epsilon_{t-i}\right)^2 = \sum_{i=0}^{q}\theta_i^2\epsilon_{t-i}^2 + \sum_{i=0}^{q}\sum_{j\neq i}\theta_i\theta_j\epsilon_{t-i}\epsilon_{t-j}$$

It should be obvious, from the "no-memory" property of white noise processes, that the expected value of the second summation is zero, so that

$$E(y_t^2) = E\left[\sum_{i=0}^{q}\theta_i^2\epsilon_{t-i}^2\right] = \sum_{i=0}^{q}\theta_i^2 E(\epsilon_{t-i}^2) = \sum_{i=0}^{q}\theta_i^2\sigma^2 = \sigma^2\sum_{i=0}^{q}\theta_i^2 \tag{2.8}$$

that has a finite value if $\sum_{i=0}^{q}\theta_i^2 < \infty$, which is always true if q is finite.

Finally, the autocovariances can be worked out by a completely analogous argument: the k-order autocovariance is

$$E(y_t y_{t+k}) = E\left[\left(\sum_{i=0}^{q}\theta_i\epsilon_{t-i}\right)\left(\sum_{j=0}^{q}\theta_j\epsilon_{t-j+k}\right)\right] = \sum_{i=0}^{q}\theta_i\left(\sum_{j=0}^{q}\theta_j E(\epsilon_{t-i}\epsilon_{t-j+k})\right). \tag{2.9}$$

Exploiting again the properties of white noise, we get $E(\epsilon_{t-i}\epsilon_{t-j+k}) = \sigma^2$ for $j = i + k$ and 0 in all other cases, so that the previous expression reduces to:

$$\gamma_k = E(y_t y_{t+k}) = \sigma^2\sum_{i=0}^{q}\theta_i\theta_{i+k},$$

where $\theta_i = 0$ for $i > q$.

Note that:

- the equation for the variance is a special case of the previous formula, setting $k = 0$;
- for $k > q$, all autocovariances are zero.

An MA(q) process, therefore, is a process obtained by combinining different elements of the same white noise sequence. The higher its order, the stronger is its persistence. Note that the order q may be infinite; in this case, however, the existence of the second moments (and therefore stationarity) is guaranteed only if $\sum_{i=0}^{q}\theta_i^2 < \infty$.

Example 2.3.1 *Consider the MA(1) process $x_t = \epsilon_t + \theta\epsilon_{t-1}$ and calculate its autocovariances: its variance is given by*

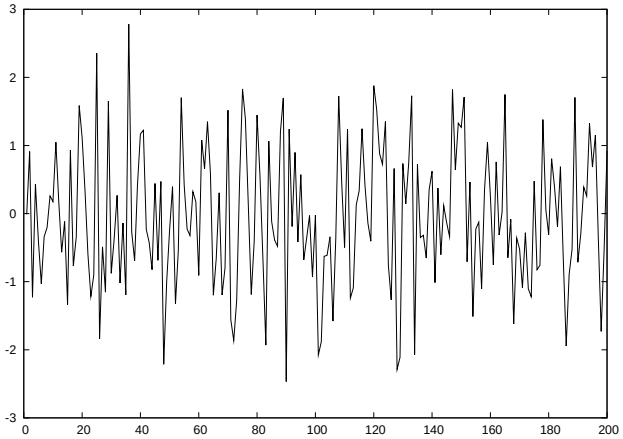
$$E(x_t^2) = E\left(\epsilon_t + \theta\epsilon_{t-1}\right)^2 = E(\epsilon_t^2) + \theta^2 E(\epsilon_{t-1}^2) + 2\theta E(\epsilon_t\epsilon_{t-1}) = (1+\theta^2)\sigma^2.$$

According to its definition, the autocovariance of order 1 is

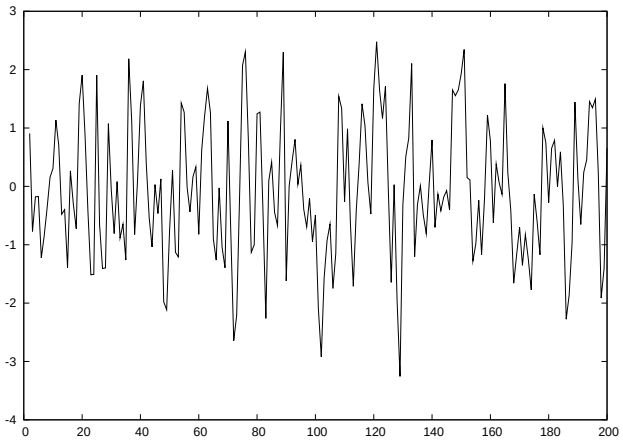
$$E(x_t x_{t-1}) = E\left[\left(\epsilon_t + \theta\epsilon_{t-1}\right)\left(\epsilon_{t-1} + \theta\epsilon_{t-2}\right)\right].$$

[5]Note: by the argument I made a few pages back, $[C(L)\epsilon_t]^2$ is *different* from $C(L)^2\epsilon_t^2$. Think of the simple case $C(L) = L$ and you will be immediately convinced.

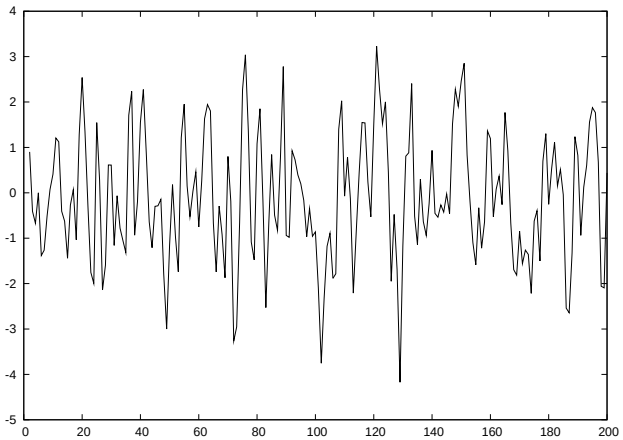
Figure 2.1: MA(1) with different parameter θ
(a) $\theta = 0$ (white noise)



(b) $\theta = 0.5$



(c) $\theta = 0.9$



By expanding the product, we obtain:

$$E(\epsilon_t\epsilon_{t-1}) + \theta E(\epsilon_t\epsilon_{t-2}) + \theta E(\epsilon_{t-1}^2) + \theta^2 E(\epsilon_{t-1}\epsilon_{t-2}) = \theta\sigma^2.$$

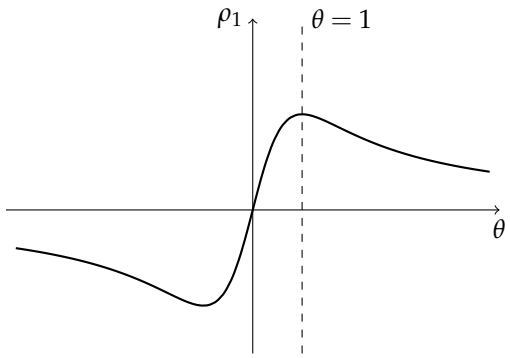
Consequently, the autocorrelation of order 1 is

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta}{1+\theta^2}.$$

Similarly, it can be shown that the higher-order autocovariances are all zero.

To illustrate this more concretely, let us take an example of an MA(1) process and plot it: if the process is $y_t = \epsilon_t + \theta\epsilon_{t-1}$, the behaviour of y_t for different values of θ is shown in Figure 2.1. Of course, when $\theta = 0$ (as in Figure 2.1(a)) the process is a white noise. As can be seen, as θ increases, the persistence characteristics become more visible (the series 'smoothes out') and its variance (roughly measured by the order of magnitude of the ordinates) increases. If we had simulated a higher-order MA process, this would have been even more evident.

Figure 2.2: MA(1): Autocorrelation of order 1 as a function of θ



Some interesting considerations can be made by digging a little deeper. As the example shows, the autocorrelation of order 1 of an MA(1) process is given by the formula $\rho_1 = \frac{\theta}{1+\theta^2}$, plotted in Figure 2.2. It can be noted that the maximum value reached by ρ_1 is 0.5, corresponding to $\theta = 1$; an analogous argument, with inverted signs, holds for the minimum. Furthermore, we know from previous arguments (see equation (2.9)), that all autocorrelations of order greater than 1 are zero. This means that the correlogram of an MA(1) process has only one interesting spike (the first one), and even that is constrained to lie between $-1/2$ and $1/2$.

Let's now turn to an inferential problem: suppose we were certain that a certain observed time series is the realisation of an MA(1) process. Which statistic could we use to obtain estimates of the process parameters (in this case, the parameter θ)? Obviously, this procedure would be justifiable only if the empirical correlogram had moderate values for the first-order autocorrelation and negligible ones for the others. If this were the case, we could adopt the following strategy:

- if the process that generated the data is in fact an MA(1), then,
- it is stationary and ergodic, then,
- the sample autocorrelation converges in probability to the theoretical one, then,
- in an infinitely-sized sample, $\hat{\rho}_1$ should be exactly equal to ρ_1.

More formally,

$$\hat{\rho}_1 \xrightarrow{\text{p}} \rho_1 = \frac{\theta}{1+\theta^2};$$

since this is a continuous function of θ, it can be inverted to give a consistent estimator of θ via the method of moments. This amounts to finding the value $\hat{\theta}$ which satisfies the equation

$$\hat{\rho}_1 = \frac{\hat{\theta}}{1+\hat{\theta}^2}; \tag{2.10}$$

It can be easily seen that the solution to (2.10) is[6]

$$\hat{\theta} = \frac{1}{2\hat{\rho}_1}\left(1 - \sqrt{1-4\hat{\rho}_1^2}\right).$$

Note that, for the existence of the estimator, it is necessary that $|\hat{\rho}_1| \leq 0.5$, but in this case there is no problem, since we are assuming that we are dealing with a series in which the first-order autocorrelation is not too strong.

For example: if the sample autocorrelation is 0.4, and I am convinced that the process that generated the data is an MA(1), then I will choose that value of θ such that the theoretical autocorrelation is also 0.4, namely $\hat{\theta} = 0.5$. Clearly, this strategy is perfectly justified because the time series actually has the required covariance characteristics, *i.e.* a not-too-large first-order autocorrelation and negligible subsequent autocorrelations.

Now, we know that things are not always this easy: just take a look at Figures 1.2 on page 7 and 1.4 on page 9. It is true, however, that a higher-order MA process has more complex autocovariances, and therefore one may conjecture that the same strategy could be feasible, at least in theory, provided that a sufficiently high order of the polynomial $C(L)$ is specified.

In fact, one might wonder whether the conjecture holds for any autocovariance structure. The answer lies in the never-sufficiently-celebrated **Wold representation theorem**, of which I provide only the statement.

Theorem 1 (Wold representation theorem) *Given any stochastic process y_t, which is covariance-stationary and has a mean of 0, it is always possible to find a sequence (not necessarily finite) of coefficients θ_i such that*

$$y_t = \sum_{i=0}^{\infty} \theta_i \epsilon_{t-i},$$

[6] In fact, there are two values, because the actual solution would be $\hat{\theta} = \frac{1 \pm \sqrt{1-4\hat{\rho}_1^2}}{2\hat{\rho}_1}$ (note the $\pm$ symbol), but to kkep things simple let's agree on choosing the solution inside the interval $[-1,1]$, *i.e.* the one reported in the text.

where ϵ_t is a white noise. Moreover, the sum of the squared coefficients is finite: $\sum_{i=0}^{\infty} \theta_i^2 < \infty$.

In other words: for any stochastic process, provided it is stationary, there exists a moving average process that possesses the same autocovariance structure. This result is of enormous importance: it tells us, essentially, that whatever the 'true' form of a stationary stochastic process, we can always represent it as an MA process (possibly, of infinite order). It is for this reason that, by studying MA processes, we are in fact studying *all* possible stationary processes, at least as far as their mean and covariance characteristics are concerned.

The picture I have just given of Wold's theorem is not really accurate: if you look at 'proper' textbooks, you will notice that in fact the theorem does not say exactly what I wrote above. To be precise, one should say that every second-order stationary process can be decomposed into a 'deterministic' part (i.e. perfectly predictable given the past) plus a moving average part. I have cunningly simplified the definition by adding the words 'with a mean of 0', thus excluding the existence of the deterministic part but the message remains essentially the same.

2.4 AR processes

The class **AR (AutoRegressive)** processes is another important one. These processes provide representation of a persistent time series that some may find more intuitive than the one implied by MA processes, because in AR processes the main idea is that the value of the series at time t is a linear function of its own past values, plus white noise. The name derives precisely from the form of the AR model that is very similar to a regression model in which the explanatory variables are the past values of the dependent variable.

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t \tag{2.11}$$

Note that, in this context, the white noise process ϵ_t is like the disturbance term in a regression model, that is, as the difference between y_t and its conditional mean; in this case, the random variables that constitute the information set are simply the past of y_t.

AR processes are, to some extent, a natural counterpart to to MA processes because an MA process is what you get by filtering a white noise process, whereas an AR process is a process that yields white noise if suitably filtered. In symbols

$$A(L)y_t = \epsilon_t,$$

where $A(L)$ is, as usual, a polynomial in L (with degree p) with $A(0) = 1$ and $a_i = -\varphi_i$.

To describe these processes, let us begin by considering the simplest case, where $p = 1$:

$$y_t = \varphi y_{t-1} + \epsilon_t \longrightarrow (1 - \varphi L)y_t = \epsilon_t.$$

What are the characteristics of this process? Let's start from its moments. These can be obtained in several ways: one rather intuitive way is to assume the stationarity of the process (so we know they exist), and then focus on the consequences of this hypothesis. Let us assume, therefore, that the process has a constant mean μ. This hypothesis implies

$$\mu = E(y_t) = \varphi E(y_{t-1}) + E(\epsilon_t) = \varphi\mu.$$

There are two ways this expression can be true: either $\mu = 0$, in which case it is true for any value of φ, or in the case $\varphi = 1$, and then the expression is true for any value of μ, and the mean of the process is undefined. In this second case, the process is characterised by a **unit root**, because the value of z for which $A(z) = 0$ is precisely 1; the analysis of this situation, in which bizarre things happen, kept the best econometricians busy during the final two decades of the 20th century, and for a long time it was considered by applied economists as rough terrain, upon which it is inadvisable to venture without a native guide. We shall discuss this case in Chapters 3 and **??**. For the moment, let's rule out the $A(1) = 0$ case. It follows that — in the cases we analyse here — the process has a zero mean.

Another method to obtain $E(y_t)$ is to re-express y_t as a MA process (under stationarity, Wold's theorem guarantees this is legitimate). To do this, we use the results reported above on the manipulation of polynomials. If we focus on the cases where $|\varphi| < 1$ (so unit roots are excluded), we have

$$A(L)^{-1} = (1 - \varphi L)^{-1} = 1 + \varphi L + \varphi^2 L^2 + \cdots = C(L),$$

therefore the MA rapresentation of y_t is

$$y_t = \left(1 + \varphi L + \varphi^2 L^2 + \cdots\right)\epsilon_t = C(L)\epsilon_t,$$

that is, an MA process with $\theta_i = \varphi^i$, which has zero mean[7]; therefore, $E(y_t) = 0$.

As for the second moments, we proceed as above; let us call $V(\epsilon_t) = \sigma^2$. If we call ω the variance of y_t, and assume that it exists and is constant over time, it follows that

$$\omega = E(y_t^2) = E\left[(\varphi y_{t-1} + \epsilon_t)^2\right] = \varphi^2\omega + \sigma^2 + 2\varphi E(y_{t-1}\epsilon_t).$$

The last element of the sum is 0, since $y_{t-1} = C(L)\epsilon_{t-1}$, and therefore $E(y_{t-1}\epsilon_t)$ is a linear combination of autocovariances of a white noise (all of which are zero by definition). It follows that

$$\omega = \varphi^2\omega + \sigma^2 \Longrightarrow \omega = \frac{\sigma^2}{1 - \varphi^2}.$$

The same result could also have been obtained from the MA representation, noting that

$$\omega = \sigma^2 \sum_{i=0}^{\infty} \theta_i^2 = \sigma^2 \sum_{i=0}^{\infty} \varphi^{2i} = \sigma^2 \sum_{i=0}^{\infty} (\varphi^2)^i = \frac{\sigma^2}{1 - \varphi^2}.$$

[7] The moving average representation of an AR(1) process can also be obtained by the so-called method of 'recursive substitutions', which is cruder and less elegant, but perhaps more intuitive. Let us consider that, if $y_t = \varphi y_{t-1} + \epsilon_t$, then we will also have $y_{t-1} = \varphi y_{t-2} + \epsilon_{t-1}$; we substitute the second expression into the first and proceed iteratively.

The expression $\omega = \frac{\sigma^2}{1-\varphi^2}$ provides several insights. First, it tells us that only if $|\varphi| < 1$ the variance is stable over time (for $|\varphi| \geq 1$ the last equality no longer holds); this condition rules out from the category of stationary AR(1) processes not only those with a unit root, but also those with a so-called explosive root ($|\varphi| > 1$).

Second, by comparing ω, which is the *unconditional* variance of y_t, with σ^2, which is the variance of $y_t|\Im_{t-1}$, it can be seen that ω is always greater than σ^2. Furthermore, the difference is greater the closer φ is to 1: the more persistent the process is, the smaller its variance *conditional* on its own past will be compared to its unconditional variance. That is to say, knowing the value of y_{t-1} reduces the uncertainty about the value of y_t the more persistent the series is.

All it remains to do is to compute the autocovariances: the autocovariance of order 0 is ω, which we already know; the autocovariance of order 1 is given by

$$\gamma_1 = E(y_t y_{t-1}) = E\left[(\varphi y_{t-1} + \epsilon_t) y_{t-1}\right] = \varphi\omega$$

and, more generally,

$$\gamma_k = E(y_t y_{t-k}) = E\left[(\varphi y_{t-1} + \epsilon_t) y_{t-k}\right] = \varphi\gamma_{k-1},$$

thus obtaining

$$\gamma_k = \varphi^k \frac{\sigma^2}{1-\varphi^2}.$$

Alternatively, starting from the MA representation, we get

$$E(y_t y_{t+k}) = \sigma^2 \sum_{i=0}^{q} \theta_i \theta_{i+k} = \sigma^2 \sum_{i=0}^{q} \varphi^i \varphi^{i+k} = \sigma^2 \sum_{i=0}^{q} \varphi^{2i+k}$$

which is equal to

$$\gamma_k = \varphi^k \sigma^2 \sum_{i=0}^{q} \varphi^{2i} = \varphi^k \frac{\sigma^2}{1-\varphi^2}.$$

As a consequence, the autocorrelations take a very simple form:

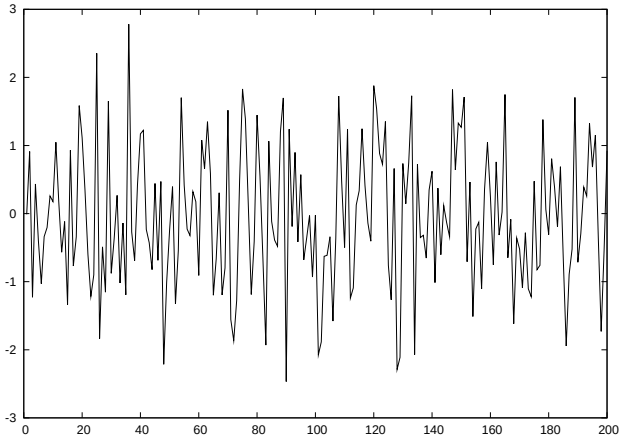
$$\rho_k = \varphi^k.$$

In this case too, it is possible to give an intuitive interpretation of the result: the autocorrelations, which we take as an index of the memory of the process, are larger (in absolute value) the larger (in absolute value) φ is, thus confirming the interpretation of φ as a persistence parameter. In addition the autocorrelations are nonzero for all k, although $\lim_{k\to\infty} \gamma_k = 0$. In practice, the memory of the process is infinite, but the very remote past plays a negliglible role.

