Time Series Analysis for Economists

Zem Wang

1/1/24

Table of contents

Preface

Empirical macroeconomics frequently involves the analysis of time series data, encompassing variables such as GDP, inflation, and interest rates, employing methodologies distinct from those utilized in cross-sectional studies. The goal of this book is to bridge the gap between introductory time series textbooks and theoretical econometrics. In the realm of empirical research, a rudimentary comprehension of the subject matter often proves insufficient. While computational tasks can be executed through simple computer commands, practitioners must go beyond surface-level in order to understand the intricacies and limitations of the involved techniques. Conversely, an exhaustive exploration of advanced econometric theories would be excessive for practical purposes. For instance, introductory textbooks would caution against applying OLS on non-stationary time series, citing the potential risk of spurious regression. Students often accept this as a rule of thumb without a grasp of its underlying rationale. Yet, delving into intricate topics such as Itô calculus is deemed unnecessary for empirical researchers.

This book seeks to acquaint readers with essential time series topics crucial for understanding and conducting empirical research, with a specific focus on macroeconomic applications. In addition to introducing basic concepts and applications, such as running a regression and interpreting the result, the book endeavors to elevate comprehension to a deeper level by elucidating the "why" alongside the "what" and "how." However, the objective is not to provide an exhaustive treatment replete with formal proofs; rather, emphasis is placed on providing intuitive explanations. Consequently, readers may encounter instances of non-rigorous proofs where a more formal approach is deemed unnecessary for an understanding required by applied works. This book can be read as intermediary materials 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, which are listed in the References at the end of the book without being cited 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 A and B. Vectors and matrices are sometimes written in bold for emphasizing, such as $\boldsymbol{X}$ and $\boldsymbol{y}_t$, or in plain format as scalars if that does not lead to confusion. 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.

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.

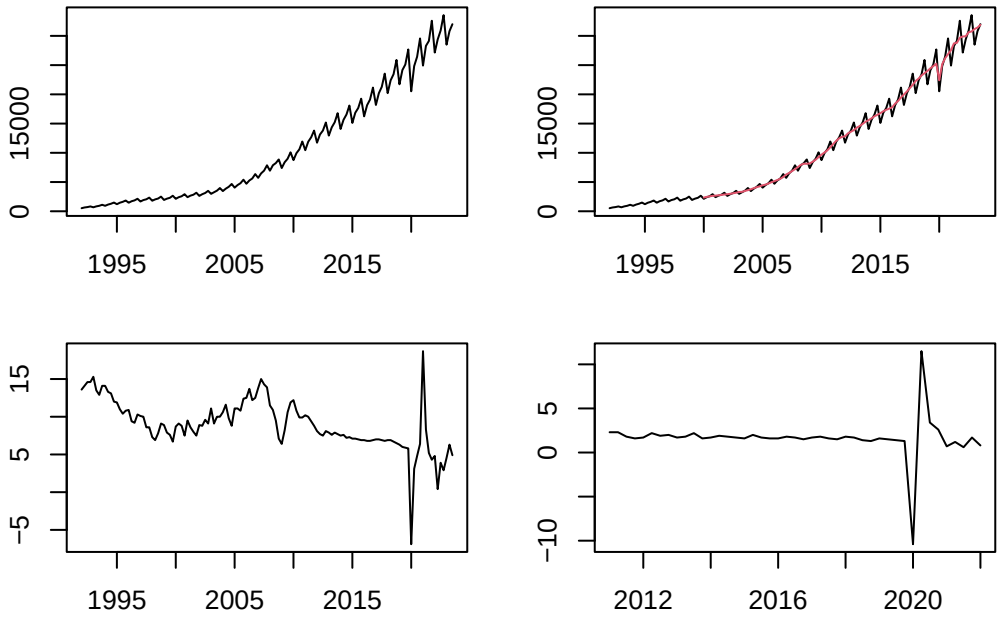


Figure 1.1: Quarterly GDP Time Series (Unit: RMB Billion or %)

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.

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

Exercise

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

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

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

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

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,

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

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.

```
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)
```

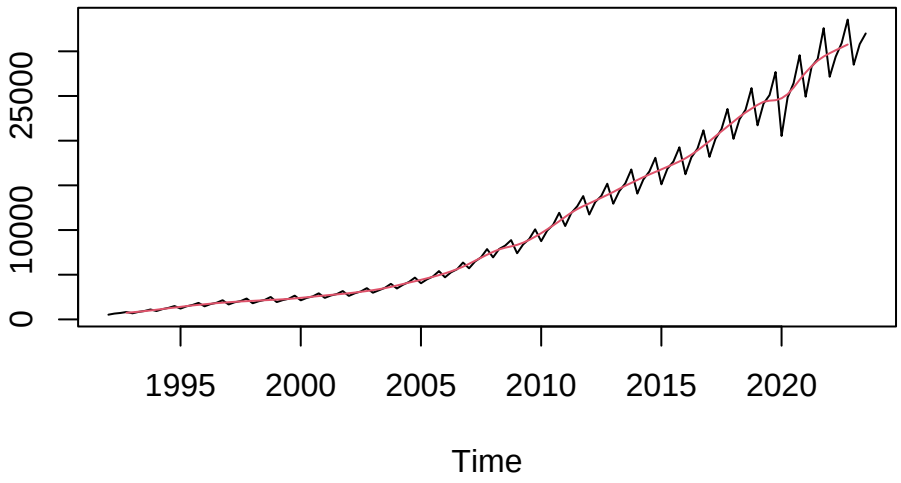


Figure 2.1: Quarterly GDP with 2x4-MA estimate of the trend-cycle

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()
```

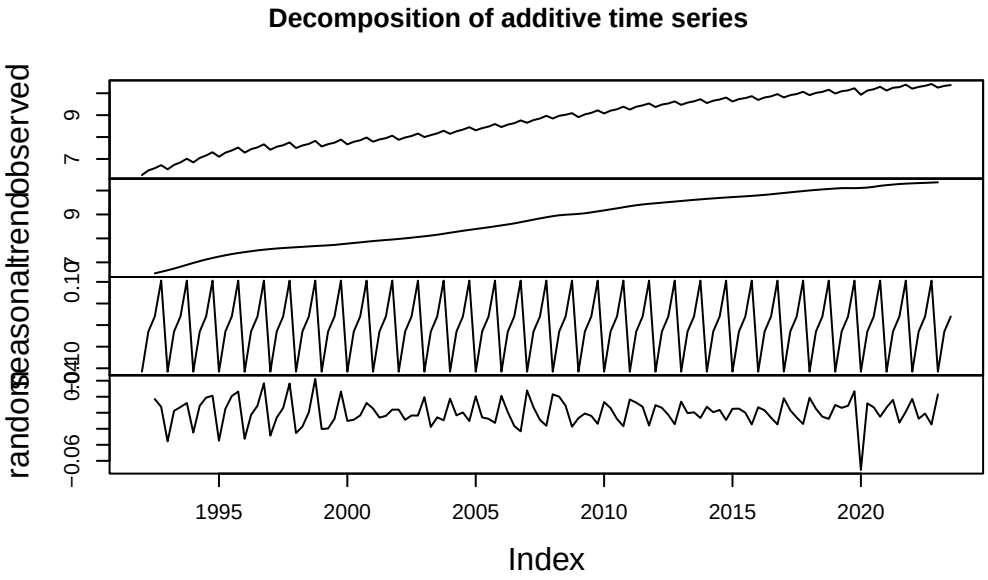


Figure 2.2: Classical multiplicative decomposition of quarterly GDP

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)
```

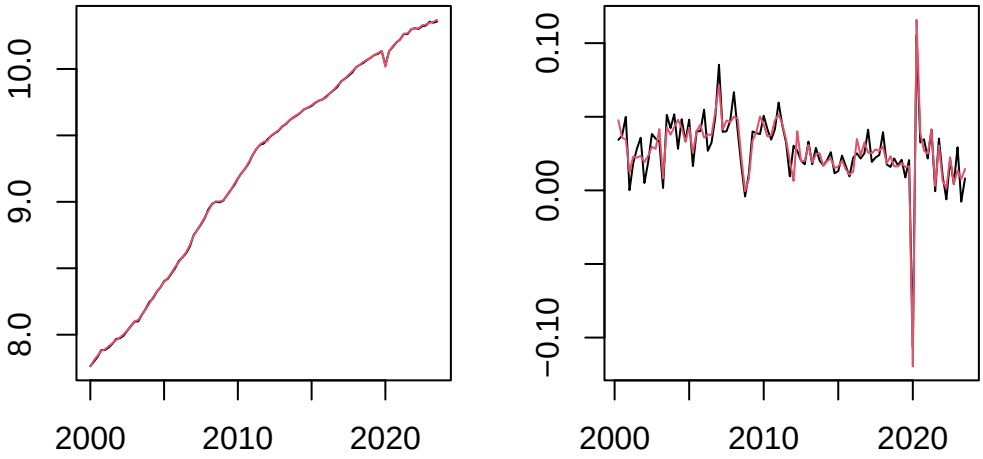


Figure 2.3: Comparing classical decomposition and X-13

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.

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

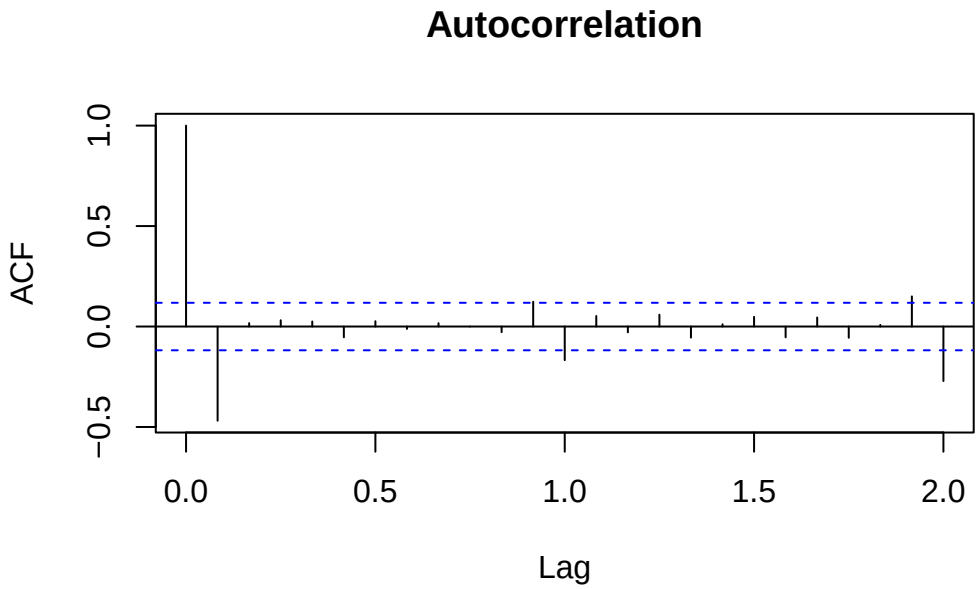


Figure 3.1: ACF for monthly export growth

3.2 Partial Autocorrelation

ACF measures the correlation between y_t and y_{t-k} regardless of their relationships with the intermediate variables $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$. 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 $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$.

$$\phi_k = \begin{cases} \text{corr}(y_t, y_{t-1}) = \rho_1, & \text{if } k = 1; \\ \text{corr}(r_{y_t|y_{t-1},\ldots,y_{t-k+1}}, r_{y_{t-k}|y_{t-1},\ldots,y_{t-k+1}}), & \text{if } k \geq 2; \end{cases}$$

