Time Series Analysis for Economists

Zem Wang

1/1/24

Table of contents

Preface

Empirical macroeconomics often involves dealing with time series data, such as GDP, inflation, and interest rates, which differ from the techniques employed in cross-sectional studies. This book aims to bridge the gap between introductory time series textbooks and theoretical econometrics. In empirical research, a basic understanding of the topic is often insufficient. As it could be potentially dangerous to apply advanced techniques without understanding the underlying theories, the limitations and drawbacks of the techniques. On the other hand, for practical purposes, an in-depth study of advanced econometric theories would be excessive. For example, introductory textbooks would emphasize the risk of applying OLS to non-stationary time series, as it might lead to spurious regression. Students usually accept this as a rule of thumb without fully understanding the reasons behind it. However, a thorough study of Itô calculus and stochastic processes is not necessary for practitioners.

The goal of this book is to introduce the various time series topics that are important for understanding and conducting empirical research. It is written with macroeconomic applications in mind, although the techniques discussed can also be applied to other disciplines. In addition to introducing basic concepts and applications (running a regression and interpreting the results), the book strives to offer a higher level of understanding regarding why certain methods work and others do not. However, it does not seek to provide a rigorous treatment with fully formal proofs; instead, it focuses on providing intuitive explanations. As a result, readers will occasionally encounter non-rigorous proofs when a more formal treatment is deemed unnecessary for a understanding required in applied works. This book can be seen as an intermediate reading between undergraduate econometrics and more rigorous treatments of the subject, such as Hamilton's *Time Series Analysis*.

The materials presented are drawn from or influenced by various sources. I list all of them in the References at the end of the book without citing them individually in the context.

Regarding notations, I use lowercase letters for random variables, such as x_t and y_t. Realizations of random variables are expressed as x_1, x_2, and so on. The context will make it clear whether I am referring to a random variable or its realizations. Capital letters are reserved for matrices, such as $\mathbf{A}$ and $\mathbf{B}$. Vectors are denoted in bold, such as $\mathbf{x}_t$ and $\mathbf{y}_t$, although sometimes I do not explicitly distinguish between vectors and scalars. Greek letters are preferred for parameters, such as α and β. Estimators are indicated with a hat, such as $\hat{\alpha}$ and $\hat{\beta}$.

I use the statistical language R whenever programming is involved. I am aware that there are many time series solutions available in R. To avoid burdening readers with excessive packages, I stick to base R as much as possible with a little help from the *zoo* package.

I would like to emphasize that my knowledge and understanding of the subject are limited, and I acknowledge that there may be mistakes or areas where I could have provided a more accurate explanation. I deeply appreciate any feedback or corrections from readers that could improve the accuracy and clarity of this book. Your input is invaluable, and I am grateful for your assistance in making this book as informative and reliable as possible.

Part I

The Basics

1 Time Series Data

Raw data: The raw values without any transformation. We are not so interested in the raw data, as it is hard to read information from it. Take the GDP plot as an example (Figure 1.1, upper-left subplot). There is an overall upward trend. But we are more interested in: how much does the economy grow this year? Is it better or worse than last year? The answers are not obvious from the raw data. Besides, there are obvious seasonal fluctuations. Usually the first quarter has the lowest value in a whole year, due to the Spring Festival, which significantly reduces the working days in the first quarter. The seasonal fluctuations prohibit us from sensibly comparing two consecutive values.

Growth rate: The headline GDP growth is usually derived by comparing the current quarter with the same quarter from last year. $g = \frac{x_t - x_{t-4}}{x_{t-4}} \times 100$. This makes sense. As mentioned above, due to seasonal patterns, comparing two consecutive quarters directly does not make sense. The year-on-year growth rate directly tells us how fast the economy grows. However, by dividing the past values, it loses the absolute level information. For instance, it is hard to tell after the pandemic, whether or not the economy recovers from its pre-pandemic output level. Besides, it is sensitive to the values of last year. For example, due to the pandemic, the GDP for 2020 is exceptionally low, which makes growth rate for 2021 exceptionally high. This is undesirable, because it does not mean the economy in 2021 is actually good. We would like a growth rate that shirks off past burdens.

That's why we sometimes prefer (annualized) quarterly growth rate. $g = \frac{x_t - x_{t-1}}{x_{t-1}} \times 400$. Due to seasonally patterns, two consecutive quarters are not comparable directly. A first quarter value is usually much lower than the fourth quarter of last year due to holidays, which does not necessarily mean the economy condition is getting worse. Since this pattern is the same every year, it is possible to remove the seasonal fluctuations. This is called *seasonally adjustment*. We won't cover seasonally adjustment in detail, but the next section will give some intuitions on how this can possibly be done. After seasonally adjusting the time series, we can calculate the growth rate based on two consecutive values (annualized by multiplying 4). The bottom-right panel of Figure 1.1 is the seasonally-adjusted quarterly growth. Note that it is no longer biased upward in 2021 as the YoY growth.

Seasonally-adjusted series: This is usually the data format we prefer in time series analysis. FRED reports both seasonally-adjusted and non-seasonally-adjusted series. Seasonal adjustment algorithm is a science in itself. Popular algorithms include X-13-ARIMA developed by the United States Census Bureau, TRAMO/SEATS developed by the Bank of Spain, and so on.

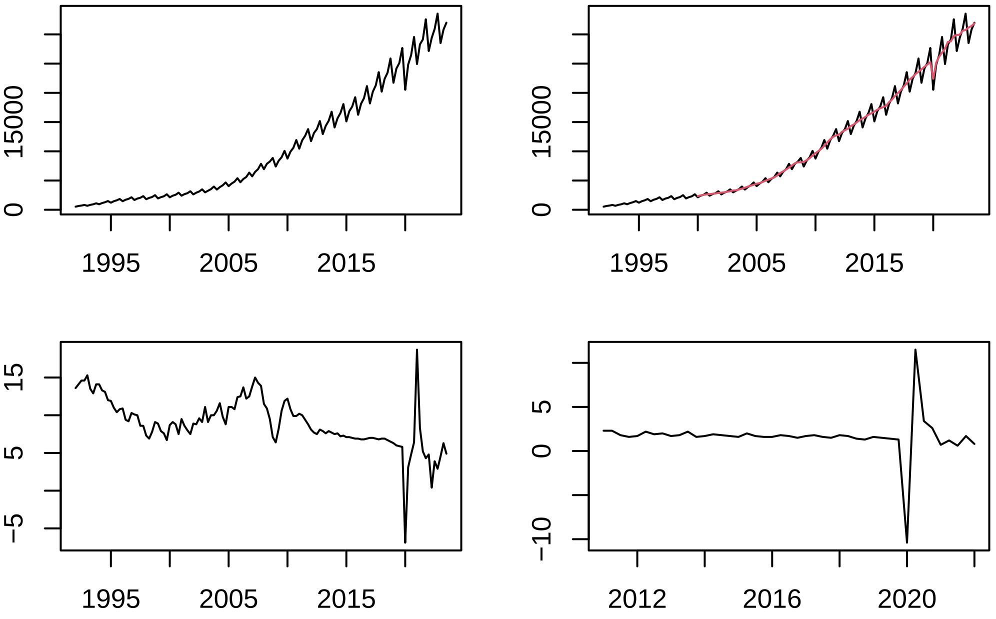

Figure 1.1: Quarterly GDP Time Series (Unit: RMB Billion or %)

Log levels and log growth rates: We like to work with log levels. A lot of economic time series exhibit exponential growth, such as GDP. Taking logs convert them to linear. Another amazing thing about logs is the difference of two log values can be interpreted as percentage growth. We know from Taylor expansion that for small values of Δx : $\ln(\Delta x + 1) \approx \Delta x$. Therefore,

$$\ln x_t - \ln x_{t-1} = \ln\left(\frac{x_t}{x_{t-1}}\right) = \ln\left(\frac{x_t - x_{t-1}}{x_{t-1}} + 1\right) \approx \frac{x_t - x_{t-1}}{x_{t-1}}.$$

So it is very handy to just difference the log levels to get the growth rates. Log difference can also be interpreted as the continuously compounded rate of change, if assuming

$$\frac{x_t}{x_{t-1}} = e^g \implies g = \ln x_t - \ln x_{t-1}.$$

Log difference also has the property of summability: summing up a series of log differences gives the log level provided the initial level. It is not as handy if you want to recover the level values from a series of percentage growth.

$$\ln x_t = x_0 + \sum_{j=1}^{t} (\ln x_j - \ln x_{j-1}).$$

Tip

Buying vs. renting a home, which is better? Compute the NPV:

$$\text{NPV} = \sum_{t=0}^{T} \frac{C_t}{(1+r)^t} = \int_0^T C(t)e^{-rt}dt.$$

2 Decomposition

2.1 Time Series Components

It is helpful to think about a time series as composed of different components: a trend component, a seasonal component, and a remainder.

$$x_t = T_t + S_t + R_t.$$

The formula assumes the “additive” composition. This assumption is appropriate if the magnitude of the fluctuations does not vary with the absolute levels of the time series. If the magnitude of fluctuations is proportional to the absolute levels, a “multiplicative” decomposition is more appropriate:

$$x_t = T_t \times S_t \times R_t.$$

Note that a multiplicative decomposition of a time series is equivalent to an additive decomposition on its log levels:

$$\ln x_t = \ln T_t + \ln S_t + \ln R_t.$$

Decomposing a time series allows us to extract information that is not obvious from the original time series. It also allows us to manipulate the time series. For example, if the seasonal component can be estimated, we can remove it to obtain seasonally-adjusted series, $x_t^{SA} = x_t - S_t$, or $x_t^{SA} = x_t / S_t$. The question is how to estimate the components given a time series.

2.2 Moving Averages

Moving averages turn out to be handy in estiming trend-cycles by averaging out noisy fluctuations. A moving average of order m (assuming m is an odd number) is defined as

$$\mathrm{MA}(x_t, m) = \frac{1}{m} \sum_{j=-k}^{k} x_{t+j},$$

where $m = 2k + 1$. For example, a moving average of order 3 is

$$\mathrm{MA}(x_t, 3) = \frac{1}{3}(x_{t-1} + x_t + x_{t+1}).$$

Note that x_t is centered right in the middle and the average is symmetric. This also means, if we apply this formula to real data, the first and last observation will have to be discarded. If the order m is an even number, the formula will no longer be symmetric. To overcome this, we can estimate a moving average over another moving average. For example, we can estimate a moving average of order 4, followed by a moving average of order 2. This is denoted as 2×4-MA. Mathematically,

$$\begin{aligned}
\mathrm{MA}(x_t, 2 \times 4) &= \frac{1}{2}[\mathrm{MA}(x_{t-1}, 4) + \mathrm{MA}(x_t, 4)] \\
&= \frac{1}{2}\left[\frac{1}{4}(x_{t-2} + x_{t-1} + x_t + x_{t+1}) + \frac{1}{4}(x_{t-1} + x_t + x_{t+1} + x_{t+2})\right] \\
&= \frac{1}{8}x_{t-2} + \frac{1}{4}x_{t-1} + \frac{1}{4}x_t + \frac{1}{4}x_{t+1} + \frac{1}{8}x_{t+2}.
\end{aligned}$$

Note that how the 2×4-MA averages out the seasonality for time series with seasonal period 4, e.g. quarterly series. The formula puts equal weight on every quarter — the first and last terms refer the same quarter and their weights combined to $\frac{1}{4}$.