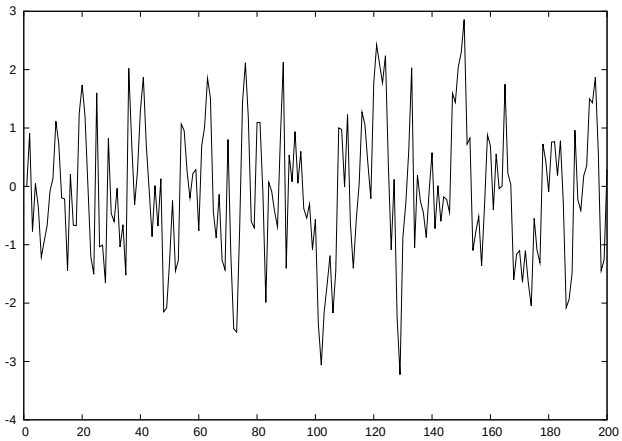
Again, let's look at an example. Figure 2.3(a) shows exactly the white noise already presented in Figure 2.1(a) as an example of MA(1) processes. Let us apply the operator $(1-\varphi L)^{-1}$ to this white noise, with $\varphi = 0.5$ and $\varphi = 0.9$. In this case too, an increase in the persistence characteristics can be observed as

Figure 2.3: AR(1) with different parameter φ

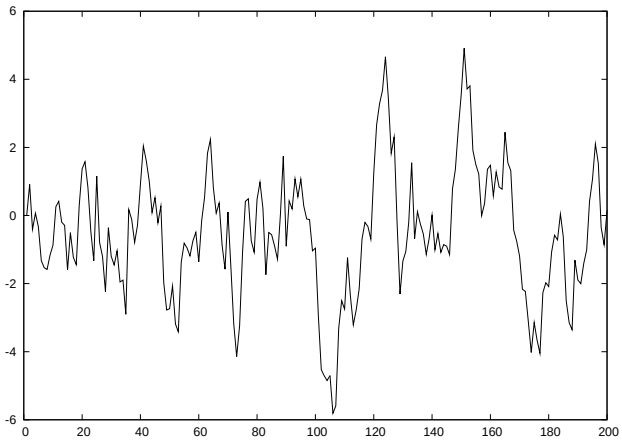
(a) $\varphi = 0$ (white noise)



(b) $\varphi = 0.5$



(c) $\varphi = 0.9$



the parameter (φ in this case) increases, even though here the effect is much easier to spot.

Like with MA processes, it is easy to generalise AR processes to the case of non-zero mean: let us suppose we add an 'intercept' to the AR(1) model:

$$y_t = \mu + \varphi y_{t-1} + \epsilon_t \rightarrow (1 - \varphi L) y_t = \mu + \epsilon_t.$$

Inverting the polynomial $A(L)$, we get

$$y_t = \left(1 + \varphi L + \varphi^2 L^2 + \cdots\right)(\mu + \epsilon_t) = C(L)(\mu + \epsilon_t)$$

since applying L to a constant leaves it unchanged, it follows that

$$y_t = (1 + \varphi + \varphi^2 + \cdots)\mu + C(L)\epsilon_t = \frac{\mu}{1 - \varphi} + C(L)\epsilon_t,$$

therefore $E(y_t) = \frac{\mu}{1-\varphi}$.

Generalising to the AR(p) case is very boring, since the necessary algebraic manipulations, stemming from the necessity to invert a p-th degree polynomial to obtain the moving average representation, are a pain. We have

$$C(L) = A(L)^{-1} = \prod_{j=1}^{p} (1 - \lambda_j L)^{-1}$$

where the λ_j are the reciprocals of the roots of $A(L)$. On the other hand, this generalisation does not bring great advantages to the intuitive understanding of the salient characteristics of these processes. The fundamental point is that if $|\lambda_j| < 1$ for all j, then the expression above makes sense because the $A(L)$ polynomial does in fact have an inverse that we can write as $C(L)$ (in this case, $A(L)$ is said to be "stable"). In this case, the AR(p) process is stationary. A rigorous proof is omitted here, but the curious reader should note that all is needed is equation (2.27) in the Appendix of this chapter.

Note: λ_j can be a complex number: in this case, allow me to remind the reader that that its absolute value is given by the formula $|a + bi| = \sqrt{a^2 + b^2}$. For those unfamiliar with the topic, a brief overview is provided in the Appendix to this chapter, on page 54.

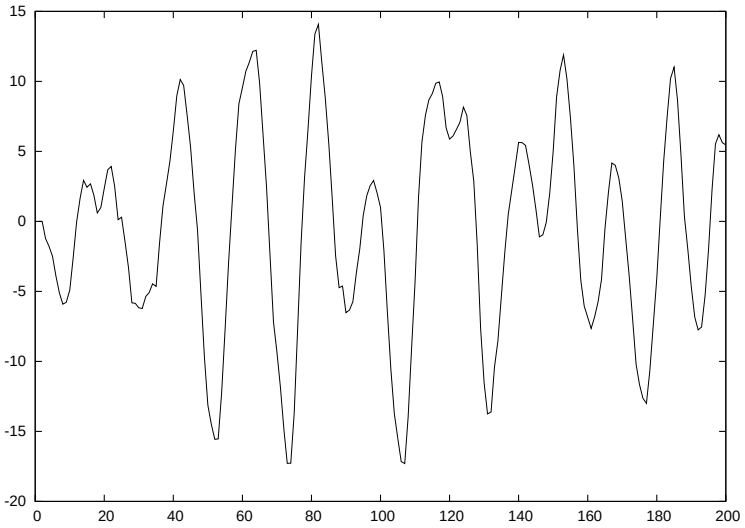
Other interesting properties (I won't prove these either) are that an AR(p) process

- has infinite memory, but the autocorrelations decrease as k increases in a geometric progression: this is often called "exponential decay";
- in the case of an 'intercept' different from 0, it has an expected value of $\frac{\mu}{A(1)}$, where $A(1)$ is the polynomial $A(z)$ evaluated at $z = 1$ instead of $z = L$ as usual; in practice, $A(1) = \sum_{i=0}^{p} a_i$.

The only aspect worth emphasising when the autoregressive order p is greater than 1 is that AR(p) processes can have cyclical trends: this occurs if and only the roots of the polynomial $A(z)$ contain one or more pairs of complex conjugate numbers. In this case, the process assumes a cyclical trend in which the amplitude of the oscillations varies around a mean value. Clearly, processes of this type are the natural candidates to model economic phenomena characterised by cyclical phases.

The reason why complex numbers and cyclical trends are linked would be wonderful to explain, but unfortunately the average student in an Economics faculty considers complex numbers and trigonometric functions to be arcane witchcraft, so I'll pass. For the eccentric few who like these things, there's an Appendix at the end of this chapter so that the majority are left unfazed.

Figure 2.4: AR(2): $\varphi_1 = 1.8$; $\varphi_2 = -0.9$



Let us take a look at an example: consider the white noise in Figure 2.3(a) and use it to construct an AR(2) process in which the polynomial $A(z)$ has no real roots. In this specific case,

$$y_t = 1.8y_{t-1} - 0.9y_{t-2} + \epsilon_t,$$

where the roots are

$$\lambda = \frac{1.8 \pm \sqrt{3.24 - 3.6}}{1.8} = 1 \pm \frac{i}{3},$$

both greater than 1 in absolute value (they are "one plus something"). As can be seen in Figure 2.4, there is a more or less regular alternation of 'peaks' and 'troughs'

2.5 ARMA processes

The class of ARMA processes includes both AR processes and MA processes as special cases. An ARMA(p,q) process is defined as

$$A(L)y_t = C(L)\epsilon_t, \tag{2.12}$$

where p is the order of the polynomial $A(L)$ and q is the order of the polynomial $C(L)$. Both are finite numbers. AR or MA processes are therefore special cases ($q = 0$ and $p = 0$ respectively).

If all the roots of $A(L)$ are greater than 1 in modulus, then y_t can also be represented in MA form

$$y_t = A(L)^{-1}C(L)\epsilon_t = C^*(L)\epsilon_t,$$

where $C^*(L)$ is a polynomial of infinite order if $p > 0$. This condition on $A(L)$ is necessary and sufficient for the process to be stationary. Similarly, if the polynomial $C(L)$ is invertible, then y_t admits an autoregressive representation (of infinite order if $q > 0$)

$$C(L)^{-1}A(L)y_t = A^*(L)y_t = \epsilon_t.$$

In this case, the *process* is said to be invertible.

The characteristics of the moments of an ARMA(p,q) process can be derived in a conceptually simple (but algebraically cumbersome) manner from its moving average representation, and I shall not report them here. The only one worth mentioning is that, if we add an intercept, it can be easily shown[8] that the mean of the process is still $\frac{\mu}{A(1)}$. Optionally, you can add a touch of further sophistication by allowing for the possibility of a non-zero, time-varying mean, namely something like

$$A(L)y_t = \mu(x_t, \beta) + C(L)\epsilon_t,$$

which is similar to a regression model. An alternative way of writing this is to consider a model of the form

$$y_t = \mu_t + u_t,$$

where $\mu_t = \frac{1}{A(L)}\mu(x_t, \beta)$ and $u_t = \frac{C(L)}{A(L)}\epsilon_t$. This is a sort of a regression model where the errors are given by an ARMA(p,q) process. As can be seen, it is easy to switch from one representation to the other.

Why do we need ARMA processes? In theory, we don't: Wold's representation theorem ensures that any stationary process can be represented as an MA process. In practice, however, Wold's representation is, in general, of infinite order. This is a serious problem for inference: the observed time series is conceived as the realisation of a stochastic process, whose parameters are the coefficients of the polynomials in the L operator that determine its persistence characteristics (plus the variance of the white noise). If an observed time series is considered as a realisation of some stationary process, using an MA process to summarise its mean and covariance characteristics therefore involves the inferential problem of estimating a potentially infinite number of parameters. If we assume that y_t can be represented in MA form as

$$y_t = B(L)\epsilon_t,$$

[8] Blitz proof: $A(L)y_t = \mu + C(L)\epsilon_t \Longrightarrow E[A(L)y_t] = \mu + E[C(L)\epsilon_t]$. By the linearity of the operators E e L, we get $A(L)E[y_t] = \mu + C(L)E[\epsilon_t] = \mu$. However, if y_t is stationary, then $E[y_t]$ exists, is finite and constant, so that $A(L)E[y_t] = A(1)E[y_t]$, and $E[y_t] = \frac{\mu}{A(1)}$.

nothing prevents the polynomial $B(L)$ from being of infinite order.

However, one could consider using an approximation of $B(L)$: in applied mathematics, it is well known that sometimes functions can be approximated very well by "rational functions", that is, the ratio of two polynomials. So, it may be possible to find two polynomials of finite (and possibly low) order, $A(L)$ and $C(L)$, such that

$$B(z) \simeq \frac{C(z)}{A(z)}.$$

If the equality were exact, one could then write

$$A(L)y_t = C(L)\epsilon_t.$$

If the equality holds only approximately, then one will have

$$A(L)y_t = C(L)\epsilon_t^*,$$

where

$$\epsilon_t^* = \frac{A(L)}{C(L)}B(L)\epsilon_t.$$

Strictly speaking, the process ϵ_t^* is not white noise; however, if its autocovariances are "small enoughh", it can be treated as such for all practical purposes. Alternatively, one could argue that treating ϵ_t^* as a white noise process is a only slightly more misleading metaphor than the one based on ϵ_t (namely, Wold's representation), whilst offering the distinct advantage of relying on a finite number of parameters.

In practice, an ARMA model is built by making an *a priori* assumption about the degrees of the two polynomials, $A(L)$ and $C(L)$, and then — once the polynomial coefficients have been estimated — by examining the sample autocorrelations of the time series corresponding to ϵ_t^*. If these are not too large, then ϵ_t^* can be treated as white noise without adverse consequences[9].

The need to keep the number of parameters under control leads, in certain cases, to the usage of models known as **multiplicative ARMA** models, used primarily for time series characterised by seasonal persistence, so that they are also known as **seasonal ARMA**, or **SARMA**.

For example, let us consider the monthly time series of hotel arrivals for some skiing holiday destination (say, Gstaad or Cortina). We would expect strong seasonality, in the sense that the data for December likely resembles those of December from the previous year much more than that of May from the same year, which is closer in time but qualitatively different. For simplicity, let us imagine that we wish to use a pure autoregressive model, that is, without an MA component. A naive application of the ideas presented so far would evidently lead to the use of a polynomial of (at least) order 12, and thus to a structure with a fairly large number of parameters. Many of these, however, are likely redundant: while the conditional mean for the month of December is likely to depend on previous year's December, there is no obvious reason why the data for August or October should be equally relevant. This consideration would simply lead us to use a polynomial $A(L)$

[9]For estimation techniques, see Section 2.7.

with "gaps" — that is, with coefficients equal to 0. A more elegant and efficient approach is to write the polynomial by separating the seasonal effects from the ordinary ones. Let us consider the polynomial given by

$$A(L) = (1 - \varphi L)(1 - \Phi L^s) = 1 - \varphi L - \Phi L^s + (\varphi \cdot \Phi) L^{s+1},$$

where s is the number of sub-periods (*i.e.* 12 for the months in a year, and so on). Obviously, in this case, $A(L)$ is a polynomial of order $s + 1$, whose only non-zero coefficients are those of order 1, s, and $s + 1$. The number of parameters that characterise it, however, is only 2, because the coefficient of order $s + 1$ is the product of the other two (with reverse sign), so that it is possible to model an even quite long seasonal pattern while keeping the number of parameters necessary to do so under control.

Evidently, a similar trick can also be applied to the polynomial $C(L)$, so you get a remarkable degree of flexibility without having the number parameters explode. By generalising the above expression in an obvious way, we obtain a model that can be written as

$$A(L)B(L^s)y_t = C(L)D(L^s)\epsilon_t$$

which contains the seasonal autoregressive part $B(L^s)$ and the seasonal moving average part $D(L^s)$. If the order of the polynomials $B(\cdot)$ and $D(\cdot)$ is zero, we go back to the simple ARMA case.

2.6 Using ARMA models

If the parameters of an ARMA process are known, the model can be used for two purposes: forecasting its future values and/or analysing its dynamic characteristics.

2.6.1 Forecasting

We'll build on a simple and intuitive argument: if there is persistence, and the past contains information on the present, then the present contains information on the future, and we may be able to say something sensible on the distribution of future values $y_{T+1}, y_{T+2}, \ldots$ on the basis of the available data. Clearly, we'd like to follow a rule that can be considered optimal in some sense.

The best way to forecast the future values of y_t can be defined as follows: let us define a **forecast function** for y_t as some function of the variables contained in the information set $\Im_{T-1}$. That is, a forecast function is some rule yielding the forecast we make for y_t given its previous values, which we assume known. We call this value $\hat{y}_t = f(y_{t-1}, y_{t-2}, \ldots)$. Our problem is choosing a forecast function that works "well".

To make this idea less fuzzy, let's define the *forecast error*, as

$$e_t = y_t - \hat{y}_t.$$

Clearly, we'd like e_t to be zero. But e_t is a random variable, so we must consider its distribution. If y_t is an ARMA process (or representable as such), once we have the model in the form $A(L)y_t = C(L)\epsilon_t$, an assumption about the distribution of ϵ_t makes it possible, at least in principle, to determine the distribution of y_t conditional on $\Im_{T-1}$. In turn, this determines the conditional distribution $e_t|\Im_{T-1}$, which is crucial for our choice of a forecast function.

Strictly speaking, an optimal choice should be made according to the following criterion:

1. first, choose a function $c(e_t)$ (the so-called *loss* function), which associates a cost with the forecast error. In general, we have that $c(0) = 0$ (the cost of a perfect forecast is 0) and $c(e_t) \geq 0$ for $e_t \neq 0$;

2. then, define the expected loss as

$$c^* = E\left[c(e_t)|\Im_{T-1}\right] = E\left[c(y_t - \hat{y}_t)|\Im_{T-1}\right].$$

   The quantity c^* is the cost that, on average, we incur because of incorrect forecasts. Clearly, we want it to be as small as possible;

3. since c^* is a function of $\hat{y}_t$, we choose $\hat{y}_t$ in such a way as to minimise c^*; that is, we *define* $\hat{y}_t$ as that function which minimises the expected cost of the forecast error.

It should be clear that the best forecast function depends on what the loss function looks like: for example, if this function is asymmetric, the optimal forecast may be unintuitive. The example I often use is booking a restaurant: since the loss function in this case is asymmetric (it is better to have empty chairs than people left standing), it is always advisable to book for a slightly higher number of people than actually expected.

Fortunately, the matter becomes far less intricate if the loss function is quadratic, that is, if $C(e_t) = \kappa e_t^2$ for any positive κ. In this case (which can often be taken as a satisfactory approximation to the most appropriate cost function), it can be shown that $\hat{y}_t$ coincides with the conditional expectation:

$$C(e_t) = \kappa e_t^2 \Longrightarrow \hat{y}_{T+1} = E(y_{T+1}|\Im_T).$$

This property is so convenient that, in the vast majority of cases, the conditional mean is adopted as "the" forecast function with no justification necessary.

Therefore, for a set of observations ranging from 1 to T, from now on, to forecast y_{T+1} we'll use its expected value, conditional on the information set we have available:

$$\hat{y}_{T+1} = E(y_{T+1}|\Im_T). \tag{2.13}$$

In the case of a pure AR model, the solution is trivial, since all values of y_t up to time T are known, and therefore $E(y_{t-k}|\Im_T) = y_{t-k}$ for any $k \geq 0$:

$$E(y_{T+1}|\Im_T) = \varphi_1 y_T + \cdots + \varphi_p y_{T-p+1} + E(\epsilon_{T+1}|\Im_T),$$

but the value of $E(\epsilon_{T+1}|\Im_T)$ is evidently 0, since the memoryless property of white noise ensures that there is no information available at present regarding the future of ϵ; consequently, $E(\epsilon_{T+1}|\Im_T) = E(\epsilon_{T+1}) = 0$. Therefore, the forecast of y_{T+1} is

$$\hat{y}_{T+1} = \varphi_1 y_T + \cdots + \varphi_p y_{T-p+1}. \tag{2.14}$$

For the two-step-ahead forecast, start from the expression

$$\hat{y}_{T+2} = E(y_{T+2}|\Im_T) = \varphi_1 E(y_{T+1}|\Im_T) + \cdots + \varphi_p y_{T-p+2} + E(\epsilon_{T+2}|\Im_T),$$

which, using the same logic, can be easily shown to be equal to

$$\hat{y}_{T+2} = \varphi_1 \hat{y}_{T+1} + \cdots + \varphi_p y_{T-p+2}.$$

More generally,

$$\hat{y}_{T+k} = \varphi_1 \hat{y}_{T+k-1} + \cdots + \varphi_p \hat{y}_{T+k-p},$$

where, $\hat{y}_{T+k} = y_{T+k}$ for $k \leq 0$. Note the intriguing parallelism between $A(L)y_t = \epsilon_t$ and $A(L)\hat{y}_t = 0$, which is easily arrived at by considering the expected value (conditional on $\Im_{t-1}$) of the first of the two expressions.