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')
```

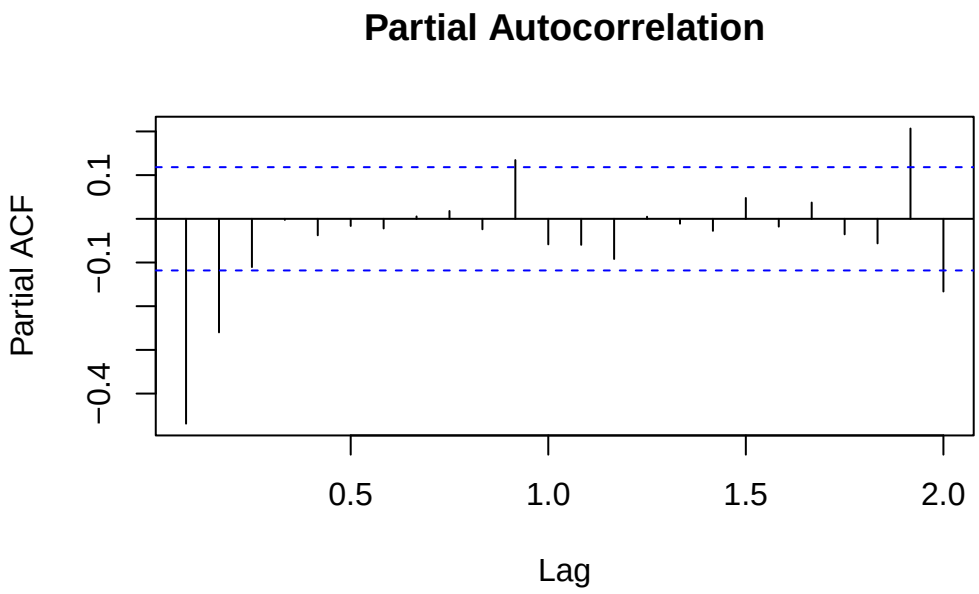


Figure 3.2: PACF for monthly export growth

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 $k = 1, 2, 3, \ldots$

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

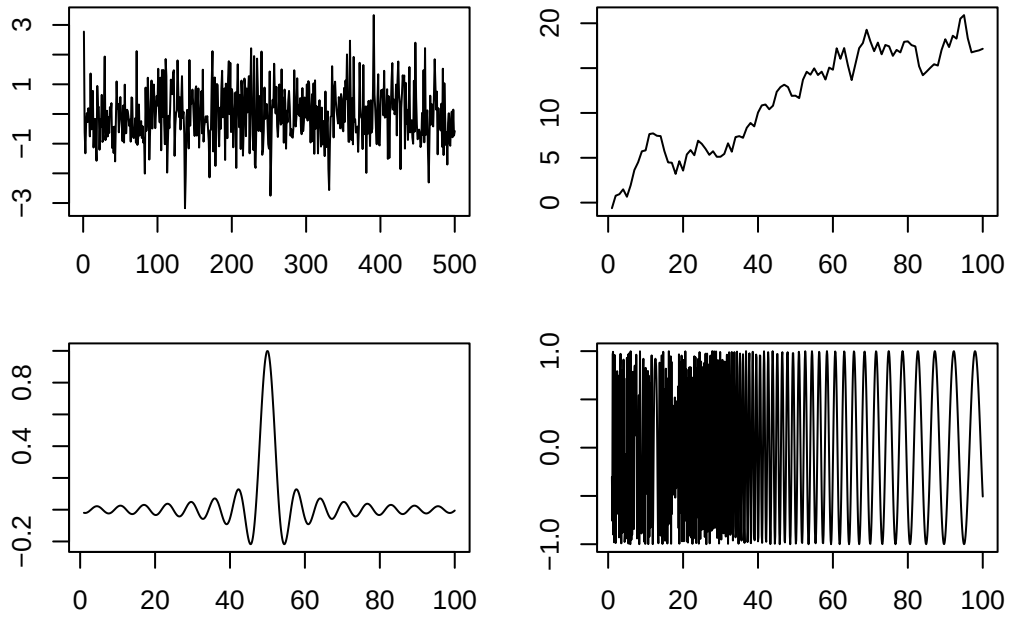


Figure 4.1: Stationary and non-stationary time series

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

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

□

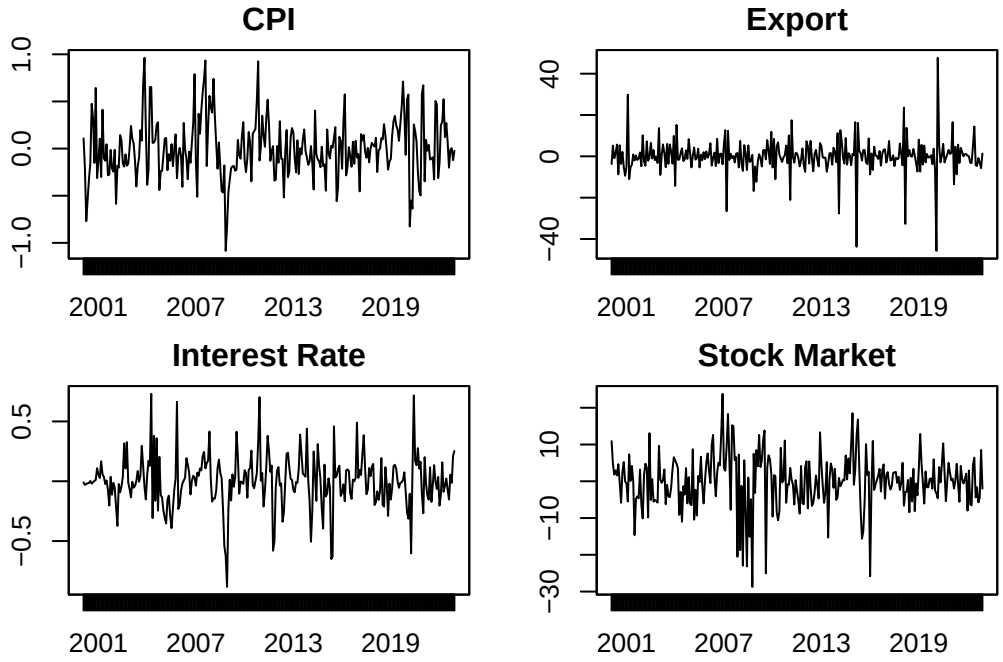


Figure 4.2: Stationary and non-stationary time series (real life)

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

```
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))
```

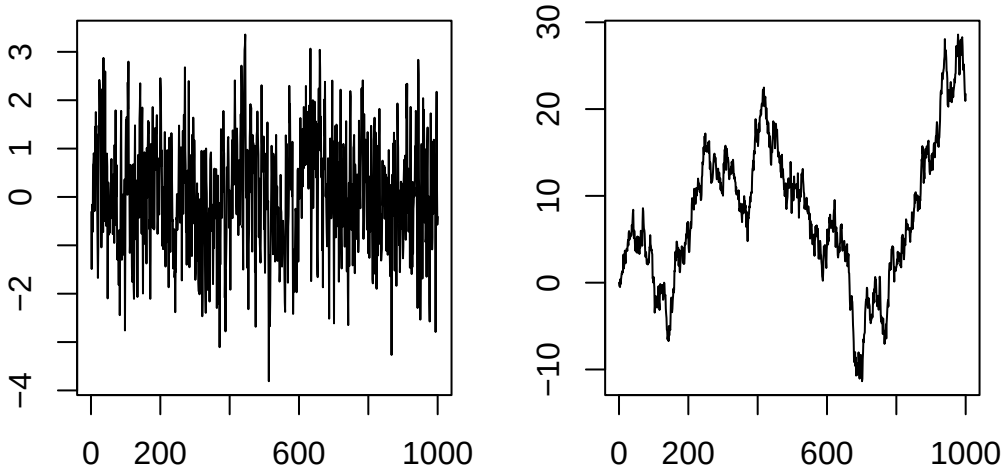


Figure 6.1: Simulation of AR(1) processes

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,

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

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

$$
\vdots
$$

$$
\gamma_j = \frac{\phi^j \sigma^2}{1 - \phi}.
$$

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)
```

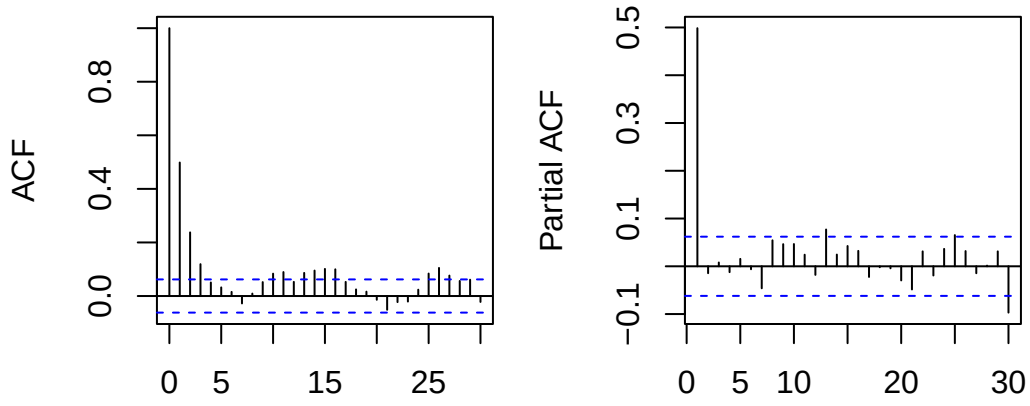


Figure 6.2: ACF and PACF for AR(1) process

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

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

This is called the Yule-Walker equation. The first p unknowns $\gamma_0, \ldots, \gamma_{p-1}$ 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)
```

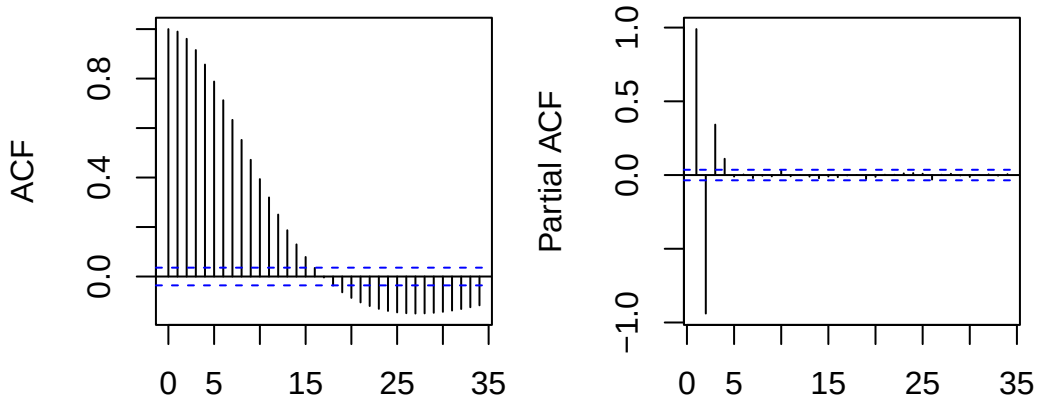


Figure 6.3: ACF and PACF for AR(p) process

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 $y_{t-j}, \ldots, y_{t-1}$). 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:

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

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 $\epsilon_{t-k}, \epsilon_{t-k-1}, ...$ 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)
```

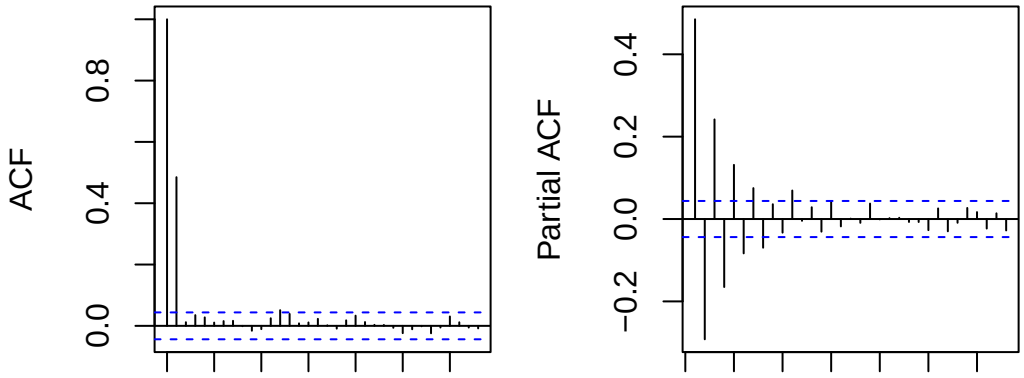


Figure 7.1: ACF and PACF of MA(1) process

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.