In general, we can use m-MA to estimate the trend if the seasonal period is an odd number, and use $2 \times m$-MA if the seasonal period is an even number.

```r
data = readRDS("data/gdp.Rds")  # a `zoo` object
gdp2x4MA = ma(ma(data$GDP,4),2) # from `forecast` package
ts.plot(cbind(data$GDP, gdp2x4MA), col=1:2)
```

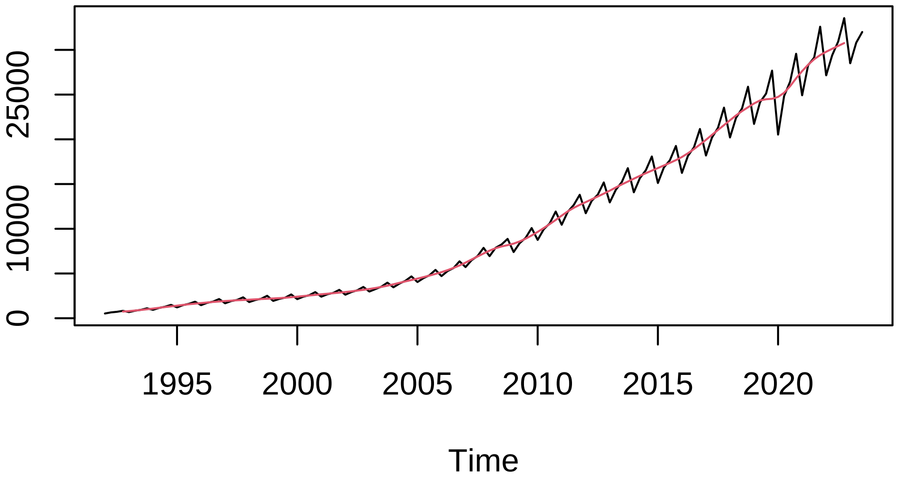

Figure 2.1: Quarterly GDP with 2x4-MA estimate of the trend-cycle

2.3 Classical Decomposition

Moving averages give us everything we need to perform classical decomposition. Classical decomposition, invented 1920s, is the simplest method decompose a time series into trend, seasonality and remainder. It is outdated nowadays and has been replaced by more advanced algorithms. Nonetheless, it serves as a good example for introductory purpose on how time series decomposition could possibly be achieved.

The algorithm for additive decomposition is as follows.

1. Estimate the trend component T_t by applying moving averages. If the seasonal period is an odd number, apply the m-th order MA. If the seasonal period is even, apply the $2 \times m$ MA.
2. Calculate the detrended series $x_t - T_t$.
3. Calculate the seasonal component S_t by averaging all the detrended values of the season. For example, for quarterly series, the value of S_t for Q1 would be the average of all values in Q1. This assumes the seasonal component is constant over time. S_t is then adjusted to ensure all values summed up to zero.
4. Subtracting the seasonal component to get the remainder $R_t = x_t - T_t - S_t$.

```
log(data$GDP) |> decompose() |> plot()
```

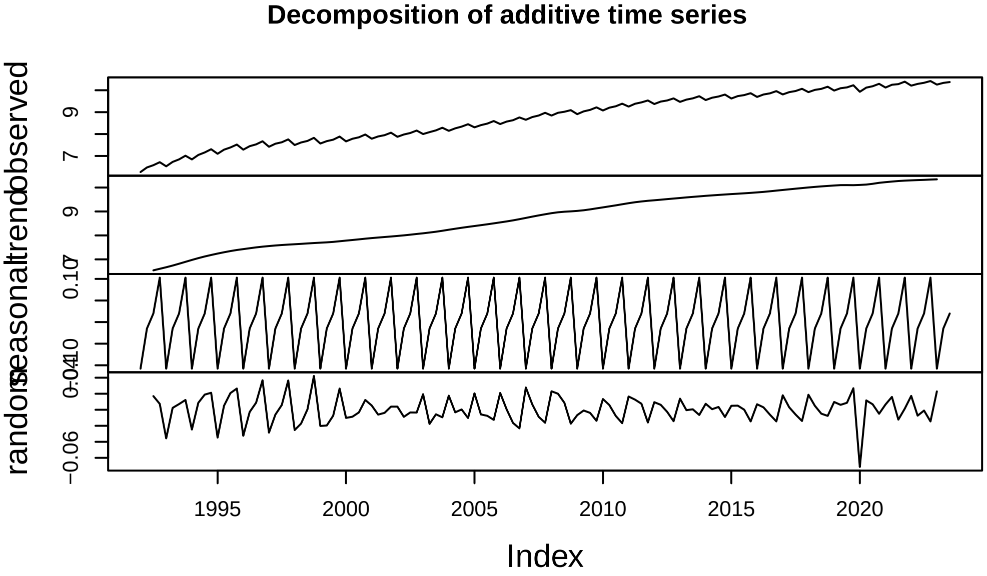

Figure 2.2: Classical multiplicative decomposition of quarterly GDP

The example performs additive decomposition to the logged quarterly GDP series. Note how the constant seasonal component is removed, leaving the smooth and nice-looking up-growing trend. The remainder component tells us the irregular ups and downs of the economy around the trend-cycle. Isn't it amazing that a simple decomposition of the time series tells us a lot about the economy?

2.4 Seasonal Adjustment

By decomposing a time series into trend, seasonality and remainder, it readily gives us a method for seasonal adjustment. Simply subtracting the seasonal component from the original data, or equivalently, summing up the trend and the remainder components, would give us the seasonally-adjusted series.

The following example compares the seasonally-adjusted series using the classical decomposition with the state-of-the-art X-13ARIMA-SEATS algorithm. Despite the former is far more rudimentary than the latter, they look quite close if we simply eye-balling the plot. By taking first-order differences, we can see the series based on classical decomposition is more volatile, suggesting the classical decomposition is less robust to unusual values.

```
logdata = log(data) |> window(start=2000)
seasadj = as.ts(logdata$GDP) - decompose(logdata$GDP)$seasonal
```

```
par(mfrow=c(1,2), mar=rep(2,4))
ts.plot(cbind(seasadj, logdata$GDPSA), col=1:2)
ts.plot(diff(cbind(seasadj, logdata$GDPSA)), col=1:2)
```

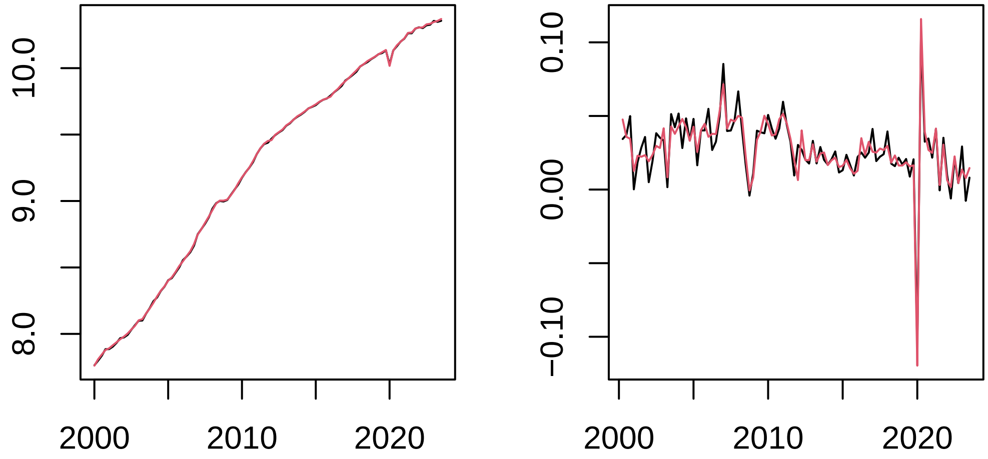

Figure 2.3: Comparing classical decomposition and X-13

3 ACF and PACF

A time series is notationally represented by $\{\dots, y_{t-1}, y_t, y_{t+1}, y_{t+2}, \dots\}$, which is a sequence of random variables. We think of each variable at a time point t as a random variable, whose realized value is drawn from some distribution.

A distinguishing feature of this sequence is temporal dependence. That is, the distribution of y_t conditional on previous value of the series depends on the outcome of those previous observations. It is of particular interest how observations are correlated across time. A big part of the time series analysis is to exploit this correlation.

3.1 Autocorrelation

The temporal dependence is characterized by the correlation between y_t and its own lags y_{t-k}.

Definition 3.1. The k-th order autocovariance of y_t is defined as

$$\gamma_k = \text{cov}(y_t, y_{t-k}).$$

The k-th order autocorrelation is defined as

$$\rho_k = \frac{\text{cov}(y_t, y_{t-k})}{\text{var}(y_t)} = \frac{\gamma_k}{\gamma_0}.$$

If we plot the autocorrelation as a function of the lag length k, we get the autocorrelation function (ACF). Here is an example of the ACF of China's monthly export growth (log-difference). The lag on the horizontal axis is counted by seasonal period. Because it is monthly data, 1 period is 12 months. We can see the autocorrelation is the strongest for the first two lags. Longer lags are barely significant. There are spikes with 12-month and 24-month lags, indicating the seasonality is not fully removed from the series.

```
data = readRDS("data/md.Rds")
acf(data$Export, main='Autocorrelation')
```

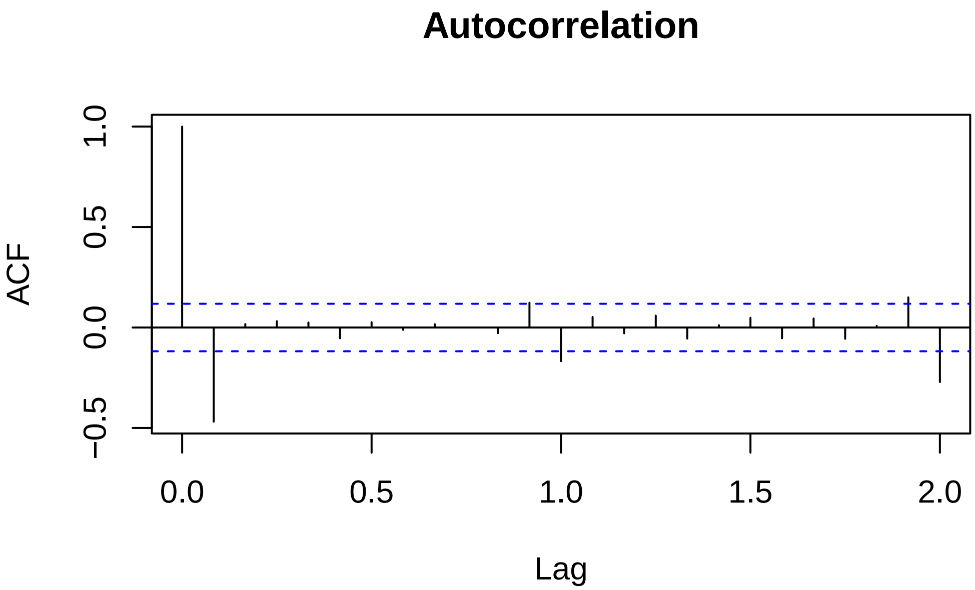

Figure 3.1: ACF for monthly export growth

3.2 Partial Autocorrelation

ACF measures the correlation between y_t and y_{t-k} regardless of their relationships with the intermediate variables $y_{t-1}, y_{t-2}, \cdots, y_{t-k+1}$. Even if y_t is only correlated with the first-order lag, it is automatically made correlated with the k-th order lag through intermediate variables. Sometime we are interested in the correlation between y_t and y_{t-k} partialling out the influence of intermediate variables.