Example 2.6.1 *Given the following AR(2) process*

$$y_t = 0.9y_{t-1} - 0.5y_{t-2} + \epsilon_t,$$

suppose we observe a realisation of this process, and that the last two observations are equal to: $y_{T-1} = 2$ *and* $y_T = 1$. *The best forecast for* y_{T+1} *is*

$$\hat{y}_{T+1} = 0.9 \times 1 - 0.5 \times 2 = -0.1,$$

while the forecast for y_{T+2} *is*

$$\hat{y}_{T+2} = 0.9 \times (-0.1) - 0.5 \times 1 = -0.59,$$

and going on is easy. For the record, the following five values are -0.481, -0.1379, 0.11639, 0.173701, 0.098136.

Of course, evaluating the conditional mean provides a point forecast, but says nothing about the reliability of the forecast — that is, about the variance of the forecast error. This is not a topic on which I wish to dwell for too long. For the interested reader, I will simply suggest, in addition to the usual bibliographic references at the end, that a useful exercise would be to prove that, in the case of an AR(1) process,

$$V(\hat{y}_{T+k}) = \sigma^2 \frac{1 - \varphi^{2k}}{1 - \varphi^2}.$$

It may be interesting to note that the forecast error variance is always smaller than the unconditional variance of y_t: leveraging on the persistence characteristics of the time series makes its future behaviour less uncertain. Moreover, as $k \to \infty$, the two variances tend to coincide; this occurs because, in stationary AR(1) processes, persistence is always of a short-term nature. Knowledge of the current state of the system is no more informative about its remote future than its unconditional distribution is: for a sufficiently large k, y_t and y_{t+k} are virtually uncorrelated (and therefore, if Gaussian, virtually independent).

Note that here we are treating the true parameters of the process as known, whereas in fact all we usually have are estimates; therefore, the variance of the forecast error in pricniple depends not only on the intrinsic variability of the process σ^2, but also on the uncertainty on the parameters of the process itself.
For example, take an AR(1) process for which we have a parameter estimate $\hat{\varphi}$; the argument above would have us analyse

$$\hat{y}_{T+k} = E(y_{T+k}|\Im_T) = E(\hat{\varphi} \cdot y_{T+k-1}|\Im_T),$$

where $\hat{\varphi}$ cannot be "taken out" of the expectation operator, because it is an estimator and not a constant.
That said, this distinction is important in theory, but in practice nobody bothers, and people normally uses the estimates as if they were the true parameters. This is justified because, with consistent estimators, for very large T the distinction becomes immaterial.

In the more general case of processes with an MA part, forecasting can still be carried out by applying the same concept. In particular, it should be noted that if $\Im_{t-1}$ has no lower bound in time, it includes not only all past values of y_t but also those of ϵ_t. This is quite obvious by considering that ϵ_t can be *defined* as

$$\epsilon_t = y_t - E(y_t|\Im_{t-1}) :$$

if we assume that the the process y_t is invertible, it can be written as

$$C(L)^{-1}A(L)y_t = G(L)y_t = \epsilon_t$$

from which we obtain

$$y_t = \underbrace{g_1 y_{t-k-1} + g_2 y_{t-k-2} + \cdots}_{E(y_t|\Im_{t-1})} + \epsilon_t$$

and therefore, given $\Im_t$, all the elements of the sequence ϵ_t up to time t can be considered known (as in "available from $\Im_t$")

With this proviso, the conditional expectation operator can be applied to every component of an ARMA model. The fact that the available information set is not infinite represents only a minor issue. Indeed, if we only have observations over the time span $\{0 \cdots T\}$, a very convenient solution is to extend our information set backwards by using the unconditional mean values of y_{-1}, y_{-2}, $\cdots$ *etc*. If the process is stationary and ergodic, as the sample size increases, the difference is negligible.[10]

I shall illustrate this point using an ARMA(1,1) process: once the concept is clear, its generalisation is trivial. Let us take, therefore, a process like

$$y_t = \varphi y_{t-1} + \epsilon_t + \theta \epsilon_{t-1}.$$

Let us place ourselves at time 0, when we have no observations. What is the best forecast we can make for y_1? Since we have no data, the conditional mean coincides with the marginal mean, and therefore $\hat{y}_1 = E(y_1) = 0$. One period passes, and we observe the actual value y_1. At this point, we can calculate the forecast error for period 1, namely $e_1 = y_1 - \hat{y}_1$; since $\hat{y}_1$ is 0, for the reasons

[10] The *exact* calculation can be carried out, if desired. There are many ways to do this, but the most common — also because it is easily automated — is to use a tool called the **Kalman filter**. For those who wish to know more, further details can be found in the literature.

just mentioned, it follows that $e_1 = y_1$. Our forecast for $\hat{y}_2$ is given by the following formula:

$$\hat{y}_2 = E(y_2|\Im_1) = E(\varphi y_1 + \epsilon_2 + \theta\epsilon_1|\Im_1) = \varphi E(y_1|\Im_1) + E(\epsilon_2|\Im_1) + \theta E(\epsilon_1|\Im_1).$$

Let's consider these terms one by one, bearing in mind that $\Im_1 = \{y_1\}$: clearly, the first two terms create no problems, since $E(y_1|\Im_1) = y_1$ (this is obvious) and $E(\epsilon_2|\Im_1) = 0$ (by assumption). But what about $E(\epsilon_1|\Im_1)$? Since ϵ_1 can also be interpreted as the forecast error that would be made at time 0 if the information set were infinite, then *the best possible forecast we can make regarding the forecast error at time 1 is exactly the forecast error that we actually made*. Based on this concept, we can formulate our forecast for y_2 as

$$\hat{y}_2 = \varphi y_1 + \theta e_1.$$

Let another period pass, and we observe y_2; from this, we calculate e_2, and the process continues, in the sense that at this stage we have everything we need to calculate $\hat{y}_3 = \varphi y_2 + \theta e_2$, and so forth. In practice, one-step-ahead forecasts for processes of the type

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

are computed as follows:

$$\hat{y}_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}, \tag{2.15}$$

that is to say, by using the actually observed values of y_t and the values of past forecast errors in place of ϵ_{t-i}.

A brief digression. One might legitimately wonder what the practical value is of forecasts made in this way, given that ARMA processes are merely a stylised and approximate representation of the time series we observe. In other words, the *historical* persistence characteristics of the time series are summarised by pretending that our data series is a realisation of suitably-chosen ARMA process.

There is no *logical* reason, however, why an approximation that worked well for the past should continue to work well for the future. Yes, we assume that the data-generating proces is stationary, but this may well be just wishful thinking. The stationarity assumption means, more or less implicitly, that the set of circumstances that have hitherto ensured that a given process is a good approximation of the behaviour of that given time series will continue to hold over the forecast horizon of interest.

This condition is often plausible when the series is a description of a reasonably stable physical phenomenon (for example, the temperature recorded daily at a certain geographical location at a certain hour). However, in the case of economic phenomena, this may be a rather bold assumption, as the causal chain of events that combine to determine the value of the series at any given moment is likely to be more unstable. I would consider it unprofessional to make a forecast for the price of crude oil based exclusively on an ARMA process that does not, for instance, take into account the political situation in Venezuela, or the state of affairs between Iran and the US.

More precisely, an ARMA model forecast should be accepted as a conditional forecast for a specific scenario: *if and only if* the political situation in Venezuela (and in Iran, and in the United States, *etc.*) remains more or less as it is today, *then* it can be said that, and so on and so forth.

2.6.2 Process dynamics

This aspect is generally investigated by making use of the so-called **impulse response function**. What is the impulse response function? The answer to this question involves a consideration that can be made in light of what was stated in the previous subsection: let us consider the equation

$$y_t = E\left[y_t|\Im_{t-1}\right] + \epsilon_t = \hat{y}_t + \epsilon_t,$$

that comes from equation (2.13).

The value of y_t can therefore be interpreted as the sum of two components: one ($\hat{y}_t$) which, at least in principle, is perfectly predictable given the past; the other (ϵ_t) which is absolutely unpredictable. In other words, the value of y_t can be thought of as depending on a persistence component to which a random disturbance or, as it is commonly known, a random **shock** is added, summarising everything unpredictable that occurred between time $t-1$ and time t. The effect of this component, however, also propagates into the future of the series y_t through the persistence effect. This is why the white noise ϵ_t is often called, in a more neutral form, the *one-step-ahead forecast error* or *innovation*.

Therefore, by writing the process in MA form (or, if you like, by considering its Wold representation)

$$y_t = A(L)^{-1}C(L)\epsilon_t = B(L)\epsilon_t$$

the i-th coefficient of the polynomial $B(L)$ can be thought of as the effect that a shock occurring i periods ago has on the current value of y_t, or, equivalently, as the impact that today's events will have on the series under study in i periods' time:

$$b_i = \frac{\partial y_t}{\partial \epsilon_{t-i}} = \frac{\partial y_{t+i}}{\partial \epsilon_t}. \tag{2.16}$$

In short, the impulse response function is simply given by the coefficients of the MA representation of the process, and it is generally examined via a plot with the values of i on the horizontal axis and the values of b_i on the vertical axis.

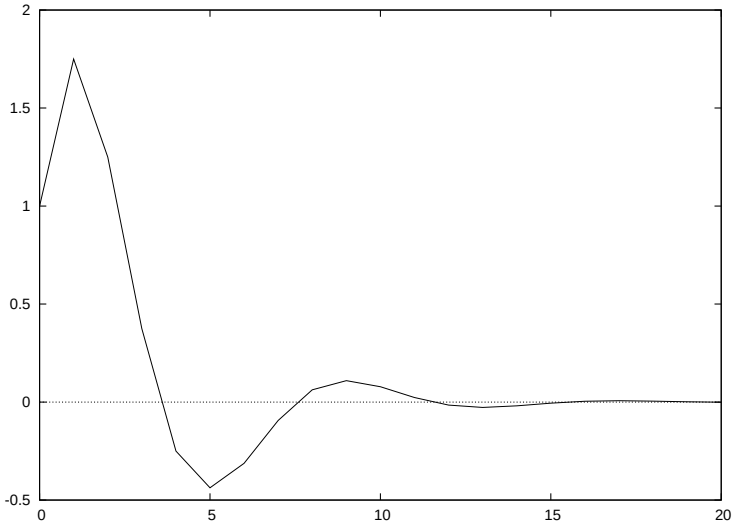
In principle, to calculate the Wold representation of an ARMA process, we need the inverse polynomial of $A(L)$. This can be rather tedious, especially if the order of the autoregressive component is high. Fortunatly, the same coimputation can be performed recusrively, by a simple algorithm which can also be implemented on a standard spreadsheet, as follows:

1. define a series e_t that contains all zeros except for one period, in which it equals 1. In other words, define a series e_t for which $e_0 = 1$ and $e_t = 0$ for $t \neq 0$;

2. define a series i_t, which you constrain to be equal to 0 for $t < 0$; for $t \geq 0$, instead, let $A(L)b_t = C(L)e_t$ hold.

The values obtained for the series b_t are exactly the Wold representation coefficients, that is the values of the impulse response function.

Figure 2.5: Impulse response function for $y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$



Example 2.6.2 *Let us take, for example, the following ARMA(2,1) process:*

$$y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$$

and say that, at time t, an "unpredictable event" equal to 1 has occurred (that is to say, $\epsilon_t = 1$). What effect does this have on the values of y from time t onwards? Let's go one step at a time. At time t, clearly, the effect is 1, since ϵ_t acts directly on y_t and does not affect its other components. At time $t+1$, we will have that

$$y_{t+1} = y_t - 0.5y_{t-1} + \epsilon_{t+1} + 0.75\epsilon_t,$$

and the effect of ϵ_t on y_{t+1} will be twofold: on the one hand, it appears directly, associated with a coefficient of 0.75; on the other hand, one must take into account the fact that the effect of ϵ_t is also contained within y_t, which is associated with a coefficient equal to 1. Therefore, the total effect will be 1.75. Moving forward by another period, the direct effect disappears, and only the effect generated by the lagged values of y remains. Doing some calculations, the effect of ϵ_t on y_{t+2} is found to be 1.25. The following small table may be helpful:

t	e_t	i_t
-2	0	0
-1	0	0
0	1	$b_{-1} - 0.5b_{-2} + e_0 + 0.75e_{-1} = 1$
1	0	$b_0 - 0.5b_{-1} + e_1 + 0.75e_0 = 1.75$
2	0	$b_1 - 0.5b_0 + e_2 + 0.75e_1 = 1.25$
3	0	$b_2 - 0.5b_1 + e_3 + 0.75e_2 = 0.375$
⋮	⋮	⋮

Those stubborn enough to continue up to 20 periods will be able to draw a plot like the one in Figure 2.5, from which it can be seen quite clearly that, after 8 periods, the func-

tion reproduces (with obvious decay) more or less the same dynamics. Consequently, it is reasonable to expect that a realisation of this process will exhibit cyclical patterns with a wavelength of approximately 8 periods.

A final remark: one might be led to believe that the Wold representation and the impulse response function are the same thing. Well, not necessarily: the impulse response function is defined as in equation (2.16), and can always be calculated, even if the process is not covariance-stationary, and, therefore, does not possess a Wold representation. However, if the process is in fact stationare, then the impulse response and the Wold representation coefficients coincide.

2.7 Estimation of ARMA models

Up until now, we have been pretending that the parameters of the stochastic process of our interest were known. But, in practice, they never are, so we have to use estimates. The standard technique for estimating ARMA parameters is maximum likelihood, where the distribution of the data is usually assumed in such a way that the problem is tractable. What everybody does[11] is assume that the white noise sequence $\ldots,\epsilon_{t-1},\epsilon_t,\epsilon_{t+1},\ldots$ has a joint Gaussian distribution, so, by virtue of the Wold representation, y_t is Gaussian too.

It may be useful to briefly recall what is meant by the likelihood function. The likelihood is the sample density function, calculated for the observed sample. It will depend on a vector ψ of unknown parameters, which determine its shape. For this reason, we write it as $L(\psi)$. Maximising this function with respect to ψ yields the maximum likelihood estimate.

Example 2.7.1 *If we toss a coin and obtain 'heads', we have a realisation of a random variable that takes the value 1 (heads) with probability p and 0 with probability $1-p$; in this case, the likelihood is the probability of observing the sample that was actually observed, given the parameter $p \in [0,1]$, namely $L(p) = p$; the maximum likelihood estimate in this example is 1. If we had obtained 'tails', the likelihood would have taken the form $L(p) = 1-p$, and the maximum likelihood estimate would have been 0.*

If we toss 2 coins, we would have the following possible outcomes:

Sample	*L(p)*	*Maximum*
TT	p^2	*1*
TC	$p(1-p)$	*0.5*
CT	$(1-p)p$	*0.5*
CC	$(1-p)^2$	*0*

and so forth.

When we observe a realisation of a stochastic process (or, more pedantically, a time series that we can think of as such) $x_1,\ldots,x_T$, the likelihood function is nothing but the joint density function of the observed part of the process, namely the marginal density function of the random vector $x_1,\ldots,x_T$,

[11] Well, except vor very special cases we don't concern us with here.

calculated for the observed values; in the case of an ARMA process of the type

$$A(L)x_t = \mu + C(L)\epsilon_t$$

it will depend on the parameter vector $\psi = \{\mu; \varphi_1 \dots \varphi_p; \theta_1 \dots \theta_q; \sigma^2\}$.

If the process is Gaussian, the likelihood function is simply the density function of a multivariate normal:

$$L(\psi) = f(x;\psi) = (2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(x-k)'\Sigma^{-1}(x-k) \right\},$$

where x is the vector $(x_1, \dots, x_T)$ of the T observations; k and Σ are its first and second moments, which depend on ψ. For example, the ij-th element of the matrix Σ is simply $\gamma_{|i-j|}$, the autocovariance of order $|i-j|$ which, as we know, is a function of the parameters of the ARMA process. The ML estimator is therefore

$$\hat{\psi} = \underset{\psi}{\text{Argmax}}\ L(\psi). \tag{2.17}$$

It can be shown that the maximum likelihood estimators of Gaussian ARMA processes are consistent, asymptotically normal, and asymptotically efficient. Furthermore, under fairly mild conditions, the properties of consistency and asymptotic normality are preserved even when the true distribution of the process is not normal (in this case, we refer to quasi-maximum likelihood estimates). This is the reason why, in many cases, people don't bother checking for normality.

From a theoretical perspective, this is all. From a practical point of view, the problems are only just beginning. First of all, it should be noted that, as usual, we do not work with the function $L(\psi)$, but with its logarithm

$$\ell(\psi) = \log L(\psi) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\left[\log|\Sigma| + (x-k)'\Sigma^{-1}(x-k) \right],$$

but this is of no consequence, since the optimal vector ψ is the same for both functions. There are three main problems:

1. the system of equations resulting from setting the score vector $s(\psi) = \frac{\partial l(\psi)}{\partial \psi}$ to 0 is non-linear, and in general cannot be solved analytically, so that no closed-form expressions are available to calculate the elements of $\hat{\psi}$ as simple functions of the data; therefore, numerical optimisation techniques are needed.

2. the order of the polynomials $A(L)$ and $C(L)$ suitable for adequately representing the process of which we believe x_t to be a realisation is unknown;

3. to calculate $\ell(\psi)$, it is necessary to know the matrix Σ; now, the matrix Σ is a $T \times T$ matrix, and when T is large (even in the tens, but time series samples can reach sizes in the order of *tens of thousands* of observations), even the simple calculation of its elements as a function of ψ may be a problem. The same applies to its determinant and its inverse.[12]

[12] It is true that Σ has a very special structure: it's a "symmetric Toeplitz" matrix. This simplifies the problem to some extent, but doesn't get rid of it.

2.7.1 Numerical techniques

The first of the problems listed above deserves an in-depth examination, but this is really not the place for it. Suffice it to say that this problem becomes negligible if adequate computing resources are available. Numerical algorithms exist that allow the maximum of likelihood functions to be found without great difficulty, and all econometric packages are equipped with them.

To oversimplify as much as possible, one can say that these algorithms serve to find the maximum of a function once we are able to calculate this function, and its first derivative, for any point in the parameter space. If the function is smooth, its derivative (the *gradient*) is 0 at the maximum.