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

□

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

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

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

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.

```
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)
```

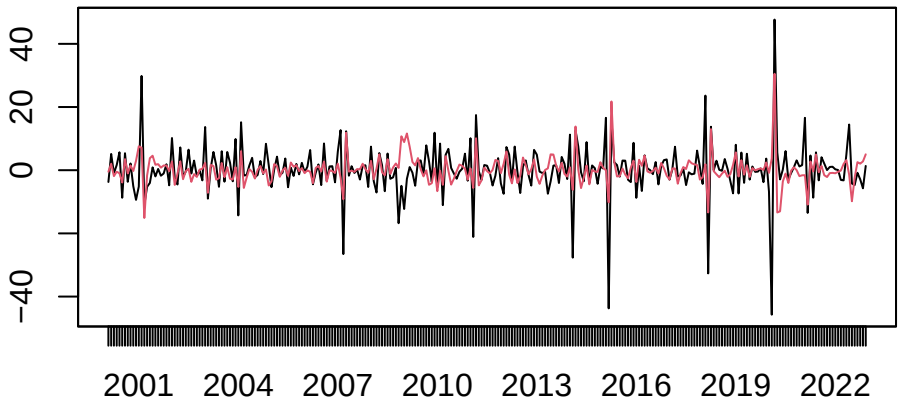


Figure 9.1: Monthly export modelled with ARIMA(2,0,1)

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:

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

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* $\bar{x}_n \xrightarrow{p} \mu$ *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,

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

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, x_i) follows a linear relationship,

$$y_i = x_i'\beta + \epsilon_i.$$

Expressed in terms of data matrix,

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

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 β from sample data, OLS seeks to minimize the squared residuals

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

The first-order condition implies,

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

Under the Gauss-Markov assumptions, particularly $\mathbb{E}(\epsilon_i|x_j) = 0$ and $\text{var}(\epsilon|X) = \sigma^2 I$ (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,

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1} \left(\sqrt{n}\frac{1}{n}\sum_i x_i \epsilon_i\right)$$

$$\to [\mathbb{E}(x_i x_i')]^{-1} \mathrm{N}(0, \mathbb{E}(x_i \epsilon_i \epsilon_i' x_i'))$$

$$\to \mathrm{N}(0, \sigma^2 [\mathbb{E}(x_i x_i')]^{-1}).$$

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

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

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

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

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

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

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.

```
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)  0.010406   0.029505  0.3527   0.7244
`lag(y, -1)` 0.506038   0.030291 16.7060   <2e-16 ***
`lag(y, -2)` 0.270190   0.028169  9.5917   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12 MLE for ARMA Models

OLS can only be used to estimate AR models, but not MA models. MA models or ARMA models in general can be estimated using maximum likelihood approach. Maximum likelihood estimation (MLE) starts with an assumed distribution of the random variables. The parameters are chosen to maximize the likelihood of observing the data under the distribution.

Consider an ARMA(p, q) model

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

Write in the form of data matrix:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_T \end{bmatrix} = \underbrace{\begin{bmatrix} y_0 & y_{-1} & \cdots & y_{1-p} \\ y_1 & y_0 & \cdots & y_{2-p} \\ y_2 & y_1 & \cdots & y_{3-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_T & y_{T-1} & \cdots & y_{T-p} \end{bmatrix}}_{\boldsymbol{X}} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_p \end{bmatrix} + \underbrace{\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \theta_1 & 1 & 0 & \ldots & 0 \\ \theta_2 & \theta_1 & 1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & \ldots & \theta_2 & \theta_1 & 1 \end{bmatrix}}_{\boldsymbol{\Gamma}} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_T \end{bmatrix}$$

Or compactly,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{\Gamma}\boldsymbol{u}.$$

We assume the innovations are jointly normal $\boldsymbol{u} \sim N(0, \sigma^2 \boldsymbol{I})$. We also assume the first p observations are known initial values $y_0, y_{-1}, \ldots, y_{1-p}$ and $u_0 = u_{-1} = \cdots = u_{1-q} = 0$. Therefore, the observed data are jointly normal given the initial condition,

$$\boldsymbol{y}|\boldsymbol{y_0} \sim N(\boldsymbol{X}\boldsymbol{\phi}, \sigma^2 \boldsymbol{\Gamma}\boldsymbol{\Gamma'}).$$

The probability density function for multivariate normal is

$$f(\boldsymbol{y}|\boldsymbol{y_0}, \boldsymbol{\phi}, \boldsymbol{\Gamma}, \sigma^2) = (2\pi)^{-T/2} |\sigma^2 \boldsymbol{\Gamma}\boldsymbol{\Gamma'}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\phi})'(\sigma^2 \boldsymbol{\Gamma}\boldsymbol{\Gamma'})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\phi})\right)$$

To simplify computation, take logarithm to get the log-likelihood function

$$\ell(\boldsymbol{\phi}, \boldsymbol{\Gamma}, \sigma^2 | \boldsymbol{y}, \boldsymbol{y_0}) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln|\sigma^2 \boldsymbol{\Gamma}\boldsymbol{\Gamma}'| - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\phi})'(\boldsymbol{\Gamma}\boldsymbol{\Gamma}')^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\phi}).$$

The parameters are then chosen to maximize this log-likelihood function, i.e. the probability of observing the data under the assumed distribution. This can be done by conducting a grid search over the parameter space using a computer. To reduce the seach dimensions, we may concentrate the log-likelihood by computing the first-order conditions:

$$\frac{\partial \ell}{\partial \boldsymbol{\phi}} = 0 \implies \hat{\boldsymbol{\phi}} = (\boldsymbol{X}'(\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}')^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}')^{-1}\boldsymbol{y}$$

$$\frac{\partial \ell}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\phi}})'(\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}')^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\phi}})$$

Thia allows us to focus our search only on $\boldsymbol{\phi}$.

13 Forecasting

So far we have introduced basic univariate time series models and their estimation. One common application of univariate time series analysis is forecasting. Forecasting is a rather complex topic, with a wide range of techniques from basic ARMA models to machine learning. This book is not specialized in forecasting. We only devote this section to briefly cover forecasting based on ARMA models. We will start with some intuition. Then justify the intuition with a bit of formal theory.

13.1 Intuitive Approach

Suppose we have an AR(1) process,

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

What would be the reasonable forecast for y_{T+1} given $y_1, ..., y_T$? It seems sensible to simply drop the white noise, as it is something completely unpredictable and it has mean zero. Thus,

$$\hat{y}_{T+1|T} = \phi y_t.$$

This is 1-period ahead forecast. But how do we forecast k-period ahead? Heuristically, we can simply iterate over to the future:

$$\begin{aligned}\hat{y}_{T+2|T} &= \phi \hat{y}_{T+1|T} = \phi^2 y_T, \\ \hat{y}_{T+h|T} &= \phi \hat{y}_{T+h-1} = \cdots = \phi^h y_T.\end{aligned}$$

We will leave the heuristic solutions here and justify them later. If we accept this heuristic approach, we can easily generalize it to AR(p) processes:

$$
\begin{aligned}
y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t. \\
\hat{y}_{T+1|T} &= \phi_1 y_T + \phi_2 y_{T-1} + \cdots + \phi_p y_{T-p+1}, \\
\hat{y}_{T+2|T} &= \phi_1 \hat{y}_{T+1|T} + \phi_2 y_T + \cdots + \phi_p y_{T-p+2}, \\
&\vdots \\
\hat{y}_{T+h|T} &= \phi_1 \hat{y}_{T+h-1|T} + \phi_2 \hat{y}_{T+h-2|T} + \cdots + \phi_p y_{T-p+h}.
\end{aligned}
$$

For MA(q) processes

$$
y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}
$$

Suppose we know the past innovations until T: $\epsilon_T, \epsilon_{T-1}, ...$ The best way to forecast $\hat{y}_{T+h|T}$ looks to simply discard $\epsilon_{T+1}, ..., \epsilon_{T+h}$. Since we have no knowledge about future innovations given the information at time T. Therefore,

$$
\begin{array}{lllll}
\hat{y}_{T+1|T} & = \theta_1 \epsilon_T + & \theta_2 \epsilon_{T-1} + & \theta_3 \epsilon_{T-2} + \cdots \\
\hat{y}_{T+2|T} & = & \theta_2 \epsilon_T + & \theta_3 \epsilon_{T-1} + \cdots \\
& \vdots & & \\
\hat{y}_{T+h|T} & = & & \theta_h \epsilon_T + \theta_{h+1} \epsilon_{T-1} + \cdots
\end{array}
$$

13.2 Best Linear Predictor

We now justify our heuristic solutions by the theory of best linear predictor. Suppose we want to forecast y give the information set X.

Definition 13.1. The best linear predictor (BLP) is defined as

$$
\mathcal{F}(y|X) = x'\beta^*
$$

which is a linear function of $X = (x_1, x_2, ..., x_p)$ such that

$$
\beta^* = \text{argmin } \mathbb{E}(y - x'\beta)^2.
$$

Taking first-order condition with respect to β gives

$$
\beta^* = [\mathbb{E}(xx')]^{-1}\mathbb{E}(xy).
$$

Therefore, the BLP is given by

$$\hat{y} = \mathcal{F}(y|X) = x'\beta^* = x'[\mathbb{E}(xx')]^{-1}\mathbb{E}(xy).$$

The prediction error is

$$r_{y|X} = y - \hat{y} = y - x'[\mathbb{E}(xx')]^{-1}\mathbb{E}(xy).$$

The BLP is the linear projection of y onto X. Because $\mathbb{E}[x(y - x'\beta)] = 0$. The forecast error is orthogonal to X.

Proposition 13.1. *BLP has the following properties:*

1. $\mathcal{F}[ax + by|z_1...z_k] = a\mathcal{F}[x|z_1...z_k] + b\mathcal{F}[y|z_1...z_k]$;
2. *If* $x = a_1z_1 + \cdots + a_kz_k$ *is already a linear combination of* $z_1...z_k$, *then* $\mathcal{F}[x|z_1...z_k] = x$;
3. *If for all* $1 \leq j \leq k$, $cov(x, z_j) = \mathbb{E}(xz_j) = 0$, *then* $\mathcal{F}[x|z_1...z_k] = 0$.

13.3 Forecasting with ARMA Models

ARMA model is a basic yet powerful tool for forecasting. Given all stationary time series can be approximated by ARMA processes, it makes sense to model a stationary time series with ARMA, and then make forecast based on that model. We will see our heuristic solutions in the first part can be easily justified with the theory of BLP.

13.3.1 Forecasting with AR(p)

We have said that, for an AR(p) process

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

The one-step-ahead forecast is simply

$$\hat{y}_{T+1|T} = \phi_1 y_T + \phi_2 y_{T-1} + \cdots + \phi_p y_{T-p+1}.$$

This is the BLP immediately from Property 2 of Proposition 13.1. We can also justify the iterated h-step-ahead forecast by Property 1 (assuming $h < p$):

$$\begin{aligned}\hat{y}_{T+h|T} &= \phi_1\hat{y}_{T+h-1|T} + \phi_2\hat{y}_{T+h-2|T} + \cdots + \phi_h y_T + \cdots + \phi_p y_{T+h-p} \\ &= \phi_1\mathcal{F}[y_{T+h-1}|y_T, y_{T-1}...] + \cdots + \phi_p\mathcal{F}[y_{T+h-p}|y_T, y_{T-1}...] \\ &= \mathcal{F}[\phi_1 y_{T+h-1} + \cdots + \phi_p y_{T+h-p}|y_T, y_{T-1}, ...] \\ &= \mathcal{F}[y_{T+h}|y_T, y_{T-1}, ...]\end{aligned}$$

This is assuming all the forecast before h are BLPs, which can be justified recursively. Also note that for the values readily observed: $y_T, y_{T-1}, \ldots$, the BLP is the value itself.

13.3.2 Forecasting with MA(q)

For the MA(q) process

$$y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$$

The BLP for h-step-ahead forecast is (assuming $h < q$)

$$\begin{aligned}\hat{y}_{T+h|T} &= \mathcal{F}(y_{T+h}|\epsilon_T, \epsilon_{T-1}...) \\ &= \mathcal{F}(\epsilon_{T+h}|\epsilon_T, \epsilon_{T-1}...) + \theta_1\mathcal{F}(\epsilon_{T+h-1}|\epsilon_T, \epsilon_{T-1}...) + \cdots + \theta_q\mathcal{F}(\epsilon_{T+h-q}|\epsilon_T, \epsilon_{T-1}...) \\ &= 0 + \cdots + 0 + \theta_h\epsilon_T + \cdots + \theta_q\epsilon_{T+h-q}\end{aligned}$$

We make use of Property 3 of Proposition 13.1 with the knowledge that $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. This result is also consistent with our intuition, because we have no knowledge of future innovations, the best thing we can do is assuming they are zeros. If $h > q$, then all $\mathcal{F}(\epsilon_{T+h-q}|\epsilon_T, \epsilon_{T-1}...)$ are zero, which yields $\hat{y}_{T+h|T} = 0$.

In practice, we do not observe $\{\epsilon_t\}$. If we have an estimated MA model and we want to make forecast based on the model, we need to back out $\{\epsilon_t\}$ from $\{y_t\}$ by inverting the MA process: $\epsilon_t = \theta^{-1}(L)y_t$.

With the MA specification, we can easily compute the **Mean Squared Forecast Error (MSFE)** as follows

$$Q_{T+h|T} = \mathbb{E}(y_{T+h} - \hat{y}_{T+h|T})^2 = \mathbb{E}\left(\sum_{j=0}^{h-1}\theta_j\epsilon_{T+h-j}\right)^2 = \sigma^2\sum_{j=0}^{h-1}\theta_j^2.$$

13.3.3 Forecasting with ARMA(p,q)

Consider the ARMA(p, q) process

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

We assume the process is causal and invertible. We can transform it to an AR(∞) process or MA(∞) process.

Causal form:

$$y_t = \phi^{-1}(L)\theta(L)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

Invertible form:

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

or

$$y_t = -\sum_{j=1}^{\infty} \pi_j y_{t-j} + \epsilon_t$$

As we have seen so far, it is relatively easier to compute the mean forecast with AR models, and the MSFE with MA models. So we make forecast with the AR representation:

$$\hat{y}_{T+h|T} = -\sum_{j=1}^{h-1} \pi_j \hat{y}_{T+h-j} - \sum_{j=h}^{\infty} \pi_j y_{T+h-j}$$

However, we do not observe infinite past values in real world. We can only use the truncated values, discarding past values that we do not observe $y_0, y_{-1}, y_{-1}, ...$

$$\hat{y}_{T+h|T} = -\sum_{j=1}^{h-1} \pi_j \hat{y}_{T+h-j} - \sum_{j=h}^{T+h-1} \pi_j y_{T+h-j}$$

We compute the MSFE with the MA representation:

$$Q_{T+h|T} = \mathbb{E}(y_{T+h} - \hat{y}_{T+h})^2 = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2$$

If we can compute the prediction interval if we assume some probability distributions for the innovations. If we assume $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$, then $(y_1, ..., y_{T+h})'$ is jointly normal. Therefore,

$$y_{T+h} - \hat{y}_{T+h|T} \sim N(0, Q_{T+h|T})$$

The prediction interval is thus given by $\hat{y}_{T+h|T} \pm z_{\alpha/2}\sqrt{Q_{T+h|T}}$.

13.4 Applications

The following examples use ARMA models to forecast inflation rate and stock market index. The parameters of the ARMA models are chosen automatically. We can see for the inflation rate, the model produces some patterns in the forecast. But for the stock market index, the forecast is an uninformative flat line, indicating there is no useful patterns in the past data can be extrapolated by the ARMA model.