Definition 3.2. The partial autocorrelation function (PACF) considers the correlation between the remaining parts in y_t and y_{t-k} after partialling out the intermediate effect of $y_{t-1}, y_{t-2}, \cdots, y_{t-k+1}$.

$$\phi_k = \begin{cases} \text{corr}(y_t, y_{t-1}) = \rho_{_1}, & \text{if } k = 1; \\ \text{corr}(r_{y_t|y_{t-1},\dots,y_{t-k+1}}, r_{y_{t-k}|y_{t-1},\dots,y_{t-k+1}}), & \text{if } k \geq 2; \end{cases}$$

where $r_{y|x}$ means the remainder in y after partialling out the intermediate effect of x.

In practice, ϕ_k can be estimated by the regression

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

The estimated coefficient $\hat{\phi}_k$ is the partial autocorrelation after controlling the intermediate lags.

```
pacf(data$Export, main='Partial Autocorrelation')
```

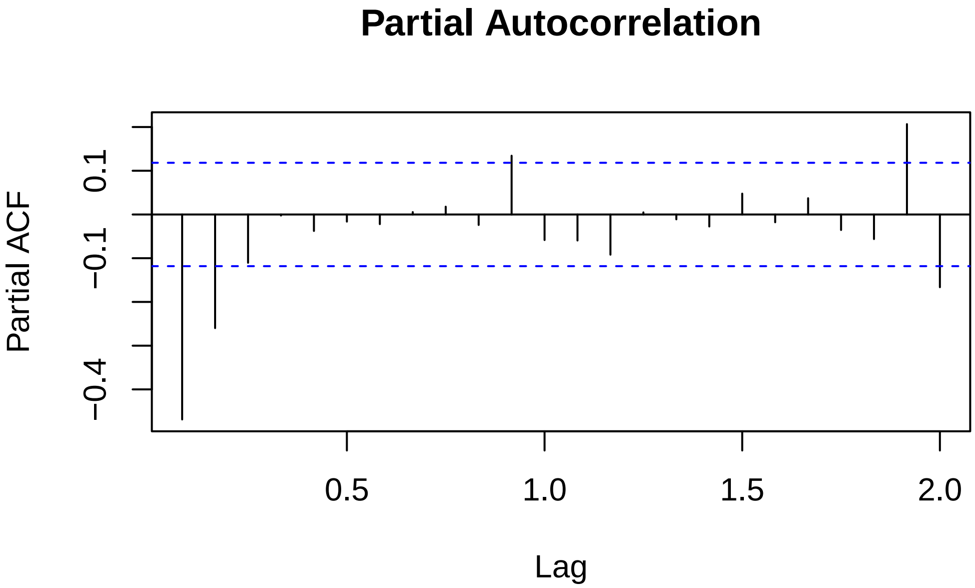

Figure 3.2: PACF for monthly export growth

4 Stationarity

4.1 Stationary Process

Definition 4.1. A stochastic process is said to be **strictly stationary** if its properties are unaffected by a change of time origin. In other words, the joint distribution at any set of time is not affect by an arbitrary shift along the time axis.

Definition 4.2. A stochastic process is called **covariance stationary** (or **weak stationary**) if its means, variances, and covariances are independent of time. Formally, a process $\{y_t\}$ is covariance stationary if for all t it holds that

- $\mathbb{E}(y_t) = \mu < \infty$;
- $\text{var}(y_t) = \gamma_0 < \infty$;
- $\text{cov}(y_t, y_{t-k}) = \gamma_k$, for $k = 1, 2, 3, ...$

Stationarity is an important concept in time series analysis. It basically says the statistical properties of a time series are stable over time. Otherwise, if the statistical properties vary with time, statistics estimated from past values, such autocorrelations, would be much less meaningful. Strict stationarity requires the joint distribution being stable, that is moments of any order would be stable over time. In practice, mostly we only care about the first- and second-order moments, that is means and variances and covariances. Therefore, covariance stationary is sufficient.

Figure 4.1 shows some examples of stationary and non-stationary time series. Only the first one is stationary (it is generated from *i.i.d* normal distribution). The second one is not stationary as its mean is not constant over time. The third one is not stationary as its variance is not constant. The last one is not stationary either, because its covariance is not constant.

Real-life time series are rarely stationary. But they can be transformed to (quasi) stationary by differencing. Figure 4.2 shows some examples of the first-order (log) differences of real-life time series. They more or less exhibit some properties of stationarity, but not perfectly stationary. The series can be further “stationarized” by taking a second-order difference. But

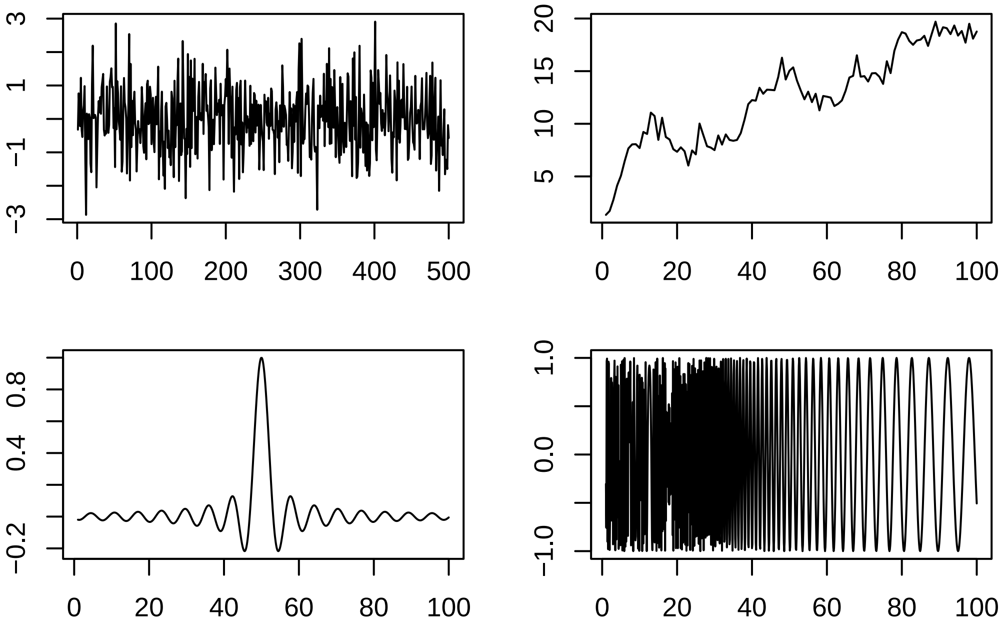

Figure 4.1: Stationary and non-stationary time series

these examples are acceptable to be treated as stationary in our models. Even if they are not perfectly stationary, the model can be thought of being used to “extract” their stationary properties.

Proposition 4.1. *For stationary series, it holds that* $\gamma_k = \gamma_{-k}$*.*

Proof. By definition,

$$\gamma_k = \mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)],$$

$$\gamma_{-k} = \mathbb{E}[(y_t - \mu)(y_{t+k} - \mu)].$$

Since y_t is stationary, γ_k is invariant with time. Let $t' = t + k$, we have

$$\begin{aligned}\gamma_k &= \mathbb{E}[(y_{t'} - \mu)(y_{t'-k} - \mu)]\\ &= \mathbb{E}[(y_{t+k} - \mu)(y_t - \mu)]\\ &= \gamma_{-k}.\end{aligned}$$

□

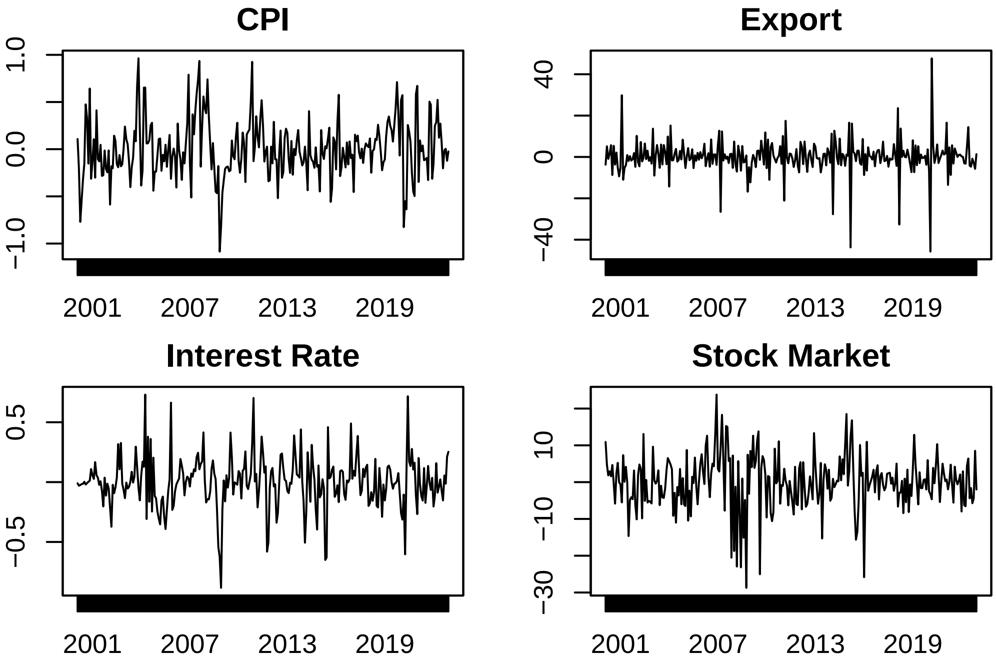

Figure 4.2: Stationary and non-stationary time series (real life)

4.2 Ergodicity

Temporal dependence is an important feature of time series variables. This dependence is both a bless and a curve. Autocorrelation enables us to make predictions based on past experiences. However, as we will see in later chapters, it also invalidates theorems that usually require *iid* assumptions. Ideally, we would like the temporal dependence to be not too strong. This is the property of ergodicity.