Roughly speaking, the procedure is as follows:

1. Start from a "random" point, ψ_0;
2. calculate the score $s(\psi_0) = \frac{\partial \ell(\psi)}{\partial \psi}$;
3. if $s(\psi_0)$ is "small enough", stop. Otherwise, use it to figure out a direction $d(s(\psi_0))$ in which to move;
4. calculate $\psi_1 = \psi_0 + d(s(\psi_0))$;
5. check that $\ell(\psi_1) > \ell(\psi_0)$;
6. replace ψ_0 with ψ_1 and start again from step 2.

There are many algorithms of this type: essentially, each one calculates the direction vector $d(s(\theta_0))$ in its own way, aiming to ensure that $\ell(\psi_1) > \ell(\psi_0)$ at each iteration, so that sooner or later the maximum is reached.

Some readers may have noticed that the algorithm above crucially depends on what happens at step 3. As long as $s(\psi_0)$ is not 'small enough', the algorithm keeps going. Hence, we must decide what it means for a vector to be 'small', to avoid the risk of the algorithm running forever and our CPU turning into a fireball.
This is one of those cases where a question is only seemingly simple: there are various ways to answer it, and as usual, none of them is 'right'. One of the most common criteria is to decide that $s(\psi_0)$ is 'zero' when none of its elements exceeds a pre-specified threshold in absolute value, for example, something like 1.0E-07. However, it is not the only one, nor is it necessarily the best.

Usually, the likelihood functions encountered in estimating ARMA models are smooth, and the possibility of having multiple maxima is negligible. Consequently, it is sufficient to be able to calculate the first derivatives of the likelihood for any vector ψ to reach — sooner or later — the maximum.

2.7.2 Lag Selection

As for the second problem at the end of section 2.7, choosing the orders of the polynomials $A(L)$ and $C(L)$ (call them p and q, respectively) is a matter of highly skilled craftsmanship, which requires theoretical knowledge, experience, and a keen eye.

One may be tempted to think that p and q can be estimated along with the rest of the parameters, but things are not this simple, because of the so-called so-called **common factors** problem: start from an ARMA(p,q) process

$$A(L)x_t = C(L)\epsilon_t,$$

and apply the filter $(1-\lambda L)$ to both sides of the equality: call

$$A_\lambda(L) = (1-\lambda L)A(L)$$

and

$$C_\lambda(L) = (1-\lambda L)C(L),$$

so the resulting representation is

$$A_\lambda(L)x_t = C_\lambda(L)\epsilon_t, \tag{2.18}$$

that is, an ARMA($p+1$,$q+1$) process. The Wold representation for x_t you get with both representations is exactly the same, since

$$B(L) = \frac{C_\lambda(L)}{A_\lambda(L)} = \frac{(1-\lambda L)C(L)}{(1-\lambda L)A(L)} = \frac{C(L)}{A(L)}$$

and the $(1-\lambda L)$ factors cancel out. The process x_t, therefore, has a completely equivalent ARMA($p+1$,$q+1$) representation *for all possible values of* λ. Therefore, for every ARMA(p,q) representation, you have infinitely many ARMA($p+1$,$q+1$) representations that are absolutely equivalent, because the shape of the autocorrelation function is the same for every λ. As we say in econometrics, the model is under-identified.[13]. Therefore, the value of the likelihood function is the same for any λ, and thus there is no unique maximum: there are infinitely many maxima, one for each value of λ.

Example 2.7.2 *A simple example: what kind of process is*

$$y_t = 0.5y_{t-1} + \epsilon_t - 0.5\epsilon_{t-1}?$$

Easy, you will say: it is an ARMA(1,1). Quite right. However, it is also a white noise; in fact

$$y_t = \frac{1-0.5L}{1-0.5L}\epsilon_t = \epsilon_t.$$

In practice, I wrote a white noise process as an ARMA(1,1). This latter representation is redundant, but not wrong. *The important aspect to note is that my choice $\lambda = 0.5$ is completely irrelevant: I could have used 0.7, 0.1, or whatever. There are infinitely many "not wrong" representations.*

[13] I will briefly recap what is meant by the identification of a model in econometrics: a model is said to be under-identified if there is more than one data representation consistent with what is observed. In practice, it is not possible to decide on the basis of the data whether representation A or representation B is more correct; in these cases, the expression "observational equivalence" is used. If the model is parametric (as in the majority of cases), it is identified if the likelihood function has a unique global maximum; consequently, a necessary condition for identification is the non-singularity of the Hessian matrix at the maximum point. Identification is, clearly, itself a necessary condition for the existence of a consistent estimator.

From a practical point of view, modelling an ARMA(p,q) as an ARMA($p+1$,$q+1$) leads to all sorts of problems. First, the numerical algorithm often struggles to converge (which is unsurprising, given that there is no unique maximum). Second, even assuming that convergence is eventually achieved somehow, the maximum point we find is only one of the infinitely many possible representations of the model[14].

Therefore, before we embark on likelihood maximisation, we must make a preliminary choice on p and q. Years ago, when CPUs were toys, RAM was precious and software was scarce, textbooks used to recommend a rather involved procedure,[15] based on the fact that there are highly specific relationships between the order of the polynomials and the autocorrelation structure. By examining of the sample autocorrelations, an initial hypothesis can be made about the orders of the polynomials. If, for example, it is noted that the sample autocorrelations cut off abruptly beyond a certain order q, one might consider using an MA(q) model, whose theoretical autocorrelations exhibit the same characteristic. If, on the other hand, the autocorrelations decay smoothly, an AR process might be better.

At this stage, statistics known as partial autocorrelations (which, in practice, are never used for any other purpose) were also employed. Defining partial autocorrelations rigorously is somewhat boring. It is quicker to say how they are calculated: the partial autocorrelation of order p is calculated by running a regression of y_t on a constant and $y_{t-1} \dots y_{t-p}$. The resulting coefficient associated with y_{t-p} is the partial autocorrelation of order p. These quantities cut off abruptly in the case of pure AR models, and decay gradually in the case of pure MA models.

Even today, some teachers like to inflict on their students with these 'old-fashioned' techniques for choosing p and q by making elaborate considerations about the shape of the autocorrelation function, while conveniently glossing over that, in the majority of practical cases, you either have a highly trained eye or you are left completely in the dark. The fact is that these techniques were invented in an era when ML estimation of ARMA models was difficult and expensive; as a result, it was essential to have as precise an idea as possible of the potential values of p and q before attempting estimation. Today, estimating an ARMA model is ridiculously easy, and we leave the art of interpreting correlograms to vintage fans.

The modern approach to the choice of p and q is based on special statistics called **information criteria**. These are based on concepts taken from information theory, and I am happy to refer to the literature for further details. Information criteria start from the value of the log-likelihood (multiplied

[14] It is true that all these representations have the same Wold representation, so the forecasts they lead to and the impulse responses they generate are identical, but we have the problem that any kind of testing in this case is precluded. Indeed, the maximum point we find is only one of the infinitely many possible ones, and therefore the Hessian of the likelihood function is singular. Since all test statistics are based in some way on the curvature of the likelihood function, it is clear that tests cannot be performed.

[15] In the statistical literature this stage is known as the identification phase. This term sometimes causes a bit of confusion, because in econometrics the word 'identification' usually means something else.

by -2) at the maximum and add a *penalty function*, which is increasing in the number of parameters.

Several criteria exist, which differ on the form of the penalty function. The most often used ones, in the contex of ARMA modelling are the Akaike IC (AIC), the Schwartz IC (BIC, where B is for "Bayesian") and the one by Hannan and Quinn (HQC):

$$\begin{aligned} \text{AIC} &= -2\ell(\hat{\psi}) + 2k \quad (2.19) \\ \text{BIC} &= -2\ell(\hat{\psi}) + k\log T \quad (2.20) \\ \text{HQC} &= -2\ell(\hat{\psi}) + 2k\log\log T \quad (2.21) \end{aligned}$$

where k is the dimension of the vector ψ (the number of estimated parameters) and T is the sample size. The rule is to pick the model that minimises information criteria, so you strike a balance between model fit (given by the value of the log-likelihood) and the model size. Therefore, when choosing between two models, the "better" one should have a lower IC value.

In the light of what we said earlier, the logic should be clear. If p and/or q are too small, the fit will be poor, and the IC will be large; if, conversely, p or q are too large, the model will be over-identified and will give nearly the same fit as a correctly identified one, while using more parameters. By choosing the model with minimum IC, you try to pick a model that fits the data nicely without being too complex.

Therefore, the modern approach is simply a brute-force one: you select "maximal" orders p and q on the basis of economic/statistical common sense and try all the possible combinations. The model that minimises your IC of choice is the winner. Sometimes, the three criteria may not agree, but even if they don't your problem is reduced to picking one out of two or three at most.

Once the estimation has been carried out, one usually checks whether the 'residuals' (the one-set-ahead forecast errors) look like white noise process. The approach usually adopted is based on the use of some testing procedures which allow us to determine whether a process is a white noise or possesses persistence; the most widespread is the so-called **Ljung-Box** test, which is based on the property that in large samples, sample autocovariances tend to 0 in the case of a white noise: the actual test statistic is

$$LB(p) = T(T+2)\sum_{i=1}^{p}\frac{\hat{\rho}_i^2}{T-i};$$

it will be noted that it is substantially a weighted sum of the squared sample autocorrelations up to order p. The smaller these are, the smaller the test statistic will be; under the null hypothesis the true value of all autocorrelations up to order p is 0, and the critical values are those of the χ_p^2 distribution.

2.7.3 Likelihood calculation

The third problem I hinted at near the end of section 2.7 is more intriguing: essentially, we need to write the likelihood function in such a way that the calculation of prohibitively large matrices is not required. The literature studied the specific problem of evaluating log-likelihood functions for Gaussian

ARMA models in great detail between the 1970s and the 1980s, so that the built-in algorithms found in econometric packages today are very stable and efficient. Here, however, I want to illustrate a rather interesting technique (the so-called sequential factorisation) because of its great ingenuity and its generality (it is used in much more general cases than ARMA models).

To illustrate sequential factorisation, start from the definition of conditional probability, which is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

from which

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

If we apply this rule to the density function of a bivariate random variable, we obtain

$$f(x,y) = f(y|x)f(x) = f(x|y)f(y). \tag{2.22}$$

This little trick can also be repeated with a trivariate random variable, thus obtaining

$$f(x,y,z) = f(x|y,z)f(y,z) = f(y|x,z)f(x,z) = f(z|x,y)f(x,y). \tag{2.23}$$

By combining (2.22) and (2.23), one can clearly write, for example,

$$f(x,y,z) = f(z|x,y)f(x,y) = f(z|x,y)f(y|x)f(x)$$

and therefore a joint density function of n random variables can be written as

$$f(x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i|x_1,\ldots,x_{i-1})$$

thereby transforming a function of many variables into a product of functions of one variable. In the simple case when the variables x_i are independent $f(x_i|x_1,\ldots,x_{i-1}) = f(x_i)$ and (as is well known), the joint density function is the product of the marginal ones.

Since the likelihood function is nothing but the density function of the sample, this same decomposition can be applied to the likelihood function, and in the case of ARMA models it takes an especially nice form. Since in ARMA models

$$y_t = E(y_t|\Im_{t-1}) + \epsilon_t$$

(see equation (2.13)) we have that, conditionally on $\Im_{t-1}$, the distribution of y_t is the same as that of ϵ_t, and therefore the likelihood can be written equivalently in terms of the one-step-ahead forecast errors rather than the y_t values.

In the case of pure AR processes, the result is straightforward: in this case, if the process is

$$y_t = \mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t$$

and ϵ_t is normal, the density function $f(y_t|\Im_{t-1})$ is simply a normal distribution:

$$y_t|\Im_{t-1} \sim N\left(\mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p}, \sigma^2\right).$$

In the case of an AR(1) process, we will have that

$$y_t|\Im_{t-1} \sim N\left(\mu + \varphi y_{t-1}, \sigma^2\right)$$

and therefore $f(y_t|\Im_{t-1}) = f(y_t|y_{t-1})$: the only element of $\Im_{t-1}$ that matters is the most recent one. Consequently, the likelihood can be written as

$$L(\mu, \varphi, \sigma^2) = f(y_1)f(y_2|y_1)f(y_3|y_2)\cdots = f(y_1) \times \prod_{t=2}^{T} f(y_t|y_{t-1}).$$

Taking logarithms, we obtain

$$\begin{aligned} l(\mu, \varphi, \sigma^2) &= \log f(y_1) + \tilde{l}(\mu, \varphi, \sigma^2) = \\ &= \log f(y_1) - \frac{1}{2}\sum_{i=2}^{T}\left(\log 2\pi + \log\sigma^2 + \frac{(y_t - \mu - \varphi y_{t-1})^2}{\sigma^2}\right) \\ &= \log f(y_1) - \frac{1}{2}\sum_{i=2}^{T}\left(\log 2\pi + \log\sigma^2 + \frac{e_t^2}{\sigma^2}\right), \end{aligned}$$

where I adopted the notation $e_t = y_t - E(y_t|\Im_{t-1})$.

If the first term were zero, the remainder would be equal to a completely standard log-likelihood function for a linear model in which the dependent variable is y_t, the explanatory variable is y_{t-1} (plus the intercept), and the disturbance is normal with mean 0 and variance σ^2.

As is well known, in this case OLS is the maximum likelihood estimator, so that, if it were not for the $\log f(y_1)$ term, we could simply run a regression to get the job done. But then, for very large samples, this term becomes negligible in determining the total loglikelihood $l(\mu, \varphi, \sigma^2)$, since (unlike the second term), it does not grow as T increases. The OLS statistic, which maximise $\tilde{l}(\mu, \varphi, \sigma^2)$, therefore converge towards the ML estimator, and asymptotically share their properties[16].

This line of reasoning also holds for AR(p) models: in this case, the first term of the likelihood becomes $\log f(y_1, \ldots, y_p)$, but the argument remains unchanged. Moreover, it is true that, under fairly general conditions, the OLS estimator of the parameters of a stationary autoregressive process is a consistent and asymptotically normal estimator even if the process ϵ_t is not Gaussian; in these cases, the OLS estimator as asymptotically equivalent to the quasi-maximum likelihood estimator, and it's OK to use it (with the appropriate corrections to standard errors). This is the case of *dynamic regression*, where the asymptotic properties of OLS estimators can be proven via limit theorems that are usually dealt with elsewhere in an Econometrics course, so that I don't have to do it here.

In the case of models with a moving average component, the discussion becomes only slightly more intricate. Ss I have already pointed out in section

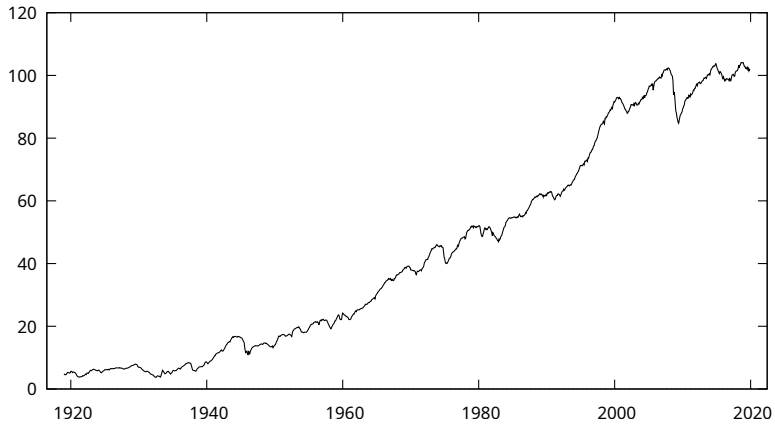
[16] In cases such as this, one speaks of **conditional** likelihood, to indicate that the likelihood function we are using treats the first p observations as fixed, and therefore is based on the distribution of $y_{p+1} \ldots y_T$ *conditional on* $y_1 \ldots y_p$. There are also techniques that maximise the so-called **exact** likelihood, that is, the one which also takes into account the distribution of the first p observations, but asymptotically it makes no difference.

2.6.1, in an ARMA model, the white noise process driving the y_t process can be interpreted as the difference between y_t and its expected value conditional on $\Im_{t-1}$ (see eq. (2.13)). Consequently, if we assume that the distribution of y_t conditional on $\Im_{t-1}$ is normal, it follows that the one-step-ahead forecast errors are a sequence of uncorrelated (and therefore independent) normal variables with mean 0 and constant variance, so that the likelihood can be calculated very simply using the forecast errors.

2.8 In practice

To illustrate how everything works, let us take the case where we want to set up a model for a time series that we have already spent some time on in Section 1.4, namely US industrial production[17]. Let us say that the available time series is the one plotted in figure 2.6.

Figure 2.6: USA industrial production: seasonally adjusted index (1921–2019)



To begin with, you will notice that the time series is very long: it runs from January 1919 to September 2019. 'Wow, over a thousand observations!', some might say, 'when it comes to sample size, the more the better!'. Well, not really. Let me remind the reader that what we are trying to do is find a stochastic process such that these data can be thought of as a single realisation.

Now, the "true" stochastic process that generated this time series (assuming it exists) is something that has surely changed a lot over our observation window. Within the timespan of our data, we have Charlie Chaplin, World War II, the invention of the ballpoint pen, Marilyn Monroe, personal computers, *Smells Like Teen Spirit*, and YouTube.

I find it somewhat bold to postulate the existence of a data representation that survives all this intact. In the terms of Chapter 1, we could say with a certain degree of confidence that the "true" stochastic process that generated

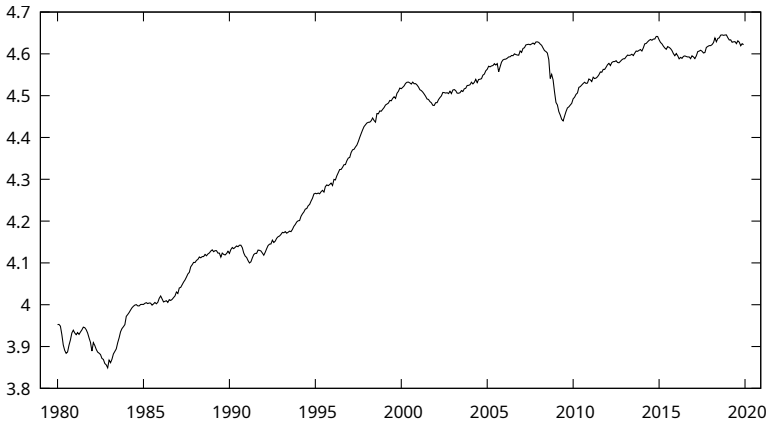
[17] For those who are familiar with these matters: I have taken a shortcut and used the seasonally adjusted series. These people will easily guess why.

the data is non stationary. If we really want to force this series into the realisation of a stationary process, it is advisable to shorten the sample to make our fictional assumption a little less unrealistic. In this case, economists like to say that we are excluding the so-called *strucural breaks*[18]; it should be noted that this line of reasoning is possible *without even looking at the data*.

I'll adopt a completely arbitrary procedure (this is just an example, after all) and I declare that the world we live in today began in January 1980. I also decided to use the log-transformed series. This is a very common procedure, and it allows for a more natural interpretation, since first differences are (approximately) percentage changes[19].