```
library(forecast)
data = readRDS("data/md.Rds")
data$CPI |>
  auto.arima() |>
  forecast(h=20) |>
  autoplot()
```

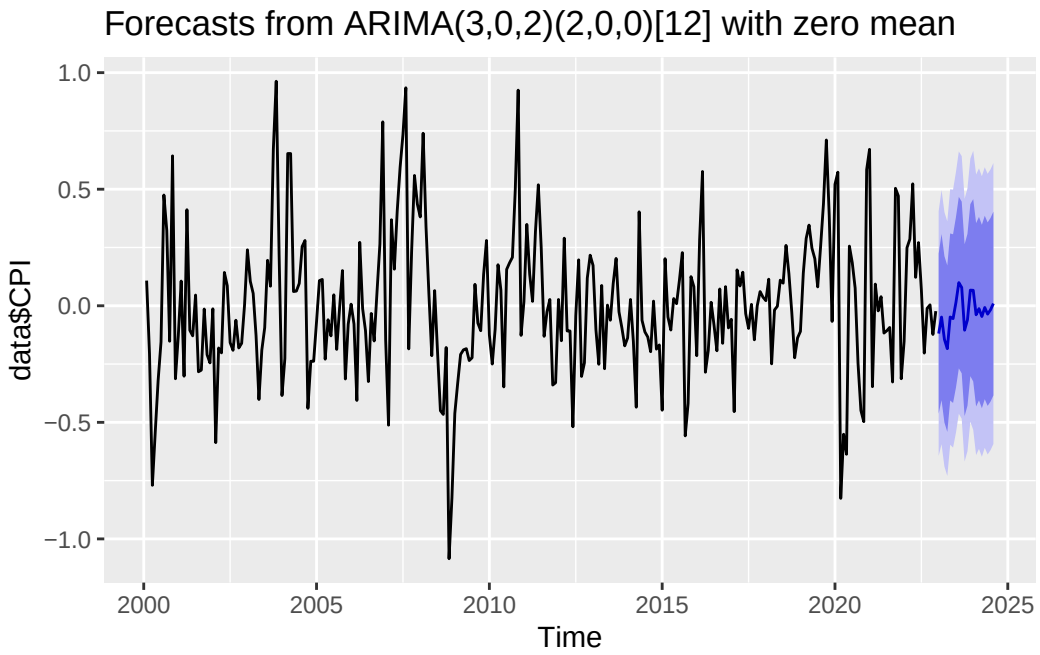


Figure 13.1: ARMA forecast for monthly inflation