Definition 4.3. A stationary process $\{y_t\}$ is **ergodic** if

$$\lim_{n\to\infty} |\mathbb{E}[f(y_t \cdots y_{t+k})g(y_{t+n} \cdots y_{t+n+k})]| = |\mathbb{E}[f(y_t \cdots y_{t+k})]||\mathbb{E}[g(y_{t+n} \cdots y_{t+n+k})]|.$$

Heuristically, ergodicity means if two random variables are positioned far enough in the sequence, they become almost independent. In other words, ergodicity is a restriction on dependency. An ergodic process allows serial correlation, but the serial correlation disappears if the two observations are far apart. Ergodicity is important because as we will see in later chapters, the Law of Large Numbers or the Central Limit Theorem will not hold without it.

Theorem 4.1. *A stationary time series is ergodic if* $\sum_{k=0}^{\infty} |\gamma_k| < \infty$.

Proof. A rigorous proof is not necessary. It is enough to give an intuition why autocorrelation disappears for far apart variables. Note that $\sum_{k=0}^{\infty} |\gamma_k|$ is monotonic and increasing, it converges. Therefore, $\gamma_k \to 0$ by Cauchy Criterion.

□

4.3 White Noise

White noise is a special stationary process that is an important building block of many time series models.

Definition 4.4. A stochastic process w_t is called **white noise** if its has constant mean 0 and variance σ^2 and no serial correlation $\text{cov}(w_t, w_{t-k}) = 0$ for any $k \neq 0$. The white noise process is denoted as

$$w_t \sim \text{WN}(0, \sigma^2).$$

This is the weakest requirement for while noise. It only requires no serial correlation. We may impose further assumptions. If every w_t is independent, it becomes independent white noise $w_t \sim\perp \text{WN}(0, \sigma^2)$. Independence does not imply identical distribution. If every w_t is independently and identically distributed, it is called *i.i.d* white noise, $w_t \overset{iid}{\sim} \text{WN}(0, \sigma^2)$. If the distribution is normal, it becomes the most perfect white noise, that is *i.i.d* Gaussian white noise, $w_t \overset{iid}{\sim} N(0, \sigma^2)$. The first plot of Figure 4.1 is a demonstration of the *i.i.d* Gaussian white noise. In most cases, the weakest form of white noise is sufficient.

Exercise

Prove that a while noise process is stationary.

Part II

ARIMA Model

5 Model vs Spec

5.1 Classification

Time series models can be broadly sorted into four categories based on whether we are dealing with stationary or non-stationary time series, or whether the model involves only one variable or multiple variables.

Table 5.1: Time series model classification

	Stationary	Nonstationary
Univariate	ARMA	Unit root
Multivariate	VAR	Cointegration

5.2 Model vs Spec

We use the word "model" rather loosely in economics and econometrics. Anything that deals with the quantified relationships between variables can be called a model. A general equilibrium model is a model. A regression is also a model.

To make things less confusing, we would use the word "model" more restrictively in this chapter. We reserve the word **model** to those representing the **data generating processes** (DGPs). That is, when we write down a model in an equation, we literally mean it. If we say y_t follows an AR(1) model:

$$\begin{aligned} y_t &= \phi y_{t-1} + \epsilon_t, \\ \epsilon_t &\sim N(0, \sigma^2). \end{aligned}$$

We literally mean y_t is determined by its previous value and an contemporary innovation drawn from a Gaussian distribution.

A model is distinguished from a **specification**. Suppose $\{y_t\}$ represent the GDP series, we can estimate a regression:

$$y_t = \phi y_{t-1} + e_t$$

This is a specification not a model. Because the DGP of GDP data is unknown, definitely not an AR(1). We can nontheless fit this spec with the data and get an estimated $\hat{\phi}$. If e_t satisfies some nice properties, for example, uncorrelated with the regressor, then we know this $\hat{\phi}$ is consistent.

When we run regressions with real-life data, we are actually working with specifications. They are not the DGPs of the random variables. But they allow us to recover some useful information from the data when certain assumptions are met. Mostly we are interested in the relationships between variables. A specification describes this relationship, even though it does not describe the full DGP.

This chapter deals with models in the abstract sense. The next chapter will discuss how to fit a model or a spec with real data.

6 AR Models

6.1 AR(1) Process

We start with the simplest time series model — autoregressive model, or AR model. The simplest from of AR model is AR(1), which involves only one lag,

$$y_t = \mu + \phi y_{t-1} + \epsilon_t, \tag{6.1}$$

where $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$. The model can be extended to include more lags. An AR(p) model is defined as

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t.$$

We focus on AR(1) first. The model states that the value of y_t is determined by a constant, its previous value, and a random innovation. We call the last term ϵ_t *innovation*, not an error term. It is not an error, it is a random contribution that is unknown until time t. It should also not be confused with the so-called "structural shock", which is attached with a structural meaning and will be discussed in later chapters.

The model is *probabilistic*, as oppose to *deterministic*, in the sense that some information is unknown or deliberately omitted, so that we do not know the deterministic outcome, but only a probability distribution.

Note

Think about tossing a coin: if every piece of information is incorporated in the model, including the initial speed and position, the air resistance, and so on; then we can figure out the exact outcome, whether the coin will land on its head or tail. But this is unrealistic. Omitting all these information, we can model the process as a Bernoulli distribution. The probability model will not give a deterministic outcome, but only a distribution with each possible value associated with a probability.

Note

The assumption that a process is only determined by its past values and a white noise innovation seems very restrictive. But it is not. Think about the three assumptions for technical analysis of the stock market (there are still many investors believing this): (1) The market discounts everything, (2) prices move in trends and counter-trends, and (3) price action is repetitive, with certain patterns reoccurring. Effectively, it is saying we can predict the stock market by the past price patterns. If we were to write a model for the stock market based on these assumptions, AR(p) isn't a bad choice at all.

Note that the model can be rewritten as

$$y_t - \frac{\mu}{1-\phi} = \phi\left(y_{t-1} - \frac{\mu}{1-\phi}\right) + \epsilon_t,$$

assuming $\phi \neq 1$. If we define $\tilde{y_t} = y_t - \frac{\mu}{1-\phi}$, we can get rid of the constant term:

$$\tilde{y}_t = \phi \tilde{y}_{t-1} + \epsilon_t. \tag{6.2}$$

It can be easily shown, if y_t is stationary, $\frac{\mu}{1-\phi}$ is the stationary mean. Because this mechanical transformation can always be done to remove the constant. We can simply ignore the constant term without lost of generality.

Note

Working with demeaned variables greatly simplify the notation. For example, assuming $\mathbb{E}(y_t) = 0$, the variance is simply the second-order moment $\mathbb{E}(y_t^2)$; the covariance can be written as $\mathbb{E}(y_t y_{t-k})$.

For a constant-free AR(1) model, we can rewrite the model as follows:

$$
\begin{aligned}
y_t &= \phi y_{t-1} + \epsilon_t \\
&= \phi(\phi y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
&= \phi^2 y_{t-2} + \phi\epsilon_{t-1} + \epsilon_t \\
&= \phi^2(\phi y_{t-3} + \epsilon_{t-2}) + \phi\epsilon_{t-1} + \epsilon_t \\
&= \phi^3 y_{t-3} + \phi^2\epsilon_{t-2} + \phi\epsilon_{t-1} + \epsilon_t \\
&\vdots \\
&= \phi^t y_0 + \sum_{j=0}^{t-1} \phi^j \epsilon_{t-j} \\
&= \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}.
\end{aligned}
\tag{6.3}
$$

The exercise shows an AR(1) process can be reduced to an MA process, which will be discussed in the next section. It says the value of y_t is determined by its initial value (if it has one) and the accumulated innovations in the past. It is our deeds in history that shapes our world today.

> **Note**
>
> The property that an AR process can be rewritten as an infinite MA process with absolute summable coefficients $\sum_{j=0}^{\infty} |\phi^j| < \infty$ is called *causal*. This must not be confused with the causal effect in econometrics (defines in the *ceteris paribus* sense). To avoid confusion, we avoid use this term as much as possible.

Now we focus our attention on the critical parameter ϕ. If $|\phi| > 1$, the process is explosive. We are not interested in explosive processes. If a real-world time series grows exponentially, we take logarithm to transform it to linear. So in most of our discussions, we rule out the case of explosive behaviour.

If $|\phi| < 1$, $\phi^j \to 0$ as $j \to \infty$. This means the influence of innovations far away in the past decays to zero. We will show that the series is stationary and ergodic.

If $|\phi| = 1$, we have $y_t = \sum_{j=0}^{\infty} \text{sgn}(\phi)^j \epsilon_{t-j} = \sum_{j=0}^{\infty} \tilde{\epsilon}_{t-j}$. This means the influence of past innovations will not decay no matter how distant away they are. This is known as a *unit root process*, which will be covered in later chapters. But it is clear that the process is not stationary. Consider the variance of y_t conditioned on an initial value:

$$
\text{var}(y_t|y_0) = \text{var}(\sum_{j=0}^{t-1} \epsilon_{t-j}) = \sum_{j=0}^{t-1} \text{var}(\epsilon_{t-j}) = \sum_{j=0}^{t-1} \sigma^2 = t\sigma^2.
$$

The variance is increasing with time. It is not constant. Figure 6.1 simulates the AR(1) with $\phi = 0.5$ and $\phi = 1$ respectively.

```r
y = arima.sim(list(ar=0.5), n=1000)
z = arima.sim(list(order=c(0,1,0)), n=1000)
plot(cbind(y,z), plot.type="multiple", nc=2, ann=F,
     mar.multi=rep(2,4), oma.multi = rep(0,4))
```

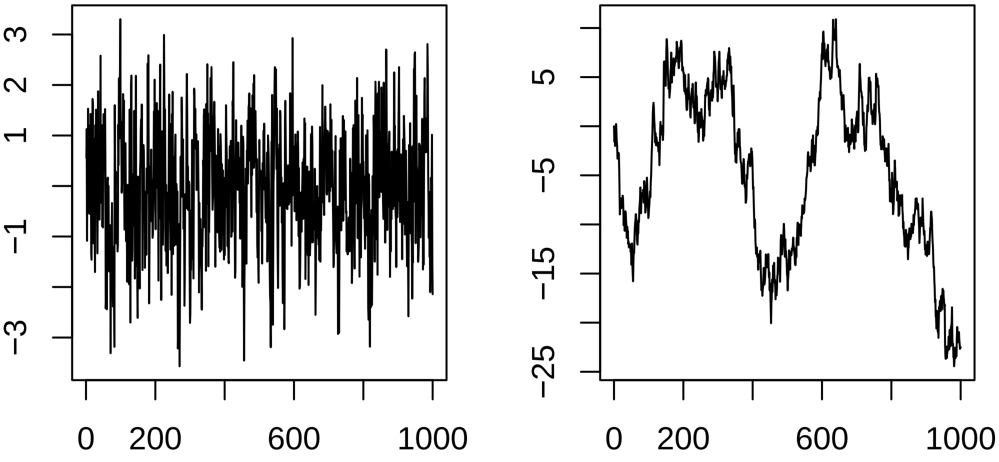

Figure 6.1: Simulation of AR(1) processes

Proposition 6.1. *An AR(1) process with $|\phi| < 1$ is covariance stationary.*

Proof. Let's compute the mean, variance and covariance for the AR(1) process.