The result is the time series in Figure 2.7, which is long enough to allow us to say something interesting (480 observations), but at the same time tells us a reasonably homogeneous story.

Figure 2.7: Logarithm of US monthly industrial production



Can we now claim to be observing a realisation of a stationary process? Again, the answer is "probably not". Here, however, the problem does not arise from the heterogeneity of the sample, but from the upward you see in Figure 2.7, which evidently precludes us from thinking that the process has a constant mean. One might consider a process that is stationary around a deterministic trend, that is, something of the type $Y_t = (a + bt) + u_t$, where u_t is some ARMA process.

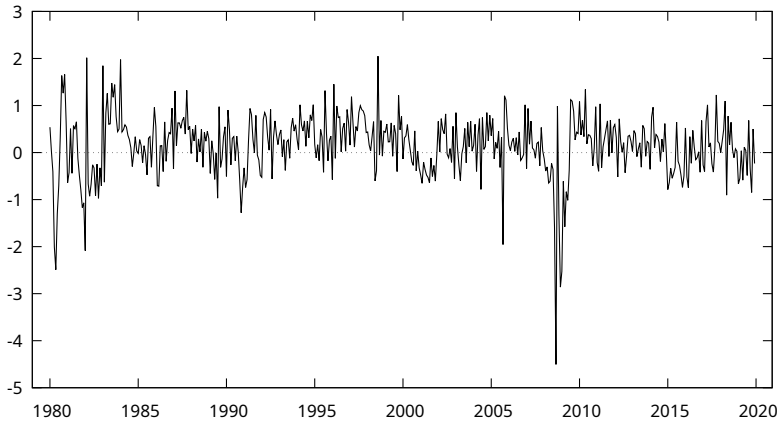
This wouldn't be unreasonable, especially because the parameter b could be interpreted as the long-run exogenous rate of technical progress. However, this is not such a good idea, for reasons that I will explain in Chapter 3. For now, will simply hint at the fundamental problem: even after removing a deterministic trend, this time series is too persistent to conclude that its generating process is stationary. Sorry, be patient.

Another possibility is to transform the time series in such a way to get rid

[18] Some methods exist for working with time series containing structural breaks, but these are a bit too esoteric for these lecture notes.

[19] Let me recall that $\log(y_t) - \log(y_{t-1}) = \log(1 + \frac{\Delta y_t}{y_{t-1}}) \simeq \frac{\Delta y_t}{y_{t-1}}$

Figure 2.8: Industrial production growth rate



of the problem, but retaining economic interpretability. We brilliantly solve the problem by taking first differences and considering $y_t = 100\,\Delta Y_t$; you can admire the result in Figure 2.8 and, like I said earlier, is roughly the monthly growth rate in percentage points.

Figure 2.9: Industrial production growth rate — correlogram

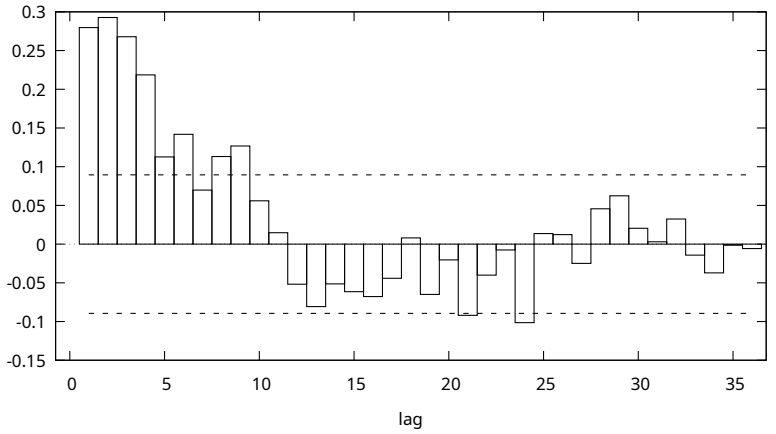


Figure 2.9, on the other hand, shows the correlogram. The two horizontal dashed lines surrounding the actual correlogram show the constants $\pm 1.96/\sqrt{T}$, where T is the sample size; since $T = 480$ observations, this value is around 0.089. These two lines are often included in correlograms to provide a quick-and-dirty hypothesis test. The sample autocorrelations $\hat{\rho}_k$ are consistent estimators of the true autocorrelations ρ_k; if $\rho_k = 0$, then it can be shown that $\sqrt{T}\hat{\rho}_k \xrightarrow{\text{d}} N(0,1)$. Consequently, the horizontal band $\pm 1.96/\sqrt{T}$ is the 95% non-rejection region for the hypothesis $\rho_k = 0$; in practice, autocorrelations outside the band are "statistically significant".

As can be seen in Figure 2.9, there are at least six or seven significant bars,

so we can reasonably rule out the hypothesis that y_t possesses no persistence (something which a trained eye could have spotted from Figure 2.8). Furthermore, the Ljung-Box test up to 12 lags equals 172.8037; under the null, this would be a realisation of a χ^2_{12} so the p-value is 0 and rejecting the null hypothesis is the only possible choice.

Persistence, therefore, is indeed present and an ARMA model is worth trying. Let's choose p and q by the "brute-force" approach I described in Section 2.7.2. Set (quite arbitrarily) the maximum order for p and q to 4 and try all possible combinations using use Akaike (AIC), Schwarz (BIC), and the Hannan–Quinn (HQC) information criteria.

Table 2.1: ARMA models for the USA industrial production index: information criteria

Akaike (AIC)

AR \ MA	0	1	2	3	4
0	972.925	948.139	931.396	916.234	900.387
1	935.846	900.841	899.432	900.705	901.454
2	911.245	899.924	901.075	902.668	902.372
3	900.744	900.648	901.070	**892.969**	894.739
4	899.520	899.807	899.471	902.661	896.179

Schwartz (BIC)

AR \ MA	0	1	2	3	4
0	981.272	960.660	948.091	937.103	925.429
1	948.367	**917.536**	920.300	925.747	930.671
2	927.940	920.793	926.118	931.884	935.763
3	921.613	925.691	930.287	926.359	932.303
4	924.563	929.023	932.862	940.225	937.917

Hannan-Quinn (HQC)

AR \ MA	0	1	2	3	4
0	976.206	953.061	937.959	924.437	910.230
1	940.768	907.403	907.635	910.548	912.939
2	917.807	908.127	910.919	914.152	915.497
3	908.948	910.492	912.554	**906.094**	909.504
4	909.364	911.291	912.596	917.427	912.585

As I argued back in Section 2.7.2, the choice is made by selecting the model for which the criterion is lowest. Now, take a look at Table 2.1, where rows indicate the AR order and the columns the MA order; for each table, I highlighted the minimum by framing it in a small box. As can be seen, the choice is not unique: while AIC and HQC suggest an ARMA(3,3), BIC opts for a more conservative ARMA(1,1). This is no coincidence: as a rule, AIC usually tends to be rather permissive in terms of the total number of parameters, whereas BIC has the opposite tendency. HQC generally opts for a middle ground.[20]

[20]Moreover, it is true that the Akaike information criterion, which has its own historical im-

All in all, there is no clear winner. Let's look at these two models a bit more closely. In both cases, the correlogram of the one-step-ahead forecast errors, what we would call the residuals in a regression context, is fairly flat and the relative Ljung-Box test accepts the null without any problems. I'm not reporting these statistics here to avoid interrupting the flow; the core point is that is that both these models adequately captures the persistence that is present. What remains to be seen is: is one preferable to the other?

Table 2.2: ARMA(1,1) model

	Coefficient	Std. Error	z	p-value
const	0.1353	0.0690	1.9602	0.0500
ϕ_1	0.8560	0.0431	19.8457	0.0000
θ_1	−0.6420	0.0607	−10.5816	0.0000

Mean dependent var	0.140598	S.D. dependent var	0.664580
Mean of innovations	0.001256	S.D. of innovations	0.613140
R^2	0.147043	Adjusted R^2	0.145258
Log-likelihood	−446.4203	Akaike criterion	900.8405
Schwarz criterion	917.5357	Hannan–Quinn	907.4030

		Real	Imaginary	Modulus	Frequency
AR					
	Root 1	1.1682	0.0000	1.1682	0.0000
MA					
	Root 1	1.5575	0.0000	1.5575	0.0000

The ARMA(1,1) model, shown in Table 2.2, is the more parsimonious of the two. Its goodness of fit is admittedly limited: the standard deviation of one-step-ahead forecast errors ($\hat{\sigma}$) is approximately 0.613, which is only slightly smaller than the observed time series' standard deviation, which is 0.665. The R^2 index, which has the same interpretation as in an ordinary regression model, is just 14.7%. This means that the unconditional dispersion of y_t is barely lower than the dispersion of the distribution conditional on $\Im_{t-1}$. Some persistence is certainly there, and it does in fact help us somehow to predict the series, but only to some extent (more on this later).

The ARMA(3,3) model (Table 2.3) shows a different, slightly improved scenario where the standard deviation of forecast errors is slightly smaller (0.603) and the R^2 index goes up to 17.6%. That said, one may be tempted to point out that all the parameters in the table appear to be very significant and therefore consider the model in Table 2.3 as preferable on these grounds, which is what you would do when you interpret the output from a regression.

However, consider the bottom part of the table, where the roots of the polynomials are shown. For example, the estimated AR polynomial

$$A(z) = 1 - 1.9661z + 1.7748z^2 - 0.7031z^3$$

portance because it was the first one to be invented, has the problem, compared to the others, of not being *consistent*: it can be shown that the probability of it selecting the "correct" model does not necessarily converge to 1 asymptotically, as happens with the other two; consequently, it is slightly out of favour today.

Table 2.3: ARMA(3,3) model

	Coefficient	Std. Error	z	p-value
const	0.1378	0.0669	2.0591	0.0395
ϕ_1	1.9661	0.1079	18.2249	0.0000
ϕ_2	−1.7748	0.1718	−10.3313	0.0000
ϕ_3	0.7031	0.0934	7.5298	0.0000
θ_1	−1.8041	0.1053	−17.1385	0.0000
θ_2	1.6848	0.1346	12.5167	0.0000
θ_3	−0.6221	0.0840	−7.4027	0.0000

Mean dependent var	0.140598	S.D. dependent var	0.664580
Mean of innovations	0.000454	S.D. of innovations	0.602790
R^2	0.175611	Adjusted R^2	0.166914
Log-likelihood	−438.4843	Akaike criterion	892.9687
Schwarz criterion	926.3590	Hannan–Quinn	906.0937

			Real	Imaginary	Modulus	Frequency
AR						
	Root	1	1.1744	0.0000	1.1744	0.0000
	Root	2	0.6750	−0.8692	1.1005	−0.1449
	Root	3	0.6750	0.8692	1.1005	0.1449
MA						
	Root	1	1.4838	0.0000	1.4838	0.0000
	Root	2	0.6122	−0.8417	1.0408	−0.1499
	Root	3	0.6122	0.8417	1.0408	0.1499

has three roots, that is

$$A(z) = 0 \Longleftrightarrow z = \begin{cases} \lambda_1 = 1.1744 \\ \lambda_2 = 0.6750 - 0.8692i \\ \lambda_3 = \bar{\lambda}_2 = 0.6750 + 0.8692i \end{cases}$$

so it can be written as

$$A(z) = \left(1 - \frac{z}{\lambda_1}\right)\left(1 - \frac{z}{\lambda_2}\right)\left(1 - \frac{z}{\lambda_3}\right),$$

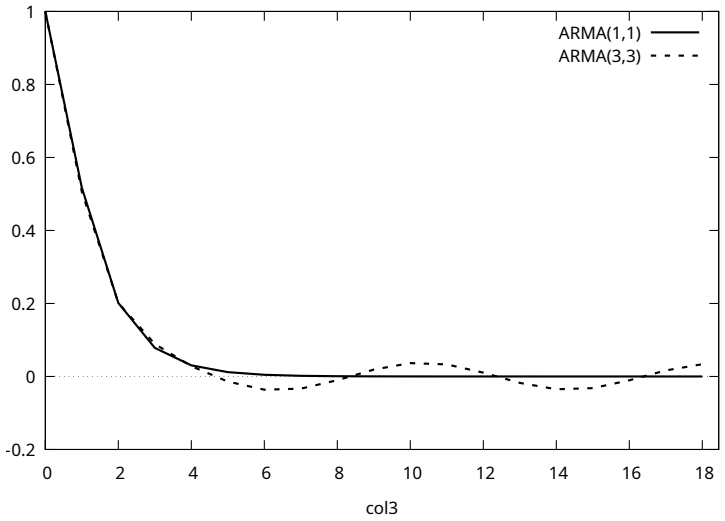
and a similar expression holds for $C(L)$.

Two of the roots of $A(L)$ are the pair of complex numbers $0.6750 \pm 0.8692i$ (which, incidentally, is suspiciously close to 1 in modulus), which is accompanied by the pair of complex numbers $0.6122 \pm 0.8417i$ among the roots of $C(L)$. One may suspect that we are are very close to the "common factors" situation here (go back to page 40 if needed). If those complex roots were the same, they could be "simplified" from the Wold representation and we'd be back to an ARMA(1,1) model.

In other words, the estimate in Table 2.3 may be an inflated estimate with parameters that serve no purpose at all. In fact, the other roots of the two polynomials are very close to those of the ARMA(1,1) model; in other words, what you see in table 2.3 could well be the exact same model from table 2.2 in disguise.

Is this the case? Well, here is where the different information criteria disagree: BIC says "they're the same model", AIC and HQC say "no, they're not" so you have to make your decision yourself.

Figure 2.10: Impulse response functions



To make a choice, it may be helpful to consider Figure 2.10, which contains, for the two models examined here, the impulse response function up to 18 steps, that is, a graphical representation of the first 18 coefficients of the polynomial $\frac{C(L)}{A(L)}$. Notice how the impulse responses for the ARMA(1,1) and ARMA(3,3) models are very close to one another, especially for the first five terms. The only noticeable difference is a slight undulating pattern for the ARMA(3,3) model, which is due to the the complex roots in the AR and MA polynomials not cancelling out exactly and seems to suggest a faint 10-months cycle. In practice, these two models are two alternative possibilities for summarising a feature of persistence in the series that is roughly equivalent.

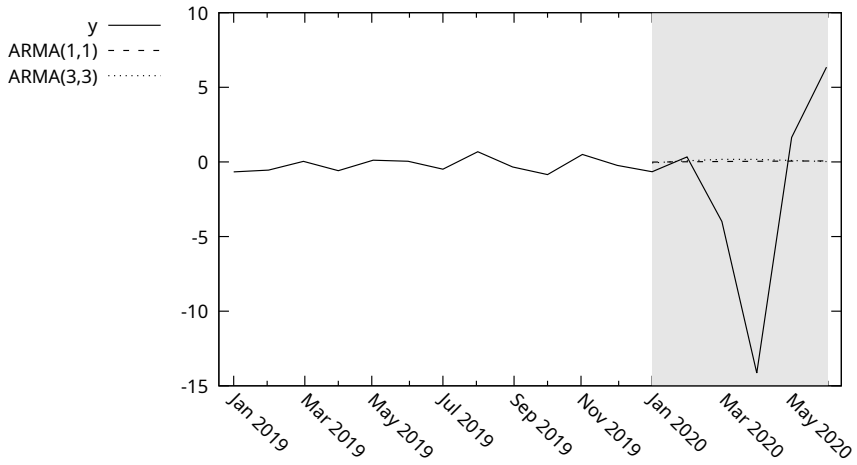
Another way to see the substantial equivalence of the two models is to consider what happens when using them as forecasting models: let's see what happens if we try to use the two models for out-of-sample forecasting in the first half of 2020: Table 2.4 displays the actual values and the forecasts for our two models. The same date are used to plot Figure 2.11, where the shaded area is the period outside $\Im_T$.

Table 2.4: Out-of-sample forecasts

	y_t	ARMA(1,1)	ARMA(3,3)
2020:1	-0.657	-0.011	-0.066
2020:2	0.335	0.010	0.077
2020:3	-3.990	0.028	0.172
2020:4	-14.142	0.044	0.170
2020:5	1.640	0.057	0.097
2020:6	6.347	0.068	0.025

The two forecasts are very similar to each other. This should come as no

Figure 2.11: Forecasts



surprise, since we already stablished that the two Wold representations are very much alike. On these grounds, the choice between the two models becomes essentially a matter of taste. Personally, I would probably go for the ARMA(3,3) model because the weak evidence for a 10-month cycle may have some some interesting economic interpretation (provided it's not an artifact), but if one chose the ARMA(1,1) model on the grounds of parsimony, I'd have no objections. For all practical purposes, the two models are essentially equivalent.

That said, the most spectacular thing to observe is that the two forecasts are both *very, very, very* wrong, because of the COVID pandemic, when industrial production underwent a veritable collapse, totally missed by both models. Why? The reason is that the only source of information we're using here is the past of the series itself: the persistence of the series, whilst non-negligible, provides little information on the future (let me remind you that the R^2 indices for the two model were rather small). This is why, in general, models of this type only predict decently a few steps ahead. There is nothing in the information set used by our two ARMA models that could have predicted the COVID outbreak and the consequent drop in industrial production.

I'm not saying that the COVID pandemic was totally unpredictable in 1/1/2020, because in fact by then several people in Wuhan had already died, but no sign of these events was in the industrial production data, and therefore in $\Im_T$. In fact, this is a much more general point: the quality of the predictions that *any* model can make, no matter how fancy and sophisticated, depends primarily on how much information the model incorporates. Forecasting is no black magic.

The moral of the story is: in economics, an ARMA model can serve, at most, for short-term forecasting during tranquil periods. If you really want a job as a soothsayer, use ARMAs, but follow the news too.

Appendix: Assorted results

Inverting polynomials

Let us begin by noting that, for any $a \neq 1$,

$$\sum_{i=0}^{n} a^i = \frac{1-a^{n+1}}{1-a}, \tag{2.24}$$

which is easy to prove: call

$$S = \sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots + a^n; \tag{2.25}$$

of course

$$a \cdot S = a + a^2 + \cdots + a^{n+1} \tag{2.26}$$

and therefore, by subtracting (2.26) from (2.25), $S(1-a) = 1 - a^{n+1}$, and hence equation (2.24).

If a is a small number ($|a| < 1$), then $a^n \to 0$, and therefore $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$. By setting $a = \alpha L$, you may say that, for $|\alpha| < 1$, the inverse of $(1 - \alpha L)$ is $(1 + \alpha L + \alpha^2 L^2 + \cdots)$, that is

$$(1 - \alpha L)(1 + \alpha L + \alpha^2 L^2 + \cdots) = 1,$$

or, alternatively,

$$\frac{1}{1-\alpha L} = \sum_{i=0}^{\infty} \alpha^i L^i$$

provided that $|\alpha| < 1$. Now consider a n-th degree polynomial $P(x)$:

$$P(x) = \sum_{j=0}^{n} p_j x^j$$

If $P(0) = p_0 = 1$, then $P(x)$ can be written as the product of n first-degree polynomials as follows:[21]

$$P(x) = \prod_{j=1}^{n} \left(1 - \frac{1}{\lambda_j} x\right) \tag{2.27}$$

where the numbers λ_j are the roots of $P(x)$: if $x = \lambda_j$, then $1 - \frac{1}{\lambda_j} x = 0$ and consequently $P(x) = 0$. Therefore, if $P(x)^{-1}$ exists, it must satisfy

$$\frac{1}{P(x)} = \prod_{j=1}^{n} \left(1 - \frac{1}{\lambda_j} x\right)^{-1};$$

but if at least one of the roots λ_j is smaller than 1 in modulus,[22] then $1/|\lambda_j|$ is larger than 1 and, as a consequence, $\left(1 - \frac{1}{\lambda_j} x\right)^{-1}$ does not exist, and neither does $P(x)^{-1}$.