```
data$SHSE |>
  auto.arima() |>
  forecast(h=20) |>
  autoplot()
```

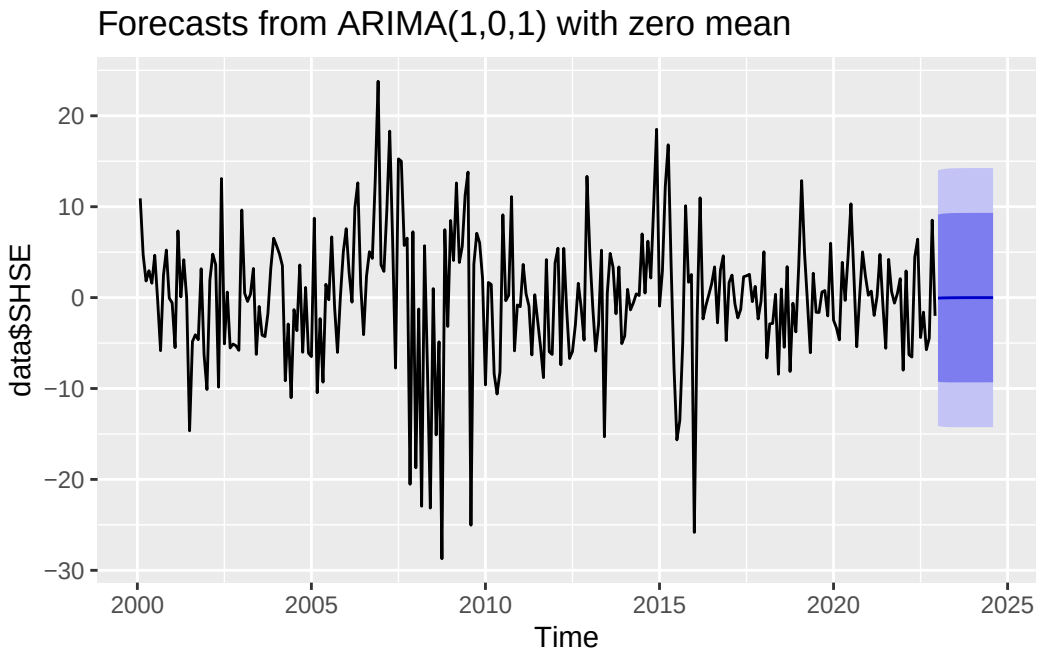


Figure 13.2: ARMA forecast for stock market index

14 Dynamic Causal Effect

As in all fields of science, we are perpetually interested in understanding the causal effect of one thing on another. In economics, we want to understand how monetary policy affects output and inflation, how exchange rate affects import and export, and so on. However, causality is something much easier said than done. In reality, there are multiple forces at work simultaneously that leads to the consequences we observed. It is challenging both conceptually and statistically to isolate the causality of a variable of particular interest.

In cross-sectional analysis, causality is defined counterfactually. That is, the causal effect of a treatment is defined as the difference between the treated outcome and the untreated outcome assuming that they would be otherwise the same without the treatment. In practice, that involves working with a large number of *iid* observations that are similar on average only differentiated by the status of the treatment. This approach, however, does not work well with many macroeconomic studies. For example, suppose we want to figure out the causal effect of monetary policy on inflation rate. The cross-sectional approach would entail finding a large number of almost identical countries, each with independent monetary policy. And a random subset of them tighten their monetary policies while others do not. Then we work out the different economic outcomes between these two groups. This is clearly infeasible. The question we posed concerns only one country with inflation and interest rates observed through time. We would need a definition of causal effect that encompasses observations over time not across individuals.

Suppose ϵ_t denote a random treatment happened at time t. Then the causal effect on an outcome variable y_{t+h}, h periods ahead, of a unit shock in ϵ is defined as

$$\mathbb{E}[y_{t+h}|\epsilon_t = 1] - \mathbb{E}[y_{t+h}|\epsilon_t = 0]. \tag{14.1}$$

We require the randomness of the treatment ϵ_t in a sense that it is uncorrelated with any other variables that could possible have an impact on the outcome. Therefore, ϵ_t happens or not does not affect other forces that shape the outcome. The difference in the outcomes is solely attributable to ϵ_t. It is this randomness that guarantees a causal interpretation.

Our example of monetary policy above clearly does not meet this requirement. The monetary authority does not set the interest rate randomly, but based on the economic conditions of the time, which makes it correlated with other economic variables that could also have an impact on inflation. A qualified random shock may be a change in weather conditions. Weather has huge impact on agricultural production, but it is determined independent of any human activity.

If ϵ_t denotes a rainy day at time t, and y_{t+h} be the agricultural production, Equation 14.1 could be a plausible causal effect. However, most variables of interest in economics are endogenously determined. How to estimate the causal effect in such cases is an art in itself. We will come back to this point later.

The conceptual definition of Equation 14.1 can not be computed directly as the counterfactual is not observed. What we have is a sample of experiments over time, in which the treatment happens randomly at some points but not others, $\{\epsilon_1 = 0, \epsilon_2 = 1, \epsilon_3 = 0, \dots\}$. We could envision that if we have long enough observations, by comparing the outcomes when the shock happens and when it does not, it gives us an reasonable estimation of the causal effect because all other factors that contributing to the outcome, despite they are changing over time, would be averaged out provided the randomness of the treatment.

Assuming linearity and stationarity, the causal effect of Equation 14.1 can be effectively captured by a regression framework,

$$y_{t+h} = \theta_h \epsilon_t + u_{t+h},$$

where u_{t+h} represents all other factors contributing to the outcome variable. Since ϵ_t is random, it holds that $\mathbb{E}(u_{t+h}|\epsilon_t) = 0$. Therefore,

$$\theta_h = \mathbb{E}(y_{t+h}|\epsilon_t = 1) - \mathbb{E}(y_{t+h}|\epsilon_t = 0).$$

Thus, θ_h captures the causal effect of one unit shock of ϵ_t on y_{t+h}. The path of the causal effects mapped out by $\{\theta_0, \theta_1, \theta_2, \dots\}$ is called the **dynamic causal effect**, in a sense that it is the causal effects through time.

15 The Structural Shock Framework

The counterfactual framework introduced in the last section defines the dynamic causal effect of any variable on another. As economists, we are more interested in understanding the causal relationships between important forces that drive the economy. We now introduce the **structural shock framework**, or the **Slutzky-Frisch paradigm**. This paradigm is explicitly or implicitly embedded in virtually every mainstream macroeconomic models or econometric models. It is not an essential component of time series analysis. But, as we would like to approach the topic from an economist's perspective, it is good to have this framework in mind for many of our applications.

The structural shock framework envisions our economy as a complex system driven by a set of fundamental structural forces and coordinated by numerous price signals that automatically balance the demand and supply of all goods and services. The structural forces could be technology progress, climate change, policy changes and so on. These structural shocks are the primitive forces underlying our economy. When a structural shock happens, it triggers a reallocation of economic resources guided by market forces. In theoretical works, we are interested in modelling the system as a whole, particularly how resources are allocated optimally by market forces. In empirical works, we are interested how to recover the underlying structural shocks and estimate their causal effect on other economic variables.

In the language of time series analysis, we can envision our economy as an MA process, in which the observable variables (output, employment, inflation, etc) are the outcomes of accumulated past and current structural shocks:

$$\boldsymbol{y}_t = \boldsymbol{\Theta}(L)\boldsymbol{\epsilon}_t,$$

or

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{bmatrix} = \sum_{j=0}^{\infty} \begin{bmatrix} \theta_{j,11} & \theta_{j,12} & \cdots & \theta_{j,1m} \\ \theta_{j,21} & \theta_{j,22} & \cdots & \theta_{j,2m} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{j,n1} & \theta_{j,n2} & \cdots & \theta_{j,nm} \end{bmatrix} \begin{bmatrix} \epsilon_{1,t-j} \\ \epsilon_{2,t-j} \\ \vdots \\ \epsilon_{m,t-j} \end{bmatrix}.$$

$\boldsymbol{y}_t$ represents the vector of economic variables of concern. The space of $\boldsymbol{y}_t$ are spanned by m structural shocks (current and past): $\{\boldsymbol{\epsilon}_{t-j}\}_{j=0}^{\infty}$.

Structural shocks are conceptual constructions that are primitive, unforeseeable, and uncorrelated underlying forces. Whether structural shocks do exist or not is an open question. But they are useful constructions that enable econometricians to disentangle different driving forces of the outcome variable.

In reality, almost every economic variable is endogenous. For example, monetary policy (interest rate) is set by the monetary authority based on their assessment of the economic conditions. But we can also imagine, there is a "genuine" component of the monetary policy, which may come from the personality of the policymaker and his mental conditions when he make the decision, that is not predictable from other variables. This genuine component is what we deem as the "monetary policy shock". It is a shock in a sense that it is not predictable. It speaks for its own sake and contribute to the economic outcomes independently.

We do not observe the structural shocks directly. The observable variables are linear combinations of the structural shocks. For example, we may think of the observed interest rate as a linear combination of the monetary policy shock together with supply-side shocks and others.

$$i_t = \theta_1(L)\epsilon_t^{\text{MP}} + \theta_2(L)\epsilon_t^{\text{SS}} + \cdots$$

Therefore, regressing inflation or output on interest rate will not give the causal effect of the monetary policy. Because interest rate does not represent the "genuine" monetary policy shock. It is determined by other economic variables and there are multiple structural forces at work. There are numerous literature that works on methods to isolate the "monetary policy shock" from the observed interest rates. Such way of constructing the structural shocks is not only a conceptual idea, but also a prerequisite for meaningful interpretation of the coefficients of econometric models.

16 Estimating Dynamic Multipliers

This section will cover the specifications commonly used to estimate dynamic causal effect. Just like in cross-sectional analysis, regression techniques can always be applied as long as the time series are covariance stationary without an emphasis on causality. However, we pay special attention to causal inferences, as we are more interested in understanding the causality rather than mere correlations in most empirical researches. We start with the case where the structural shock is directly observed and move on to the cases where the structural shocks need to be constructed.

16.1 Distributed Lags

The easiest approach to estimate dynamic causal effect is to include lags in the specification:

$$y_t = \beta_0\epsilon_t + \beta_1\epsilon_{t-1} + \cdots + \beta_p\epsilon_{t-p} + u_t,$$

where ϵ_t is the structural shock, u_t is everything that otherwise influences y_t. Since ϵ_t happens randomly, we have $\mathbb{E}(u_t|\epsilon_{t-j}) = 0$. Thus, the βs, which capture the dynamic causal effect, would be consistently estimated by OLS.

Note that we call it a specification, in a sense that the joint distribution of the random variables is unknown, which distinguishes itself from the DGP model in Chapter 6. But it does not stop us from uncovering the causal effect, as long as the exogenous condition holds.

The effect of a unit change in ϵ on y after h periods, which is β_h, is also called the h-period **dynamic multiplier**. Sometimes, we are interested in the accumulated effect over time, $\beta_0 + \beta_1 + \cdots + \beta_h$, which is called **cumulative dynamic multiplier**.

Because u_t is the linear combination of all other current and past shocks, it is likely serially correlated. So HAC standard errors are required for robust inferences.

Proposition 16.1. *Assumptions for a consistent estimation of dynamic causal effects with distributed lag models:*

1. *ϵ is an exogenous shock, $\mathbb{E}(u_t|\epsilon_t, \epsilon_{t-1}, ...) = 0$;*

2. *All variables are stationary;*
3. *Regular conditions for OLS to work.*

To reduce the serial correlations $\{u_t\}$, and also allow for slow adjustment of y_t, we can also include lagged dependent variables in the specification, which becomes an **autoregressive distributed lag (ADL)** specification:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \beta_0 \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_p \epsilon_{t-p} + u_t,$$

or

$$\phi(L) y_t = \beta(L) \epsilon_t + u_t.$$

When lags of the dependent variable are included as regressors, strict exogeneity fails for sure, because $X = \{y_{t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots\}$ is correlated with past errors u_{t-1}, despite it is uncorrelated with the contemporary error u_t. The OLS is consistent so long as $\{u_t\}$ are not serially correlated. Otherwise, u_t would be correlated with X through u_{t-1}. The serial correlation can be tested with Durbin-Watson test or Breusch-Godfrey test.

The dynamic causal effect is more convoluted with the ADL specification though,

$$\hat{\theta}(L) = \hat{\phi}^{-1}(L) \hat{\beta}(L).$$

ADL also require truncated lags. p and q are chosen as an increasing function of the sample size. In general, choosing p and q to be of order $T^{1/3}$ would be sufficient for consistency.

16.2 Local Projections

Dynamic causal effect can also be estimated by projecting future outcomes directly on the shock. Jordà (2005) named it **local projections (LP)**.

$$y_{t+h} = \theta_h \epsilon_t + u_{t+h}.$$

By assumption, $\mathbb{E}(u_{t+h} | \epsilon_t) = 0$. So $\hat{\theta}_h$ is a consistent estimate of the h-period dynamic multiplier. HAC standard errors are also required in local projections, as u_{t+h} in are usually serially correlated.

Readers may wonder, since ADL and LP both give consistent estimates of the dynamic multipliers, what is the difference between them. There are two obvious differences:

1. Lagged shocks do not appear in LP specifications as they do in distributed lag specifications.
2. The LP method requires running separate regressions for each h. The dynamic response $\{\theta_0, \theta_1, \theta_2, \dots\}$ are estimated through multiple regressions rather than one.

The error structure is also different. To see this, suppose the DGP is an MA(∞) process

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots$$

If we estimate it with a DL specification with two lags,

$$y_t = \beta_0 \epsilon_t + \beta_1 \epsilon_{t-1} + u_t,$$

where $u_t = \sum_{j=2}^{\infty} \theta_j \epsilon_{t-j}$. Exogeneity would ensure $\hat{\beta}_1 \to \theta_1$.

We can also estimate it with a local projection (suppose we are interested in the one-step-ahead dynamic multiplier):

$$y_{t+1} = \psi_1 \epsilon_t + u_{t+1}.$$

Again, we have consistency $\hat{\psi}_1 \to \theta_1$. But the error structure is different $u_{t+1} = \epsilon_{t+1} + \sum_{j=2}^{\infty} \theta_j \epsilon_{t-j}$.

Both the DL and LP specifications may include additional control variables, which can reduce the variance of the residuals and improve the efficiency of the estimators.

16.3 Example of Observable Exogenous Shocks

Directly observable exogenous shocks are rare. Here we use an example from Stock and Watson (2020), which explores the dynamic causal effect of cold weather on orange juice prices. Cold weather is bad for orange production. Orange trees cannot withstand freezing temperatures that last for more than a few hours. Florida accounts for more than 98 percent of U.S. production of frozen concentrated orange juice. Therefore, the frozen weather in Florida would reduce the supply and orange juice and raise the price. The dataset includes the number of freezing degree days in Florida and the average producer price for orange juice. Cold weather is plausibly exogenous, which allows us the utilize the regression framework above to estimate the dynamic causal effect.