$$\mathbb{E}(y_t) = \mathbb{E}\left[\sum_{j=0}^{\infty}\phi^j\epsilon_{t-j}\right] = \sum_{j=0}^{\infty}\phi^j\mathbb{E}[\epsilon_{t-j}] = 0.$$

$$\begin{aligned}\text{var}(y_t) &= \text{var}\left[\sum_{j=0}^{\infty}\phi^j\epsilon_{t-j}\right] = \sum_{j=0}^{\infty}\phi^j\text{var}[\epsilon_{t-j}] \\ &= \sigma^2\sum_{j=0}^{\infty}\phi^j = \frac{\sigma^2}{1-\phi}.\end{aligned}$$

For the covariances,

$$
\begin{aligned}
\gamma_1 &= \mathbb{E}(y_t y_{t-1}) = \mathbb{E}((\phi y_{t-1} + \epsilon_t) y_{t-1}) \\
&= \mathbb{E}(\phi y_{t-1}^2 + \epsilon_t y_{t-1}) \\
&= \phi \mathbb{E}(y_{t-1}^2) + 0 \\
&= \frac{\phi \sigma^2}{1-\phi};
\end{aligned}
$$

$$
\begin{aligned}
\gamma_2 &= \mathbb{E}(y_t y_{t-2}) = \mathbb{E}((\phi y_{t-1} + \epsilon_t) y_{t-2}) \\
&= \mathbb{E}(\phi y_{t-1} y_{t-2} + \epsilon_t y_{t-2}) \\
&= \phi \mathbb{E}(y_{t-1} y_{t-2}) \\
&= \phi \gamma_1 = \frac{\phi^2 \sigma^2}{1-\phi}; \\
&\vdots \\
\gamma_j &= \frac{\phi^j \sigma^2}{1-\phi}.
\end{aligned}
$$

All of them are independent of time t. By Definition 4.2, the process is covariance stationary.

□

So the ACF decays gradually as $\phi^j \to 0$. What about the PACF? Estimating the PACF is equivalent to regressing y_t on its lags. Since there is only one lag, the PACF should have non-zero value only for the first lag, and zeros for all other lags.

```
par(mfrow=c(1,2), mar=c(2,4,1,1))
acf(y); pacf(y)
```

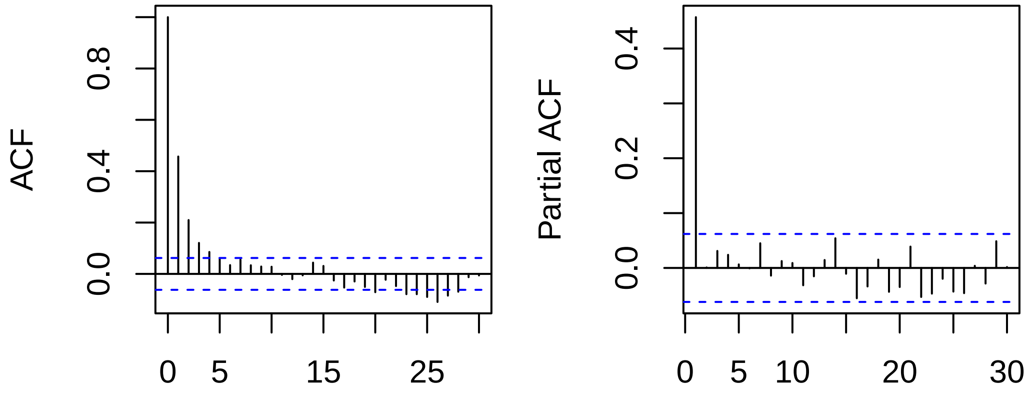

Figure 6.2: ACF and PACF for AR(1) process

6.2 Lag Operator

To facilitate easy manipulation of lags, we introduce the lag operator:

$$Ly_t = y_{t-1}.$$

The AR(1) process can be written with the lag operator:

$$y_t = \phi L y_t + \epsilon_t \implies (1-\phi L)y_t = \epsilon_t.$$

The lag operator L can be manipulated just as polynomials. It looks weird, but it actually works. Do a few exercises to convince yourself.

$$L^2 y_t = L(Ly_t) = Ly_{t-1} = y_{t-2}.$$

$$\begin{aligned}(1-L)^2 y_t &= (1-L)(y_t - y_{t-1}) \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \\ &= (1 - 2L + L^2)y_t.\end{aligned}$$

We can even inverse a lag polynomial (provided $|\phi| < 1$),

$$\begin{aligned}&(1-\phi L)y_t = \epsilon_t \\ \implies\ & y_t = (1-\phi L)^{-1}\epsilon_t = \sum_{j=0}^{\infty}\phi^j L^j \epsilon_t = \sum_{j=0}^{\infty}\phi^j \epsilon_{t-j}.\end{aligned}$$

We reach the same conclusion as Equation 6.3 with the lag operator.

6.3 AR(p) Process

We now generalize the conclusions above to AR(p) processes. With the help of the lag operator, an AR(p) process can be written as

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)y_t = \epsilon_t,$$

or even more parsimoniously,

$$\phi(L)y_t = \epsilon_t.$$

Note that we ignore the constant term, which can always be removed by redefine $\tilde{y}_t = y_t - \frac{\mu}{1-\phi_1-\phi_2-\cdots-\phi_p}$.

To derive the MA representation, we need to figure out $\phi^{-1}(L)$. By the Fundamental Theorem of Algebra, we know the polynomial $\phi(z)$ has p roots in the complex space. So the lag polynomial can be factored as

$$(1-\lambda_1 L)(1-\lambda_2 L)\dots(1-\lambda_p L)y_t = \epsilon_t,$$

where $z = \lambda_i^{-1}$ is the i-th root of $\phi(z)$. If the roots are outside the unit circle, $|\lambda_i| < 1$ means each of the left hand terms is inversible.

$$\begin{aligned} y_t &= \frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)\dots(1-\lambda_p L)}\epsilon_t \\ &= \left(\frac{c_1}{1-\lambda_1 L} + \frac{c_2}{1-\lambda_2 L} + \cdots + \frac{c_p}{1-\lambda_p L}\right)\epsilon_t \\ &= \sum_{j=0}^{\infty}(c_1\lambda_1^j + c_2\lambda_2^j + \cdots + c_p\lambda_p^j)L^j\epsilon_t \\ &= \sum_{j=0}^{\infty}\theta_j\epsilon_{t-j}, \text{ where } \theta_j = c_1\lambda_1^j + \cdots + c_p\lambda_p^j. \end{aligned}$$

It follows that this process has constant mean and variance. For the covariances, given

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t,$$

Multiply both sides by y_t and take expectation,

$$\mathbb{E}[y_t^2] = \phi_1\mathbb{E}[y_t y_{t-1}] + \phi_2\mathbb{E}[y_t y_{t-2}] + \cdots + \phi_p\mathbb{E}[y_t y_{t-p}],$$

$$\gamma_0 = \phi_1\gamma_{-1} + \phi_2\gamma_{-2} + \cdots + \phi_p\gamma_{-p}.$$

Similarly, multiply both sides by $y_{t-1}, \dots, y_{t-j}$, we have

$$\begin{aligned} \gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_{-1} + \cdots + \phi_p\gamma_{-p+1}, \\ &\vdots \\ \gamma_j &= \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p}. \end{aligned}$$

This is called the Yule-Walker equation. The first p unknowns $\gamma_0, \dots, \gamma_{p-1}$ can be solved by the first p equations. The rest can then be solved iteratively.

It can be shown all of the covariances are invariant with time. Therefore, under the condition all $|\lambda_i| < 1$, the AR(p) process is stationary.

For the PACF, a regression of y_t over its lags would recover p non-zero coefficients. Longer lags should have coefficients insignificantly different from zero.

```
y = arima.sim(list(ar=c(2.4, -1.91, 0.5)), n=3000)
par(mfrow=c(1,2), mar=c(2,4,1,1))
acf(y); pacf(y)
```

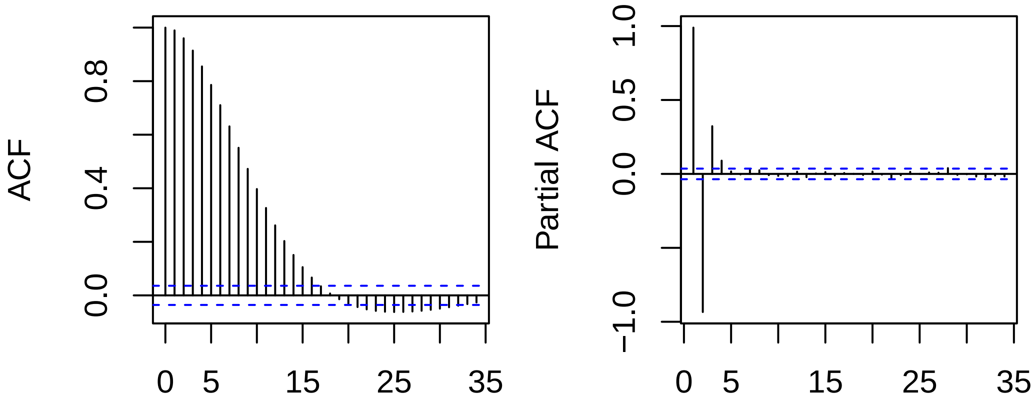

Figure 6.3: ACF and PACF for AR(p) process

Proposition 6.2. *An AR(p) process is stationary if all the roots of $\phi(z)$ are outside the unit circle.*

Proposition 6.3. *An AR(p) process is characterized by (i) an ACF that is infinite in extend but tails of gradually; and (ii) a PACF that is (close to) zero for lags after p.*

7 MA Models

7.1 MA(1) Process

Again, let's start with the simplest moving average model. A first-order moving average process, or MA(1), is defined as

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}, \tag{7.1}$$

where $\{\epsilon_t\} \sim \text{WN}(0, \sigma^2)$ are uncorrelated innovations. The MA model says the current value y_t is a moving average of past innovations (in MA(1), the weight on ϵ_{t-1} is θ). MA models directly relate the observable variable to past innovations. If we know the past innvation ϵ_{t-j}, we can easily figure out its contribution to the outcome variable (unlike AR models where the effect of a past innovation is transmitted through $y_{t-j}, \dots, y_{t-1}$). So MA models are the preferred analytic tool in many applications, despite it looks odd from the eyes of regression modelers. You may wonder how it is possible to estimate such a model. We will put off the estimation techniques to the next chapter.