[21] If you don't believe me, google for "Fundamental theorem of algebra".

[22] Warning: the roots may be complex, but this is not particularly important. If z is a complex number of the form $z = a + bi$ (where $i = \sqrt{-1}$), then $|z| = \sqrt{a^2 + b^2}$.

Example 2.8.1 *Consider the polynomial $A(x) = 1 - 1.2x + 0.32x^2$; is it invertible? Let's check its roots:*

$$A(x) = 0 \Longleftrightarrow x = \frac{1.2 \pm \sqrt{1.44 - 1.28}}{0.64} = (1.2 \pm 0.4)/0.64$$

so $\lambda_1 = 2.5$ and $\lambda_2 = 1.25$. Both are larger than 1 in modulus, so the polynomial is invertible. Specifically,

$$A(x) = (1 - \lambda_1^{-1}x)(1 - \lambda_2^{-1}x) = (1 - 0.4x)(1 - 0.8x)$$

and

$$\begin{aligned} \frac{1}{A(x)} &= (1 - 0.4x)^{-1}(1 - 0.8x)^{-1} = (1 + 0.4x + 0.16x^2 + \cdots)(1 + 0.8x + 0.64x^2 + \cdots) \\ &= 1 + 1.2x + 1.12x^2 + 0.96x^3 + 0.7936x^4 + \cdots \end{aligned}$$

In practice: if the sequence a_t is defined as the result of the application of the operator $P(L)$ to the sequence u_t, that is $a_t = P(L)u_t$, then reconstructing the sequence u_t from a_t is only possible if $P(L)$ is invertible. In this case,

$$u_t = P(L)^{-1}a_t = \frac{1}{P(L)}a_t.$$

2.8.1 Appendix: The basics of complex numbers

Many people have asked me for this, so I could not avoid it. In this appendix, you will find some very brief and highly incomplete notions on how complex numbers work and, in particular, the reason why complex roots in an autoregressive polynomial give rise to cyclical phenomena.[23]

Let us start with i, which is defined by the property $i^2 = -1$. The number i is called the *imaginary unit*, largely for historical reasons. Any real multiple of it is called an *imaginary number*. For example, $3i$ is a number whose square is -9.

A complex number is the sum of a real number and an imaginary one, that is, something of the type $z = a + bi$, where a and b are real numbers, which are called the *real part* (often written as $\Re(z)$) and the *imaginary part* (often written as $\Im(z)$), respectively.

Every complex number has its *conjugate*, which is simply the same number with the sign changed in the imaginary part. The conjugation operation is written with a bar. Therefore, if $z = a + bi$, then $\bar{z} = a - bi$. It will be noted that $z \cdot \bar{z}$ is necessarily a positive real number

$$z \cdot \bar{z} = (a + bi)(a - bi) = a^2 - (b^2 i^2) = a^2 + b^2.$$

The absolute value, or modulus, of a complex number is defined as

$$|z| = \sqrt{z \cdot \bar{z}} = \sqrt{a^2 + b^2} = \rho;$$

[23] Anyone wishing to study more deeper can easily find everything they need online; for example, on Wikipedia.

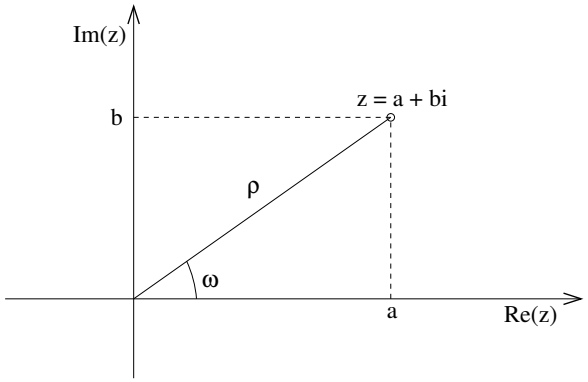
it should be noted that the definition includes that for real numbers as a special case, which occurs when $b = 0$.

In practice, a complex number is an object identified by two real numbers, so one can think of it as a point on a plane. By convention, this plane is conceived with the real part on the X-axis and the imaginary part on the Y-axis. This ensures that, alternatively, the same point can be represented in terms of its distance from the origin ρ and the angle it forms with the abscissa axis ω, that is, as $\rho(\cos\omega + i\sin\omega)$. Evidently, this translates into the two relationships

$$\cos\omega = \frac{\Re(z)}{\rho} \qquad \sin\omega = \frac{\Im(z)}{\rho}.$$

Figure 2.12 should make this quite easy to visualise. I should add that, if $\rho = 1$, the corresponding point lies on a circle of radius 1 centred at the origin, hence the expression "unit circle". Obviously, if $\rho > 1$, the point lies outside this circle, and if $\rho < 1$, it lies inside.

Figure 2.12: Graphical representation of a complex number



For reasons that I shall not go into, the following equality also holds:

$$z = a + bi = \rho\exp(i\omega),$$

where the conjugation operation translates into a change of sign in the argument of the exponential function:

$$\bar{z} = a - bi = \rho\exp(-i\omega).$$

Having two representations is convenient because the former works well for addition:

$$z_1 + z_2 = (a_1 + b_1 i) + (a_2 + b_2 i) = (a_1 + a_2) + (b_1 + b_2)i$$

and the latter for multiplication and exponentiation:

$$z_1 \cdot z_2 = [\rho_1\exp(i\omega_1)][\rho_2\exp(i\omega_2)] = \rho_1\rho_2 e^{i(\omega_1+\omega_2)}$$

$$z^k = \rho^k e^{i\omega k}.$$

Finally, since $\Re(z) = \frac{z+\bar{z}}{2} = \rho\frac{e^{i\omega}+e^{-i\omega}}{2}$, we also have that

$$\cos\omega = \frac{e^{i\omega}+e^{-i\omega}}{2}$$

and there is also a similar formula for the sine function:

$$\sin\omega = \frac{e^{i\omega}-e^{-i\omega}}{2i}.$$

More generally, it can be shown that

$$\frac{e^{i\omega k}+e^{-i\omega k}}{2} = \cos(\omega k) \quad \Longrightarrow \quad \frac{z^k+\bar{z}^k}{2} = \rho^k\cos(\omega k) \tag{2.28}$$

$$\frac{e^{i\omega k}-e^{-i\omega k}}{2i} = \sin(\omega k) \quad \Longrightarrow \quad \frac{z^k-\bar{z}^k}{2i} = \rho^k\sin(\omega k). \tag{2.29}$$

At this point, we have everything we need to understand the link between autoregressive polynomials with complex roots and cyclical phenomena: suppose we have a second-order polynomial in L with complex conjugate roots:

$$A(L) = 1 - \varphi_1 L - \varphi_2 L^2 = (1 - zL)(1 - \bar{z}L).$$

Provided that $|z|$ is less than 1 to ensure invertibility, let us calculate the form of the inverse of $A(L)$, which leads us to the Wold representation. We have

$$A(L)^{-1} = (1 - zL)^{-1}(1 - \bar{z}L)^{-1} = (1 + zL + z^2L^2 + \cdots)(1 + \bar{z}L + \bar{z}^2L^2 + \cdots)$$

As we will now see, not only does this infinite polynomial have entirely real coefficients, but these coefficients oscillate around zero: indeed

$$\begin{aligned} A(L)^{-1} \quad = \quad & 1 + \\ & (z+\bar{z})L + \\ & (z^2 + z\bar{z} + \bar{z}^2)L^2 + \\ & (z^3 + z^2\bar{z} + z\bar{z}^2 + \bar{z}^3)L^3 + \cdots \end{aligned}$$

and, more generally, if we write

$$A(L)^{-1} = 1 + \theta_1 L + \theta_2 L^2 + \cdots,$$

we also get

$$\theta_k = \sum_{j=0}^{k} z^j \bar{z}^{k-j}.$$

The result we are interested in can be found by noting that

$$\theta_{k+1} = z^{k+1} + \bar{z}\theta_k = \bar{z}^{k+1} + z\theta_k,$$

from which

$$z^{k+1} - \bar{z}^{k+1} = (z - \bar{z})\theta_k;$$

using equation (2.29) on can obtain

$$\theta_k = \rho^k \frac{\sin[\omega(k+1)]}{\sin\omega}.$$

In practice, the impulse response function is the product of two functions: $\frac{\sin[\omega(k+1)]}{\sin\omega}$, which is a periodic function in k with a period of $2\pi/\omega$ (and therefore, the smaller ω is, the longer the cycle). The other, ρ^k, attenuates the fluctuations as k increases since, by assumption, $|\rho| < 1$, and thus obviously $\rho^k \to 0$.

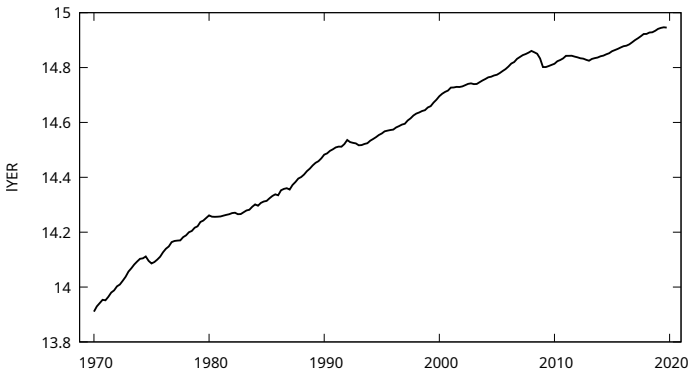
Chapter 3

Integrated processes

3.1 Macroeconomic time series characteristics

All the framework described in the previous chapters hinges on the idea that economic time series can be modelled as realisations of stationary stochastic processes. This, unfortunately, is not the standard situation with macroeconomic time series. Let us observe, just as an example, the trend over time of the logarithm of quarterly real GDP for the Euro area (seasonally adjusted), which is shown in Figure 3.1.

Figure 3.1: log(GDP)



As can be seen, the series exhibits a clear upward trend over time, which precludes the possibility of modelling it via a stationary process. In this case, it would be more appropriate to use a process whose mean changes over time. However, one might think of modelling log GDP as follows: suppose that this series follows a stable deterministic growth trend over time (arguably driven by technical progress and capital accumulation), that we can, at least as a first approximation, represent as a linear function of time. The difference between the actual series and the long-term trend would be the "business cycle", which can be thought of as representable by a stationary process, because economic

cycles are a zero-mean, short-term phenomenon by definition.

We will therefore have a relation of the type:

$$y_t = \alpha + \beta t + u_t,$$

where u_t is some stationary stochastic process with a mean of 0.

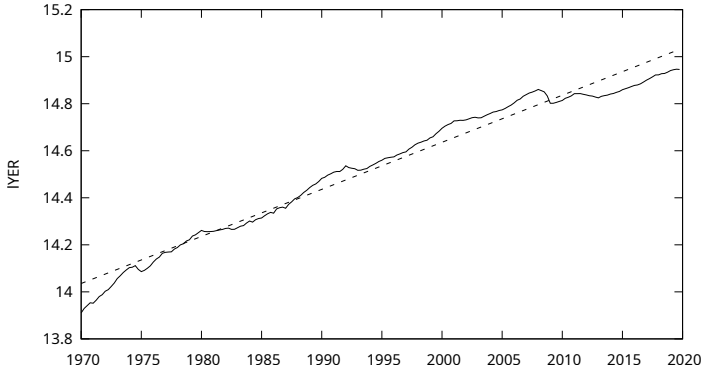
Strictly speaking, the process y_t just described is not a stationary process, since $E(y_t) = \alpha + \beta t$, and therefore the mean of y_t is not constant (for $\beta \neq 0$). However, the non-stationarity of the process is restricted to this aspect. If we consider deviations from the trend, what remains is a stationary process, which can be analysed using the techniques examined a few pages ago. It is for this reason that processes of this type are called trend-stationary processes, or **TS processes**.

If the process y_t is TS, then the parameters α and β can be consistently estimated using OLS, and standard inference applies.[1] Once $\hat{\alpha}$ and $\hat{\beta}$ are obtained, it will be easy to obtain a trend-cycle decomposition of the series: the trend (which will be a linear function of time) will be given by the series $\hat{y}_t$, while the cycle will be given by the residuals

$$\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta} t.$$

In our case, the two time series are shown in Figures 3.2 and 3.3.

Figure 3.2: log(GDP) and deterministic trend



Looking at Figure 3.3, even a beginner's eye will conclude that the residuals are not white noise. However, it is conceivable that one could find an ARMA process of some order that accounts for the persistence characteristics of $\hat{u}_t$.

An alternative option for representing the series would be to consider the series Δy_t. Given that we are working with logarithms, this series (the trend of which is shown in Figure 3.4) can be interpreted as the time series of quarterly growth rates.

[1] This case is a cornerstone of any textbook on asymptotic theory. Those who are interested should look there.

Figure 3.3: Residuals

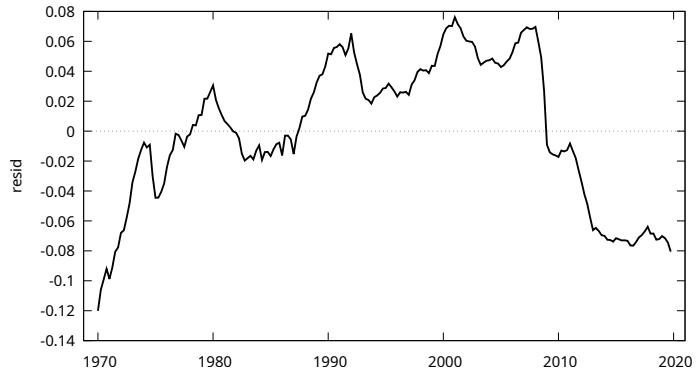
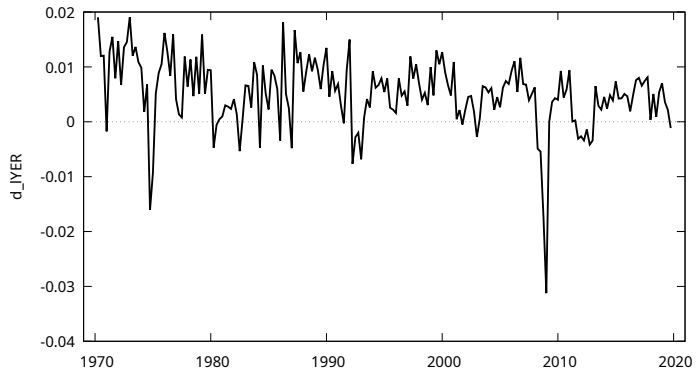


Figure 3.4: Δlog(GDP)



Since the growth rate should (arguably) fluctuate within a band, one could perhaps represent the series Δy_t as a stationary process, possibly with a non-zero mean. Note that, in this context, y_t is a unit root process: if Δy_t were indeed stationary, it would have a Wold representation of the type

$$\Delta y_t = \mu + C(L)\epsilon_t,$$

where μ is the average growth rate. This expression describes y_t as an ARMA process, where $A(L) = 1 - L$; consequently, the value of z for which $A(z) = 0$ is 1. As we know, unit root processes are not stationary, and therefore y_t is not stationary, but its first difference is. In this case, y_t is said to be difference-stationary, or DS. Another frequently used expression is that y_t is an $I(1)$ process (read as **integrated of order one**), meaning that y_t must be differenced once for the result to be stationary.

What happens when differencing an $I(0)$ process? Obviously, the resulting process is stationary, but we know that, by construction, it has a unit root in its MA representation, so it is not invertible.
In this case, some argue that it makes sense to extend the notation just illustrated to negative integers. For example, if u_t is $I(0)$, following this convention one could say that Δu_t is $I(-1)$. Like all conventions, it is good insofar as it is useful. I personally have not yet made up my mind about it, but I feel I should I tell you.

To properly appreciate the different consequences that arise from the choice of modelling a series as a TS process or as a DS process, it is necessary to analyse in detail the characteristics of unit root processes, which is the subject of the following sections.

3.2 Unit root proceses

As I have just said, an $I(1)$ process is a process that is not stationary, but its first difference is. More generally, an $I(d)$ process is defined as a process whose d-th difference is stationary. As far as we are concerned, we will focus only on the cases where d is 0 or 1, although there is a vast body of literature dedicated to more exotic cases.

The most basic $I(1)$ process we consider is the so-called **random walk**. Its definition is simple: y_t is a random walk process if Δy_t is a white noiseprocess. Therefore, $y_t = y_{t-1} + \epsilon_t$, and as a consequence, by repeatedly substituting the past values of y_{t-1}, we obtain

$$y_t = y_{t-n} + \sum_{i=0}^{n-1} \epsilon_{t-i}.$$

HERE

What are the characteristics of a process like this? To begin with, let us make things simpler: suppose that the process started at some remote time, which we call time 0, and that at that date the value of y_t was 0. In this case, the previous expression reduces to

$$y_t = \sum_{i=1}^{t} \epsilon_i. \tag{3.1}$$

Note that this expression can be considered a sort of moving average representation of y_t, where all the coefficients are equal to 1. It is clear that, at any time t, the mean of the process is 0. If it were only for the mean, therefore, the process would be stationary. The variance, however, is not constant, as y_t is the sum of t independent and identical random variables with a variance of (say) σ^2; it follows that the variance of y_t is $t\sigma^2$, and therefore increases over time. From this, and from $Cov(y_t, y_s) = \sigma^2 \min(t, s)$ (proving this is a useful exercise), it follows that y_t is not stationary.

In many respects, it is convenient to consider a random walk as a limiting case of an AR(1) process where the persistence characteristics are so extreme as to alter the qualitative features of the process. In particular, a random

walk is non-stationary, as we have already mentioned. Furthermore, one can consider its associated impulse response function: although the Wold representation does not exist, the impulse response function is perfectly defined as $IRF_k = \frac{\partial y_{t+k}}{\partial \epsilon_t}$. In this case, it is equal to 1 for all values of k, so it is flat and does not decay exponentially as in the stationary case; this means that the effect of a shock at time t persists indefinitely into the future.

This latter characteristic also implies that random walk processes do not share the property of being **mean-reverting** with stationary processes. If a process is mean-reverting, it exhibits a tendency to move preferentially towards its expected value; for a process with a mean of 0, this means that the graph of the process 'frequently' intersects the horizontal axis. More formally, the phrase is usually employed to describe a process whose impulse response function tends asymptotically to 0.