```
library(AER)
library(dynlm)
library(lmtest)
```

```
data("FrozenJuice") # load data

# compute percentage change on price
pchg = 100*diff(log(FrozenJuice[, 'price']))
sample = ts.union(fdd = FrozenJuice[,'fdd'], pchg)

# distributed lag model
mod = dynlm(pchg ~ L(fdd, 0:6), data = sample)
# confidence interval
ci = coefci(mod, vcov. = NeweyWest)

# plot dynamic multiplier
{
  plot(mod$coefficients[-1], # remove intercept
       type = "l",
       col = 2,
       ylim = c(-0.4,1),
       xlab = "Lag",
       ylab = "Dynamic Multiplier")
  abline(h = 0, lty = 2)
  lines(ci[-1,1], col = 4)
  lines(ci[-1,2], col = 4)
}
```

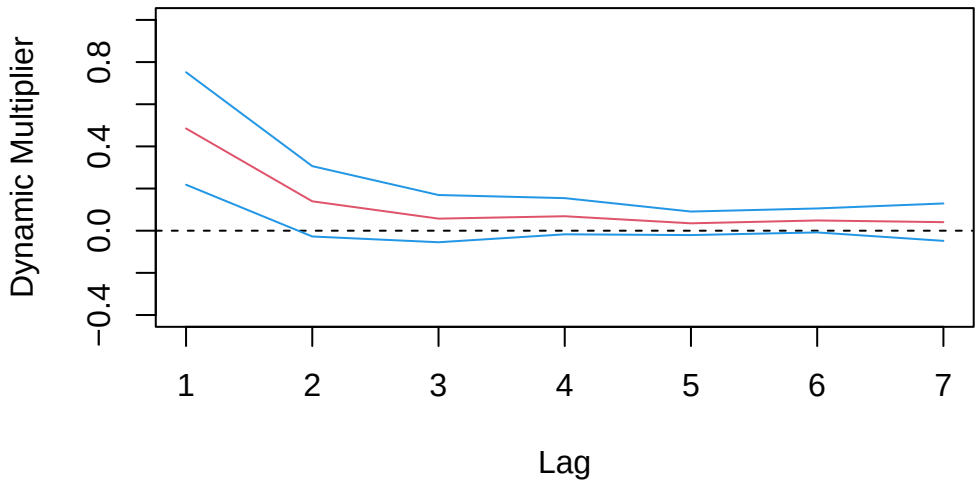


Figure 16.1: Dynamic Effect of Freezing Days on Orange Juice Price

We can also use local projections. Note that local projections require estimating multiple

regressions. The coefficients from each of the regressions constitute the dynamic multiplier.

```
# apply local projection for horizons 0-6
lps = sapply(0:6, function(h) {
  lp = dynlm(L(pchg, -h) ~ fdd, data = sample)
  ci = coefci(lp, vcov. = NeweyWest)
  c(lp$coefficients[-1], ci[-1,]) # remove intercept
}) |> t() # transpose it

# plot the LP coefficients
{
  plot(lps[,'fdd'],
       type = "l",
       col = 2,
       ylim = c(-0.4,1),
       xlab = "Horizon",
       ylab = "LP Coefficient")
  abline(h = 0, lty = 2)
  lines(lps[,'2.5 %'], col = 4)
  lines(lps[,'97.5 %'], col = 4)
}
```

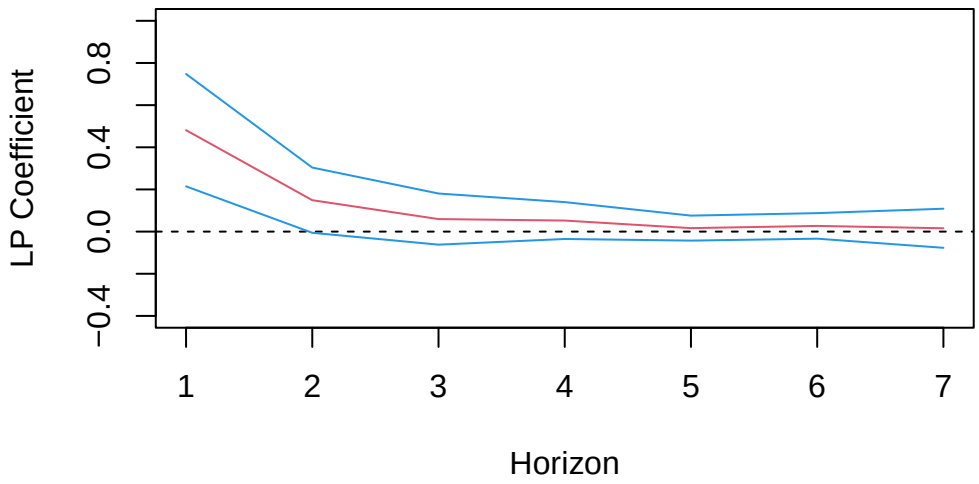


Figure 16.2: Local Projections of Freezing Days on Orange Juice Price

16.4 Example of Constructed Structural Shocks

Most structural shocks in economics are not directly observed, such as monetary policy shocks, or fiscal policy shocks, yet they are of profound interest of researchers. As we have explained before, regressing output or inflation on interest rate does not give a plausible estimation of the causal effect of monetary policy, due to the endogeneity problem. Thus, we need to isolate the exogenous part of the monetary policy from observed variables. The method to achieve this is an active research field in itself. We here demonstrate the monetary policy shock for China constructed by Das and Song (2023).

The authors utilize the high-frequency price changes of interest rate swap around the window of monetary policy announcement to approximate the monetary policy shock. The rationale of this construction is that, the price of the financial instrument reflects the expected interest rate by market participants based on the economic conditions. Therefore, the sudden change of the price in the tiny window of monetary policy announcement captures the unexpected part of the monetary policy.

```
library(zoo)

mp = readRDS("data/mpshocks.Rds") # monetary policy shock
md = readRDS("data/md.Rds") # monthly data

plot(na.exclude(cbind(mp$shock_1y, md$LoanRate1Y)),
     main="", xlab="", ylab=c("Loan Rate", "MP Shock"))
```

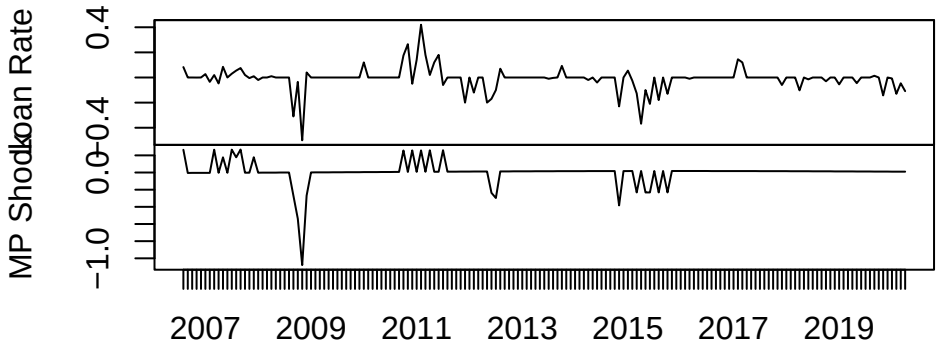


Figure 16.3: Monetary Policy Shock and Lending Rate

We estimate the dynamic causal effect of monetary policy shock on inflation using the constructed MP shocks. It shows that a tightening of monetary policy implies a gradual cooling down of inflation. The price level starts to decline roughly half a year after the initial tightening shock. However, the confidence interval is wide, suggesting an insignificant estimation of the

policy effect. The result does not provide very strong evidence underlining the effectiveness of monetary policy to control inflation.

```
sample = na.exclude(cbind(cpi=md$CPI, mp))

# distributed lag model
mod = dynlm(cpi ~ L(shock_1y, 0:12), sample)
# confidence interval
ci = coefci(mod, vcov. = NeweyWest)