It is clear that y_t has a constant mean, $\mathbb{E}(y_t) = \mu$. We can omit the constant if we work with the demeaned series $\tilde{y}_t = y_t - \mu$. Without loss of generality, we assume for the rest $\{y_t\}$ has zero mean, so the model is simplified as

$$y_t = \epsilon_t + \theta\epsilon_{t-1}. \tag{7.2}$$

Let's compute its variance and covariances:

$$\begin{aligned}
\gamma_0 &= \text{var}(\epsilon_t + \theta\epsilon_{t-1}) = \text{var}(\epsilon_t) + \theta^2\text{var}(\epsilon_{t-1}) = (1+\theta^2)\sigma^2;\\
\gamma_1 &= \text{cov}(y_t, y_{t-1}) = \text{cov}(\epsilon_t + \theta\epsilon_{t-1}, \epsilon_{t-1} + \theta\epsilon_{t-2}) = \text{cov}(\theta\epsilon_{t-1}, \epsilon_{t-1} + \theta\epsilon_{t-2}) = \theta\sigma^2;\\
\gamma_2 &= \text{cov}(y_t, y_{t-2}) = \text{cov}(\epsilon_t + \theta\epsilon_{t-1}, \epsilon_{t-2} + \theta\epsilon_{t-3}) = 0;\\
&\vdots\\
\gamma_j &= 0 \text{ for } |j| \geq 2.
\end{aligned}$$

It is clear that the MA(1) process is *stationary*. And the ACF cuts off after the first lag. Because more distant lags y_{t-k} are constituted by even more distant innovations $\epsilon_{t-k}, \epsilon_{t-k-1}, \dots$ which has no relevance for y_t given the MA(1) structure.

We have seen AR processes are equivalent to MA(∞) processes. Similar results hold for MA models. Rewrite the MA(1) process with the lag operator, assuming $|\theta| < 1$,

$$y_t = (1+\theta L)\epsilon_t \Leftrightarrow (1+\theta L)^{-1} y_t = \epsilon_t \Leftrightarrow \sum_{j=0}^{\infty} (-\theta)^j y_{t-j} = \epsilon_t.$$

That means an MA(1) is equivalent to an AR(∞) process if $(1+\theta L)$ is *invertible*. This shows AR and MA are really the same family of models. The model AR or MA is chosen by parsimonious principle. For example, an AR model with many lags can possibly be modeled by a parsimonious MA model.

Since an MA(1) is equivalent to some AR(∞) process, the PACF of an MA(1) should tail off gradually.

```
y = arima.sim(list(ma=0.8), n=2000)
par(mfrow=c(1,2), mar=c(1,4,1,1))
acf(y); pacf(y)
```

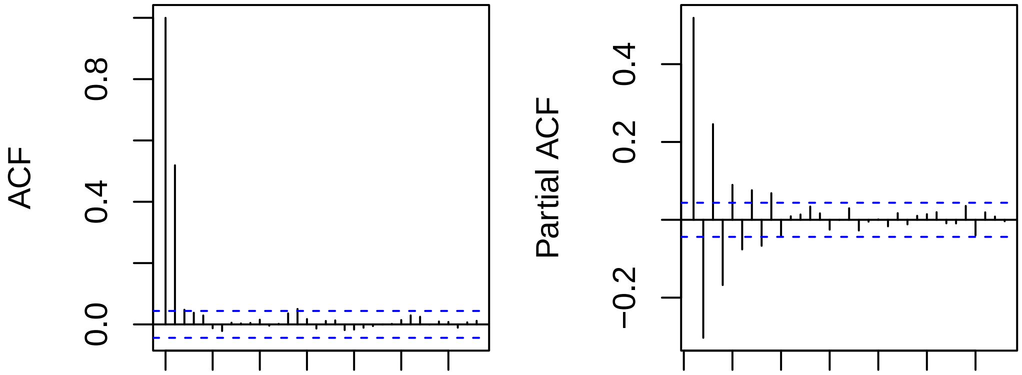

Figure 7.1: ACF and PACF of MA(1) process

Invertibility

If $|\theta| > 1$, $\theta(L)$ is not invertible. Define another MA(1) process,

$$y_t = \epsilon_t + \theta^{-1}\epsilon_{t-1}, \quad \epsilon_t \sim \text{WN}(0, \theta^2\sigma^2).$$

We can verify that its variance and covariances are exactly the same as Equation 7.2. For non-invertible MA process, as long as $\theta(L)$ avoids unit root, we can always find an invertible process that shares the same ACF. This means, for a stationary MA process, it makes no harm to just assume it is invertible.

7.2 MA(q) Process

A q-th order moving average, or MA(q) process, is written as

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \tag{7.3}$$

where $\{\epsilon_t\} \sim \text{WN}(0, \sigma^2)$.

Proposition 7.1. *An MA(q) process is stationary.*

Proof. We will show that the mean, variance and covariances of MA(q) are all invariant with time.

$$\mathbb{E}(y_t) = \mu.$$

Assume for the rest, $\{y_t\}$ is demeaned.

$$\begin{aligned}
\gamma_0 &= \mathbb{E}(y_t^2) = \mathbb{E}[(\epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q})^2] \\
&= \mathbb{E}[\epsilon^2] + \theta_1^2 \mathbb{E}[\epsilon_{t-1}^2] + \cdots + \theta_q^2 \mathbb{E}[\epsilon_{t-q}^2] \\
&= (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma^2; \\
\gamma_1 &= \mathbb{E}[y_t y_{t-1}] = \mathbb{E}[(\epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}) \\
&\qquad\qquad\qquad (\epsilon_{t-1} + \cdots + \theta_{q-1} \epsilon_{t-q} + \theta_q \epsilon_{t-q-1})] \\
&= \theta_1 \mathbb{E}[\epsilon_{t-1}^2] + \theta_2 \theta_1 \mathbb{E}[\epsilon_{t-2}^2] + \cdots + \theta_q \theta_{q-1} \mathbb{E}[\epsilon_{t-q}^2] \\
&= (\theta_1 + \theta_2 \theta_1 + \cdots + \theta_q \theta_{q-1})\sigma^2; \\
&\vdots \\
\gamma_j &= \mathbb{E}[y_t y_{t-j}] = \mathbb{E}[(\epsilon_t + \cdots + \theta_j \epsilon_{t-j} + \cdots + \theta_q \epsilon_{t-q}) \\
&\qquad\qquad\qquad (\epsilon_{t-j} + \cdots + \theta_{q-j} \epsilon_{t-q} + \cdots + \theta_q \epsilon_{t-q-j})] \\
&= \theta_j \mathbb{E}[\epsilon_{t-j}^2] + \theta_{j+1} \theta_1 \mathbb{E}[\epsilon_{t-j-1}^2] + \cdots + \theta_q \theta_{q-j} \mathbb{E}[\epsilon_{t-q}^2] \\
&= (\theta_j + \theta_{j+1} \theta_1 + \cdots + \theta_q \theta_{q-j})\sigma^2, \text{ for } j \leq q; \\
\gamma_j &= 0, \text{ for } j > q.
\end{aligned}$$

□

Proposition 7.2. *An MA(q) process is invertible iff the roots of $\theta(z)$ are outside the unit circle.*

Proposition 7.3. *An MA(q) process is characterized by (i) an ACF that is (close to) zero after q lags; and (i) a PACF that is infinite in extend but tails of gradually.*

7.3 MA(∞) Process

MA(∞) is a special case deserves attention. Partly because all ARMA processes can be reduced to MA(∞) processes. In addition to MA(q) processes, we need more conditions for MA(∞) to be stationary. Consider the variance of

$$y_t = \sum_{j=0}^{\infty} \theta_j \epsilon_{t-j},$$

$$\gamma_0 = \mathbb{E}[y_t^2] = \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \theta_j \epsilon_{t-j}\right)^2\right] = \left(\sum_{j=0}^{\infty} \theta_j^2\right)\sigma^2.$$

It only make sense if $\sum_{j=0}^{\infty} \theta_j^2 < \infty$. This property is called *square summable.*

Proposition 7.4. *An MA(∞) process is stationary if the coefficients $\{\theta_j\}$ are square summable.*

8 ARMA Models

8.1 ARMA(p,q)

ARMA(p, q) is a mixed autoregressive and moving average process.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q},$$

or

$$\phi(L) y_t = \theta(L) \epsilon_t,$$

where $\{\epsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$.

The MA part is always stationary as shown in Proposition 7.1. The stationarity of an ARMA process solely depends on the AR part. The condition is the same as Proposition 6.2.

Assume $\phi^{-1}(L)$ exist, then the ARMA(p,q) process can be reduce to MA(∞) process:

$$y_t = \phi^{-1}(L)\theta(L)\epsilon_t = \psi(L)\epsilon_t,$$

where $\psi(L) = \phi^{-1}(L)\theta(L)$.

Exercise

Compute the MA equivalence for ARMA(1,1).

8.2 ARIMA(p,d,q)

ARMA(p,q) is used to model stationary time series. If y_t is not stationary, we can transform it to stationary and model it with an ARMA model. If the first-order difference $(1-L)y_t = y_t - y_{t-1}$ is stationary, then we say y_t is **integrated** of order 1. If it requires d-th order difference to be stationary, $(1-L)^d y_t$, we say it is integrated of order d. The ARMA model involves integrated time series is called ARIMA model:

$$\phi(L)(1-L)^d y_t = \theta(L)\epsilon_t.$$

9 Wold Theorem

9.1 Wold Decomposition

So far we have spent a lot of effort with ARMA models, which are the indispensable components of any time series textbook. The following theorem justifies its importance. The Wold Decomposition Theorem basically says every covariance-stationary process has an ARMA representation. Therefore, with long enough lags, any covariance-stationary process can be approximated arbitrarily well by ARMA models. This is a very bold conclusion to make. It sets up the generality of ARMA models, which makes it one of the most important theorems in time series analysis.

Theorem 9.1 (Wold Decomposition Theorem)**.** *Every covariance-stationary time series y_t can be written as the sum of two time series, one* deterministic *and one* stochastic*. Formally,*

$$y_t = \eta_t + \sum_{j=0}^{\infty} b_j \epsilon_{t-j},$$

where $\eta_t \in I_{-\infty}$ is a deterministic time series (such as one represented by a sine wave); ϵ_t is an uncorrelated innovation sequence with $\mathbb{E}[\epsilon_t] = 0$, $\mathbb{E}[\epsilon_t \epsilon_{t-j}] = 0$ for $j \neq 0$; and $\{b_j\}$ are square summable, $\sum_{j=0}^{\infty} |b_j|^2 < \infty$.

Proof. We will prove the theorem by constructing the innovation sequence $\{e_t\}$ and showing it satisfies the conditions stated. Let $e_t = y_t - \hat{\mathbb{E}}(y_t|I_{t-1}) = y_t - a(L)y_{t-1}$, where $\hat{\mathbb{E}}(y_t|I_{t-1})$ is the best linear predictor (BLP) of y_t based on information set at $t-1$. $a(L)$ does not depend on t because y_t is covariance stationary. As the best linear predictor, $a(L)$ solves

$$\min_{\{a_j\}} \mathbb{E}(y_t - \sum_{j=1}^{\infty} a_j y_{t-j})^2.$$

The first-order conditions with respect to a_j gives

$$
\begin{aligned}
\mathbb{E}[y_{t-j}(y_t - \sum_{j=1}^{\infty} a_j y_{t-j})] &= 0, \\
\implies \mathbb{E}[y_{t-j} e_t] &= 0.
\end{aligned}
$$

We now verify that e_t satisfies the white noise conditions. Without loss of generality, we may assume $\mathbb{E}(y_t) = 0$, it follows that $\mathbb{E}(e_t) = 0$. $\text{var}(e_t) = \mathbb{E}(y_t - a(L)y_t)^2$ is a function of covariance of y_t and a_j, none of which varies with time. So $\text{var}(e_t) = \sigma^2$ is constant. Utilizing the first-order condition, $\mathbb{E}[e_t e_{t-j}] = \mathbb{E}[e_t(y_{t-j} - a(L)y_{t-j})] = 0$.