Table 3.1: Stationary AR(1) *versus* random walk

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

	$\lvert\varphi\rvert < 1$	$\varphi = 1$
Variance	Finite	Unlimited
Autocorrelations	$\rho_i = \varphi^i$	$\rho_i = \sqrt{1 - \frac{i}{t}}$
mean-reverting	Yes	No
Memory	Temporary	Permanent

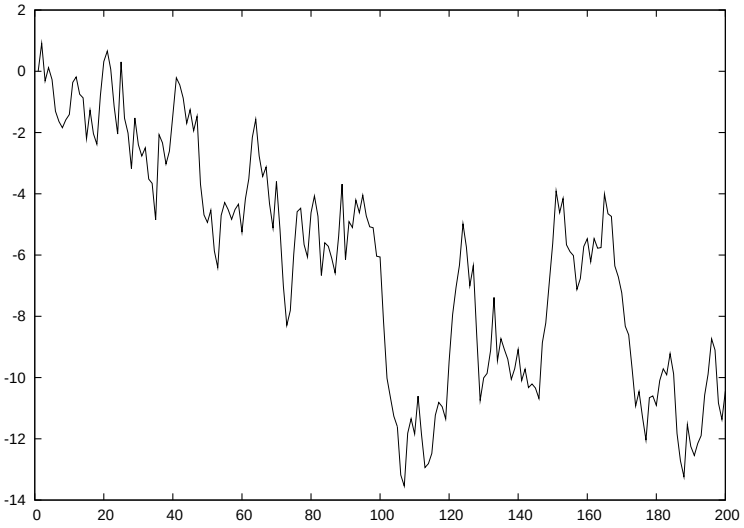
Table 3.1 **[non ho il libro, controlla la Tabella]** (adapted from **?**) highlights the differences between a stationary AR(1) process and a random walk.

It is evident here that the choice between a TS process and a DS process for modelling a variable such as, say, GDP entails significant consequences for analysing the long-run trend of that variable. If GDP were representable as a realisation of a TS process, in the long run the trajectory of the exogenous trend (technology or whatever else) matters for economic growth; a cyclical crisis may have a depressing effect, but this is only **temporary**: the system has an intrinsic tendency to return to its long-run trend.
Conversely, if the most appropriate metaphor for the GDP time series were a DS process, we would have to conclude that there are **permanent** shocks that will never be reabsorbed: the sins (or merits) of the fathers will fall upon the children of the children of the children, and even beyond.

This may well be a disturbing vision, but in certain cases, it can be entirely appropriate: who said that technology marches at an exogenous and fixed rate? The Early Middle Ages lasted a long time, but technological progress was noticeably slow.. And besides, once something has been invented, it cannot be disinvented (excluding barbarian invasions or nuclear wars): once technological progress has occurred, it remains for all those who follow. The debate on this topic is rich and flourishing, but I am satisfied to have given you the general idea.

What does a random walk look like 'by eye'? Let us recall Figure 2.3(c) on page 25 and compare it with Figure 3.5. As my astute readers will have already guessed, Figure 3.5 depicts a random walk whose increments (the ϵ_t) are nothing other than the white noise used to generate the series shown in Figure 2.3 (c). In practice, the only difference between Figure 2.3(c) and Figure 3.5 is the autoregressive coefficient, which is equal to 0.9 in the first case and

Figure 3.5: Random walk



equal to 1 in the second. Note the increase in the series' persistence induced by the unit root.

A characteristic aspect of random walk processes is that the absence of mean reversion causes periods — even very long ones — in which the series exhibits a rather marked upward or downward trend. For instance, an observer unaware that the series plotted in Figure 3.5 is purely a product of chance might easily be tempted to comment on the clear' downward trend of the initial segment, pointing to a crisis' around observation 100 followed by a 'recovery' (and I could go on). It is due to this characteristic that when discussing a random walk, or more generally $I(1)$ processes, one often speaks of **stochastic** trends, as opposed to **deterministic** trends, which are simple functions of time.

Clearly, no one rules out the possibility that there may indeed be deterministic trends combined **[overlaid?]** on purely stochastic ones. This occurs, for example, in so-called random walk processes with *drift*. These processes are simply processes for which one has

$$\Delta y_t = \mu + \epsilon_t$$

and therefore y_t is a random walk combined with a linear function of time. If the *drift*, that is, the constant μ, is positive, we will have a process that tends to rise, but with fluctuations around this trend that become progressively more marked over time. Why this occurs can be clearly seen by considering (3.1), which in this case modifies into

$$y_t = \sum_{i=0}^{t} \epsilon_i + \mu \cdot t, \tag{3.2}$$

where the second term of the sum is nothing other than a linear trend with a

slope of μ; in more general cases, one has things of the type

$$\Delta y_t = d_t + \epsilon_t,$$

where d_t is some deterministic function of time: polynomial trends of various kinds (that is, things of the type $d_t = \delta_0 + \delta_1 t + \cdots$), seasonal dummies, and so on. In this case, it is interesting to note that a result which encompasses the case of a simple *drift* as a special case holds: if d_t is a polynomial in t of degree k, then a polynomial in t of degree $k+1$ will be present within y_t. In the case of the drift, indeed, $d_t = \mu$, that is, a polynomial of order 0; similarly, it can be shown that a linear trend included in Δy_t produces a quadratic trend in y_t, and so forth. This occurs simply because, if μ_t is a polynomial in t of order p, then $\Delta\mu_t$ is a polynomial of order $p-1$ (verify this for yourself).**[non sarebbe meglio usare p invece che k anche 2 righe sopra?]**

The random walk extends rather painlessly to the case where the increments of the process are not a white noise, but more generally any stationary stochastic process: representing the latter as an ARMA process, we will have a representation of the type

$$A(L)\Delta y_t = C(L)\epsilon_t,$$

where I omit any deterministic component to keep the notation simple.

In these cases, we speak generically of $I(1)$ processes; as we will see shortly, processes of this type share many of the salient characteristics of the special case of the random walk, such as not possessing a constant second moment, not being mean-reverting, and possessing infinite memory. However, things become more complicated, because the presence of the polynomials $A(L)$ and $C(L)$ gives the process a short-run memory, in addition to its long-run memory.

Indeed, although the distinction between integrated and stationary processes is perfectly defined and absolutely unequivocal, when finite realisations of $I(1)$ processes are observed, situations can arise where the differences become more subtle. This is because the short-run memory overlaps with the long-run memory, creating curious effects: let us consider, for example, two processes defined by

$$\begin{aligned} x_t &= 0.99999x_{t-1} + \epsilon_t \\ y_t &= y_{t-1} + \epsilon_t - 0.99999\epsilon_{t-1} \end{aligned}$$

Strictly speaking, x_t is $I(0)$, while y_t is $I(1)$. However, in the case of x_t, the root of the polynomial $A(L)$ is not 1, but it is close to it; for all practical purposes, in finite samples a realisation of x_t is completely indistinguishable from that of a random walk. Conversely, y_t can be written in the form

$$(1-L)y_t = (1-0.99999L)\epsilon_t = A(L)y_t = C(L)\epsilon_t$$

thus making it evident that the polynomials $A(z)$ and $C(z)$ are very close to being able to be 'cancelled out'; from a practical point of view, any realisation of y_t exhibits no appreciable differences from a white noise.

3.3 Beveridge-Nelson decomposition

Based on what we have seen in the previous section, a random walk is a special case of an I(1) process. A process of the type $y_t = x_t + u_t$ (where x_t is a random walk and u_t is any I(0) process) is an integrated process, because it is not stationary unless it is differencited, but it is not a random walk, because Δy_t is not a white noise.

If it were possible to write an $I(1)$ process in this form, there would be an immediate advantage: one could attribute the characteristics of non-stationarity on the one hand, and short-run persistence on the other, to two distinct components. In this sense, y_t can be thought of as being split into two components: a permanent, or long-run, component given by x_t, and a transitory, or short-run, component given by u_t. Since u_t is by definition a process with a mean of 0, y_t can be thought of as a process that fluctuates around x_t, without these fluctuations ever becoming too pronounced.

The interesting thing is that *any* $I(1)$ process can be thought of as the sum of a random walk and an $I(0)$ process. This decomposition is known as the **Beveridge-Nelson decomposition**, sometimes also called the **BN decomposition**. The BN decomposition can be illustrated starting from an almost trivial property of polynomials: given a polynomial $C(z)$ of order q, it is always possible to find a polynomial $C^*(z)$ of order $q-1$ such that

$$C(z) = C(1) + C^*(z)(1-z).$$

The proof is not difficult: obviously, $D(z) = C(z) - C(1)$ is still a polynomial of order q, since $C(1)$ is a constant (the sum of the coefficients of $C(z)$).

However, it follows from the definition that $D(1) = 0$, and therefore 1 is a root of the polynomial $D(z)$. It can then also be written as $D(z) = C^*(z)(1-z)$, where $C^*(z)$ is a polynomial of degree $q-1$. In other words, the polynomial $C^*(z)$ is defined by

$$C^*(z) = \frac{C(z) - C(1)}{1-z},$$

hence the expression in the text. I shall omit the explanation as to why, but proving that

$$c_i^* = -\sum_{j=i+1}^{q} c_j$$

can be a nice little exercise.

Let us now take an arbitrary $I(1)$ process and call it y_t. The process Δy_t is consequently an $I(0)$ process, and therefore it must have a Wold representation that we can write in this way:

$$\Delta y_t = C(L)\epsilon_t.$$

Applying the polynomial decomposition just illustrated to $C(L)$, we can also write

$$\Delta y_t = [C(1) + C^*(L)(1-L)]\,\epsilon_t = C(1)\epsilon_t + C^*(L)\Delta\epsilon_t. \tag{3.3}$$

If we define a process μ_t such that $\Delta\mu_t = \epsilon_t$ holds (that is, a random walk whose increments are given by ϵ_t), we get

$$y_t = C(1)\mu_t + C^*(L)\epsilon_t = P_t + T_t, \tag{3.4}$$

where $P_t = C(1)\mu_t$ is a random walk that we call the permanent component, and $T_t = C^*(L)\epsilon_t$ is an $I(0)$ process that we call the transitory component.

Example 3.3.1 (Simple) *Let us consider an integrated process of order 1, y_t, for which the following holds:*

$$\Delta y_t = \epsilon_t + 0.5\epsilon_{t-1} = (1 + 0.5L)\epsilon_t = C(L)\epsilon_t,$$

where ϵ_t is a white noise. Since

$$C(1) = 1.5 \qquad C^*(L) = -0.5,$$

one gets

$$y_t = 1.5\mu_t - 0.5\epsilon_t.$$

Example 3.3.2 (More difficult) *Suppose that Δy_t can be represented as an ARMA(1,1) process*

$$(1 - \varphi L)\Delta y_t = (1 + \theta L)\epsilon_t,$$

therefore $C(L) = \frac{1+\theta L}{1-\varphi L}$.

$C(1)$ is easy to calculate, and is equal to $\frac{1+\theta}{1-\varphi}$. The calculation of $C^*(L)$ is a bit longer but no more difficult; one can show that

$$C^*(L) = -\frac{\varphi + \theta}{1 - \varphi}(1 - \varphi L)^{-1}.$$

The final result is

$$\begin{aligned} y_t &= P_t + T_t \\ P_t &= \frac{1+\theta}{1-\varphi}\mu_t \\ T_t &= -\frac{\varphi+\theta}{1-\varphi}(1-\varphi L)^{-1}\epsilon_t. \end{aligned}$$

Note that T_t is an autoregressive process of order 1, which is the more persistent the larger $|\varphi|$ is. Consequently, y_t can be represented as a random walk plus a stationary AR(1) process fluctuating around it.

An interesting interpretation of the quantity $C(1)$ is as a measure of the persistence of a given process, since it measures the fraction of the shock that remains in the process after an 'infinite' time. It is possible to verify that, by applying the decomposition described here to a stationary process, $C(1) = 0$, whereas $C(1) \neq 0$ in the case of $I(1)$ processes. Intuitively, this interpretation can also be motivated by observing that $C(1)$ is a coefficient that determines the weight of the random walk on the process. In the case of the $I(1)$ process examined at the end of the previous section, which looked so much like a white noise, the coefficient $C(1)$ turns out to be just 0.00001.

The utility of the BN decomposition is twofold: from a practical point of view, it is a tool that is frequently used in macroeconometrics when it comes to separating trend and cycle in a time series. In short, given a time

series that we are interested in decomposing into trend and cycle, an ARMA model is estimated on the first differences, after which the BN decomposition is applied based on the estimated parameters. The BN decomposition is not the only tool to achieve this purpose, and it is not immune to criticism[2], but on this, as usual, I refer the reader to the specialised literature.

The other use of the BN decomposition is theoretical. Under a different name (*martingale decomposition*), it plays a fundamental role in the probabilistic literature on stochastic processes when analysing certain asymptotic properties. While this is not of interest to us here, we will make systematic use of the BN decomposition in the analysis of cointegrated systems, which we will discuss later on.

3.4 Unit root tests

Integrated processes, as seen so far, possess characteristics that make them very interesting, both from a formal point of view (because they represent an example of non-stationary processes) and from a practical point of view (because their realisations are remarkably similar to the time series we usually encounter in macroeconomics).

We have not, however, yet examined the consequences of the non-stationarity of this type of process for the possibility of making inference on their realisations. Recall that, up until now, we have always assumed the stationarity of the processes for which we were interested in making inference. In the case of $I(1)$ processes, things become more complex.

Let us begin with a triviality: if y_t is $I(1)$, then Δy_t is $I(0)$ by definition, and therefore the entire wealth of knowledge accumulated so far on estimating the parameters that characterise stationary processes can be recycled without any problems by estimating a model of the type

$$A(L)\Delta y_t = C(L)\epsilon_t,$$

and therefore we will model a growth rate instead of (the logarithm of) GDP, the inflation rate instead of (the logarithm of) the price index, and so forth. It is common to refer to this type of model as ARIMA models, that is, integrated ARMA, in the statistical literature (which, on this subject, is boundless).

A strategy of this type, however, assumes that one knows exactly whether a time series is integrated or stationary[3]. In the absence of supernatural revelations, this is usually something that is not known; or to put it better, it is

[2] One, for example, is: where is it written that the long-run component must necessarily be a random walk, rather than some other type of I(1) process?

[3] This is a rather coarse simplification: apart from the fact that, by using concepts slightly more complex than those I am discussing here, one can provide examples of processes that are neither $I(0)$ nor $I(1)$, recall that integration is not a characteristic of the time series itself, but rather of the stochastic process we adopt to provide a statistical representation of it.

To be rigorous, we should say "...whether the observed time series is better represented by a stationary stochastic process or one that is integrated of order 1", and the issue, at this point, could shift to the meaning of "better". Subtleties of this kind are, moreover, completely ignored by the vast majority of contemporary macroeconomics, and therefore it is not worth losing sleep over them.

almost never possible to establish *a priori* whether a given time series can be better represented by an $I(0)$ or an $I(1)$ process.

This decision, however, can be made on the basis of the data themselves. A first idea could simply be to observe the plot of the trend of the series over time. If a process is stationary, it cannot exhibit a regular upward or downward trend, and therefore one might consider a process that oscillates around a constant value to be stationary, and non-stationary otherwise.

This rule, This rule, which at a glance and with a bit of experience is not entirely to be dismissed, is however too simplistic, and this for at least three reasons: first, because such a judgement is rather subjective and difficult to formalise; second, because it may well be that a process is stationary around a deterministic trend (as seen a few pages ago); finally, because there is also the possibility that an actually $I(1)$ process gives rise to realisations that do not exhibit a particularly marked upward or downward tendency. For all these reasons, a less arbitrary and more reliable decision rule is required. Decision rules of this type are known as **unit root tests**.

There is more than one unit root test[4]. However, the most widely used ones stem from a common approach, which I will outline briefly. Let us start with a first-order autoregressive process, which we call y_t:

$$y_t = \varphi y_{t-1} + u_t. \tag{3.5}$$

By definition, the following relation must hold:

$$\Delta y_t = \rho y_{t-1} + u_t, \tag{3.6}$$

where u_t is a white noise and $\rho = \varphi - 1$, so that the process is stationary only if $\rho < 0$. Conversely, if $\rho = 0$, we are in the presence of an $I(1)$ process.

Given that the equation just written looks suspiciously like a regression model, one might conjecture that a unit root test is nothing other than a t-test of the significance of the parameter ρ, that is, a test based on the statistic

$$t_\rho = \frac{\hat{\rho}}{\sqrt{\widehat{\mathrm{Var}(\hat{\rho})}}}, \tag{3.7}$$

where the 'hats' indicate, as usual, the OLS estimates. In this case, the test would have the non-stationarity of the process as the null hypothesis (if $\rho = 0$, then $\varphi = 1$), and stationarity as the alternative hypothesis ($\rho < 0$, and therefore $\varphi < 1$). The conjecture is indeed correct, although there are at least three observations to be made.

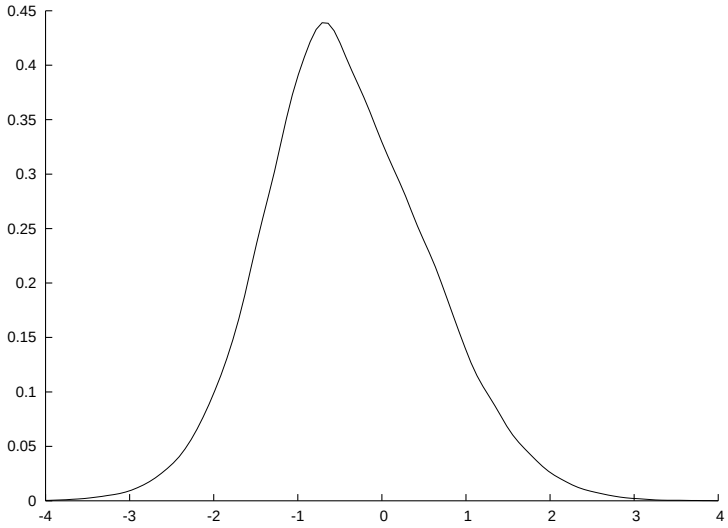
3.4.1 Distribution of the test statistic

The first observation concerns the fact that, under the null hypothesis, the distribution of the t-test for the parameter ρ being equal to zero is neither a Student's t distribution in finite samples, as occurs in the classical linear

[4] A faint understatement. There is a sea of them. Indeed, some have argued that there are even too many unit root tests. Anyone particularly interested in this topic cannot escape an examination of the relevant literature, which is vast and complex.

model, nor asymptotically Gaussian, as is the case for stationary processes. Its asymptotic distribution is instead a somewhat odd distribution, for which there is no compact expression for either the probability density function or the cumulative distribution function[5].

Figure 3.6: Density function of the DF test



A numerical estimate of the density function of this statistic is shown in Figure 3.6. The quantiles of this distribution must be calculated using numerical simulations, and the first to do so were Dickey and Fuller in 1976, which is why this distribution is sometimes called the DF, or Dickey-Fuller, distribution. For the same reason, the test is also known as the **DF test**. The immediate consequence of this is that, to accept or reject the null hypothesis, one must consult specific tables, which are however always included in modern econometrics textbooks and, even more so, in econometric software packages.