# plot dynamic multiplier
{
  plot(mod$coefficients[-1], # remove intercept
       type = "l",
       col = 2,
       ylim = c(-1,1.5),
       xlab = "Lag",
       ylab = "Dynamic Multiplier")
  abline(h = 0, lty = 2)
  lines(ci[-1,1], col = 4)
  lines(ci[-1,2], col = 4)
}
```

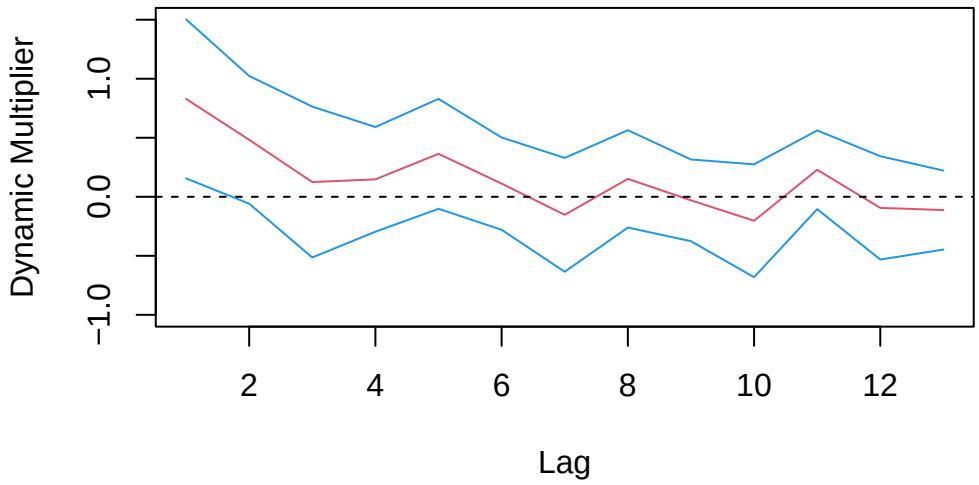


Figure 16.4: Dynamic response of inflation on monetary policy shocks

17 Instrument Variables

If a structural shock is not directly observable, neither can it be constructed through observable variables, we can identify it using an instrument variable approach if an instrument is available.

Suppose our observable space $\boldsymbol{y} = (y_1, y_2, \dots)'$ is spanned by multiple structural shocks $\epsilon = (\epsilon_1, \epsilon_2, \dots)'$. We want to identify the causal effect of structural shock ϵ_1. An instrument variable z satisfies the following conditions:

1. $\mathbb{E}(\epsilon_{1t} z_t) = \alpha \neq 0$ (relevance);
2. $\mathbb{E}(\epsilon_{2:n} z_t) = 0$ (contemporaneous exogeneity);
3. $\mathbb{E}(\boldsymbol{\epsilon}_{t+j} z_t) = 0$ for $j \neq 0$ (lead-lag exogeneity).

$\epsilon_{2:n}$ denotes all other structural shocks except ϵ_1. The lead-lag exogeneity is unique to time series. To understand this, consider an local projection: $y_{t+h} = \theta_h \epsilon_t + u_{t+h}$. As illustrated in the last section, u_{t+h} is a linear combination of the entire history of structural shocks. If z_t is to identify the causal effect of shock ϵ_{1t} alone, it must be uncorrelated with all leads and lags. The requirement that z_t be uncorrelated with future ϵ's is generally not restrictive — by definition, future shocks are unanticipated. To the contrary, the requirement that z_t be uncorrelated with past ϵ's is more restrictive and hard to meet.

Suppose we want to estimate the causal effect of $\epsilon_{1,t}$ on $y_{2,t+h}$, where $\epsilon_{1,t}$ is only observable through $y_{1,t}$. Suppose we have an instrument variable z_t that satisfies the above conditions. The local projection

$$y_{2,t+h} = \theta_{h,21} y_{1,t} + u_{t+h}$$

cannot be consistently estimated because $y_{1,t}$ and u_{t+h} are correlated. However, with the help with z_t as an instrument, we can consistently estimate the dynamic multiplier $\theta_{h,21}$:

$$\begin{aligned}
\beta_{\text{LP-IV}} &= \frac{\mathbb{E}(y_{2,t+h} z_t)}{\mathbb{E}(y_{1,t} z_t)} \\
&= \frac{\mathbb{E}[(\theta_{h,21} y_{1,t} + u_{t+h}) z_t]}{\mathbb{E}(y_{1,t} z_t)} \\
&= \frac{\theta_{h,21} \alpha}{\alpha} = \theta_{h,21}.
\end{aligned}$$

Lead-lag exogeneity implies z_t being unforecastable in a regression of z_t on lags of y_t. If the exogeneity fails, LP-IV is not consistent. This problem can be partially addressed by including control variables in the regression:

$$y_{2,t+h} = \theta_{h,21} y_{1,t} + \boldsymbol{\gamma}_{\boldsymbol{h}}' \boldsymbol{w_t} + u_{t+h}^{\perp}.$$

We could also include lagged values of y_t or other lagged variables. The IV estimator is consistent if $\boldsymbol{w}_t$ absorbs all past shocks that could potentially correlated with z_t. In a broad sense, the validity of the instrument variable with additional controls requires that the controls span the space of all structural shocks.

Part IV

Nonstationary Time Series

18 Spurious Regression

It is said, all stationary series are alike, but each non-stationary series is non-stationary in its own way (remember Leo Tolstoy's famous quote: *all happy families are alike; each unhappy family is unhappy in its own way.*)

In all previous chapters, we have been working on stationary processes. We have shown that similar regression techniques and asymptotic results hold for stationary processes as for *iid* observations, albeit not exactly the same. If a time series is not stationary, we transform it to stationary by taking differences.

This chapter is devoted to study non-stationary time series. Special attention is given to unit root processes. We will see the theories involving non-stationary processes are entirely different from those applied to stationary processes. This makes unit root analysis an rather independent topic. The obsession with unit root in academia have faded away in recent decades (I do not know if this assessment is accurate). Despite the topic posses immense theoretical interest, it does not seem to provide proportionate value for applied studies. Nonetheless, the topic is indispensable for a comprehensive understanding of time series analysis.

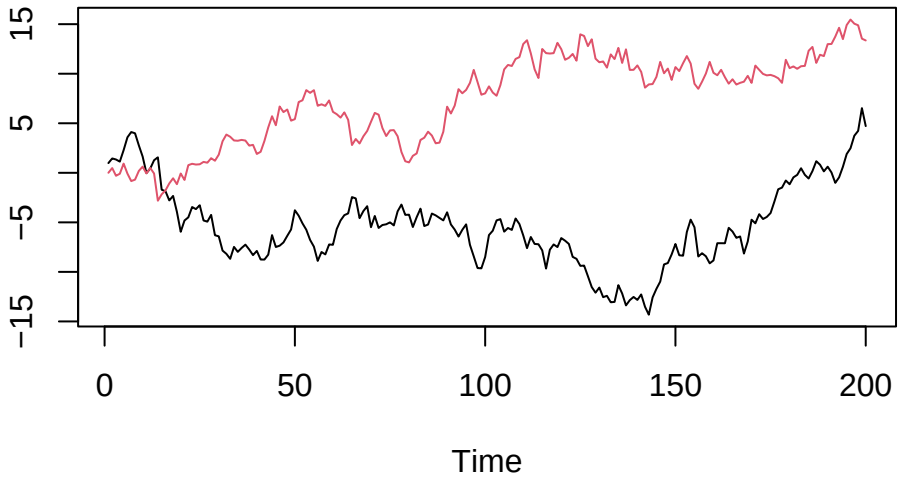
We will focus on two types of non-stationary processes: trend-stationary processes and unit root processes, which are the most common types of non-stationary series we would encounter in economic and finance. Non-stationary series with exponential growth can be transformed into linear trend, hence is not of particular interest. We will start with the relatively easy tread-stationary processes, and spend most of the paragraphs on unit root processes.

We start by pointing out that, it is very dangerous to blindly include non-stationary variables in a regression. To illustrate this, we simulate two random walks:

$$\begin{aligned} x_t &= x_{t-1} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, \sigma_X^2) \\ y_t &= y_{t-1} + \eta_t, \quad \eta_t \overset{iid}{\sim} N(0, \sigma_Y^2) \end{aligned}$$

ϵ_t and η_t are independent to each other.

```
set.seed(2024)
x = cumsum(rnorm(200))
y = cumsum(rnorm(200))
ts.plot(cbind(x,y), col=1:2)
```



We would expect the two series completely uncorrelated, as they are two independent random processes. However, if we regress y_t on x_t, we would likely find a very strong correlation. This is called a **spurious regression**.

$$y_t = \alpha + \beta x_t + u_t$$

```
coeftest(lm(y ~ x))
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  5.895196   0.510099  11.557 < 2.2e-16 ***
x           -0.267537   0.075726  -3.533 0.0005113 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that if we difference the two series to stationary, the spurious correlation disappears.

```
coeftest(lm(diff(y) ~ diff(x)))
```

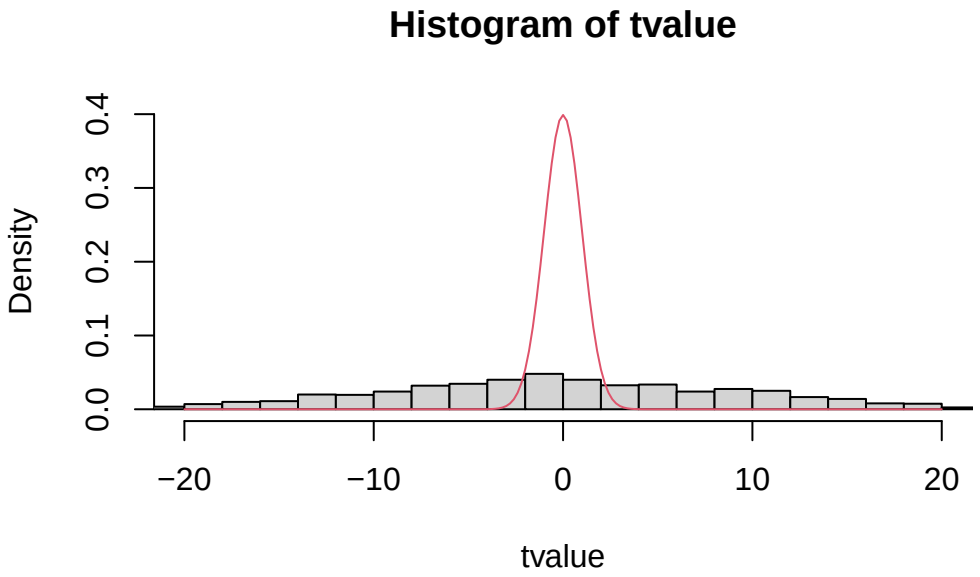
```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.068500   0.066740  1.0264   0.3060
diff(x)     -0.074427   0.065236 -1.1409   0.2553
```

If we approximate the distribution of the t-value for $\hat{\beta}$ with run Monte Carlo simulations, we would find the distribution is not Gaussian and has much heavier tails. That means we would much more likely to find significant results with spurious regression.

```
# Monte Carlo simulation
tvalue = sapply(1:1000, function(i) {
  x = cumsum(rnorm(200))
  y = cumsum(rnorm(200))
  # extract t-value
  summary(lm(y~x))$coef['x','t value']
})
# plot the density with Gaussian curve
{
  hist(tvalue, prob = TRUE, breaks = 40, xlim=c(-20,20), ylim = c(0,.4))
  range = seq(-20, 20, by = .2)
  lines(range, dnorm(range), col = 2)
}
```



Therefore, the conventional statistical inference against non-stationary series is totally misleading. The rest of the chapter will demystify the nature of spurious regression and discuss how we can properly deal with non-stationary time series.

19 Trend Stationary

Trend-stationary process is a stationary process round a deterministic trend:

$$y_t = \alpha + \delta t + \psi(L)\epsilon_t,$$

where δt is a deterministic linear time trend, $\psi(L)\epsilon_t$ is a stationary process. After de-trending $-(\alpha + \delta t)$, the result is a stationary process.

Trend stationary vs stochastic trend

Trend-stationary processes must be distinguished from *stochastic trend process* (unit root with a drift):

$$y_t = \delta + y_{t-1} + \epsilon_t = y_0 + \delta t + \sum_{j=1}^{t} \epsilon_j.$$

Both of them have a time trend component. But in the latter model, the stochastic component is not stationary. In other words, a innovation in a trend-stationary model does not have long-lasting effect, whereas the effect is persistent in a stochastic trend model.
The difference becomes clearer by comparing the variances. The variance of the trend-stationary process

$$\text{var}(y_t) = \psi^2(L)\sigma^2$$

is constant which does not depend on time. However, the variance of the stochastic trend process

$$\text{var}(y_t) = \text{var}(\sum_{j=1}^{t} \epsilon_j) = \sigma^2 t$$

is increasing over time. Therefore, stochastic trend process fluctuates more widely as time goes by.

Unlike unit root processes, trend-stationary processes can be safely estimated by OLS. The usual t and F statistics have the same asymptotic distribution as they are for stationary

processes. But they converge at a different speed, due to the presence of the trend. To see this, rewrite the regression in vector form

$$y_t = \alpha + \delta t + \epsilon_t = \begin{bmatrix} 1 & t \end{bmatrix} \begin{bmatrix} \alpha \\ \delta \end{bmatrix} + \epsilon_t = \boldsymbol{x_t'\beta} + \epsilon_t;$$

For simplicity, assume $\epsilon_t \sim IID(0, \sigma^2)$ for the following computation. The result can be generalized to ϵ_t being stationary. The OLS estimator is given by

$$\begin{aligned} \hat{\beta} = \begin{bmatrix} \hat{\alpha} \\ \hat{\delta} \end{bmatrix} &= \left(\sum_t x_t x_t' \right)^{-1} \left(\sum_t x_t y_t \right), \\ \sqrt{T}(\hat{\beta} - \beta) &= \left(\frac{1}{T} \sum_t x_t x_t' \right)^{-1} \left(\frac{1}{\sqrt{T}} \sum_t x_t \epsilon_t \right). \end{aligned}$$

The usual asymptotic results are

$$\begin{aligned} \frac{1}{T} \sum_t x_t x_t' &\to Q \\ \frac{1}{\sqrt{T}} \sum_t x_t \epsilon_t &\to N(0, \sigma^2 Q) \\ \sqrt{T}(\hat{\beta} - \beta) &\to N(0, \sigma^2 Q^{-1}) \end{aligned}$$

But this is not the case with deterministic trend if we do the computation:

$$\frac{1}{T} \sum_t x_t x_t' = \begin{bmatrix} 1 & \frac{1}{T} \sum t \\ \frac{1}{T} \sum t & \frac{1}{T} \sum t^2 \end{bmatrix}$$

does not converge. Because $\sum_{t=1}^T t = \frac{T(T+1)}{2}$, and $\sum_{t=1}^T t^2 = \frac{T(T+1)(2T+1)}{6}$. It requires stronger divider to make them converge, $T^{-2} \sum_{t=1}^T t \to \frac{1}{2}$, $T^{-3} \sum_{t=1}^T t^2 \to \frac{1}{3}$. In general,

$$\frac{1}{T^{v+1}} \sum_{t=1}^T t^v \to \frac{1}{v+1}.$$

Dividing by T^3 will make the convergence

$$\frac{1}{T^3} \sum_t x_t x_t' \to \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

However, this matrix is not invertible. We need different rates of convergence for $\hat{\alpha}$ and $\hat{\delta}$.