Repeatedly substituting for y_{t-k} gives

$$
\begin{aligned}
y_t &= e_t + \sum_{k=1}^{\infty} a_k y_{t-k} \\
&= e_t + a_1(e_{t-1} + \sum_{k=1}^{\infty} a_k y_{t-1-k}) + \sum_{k=2}^{\infty} a_k y_{t-k} \\
&= e_t + a_1 e_{t-1} + \sum_{k=1}^{\infty} \tilde{a}_k y_{t-k-1} \\
&= e_t + a_1 e_{t-1} + \eta_t^1 \\
&\vdots \\
&= \sum_{j=0}^{k} c_j e_{t-j} + \eta_t^k,
\end{aligned}
$$

where $\eta_t^k \in I_{t-k-1}$. As $k \to \infty$, we have $v_t = y_t - \sum_{j=0}^{\infty} c_j e_{t-j} \in I_{-\infty}$.

□

Let's appreciate this theorem for a while. The property of stationarity can be loosely understood as having stable patterns over time. The Wold Theorem states that any such patterns can be captured by ARMA models. In other words, ARMA models are effective in modelling stable patterns repeated over time, in so far as only 2nd-order moments are of concern. Even if the time series is not entirely stationary, if we model it with ARMA, it can be thought as extracting the stationary patterns. Figure 9.1 demonstrates the ARIMA modelling of monthly export.

```r
library(zoo)
data = read.csv.zoo("data/md.csv", FUN = as.yearmon, regular = TRUE)
y = data$Export
mod = arima(y, order = c(2,0,1))
yhat = y - mod$residuals
```

```
plot(cbind(y, yhat), plot.type = "s", col = 1:2, ann = F)
```

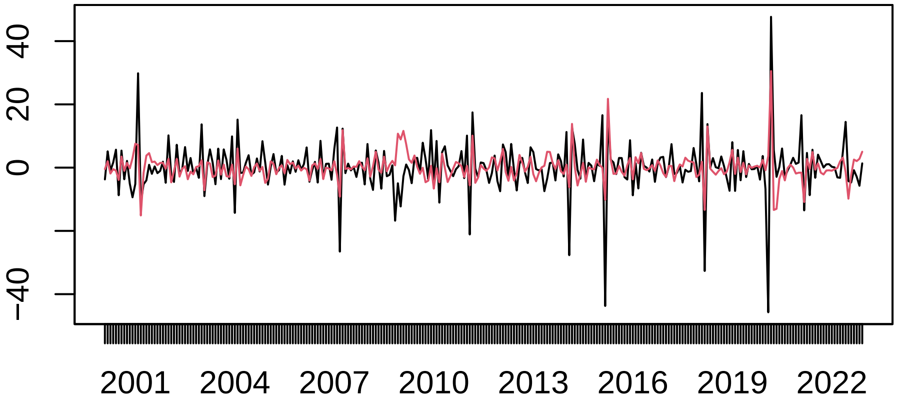

Figure 9.1: Monthly export modelled with ARIMA(2,0,1)

9.2 Causality and Invertibility*

We have seen that AR models can be rewritten as MA models and vice versa, suggesting the ARMA representation of a stochastic process is not unique. We have also seen that a non-invertible MA process can be equivalently represented by an invertible MA process. For example, the following MA(1) processes have the same ACF:

$$\begin{aligned} x_t &= w_t + \frac{1}{5}w_{t-1}, \quad w_t \sim \text{WN}(0, 25); \\ y_t &= v_t + 5v_{t-1}, \qquad v_t \sim \text{WN}(0, 1). \end{aligned}$$

The same property holds for AR processes. In Chapter 6, we state that an AR(1) process is explosive if $|\phi| > 1$. This is not entirely rigorous. Consider an AR(1) process,

$$y_t = \phi y_{t-1} + \epsilon_t, \text{ where } |\phi| > 1.$$

Multiply both sides by ϕ^{-1},

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

Rewrite it as an MA process,

$$
\begin{aligned}
y_t &= \phi^{-1} y_{t+1} - \phi^{-1}\epsilon_{t+1} \\
&= \phi^{-1}(\phi^{-1} y_{t+2} - \phi^{-1}\epsilon_{t+2}) - \phi^{-1}\epsilon_{t+1} \\
&\vdots \\
&= \sum_{j=1}^{\infty} -\phi^{-j}\epsilon_{t+j}.
\end{aligned}
$$

Given $|\phi^{-1}| < 1$, the process is stationary, expressed as discounted innovations in the future (despite this looks quite odd). In fact, for an non-causal AR process, we can find a causal AR process that generates the same ACF (remember the term *causal* means an AR process can be converted to an MA process with absolute summable coefficients).

The problem is given an ARMA equation, it is not enough to uniquely pin down a stochastic process. Both the explosive process and the stationary process can be a solution to $y_t = \phi y_{t-1} + \epsilon_t$. But for a stationary process expressed as an AR model with $|\phi| > 1$, we can always find an AR(1) process with $|\tilde{\phi}| < 1$ and a different white noise sequence $\{\tilde{\epsilon}_t\}$ that generate the same ACF.

The following theorems state the conditions for the existence of stationary solutions, and the possibility of rewriting non-causal or non-invertible ARMA representations as causal and invertible ones. Since it is always possible to do so, it loses nothing to stick with causal and invertible ARMA processes when modelling stationary time series.

Theorem 9.2. *A unique stationary solution to the ARMA process $\phi(L)y_t = \theta(L)\epsilon_t$ exists iff ϕ and θ have no common factors and the roots of $\phi(z)$ avoid the unit circle:*

$$
|\phi(z)| = 1 \implies \phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0.
$$

Theorem 9.3. *Let $\{y_t\}$ be a stationary ARMA process defined by $\phi(L)y_t = \theta(L)\epsilon_t$. If the roots of $\theta(z)$ avoid unit circle, then there are polynomials $\tilde{\phi}$ and $\tilde{\theta}$ and a white noise sequence $\tilde{\epsilon}$ such that $\{y_t\}$ satisfies $\tilde{\phi}(L)y_t = \tilde{\theta}(L)\tilde{\epsilon}_t$, and this is a causal and invertible ARMA process.*

Part III

Time Series Regression

10 Preliminaries

10.1 Chapter Overview

This chapter serves two purposes. One is to introduce the techniques for estimating time series models. The other is to explain the concept of dynamic causal effect. We join the two topics in one chapter because both of them can be done via a regression framework. Maximum likelihood estimation plays a pivotal role in estimating time series models. Nonetheless, starting with OLS always make things easier. We start with a quick review of the basic OLS concepts that are familiar to any students in econometrics, that is the regressions applied to cross-sectional *iid* observations. We then extend it to time series data. We will see it is not as straightforward as one might expect, as intertemporal dependencies between observation need additional treatment. In the second half of the chapter, we will explain the concept of dynamic causal effect, that is the causal effect of an intervention on outcome variables. Similar to cross-sectional studies, we need to define the causal effect relative to counterfactuals. With time series data, the counterfactuals have to be defined across time rather across individuals.

10.2 Asymptotic Theorems for i.i.d Random Variables

Theorem 10.1 (Law of Large Numbers)**.** *Let $\{x_i\}$ be iid random variables with $\mathbb{E}(x_i) = \mu$ and Var$(x_i) = \sigma^2 < \infty$. Define $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. Then $\bar{x}_n \overset{p}{\to} \mu$ as $n \to \infty$.*

Proof. We will give an non-rigorous proof, but nonetheless shows the tenets. It is easy to see $\mathbb{E}(\bar{x}_n) = \mu$. Consider the variance,

$$\text{Var}(\bar{x}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^n x_i\right) \overset{iid}{=} \frac{1}{n^2}\sum_{i=1}^n \text{Var}(x_i) = \frac{\sigma^2}{n} \to 0.$$

That is $\bar{x}_n$ converges to μ with probability 1 as $n \to \infty$. Note that we can move the variance inside the summation operator because x_i are *iid*, in which all the covariance terms are 0.

□

Theorem 10.2 (Central Limit Theorem)**.** *Let $\{x_i\}$ be iid random variables with $\mathbb{E}(x_i) = \mu$ and $Var(x_i) = \sigma^2 < \infty$. Define $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. Then*

$$\frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

Proof. Without loss of generality, assume x_i is demeaned and standardized to have standard deviation 1. It remains to show $\sqrt{n}\bar{x}_n \to N(0,1)$. Define the moment generating function (MGF) for $\sqrt{n}\bar{x}_n$:

$$M_{\sqrt{n}\bar{x}_n}(t) = \mathbb{E}[e^{(\sqrt{n}^{-1}\sum_{i=1}^n x_i)t}] \overset{iid}{=} \{\mathbb{E}[e^{(n^{-1/2}x_i)t}]\}^n.$$

Evaluate the MGF for each x_i:

$$\mathbb{E}[e^{(n^{-1/2}x_i)t}] = 1 + \mathbb{E}(n^{-1/2}x_i)t + \mathbb{E}(n^{-1}x_i^2)t^2 + \cdots = 1 + \frac{t^2}{2n} + o(n^{-1}).$$

Substituting back,

$$M_{\sqrt{n}\bar{x}_n}(t) = \left[1 + \frac{t^2}{2n} + o(n^{-1})\right]^n = \left[\left(1 + \frac{t^2}{2n}\right)^{\frac{2n}{t^2}}\right]^{\frac{t^2}{2}} \to e^{\frac{t^2}{2}}.$$

Note that we drop the $o(n^{-1})$ because it converges faster than $\frac{1}{n}$. $e^{\frac{t^2}{2}}$ is the MGF for standard normal distribution. Hence, the theorem is proved.

□

10.3 OLS for i.i.d Random Variables

We now give a very quick review of OLS with *iid* random variables. These materials are assumed familiar to the readers. We do not intend to introduce them in any detail. This section is a quick snapshot of some key concepts, so that we could contrast them with the time series regression introduced in the next section.

A linear regression model postulates the joint distribution of $(y_i, \mathbf{x_i})$ follows a linear relationship,

$$y_i = \mathbf{x_i}' + \epsilon_i.$$

Expressed in terms of data matrix,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11}, x_{12}, \dots, x_{1p} \\ x_{21}, x_{22}, \dots, x_{2p} \\ \ddots \\ x_{n1}, x_{n2}, \dots, x_{np} \end{bmatrix}' \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

From the perspective of dataset, the matrix matrix is fixed in the sense that they are just numbers in the dataset. But for statistical analysis, we view each entry in the matrix as random, that is as a realization of a random process.