3.4.2 Short-run persistence

A second observation concerns that, in general, y_t is not necessarily a random walk (that is, u_t is not necessarily a white noise). In other words, it is possible that Δy_t itself exhibits persistence characteristics, even if of a short-run nature. In these cases, which are indeed the most commonly encountered in practice, it is necessary to ensure that the distribution of the test is not affected by the short-run memory contained in u_t. One of the most widespread approaches is to assume that the memory characteristics of Δy_t can be satisfactorily approximated by an AR(p) process, by choosing a sufficiently large value of p.

[5]This distribution is defined in terms of integrals of Brownian motions. A Brownian motion, or Wiener process, is a continuous-time stochastic process of which I will not provide the definition, but which can essentially be thought of as a random walk where the interval between observations is infinitesimal.

Let us look at an example. Suppose we start from a formulation in levels analogous to (3.5):

$$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \epsilon_t. \tag{3.8}$$

In this case, suppose that ϵ_t is a white noise, that is, any short-run persistence is completely captured by the autoregressive component.

Exploiting the property $y_{t-k} = y_{t-1} - \sum_{i=1}^{k-1} \Delta y_{t-i}$ (it is easy to convince oneself of this by checking the case $k = 2$), equation (3.8) can be reparameterised as follows [6]:

$$\Delta y_t = \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p+1} + u_t, \tag{3.9}$$

where $\rho = (\varphi_1 + \cdots + \varphi_p) - 1$ and $\gamma_i = -(\varphi_{i+1} + \cdots + \varphi_p)$. In this case, the test is named the **ADF test** (Augmented Dickey–Fuller), and it is the t-test of the significance of the parameter ρ in regression (3.9).

If the chosen value of p is sufficiently high, and therefore the correction is effective, the distribution of the ADF test is the same as that of the DF test. What does "sufficiently high" mean? It simply means that u_t must be, at least for all practical purposes, a white noise. In practice, the same selection criteria for p are often used as for the analogous problem in the stationary framework, which I discussed in Section 2.7; that is, p is chosen so as to minimise criteria of the Akaike or Schwarz type. A similar way of resolving this problem was proposed by Phillips and Perron, and the so-called Phillips-Perron test (familiarly known as the **PP test**) is by now available alongside the ADF test in many software packages.

3.4.3 Deterministic component

Finally, it should be mentioned that the distribution of the test (whether of the ADF or PP type) is not invariant to the deterministic component included in the regression. So far, we have examined the case of a random walk without drift. In the event that a drift is actually present in the process that generated the series under examination, it must also be included in the regression used to compute the test. But how do we know whether the drift is present or not? The problem is that it is unknown. Consequently, the best approach is to include it, and therefore estimate a regression of the type

$$\Delta y_t = \mu + \rho y_{t-1} + \varphi_1 \Delta y_{t-1} + \cdots + \varphi_p \Delta y_{t-p} + u_t \tag{3.10}$$

where, at most, μ will be equal to 0 in the event that the drift is not present.

[6] Particularly motivated readers can verify that here we are doing nothing other than applying the BN decomposition: indeed, by writing

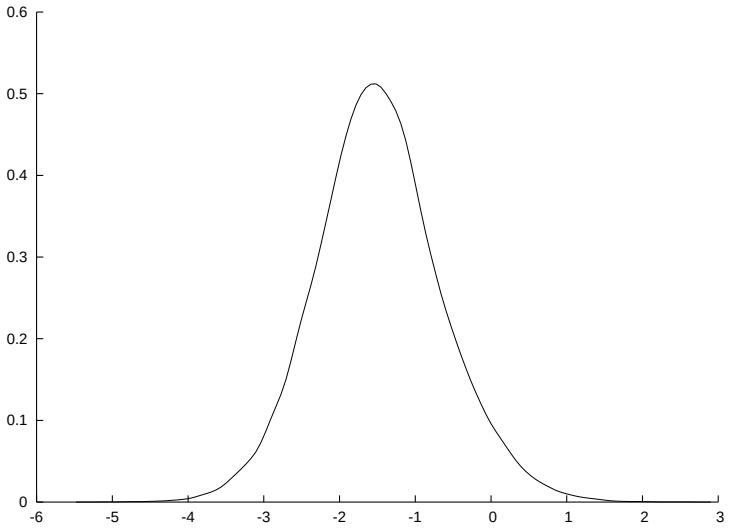
$$y_t = B(L) y_{t-1} + u_t$$

one can proceed by noting that

$$\Delta y_t = [B(L) - 1] y_{t-1} + u_t.$$

The result in the text follows by applying the BN decomposition to the polynomial $H(L) = B(L) - 1$.

Figure 3.7: Density function of the DF test with an intercept



Unfortunately, it is possible to show that, in this case, the asymptotic distribution of the test that the parameter equals zero is different from the one seen previously. As if that were not enough, there are actually two relevant distributions: one — which is non-standard, also tabulated, and shown in Figure 3.7 — in the event that the true value of μ is 0; whereas if $\mu \neq 0$, it turns out that the asymptotic distribution of the test is (perhaps surprisingly) normal, although the sample size must be very large for the approximation to be satisfactory.

As can be seen, the matter becomes a bit complicated even at this stage; furthermore, if one analyses the case where a linear deterministic trend is also added to (3.10), yet another distribution is obtained. This multiplicity of situations is perhaps one of the most puzzling aspects when first approaching unit root tests. In truth, it is possible to navigate through it, but with a great deal of patience and by making a series of distinctions; however, on this point I refer the reader to the specialised literature, as I consider the task of a popular introduction that I set myself here to be complete. A very similar problem will return in Chapter **??**, but we shall speak of it in due course.

3.4.4 Alternative tests

The ADF test assumes the existence of a unit root as the null hypothesis, and so do its variants, such as the Phillips-Perron test; conversely, there are tests that start from the null hypothesis of stationarity. The most prominent of the latter is the so-called **KPSS test**, the basic intuition of which I will explain. If y_t were stationary around a deterministic trend, then a regression of the type

$$y_t = \beta_0 + \beta_1 \cdot t + u_t$$

should produce $I(0)$ residuals. Having run the regression, one takes the OLS residuals and cumulates them, producing a new series $S_t = \frac{1}{T}\sum_{s=1}^{t} \hat{u}_s$; under the null hypothesis, this time series can be thought of (for very large samples) as a realisation of a rather strange process[7]. This is because, by construction, it holds that not only $S_0 = 0$, but also $S_T = 0$. In this case, it can be shown that the sum of squares of S_t (appropriately normalised) converges in distribution to a random variable that is always the same for any stationary process. If, on the other hand, y_t is non-stationary, the statistic diverges. Consequently, the acceptance interval runs from 0 to a certain critical value which, in this case too, has been tabulated.

The expression "appropriately normalised" that I used in the previous paragraph is deliberately a bit vague: indeed, it can be shown that the essential ingredient of this normalisation is the long-run variance of $\hat{u}_t$. The latter is defined as the sum of all its autocovariances (from minus to plus infinity). Frequently, this quantity is estimated non-parametrically by means of the statistic $\hat{\omega}^2$, which is defined as

$$\hat{\omega}^2(m) = T^{-1}\sum_{t=m}^{T-m}\left[\sum_{i=-m}^{m} w_i \hat{u}_t \hat{u}_{t-i}\right],$$

where m is known as the *window size* and the w_i terms are the so-called *Bartlett weights*, defined by $w_i = 1 - \frac{|i|}{m+1}$. It can be shown that, for a sufficiently large m, $\hat{\omega}^2(m)$ provides a consistent estimate of the long-run variance. The main problem is the choice of m, and here precise rules do not exist: asymptotic theory only says that m must be proportional to $T^{1/3}$, which in practice amounts to a license to do as one pleases. The advice I offer is to try various values of m and see when the statistic becomes stable.

The test can also be carried out without a trend, so that the $\hat{u}_s$ are simply the deviations of y_t from its mean. Evidently, in this case, the null hypothesis is that the process is stationary *tout court*. The critical values change, but these too have been tabulated.

In my view, it is always good practice to attempt to test a series' integration order in both ways. Usually, the indications coincide, in the sense that if the KPSS test does not reject, the ADF test rejects, and vice versa. However, it is not uncommon for these tests to yield inconsistent indications; that is, it frequently happens that both the ADF test and the KPSS test reject (or accept) their respective null hypotheses.

Finally, I mention that some now consider the very idea of testing hypotheses on the integration order to be outdated in a multivariate context. If we are dealing with more than one time series, one can, of course, proceed with a battery of ADF tests or similar on each of them. However, it is perhaps more intelligent to start directly from a multivariate representation (which I will discuss in Chapter **??**), which leads to the so-called Johansen test (which I will discuss in Chapter **??**).

3.4.5 Using your brain

A short comment on unit root tests: it happens very often that, when applying a unit root test to a time series which, reasonably, should fluctuate within a more or less wide band, it is not possible to reject the unit root hypothesis. This occurs, for example, almost always with unemployment rates,

[7] It is called a *Brownian bridge*, if you must know.

inflation rates, or interest rates (real or nominal). It is common, at this point, for someone to raise their hand and say: "How is it possible for the yield on Treasury bills to be $I(1)$? It was already at 12% at the time of the Babylonians!"

Two answers can be given to this objection. One is to demonstrate one's dogmatic devotion to the cult of the p-value by saying: "The test just comes out this way! What can I do about it?"; another, which in my view is more intelligent, is to point out that *within the available sample* the yield on Treasury bills clearly exhibits a degree of persistence such that it is better, from the standpoint of fitting the data, to think of it as a realisation of an $I(1)$ process rather than an $I(0)$ one.

We are not saying that the series is $I(1)$: in truth, provided that it even makes sense to think of our interest rate time series as a realisation of some stochastic process, heaven only knows what process it is; we are merely choosing, from within a limited class of processes (ARIMA models), the most appropriate parameterisation to describe the data. If we were then to have observations spanning thousands of years, I suspect that the most adequate process to represent the trajectory of interest rates over time from Hammurabi onwards would be an $I(0)$ one, but I do not believe we will ever be in a position to establish this.

It is a problem of *data representation*: with a unit root test, we are not truly deciding whether the process is $I(1)$ or $I(0)$. We are merely deciding whether it is more convenient to represent the data we have with a stationary or an integrated process.

A metaphor that I find compelling is that of the curvature of the Earth. For many centuries, it was believed that the Earth was flat simply because there was no reason to think it round: the Earth's curvature becomes a problem only when dealing with large distances, so much so that, for instance, it must be taken into account when calculating the routes of transoceanic ships. However, if one needs to build a house, or even a stadium, a car park, or a shopping centre, the scale of the problem is such that the curvature of the globe becomes negligible (and indeed, engineers neglect it without houses collapsing because of it).

In the same way, a stationary process can have a degree of persistence such that its realisations become "evidently" stationary only after a very long time. A unit root test conducted on a sample that is not so large will probably does not reject the null hypothesis. In this case, the test is simply telling us that it might be better, within our sample, to difference the series.

3.4.6 An example

Let us subject the time series seen at the beginning of the chapter to a unit root test, namely the logarithm of Italian GDP at constant prices (see Figure 3.1).

To apply the ADF test, one must choose the lag order most suitable for representing the series as an autoregressive process. There are many ways to

do this. One of the simplest is to run regressions of the type

$$y_t = b_t + \sum_{i=1}^{p} \varphi_i y_{t-i} + \epsilon_t$$

for various values of p and choose the p that minimises the Schwarz criterion; b_t is the deterministic component we deem most appropriate. In our case, we try both a constant ($b_t = a$) and a constant plus trend ($b_t = a + b \cdot t$). **[il puntino...]** The results are summarised in the following table:

p	C	C+T
1	-881.341	-877.246
2	-886.637	-884.346
3	-874.406	-872.165
4	-866.458	-863.827
5	-856.408	-853.772
6	-846.929	-844.136

As can be seen, the BIC criterion is minimised, in both cases, by the choice $p = 2$. We therefore proceed to the regression on the reparameterised model; in the case of the model with only a constant, we have:

$$y_t = a + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \epsilon_t \Longrightarrow \Delta y_t = a + \varphi y_{t-1} + \gamma_1 \Delta y_{t-1} + \epsilon_t$$

where $\varphi = \varphi_1 + \varphi_2 - 1 = -A(1)$ e $\gamma_1 = -\varphi_2$. The ADF test consists precisely in testing whether φ equals zero. I report the regression in full here:

Parameter	Coefficient	Std. Error	t-stat.
a	0.07992	0.03877	2.0614
ϕ	-0.00623	0.00316	-1.9697
γ_1	0.35290	0.08172	4.3183

The test statistic is the t-statistic relating to the parameter ϕ, namely -1.9697. Comparing it with the appropriate tables (which are *not* those of the normal or the t distribution), it turns out that the 95% critical value in this case is approximately -2.89, so we are still within the acceptance **[non rejection?]** region of the null hypothesis of non-stationarity. Some software packages do even better, calculating the p-value of the test directly, which in this case turns out to be 30.05%. Repeating the test for the model with a trend, things do not change much:

Parameter	Coefficient	Std. Error	t-stat.
a	0.42462	0.21979	1.9319
b	0.00018	0.00011	1.5930
ϕ	-0.03541	0.01858	-1.9052
γ_1	0.37087	0.08202	4.5217

In this case too, the test statistic (-1.9052) turns out to be higher than the critical value (approximately -3.45), so the null hypothesis is not rejected;

the same conclusion, of course, would have been obtained if we had used a software package that computes the p-value, given that in this case it is equal to 65.17%.

The same procedure, applied to Δy_t instead of y_t, and therefore to the growth rates of GDP (see Figure 3.4), produces instead a clear rejection of the null hypothesis. I do not report the results here for the sake of brevity; trust me on this.

Conclusion? Here, things seem rather clear: y_t has a unit root, Δy_t does not, so we conclude that the most adequate process to represent the series is $I(1)$.

3.5 Spurious regression

In the brief and incomplete presentation of unit root tests given in the previous section, it will have caught the reader's eye that, when making inference with integrated processes, many of the comfortable certainties that accompany us in the world of stationarity give way to unusual results. This state of the art is even more striking when one analyses the phenomenon of **spurious regression**.

Let us take two processes, y_t and x_t, defined as follows:

$$\begin{cases} y_t = y_{t-1} + \eta_t \\ x_t = x_{t-1} + \epsilon_t, \end{cases} \tag{3.11}$$

where η_t and ϵ_t are two mutually independent white noise. It is evident that y_t and x_t are two random walk that have very little to do with one another. One would expect to find no trace of a statistically significant relationship between y_t and x_t. And indeed, so it is, but matters are not quite that simple.

If one were to attempt to analyse the possible presence of a relationship between x_t and y_t by setting up a linear regression model, one would end up estimating an equation of the type

$$y_t = \alpha + \beta x_t + u_t. \tag{3.12}$$

At first glance, one might think that the absence of a relationship between y_t and x_t implies

1. that the R^2 index is "low";
2. that the OLS estimator of β converges in probability to 0;
3. that a t-test of the significance of β, at least in large samples, falls within the acceptance **[non rejection?]** region of the null hypothesis given by the standard normal distribution tables; put simply, that the t-statistic relating to the coefficient β is between -2 and 2 in 19 out of 20 cases.

Well, none of these three things occurs in the case under examination; on the contrary:

1. the R^2 index converges in distribution to a non-degenerate random variable;

2. the OLS estimator of β converges in distribution to a random variable;
3. a t-test of the significance of β leads, using the critical values of the standard normal distribution, to the rejection of the null hypothesis, all the more frequently the larger the sample size is (!).

It is evident that, on the basis of such a regression, an incautious researcher who does not consider the problem of the variables' integration order could "discover" relationships between variables that are absolutely non-existent in reality: hence the expression 'spurious regression'[8].

Table 3.2: Spurious regression: Monte Carlo experiment

Sample size	Rejection rate
20	47.7%
50	66.4%
100	75.9%
200	83.5%
1000	92.5%

40000 simulations for each sample size

To understand this better, take a look at Table 3.2, which highlights the results of a small Monte Carlo experiment: I have simulated a system identical to that presented in (3.11), with $E(\epsilon_t^2) = E(\eta_t^2) = 1$ for various sample sizes. Having run a regression of y_t on a constant and on x_t (like the one presented in (3.12)), I calculated the value of the t-test for the significance of β, subsequently comparing it with the 95% critical value of Student's t distribution. Following this procedure leads to a rejection rate which, as can be seen, is not only high enough to be embarrassing, but increases as the sample size grows.

Upon closer inspection, however, these results should not be all that surprising. Indeed, for $\beta = 0$, equation (3.12) reduces to $y_t = \alpha + u_t$; if y_t is $I(1)$, it is one of two things: either u_t is $I(0)$, but in this case the equation is contradictory, or u_t is also $I(1)$, and then all the limit theorems go out of the window. In other words, there is no value of β that makes equation (3.12) a correct description of the data; the value $\beta = 0$ is no more right or wrong than $\beta = 1$ or $\beta = -1$; the "true" β *does not exist*. Indeed, an examination of the equation reveals that there is no mechanism to account for the persistence of y_t; the error term must — necessarily — take charge of the latter.

In practice, this situation becomes evident by observing that the OLS estimation of equation (3.12) dumps all the persistence of y_t onto the residuals $\hat{u}_t$, which turn out to be highly autocorrelated. Indeed, it can be shown that, in the presence of this phenomenon, the Durbin-Watson statistic[9] converges in probability to 0. I will go further: a rough but effective rule of thumb to

[8] The phenomenon had already been observed in the 1920s. It was only during the 1970s and 1980s, however, that it was brought to general attention (thanks to Granger and Newbold) and analysed in depth (thanks to P. C. B. Phillips).

[9] Recall that the Durbin–Watson statistic is a venerable statistic devised in the prehistory of

signal whether a regression is spurious or not is to compare the R^2 index with the value of the DW statistic. If the former is greater than the latter, there is reason to be suspicious (although, it must be said, this should not be taken as a proper test; it is simply a heuristic suggestion contained in the original paper by Granger and Newbold).

Now, if regression is a tool that can yield misleading results if used with realisations of $I(1)$ processes, which macroeconomic time series typically strongly resemble, does this mean that regression cannot be used on macroeconomic time series? Not necessarily.

First of all, it should be noted that a large part of the seemingly paradoxical aspects just outlined is due to the fact that there is no value of β compatible with a correct description of the data, as I mentioned a moment ago. If we had estimated something of the type

$$y_t = \alpha + \varphi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

we would have found that

$$\begin{pmatrix} \hat{\varphi} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \xrightarrow{\text{p}} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

that is, the OLS estimates converge to the 'true' values of the parameters. A correct dynamic specification (that is, a specification that allows the disturbances to resemble a white noise) is a major step forward. We will discuss this result, and others, in greater detail in Section **??**. The most important thing to say, however, is that a regression between integrated variables *can* make sense, and indeed under certain circumstances represents a quick but effective way of measuring statistical relationships to which it is possible to attribute a very precise meaning from the standpoint of the economic theory underlying the estimated model. This occurs when the variables on the left- and right-hand side of the equals sign are **cointegrated**. What cointegration is, we shall see in Chapter **??**.

econometrics to check whether there was first-order autocorrelation in OLS residuals. This statistic was also, often, recklessly used as a test statistic. Today we are in the twenty-first century and we have more sophisticated tools to do the same thing better, but software packages still report it.

Index