To estimate the parameter   from sample data, OLS seeks to minimize the squared residuals

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x_i}' \ )^2.$$

The first-order condition implies,

$$\begin{aligned} \sum_i \mathbf{x_i}(y_i - \mathbf{x_i}' \ ) &= 0, \\ \sum_i \mathbf{x_i} y_i - \sum_i \mathbf{x_i}\mathbf{x_i}' \ &= 0, \\ \hat{\ } &= \left(\sum_i \mathbf{x_i}\mathbf{x_i}' \right)^{-1} \left(\sum_i \mathbf{x_i} y_i \right) \\ &= \beta + \left(\sum_i \mathbf{x_i}\mathbf{x_i}' \right)^{-1} \left(\sum_i \mathbf{x_i} \epsilon_i \right). \end{aligned}$$

Under the Gauss-Markov assumptions, particularly $\mathbb{E}(\epsilon_i|\mathbf{x_j}) = 0$ and $\mathrm{var}(\ |\mathbf{X}) = \sigma^2 \mathbf{I}$ (homoskedasticity and nonautocorrelation), the OLS estimator is **BLUE** (Best Linear Unbiased Estimator).

Under the assumption of *iid* random variables and homoskedasticity, we invoke the LLN and CLT to derive the asymptotic distribution for the OLS estimator,

$$\begin{aligned} \sqrt{n}(\hat{\ } - \ ) &= \left(\frac{1}{n} \sum_i \mathbf{x_i}\mathbf{x_i}' \right)^{-1} \left(\sqrt{n} \frac{1}{n} \sum_i \mathbf{x_i} \epsilon_i \right) \\ &\to [\mathbb{E}(\mathbf{x_i}\mathbf{x_i}')]^{-1} \mathrm{N}(0, \mathbb{E}(\mathbf{x_i}\epsilon_i\epsilon_i'\mathbf{x_i}')) \\ &\to \mathrm{N}(0, \sigma^2 [\mathbb{E}(\mathbf{x_i}\mathbf{x_i}')]^{-1}). \end{aligned}$$

Note how the *iid* assumption is required throughout the process. The following section will show how to extend the OLS to non-*iid* random variables and how it leads to modification of the results.

11 OLS for Time Series

11.1 Asymptotic Theorems for Dependent Random Variables

The asymptotic theorems and regressions that work for *iid* random variable do not immediately apply to time series. Consider the proof for Theorem 10.1, without the *iid* assumption we have

$$\begin{aligned}
\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(x_i, x_j)\\
&= \frac{1}{n^2}[\operatorname{cov}(x_1,x_1)+\operatorname{cov}(x_1,x_2)+\dots+\operatorname{cov}(x_1,x_n)+\\
&\quad\ \operatorname{cov}(x_2,x_1)+\operatorname{cov}(x_2,x_2)+\dots+\operatorname{cov}(x_2,x_n)+\\
&\quad\ \vdots\\
&\quad\ \operatorname{cov}(x_n,x_1)+\operatorname{cov}(x_n,x_2)+\dots+\operatorname{cov}(x_n,x_n)]\\
&= \frac{1}{n^2}[n\gamma_0+2(n-1)\gamma_1+2(n-2)\gamma_1+2(n-2)\gamma_2+\dots]\\
&= \frac{1}{n}\left[2\sum_{k=1}^{n}\gamma_k\left(1-\frac{k}{n}\right)+\gamma_0\right].
\end{aligned}$$

The argument for the *iid* does not work with the presence of serial correlations. If we assume absolute summability, $\sum_{j=-\infty}^{\infty}|\gamma_j|<\infty$, then

$$\lim_{n\to\infty}\frac{1}{n}\left[2\sum_{k=1}^{n}\gamma_k\left(1-\frac{k}{n}\right)+\gamma_0\right]=0.$$

In this case, we still have the LLN holds. Otherwise, as the variance may not converge. Remember Theorem 4.1, absolute summability implies the series is ergodic.

Proposition 11.1. *If x_t is a covariance stationary time series with absolutely summable auto-covariances, then a Law of Large Numbers holds.*

From the new proof of LLN one can guess that the variance in a Central Limit Theorem should also change. The serially correlated x_t, the liming variance is given by

$$\begin{aligned} \operatorname{var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\right) &= 2\sum_{k=1}^{n}\gamma_k\left(1-\frac{k}{n}\right)+\gamma_0 \\ &\to 2\sum_{k=1}^{\infty}\gamma_k+\gamma_0=\sum_{k=-\infty}^{\infty}\gamma_k=S. \end{aligned}$$

We call S the *long-run variance*. There are many CLTs for serially correlated observations. We give the two mostly commonly cited versions: one applies to MA(∞) processes, the other one is more general.

Theorem 11.1. *Let y_t be an MA process: $y_t=\mu+\sum_{j=0}^{\infty}c_j\epsilon_{t-j}$ where ϵ_t is independent white noise and $\sum_{j=0}^{\infty}|c_j|<\infty$ (this implies ergodic), then*

$$\sqrt{T}\bar{y}_t \xrightarrow{d} N(0,S),$$

where $S=\sum_{k=-\infty}^{\infty}\gamma_k$ is the long-run variance.

Theorem 11.2 (Gordin's CLT)**.** *Assume we have a strictly stationary and ergodic series $\{y_t\}$ with $\mathbb{E}(y_t^2)<\infty$ satisfying: $\sum_j\{\mathbb{E}[\mathbb{E}[y_t|I_{t-j}]-\mathbb{E}[y_t|I_{t-j-1}]]^2\}^{1/2}<\infty$ and $\mathbb{E}[y_t|I_{t-j}]\to 0$ as $j\to\infty$, then*

$$\sqrt{T}\bar{y}_t \xrightarrow{d} N(0,S),$$

where $S=\sum_{k=-\infty}^{\infty}\gamma_k$ is the long-run variance.

The Gordin's conditions are intended to make the dependence between distant observations to decrease to 0. ARMA process is a special case of Gordin series. The essence of these theorems is that we need some restrictions on dependencies for LLN and CLT to hold. We allow serial correlations as long as they are not too strong. If the observations become almost independent as they are far away in time, the can still apply the asymptotic theorems.

11.2 OLS for Time Series

Definition 11.1. Given a time series regression model

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

x_t is **weakly exogenous** if

$$\mathbb{E}(\epsilon_t | x_t, x_{t-1}, ...) = 0;$$

x_t is **strictly exogenous** if

$$\mathbb{E}(\epsilon_t | \{x_t\}_{t=-\infty}^{\infty}) = 0.$$

Strictly exogeneity requires innovations being exogenous from all past and future regressors; while weakly exogeneity only requires being exogenous from past regressors. In practice, strict exogeneity is too strong as an assumption. The weak exogenous is more practical and it is enough to ensure the consistency of the OLS estimator.

The OLS estimator is as usual:

$$\hat{\beta} = \beta + \left(\frac{1}{n} \sum_t x_t x_t' \right)^{-1} \left(\frac{1}{n} \sum_t x_t \epsilon_t \right).$$

Assuming LLN holds and x_t is weakly exogenous, we have

$$\begin{aligned} &\frac{1}{n} \sum_t x_t x_t' \to \mathbb{E}(x_t x_t') = Q, \\ &\frac{1}{n} \sum_t x_t \epsilon_t \to \mathbb{E}(x_t \epsilon_t) = \mathbb{E}[x_t \mathbb{E}[\epsilon_t | x_t]] = 0. \end{aligned}$$

Therefore, $\hat{\beta} \to \beta$. The OLS estimator is *consistent.*

Assuming the Gordin's conditions hold for $z_t = x_t \epsilon_t$, the CLT gives

$$\frac{1}{\sqrt{n}} \sum_t x_t \epsilon_t \to N(0, S),$$

where $S = \sum_{-\infty}^{\infty} \gamma_j$ is the long-run variance for z_t. Thus, we have the asymptotic normality for the OLS estimator

$$\sqrt{T}(\hat{\beta} - \beta) \to N(0, Q^{-1}SQ^{-1}).$$

Note how the covariance matrix S is different from the one in the *iid* case where $S = \sigma^2 \mathbb{E}(x_i x_i')$. The long-run variance S takes into account the auto-dependencies between observations. The auto-dependencies usually arise from the serially correlated error terms. It may also arise from x_t being autocorrelated and from conditional heteroskedasticity of the error terms. Because of the auto-covariance structure, S cannot be estimated in the same way as in the *iid* case. The estimator for S is called HAC (heteroskedasticity autocorrelation consistent) standard errors.

11.3 HAC Standard Errors

S can be estimated with truncated autocovariances,

$$\hat{S} = \sum_{j=-h(T)}^{h(T)} \hat{\gamma}_j.$$

$h(T)$ is a function of T and $h(T) \to \infty$ as $T \to \infty$, but more slowly. Because we don't want to include too many imprecisely estimated covariances. Another problem is the estimated $\hat{S}$ might be negative. The solution is weight the covariances in a way to ensure positiveness:

$$\hat{S} = \sum_{j=-h(T)}^{h(T)} k_T(j)\hat{\gamma}_j.$$

$k_T(\cdot)$ is called a kernel. The weights are chosen to guarantee positive-definiteness by weighting down high lag covariances. Also we need $k_T(\cdot) \to 1$ for consistency.

A popular HAC estimator is the Newey-West variance estimator, in which $h(T) = 0.75T^{1/3}$ and $k_T(j) = \frac{h-j}{h}$, so that

$$\hat{S} = \sum_{j=-h}^{h} \left(\frac{h-j}{h}\right) \hat{\gamma}_j.$$

11.4 Example

Note that all of our discussions in this chapter apply only to stationary time series. Without stationarity, even the autocovariance γ_j might not be well-defined. In the following example, we generate artificial data from an AR(2) process, and recover the parameters by regression y_t on its lags.

```r
library(lmtest)
y = arima.sim(list(ar = c(0.5, 0.3)), n = 1000)
mod = lm(y ~ ., data = cbind(y, lag(y,-1), lag(y,-2)))
coeftest(mod, vcov. = sandwich::NeweyWest(mod))
```

```

t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.017428   0.030538  0.5707   0.5683
`lag(y, -1)` 0.496021   0.031721 15.6367   <2e-16 ***
`lag(y, -2)` 0.298728   0.028794 10.3745   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12 Summary

In summary, this book has no content whatsoever.

References

Hamilton, James D. 1994. *Time Series Analysis.* Princeton University Press.

Hansen, Bruce. 2022. *Econometrics.* Princeton University Press.

Hayashi, Fumio. 2011. *Econometrics.* Princeton University Press.

Hyndman, Rob J, and George Athanasopoulos. 2018. *Forecasting: Principles and Practice (2nd Edition).* OTexts.com/fpp2.

Mikusheva, Anna, and Paul Schrimpf. 2007. *14.384 Time Series Analysis.* MIT OpenCourseWare.

Verbeek, Marno. 2008. *A Guide to Modern Econometrics.* John Wiley & Sons.