Define

$$\boldsymbol{\gamma}_T = \begin{bmatrix} \sqrt{T} & 0 \\ 0 & T^{3/2} \end{bmatrix}$$

Multiple this matrix with the coefficient vector would apply different convergence speed to different coefficients:

$$\begin{aligned} \begin{bmatrix} \sqrt{T}(\hat{\alpha} - \alpha) \\ T^{3/2}(\hat{\delta} - \delta) \end{bmatrix} &= \gamma_T \left(\sum_t x_t x_t' \right)^{-1} \left(\sum_t x_t \epsilon_t \right) \\ &= \left[\gamma_T^{-1} \left(\sum_t x_t x_t' \right) \gamma_T^{-1} \right]^{-1} \left[\gamma_T^{-1} \left(\sum_t x_t \epsilon_t \right) \right] \end{aligned}$$

in which

$$\begin{aligned} \gamma_T^{-1} \left(\sum_t x_t x_t' \right) \gamma_T^{-1} &= \begin{bmatrix} T^{-1/2} & \\ & T^{-3/2} \end{bmatrix} \begin{bmatrix} \sum 1 & \sum t \\ \sum t & \sum t^2 \end{bmatrix} \begin{bmatrix} T^{-1/2} & \\ & T^{-3/2} \end{bmatrix} \\ &= \begin{bmatrix} T^{-1} \sum 1 & T^{-2} \sum t \\ T^{-2} \sum t & T^{-3} \sum t^2 \end{bmatrix} \to \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix} = Q. \end{aligned}$$

Turning to the second term:

$$\gamma_T^{-1} \left(\sum_t x_t \epsilon_t \right) = \begin{bmatrix} T^{-1/2} & \\ & T^{-3/2} \end{bmatrix} \begin{bmatrix} \sum \epsilon_t \\ \sum t \epsilon_t \end{bmatrix} = \begin{bmatrix} T^{-1/2} \sum \epsilon_t \\ T^{-1/2} \sum \frac{t}{T} \epsilon_t \end{bmatrix}$$

$T^{-1/2} \sum \epsilon_t \to N(0, \sigma^2)$ by standard CLT. Observe that $\zeta_t = \frac{t}{T} \epsilon_t$ is not serially correlated,

$$\mathbb{E}(\zeta_t \zeta_{t-j}) = \frac{t(t-j)}{T^2} \mathbb{E}(\epsilon_t \epsilon_{t-j}) = 0$$

with stabilized variance

$$\operatorname{var}(T^{-1/2} \sum \zeta_t) = \frac{1}{T} \sum \operatorname{var} \left(\frac{t}{T} \epsilon_t \right) = \frac{\sigma^2}{T^3} \sum t^2 \to \frac{\sigma^2}{3}$$

Therefore, $T^{-1/2} \sum \frac{t}{T} \epsilon_t \to N(0, \frac{\sigma^2}{3})$. We also need to consider the covariance,

$$\text{cov}(T^{-1/2}\sum \epsilon_t, T^{-1/2}\sum \frac{t}{T}\epsilon_t) = \frac{1}{T}\mathbb{E}\left(\sum \epsilon_t \sum \frac{t}{T}\epsilon_t\right) = \frac{\sigma^2}{T^2}\sum t \to \frac{\sigma^2}{2}$$

Therefore, we have

$$\begin{bmatrix} T^{-1/2}\sum \epsilon_t \\ T^{-1/2}\sum \frac{t}{T}\epsilon_t \end{bmatrix} \to \begin{bmatrix} \sigma^2 & \frac{\sigma^2}{2} \\ \frac{\sigma^2}{2} & \frac{\sigma^2}{3} \end{bmatrix} = \sigma^2 Q$$

Finally, putting everything together,

$$\gamma_T(\hat{\beta} - \beta) \to N(0, \sigma^2 Q^{-1}).$$

This means the usual OLS t-test and F-test are asymptotically valid, despite at different convergence rates. After all, trend-stationary process is stationary after de-trending. But unit root process is a totally different species.

! Key Point Summary

1. Trend-stationary process vs stochastic-trend process;
2. Applying different convergence rates to OLS estimator;
3. Usual t-test and F-test are still valid.

20 Unit Root Process

A unit root process is characterized by the presence of unit roots in the character equation of its ARMA representation. The simplest unit root process is an AR(1) process with $\phi = 1$:

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

When $\phi = 1$, it makes each innovation persistent. The effect of past innovations do not fade away no matter how distant they are.

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

The persistence makes the behavior of unit root processes drastically different from stationary processes. Unit root processes hold particular significance among non-stationary processes due to the prevalence of similar behavior in economic or financial time series. For example, stock prices behave a lot like unit root processes (the Random Walk Hypothesis).

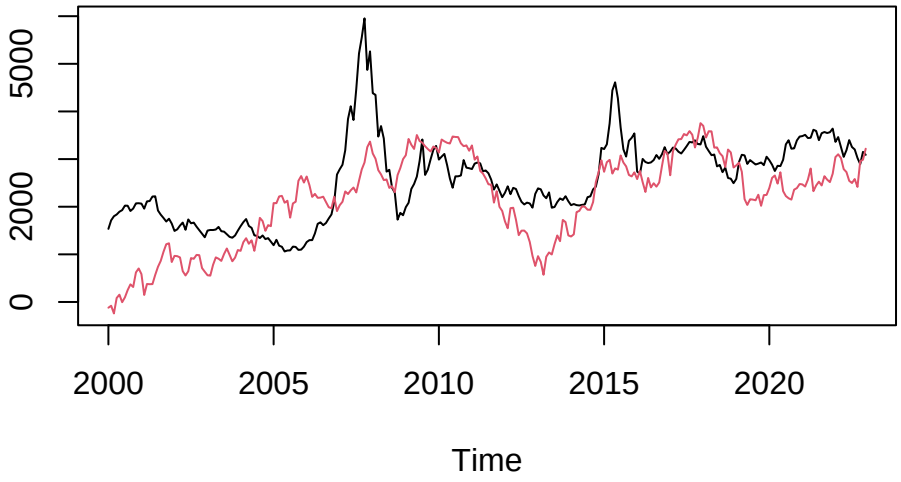


Figure 20.1: Stock market index (black) vs Simulation of a unit root process (red)

The particularity of unit root process makes it a unique class in itself in terms of analytic techniques. The usual OLS estimator and asymptotic normality does not work with unit root processes. If we regress a unit root process on its lags, the OLS estimator is given by

$$\hat{\phi} = \frac{\sum_t y_{t-1} y_t}{\sum_t y_{t-1}^2}$$

We would expect $\sqrt{T}(\hat{\phi} - 1) \to N(0, \omega^2)$. However, this is not the case. To see this, consider

$$T(\hat{\phi} - 1) = \frac{T^{-1} \sum_t y_{t-1} \epsilon_t}{T^{-2} \sum_t y_{t-1}^2}$$

Assuming Gaussian innovation $u_t \sim N(0, \sigma^2)$, we have

$$y_t = \epsilon_t + \epsilon_{t-1} + \cdots + \epsilon_1 \sim N(0, \sigma^2 t)$$

Therefore, $z_t = \frac{y_t}{\sigma\sqrt{t}} \sim N(0, 1)$. Consider the numerator,

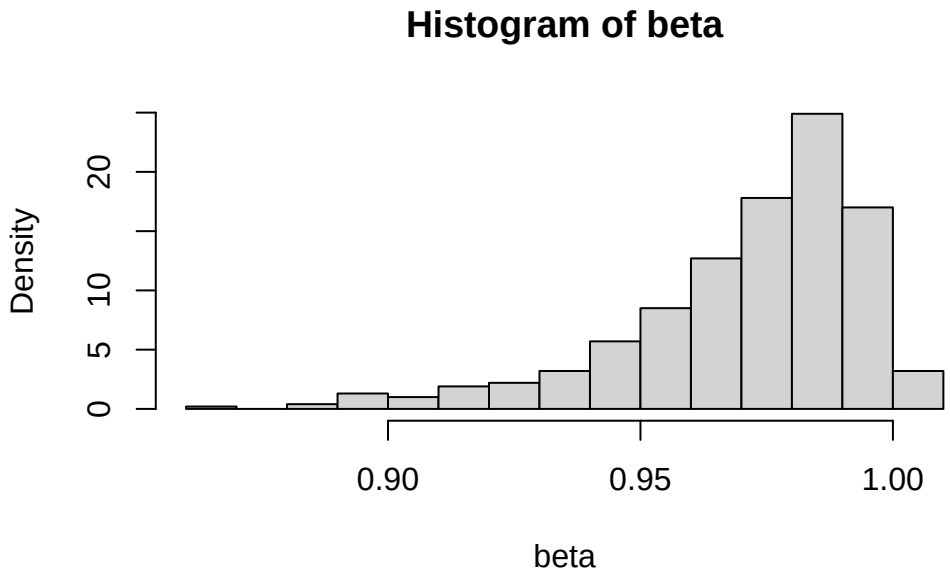
$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} y_{t-1} \epsilon_t &= \frac{1}{T} \sum_{i=1}^{T} (\epsilon_1 + \cdots + \epsilon_{t-1}) \epsilon_t = \frac{1}{T} \sum_{s<t}^{T} \epsilon_s \epsilon_t \\
&= \frac{1}{2T} \left[(\epsilon_1 + \cdots + \epsilon_T)^2 - \sum_{t=1}^{T} \epsilon_t^2 \right] \\
&= \frac{1}{2T} y_T^2 - \frac{1}{2T} \sum_{t=1}^{T} \epsilon_t^2 \\
&= \frac{\sigma^2}{2} \left(\frac{y_T}{\sigma\sqrt{T}} \right)^2 - \frac{1}{2T} \sum_{t=1}^{T} \epsilon_t^2 \\
&= \frac{\sigma^2}{2} z_T^2 - \frac{1}{2T} \sum_{t=1}^{T} \epsilon_t^2
\end{aligned}$$

Since $z_T \sim N(0, 1)$, $z_T^2 \sim \chi^2(1)$. By the LLN, $\frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2 \to \mathbb{E}(\epsilon_t^2) = \sigma^2$. Thus,

$$\frac{1}{T} \sum_{t=1}^{T} y_{t-1} \epsilon_t \to \frac{\sigma^2}{2} (\chi^2(1) - 1).$$

So the asymptotic distribution of $\hat{\phi}$ is non-Gaussian. The conventional statistical inference no longer make sense. If we simulate the distribution of $\hat{\phi}$ by Monte Carlo, we see it is left-skewed. The negative values are almost twice as likely as positive values, meaning two thirds of the time, the estimated $\hat{\phi}$ will be less than the true value 1. Therefore, the OLS estimate of a unit root process is biased.

```r
# Monte Carlo simulation
beta = sapply(1:1000, function(i) {
  y = cumsum(rnorm(200))
  x = dplyr::lag(y)
  coef(lm (y ~ x))[2]
})
hist(beta, freq = FALSE)
```



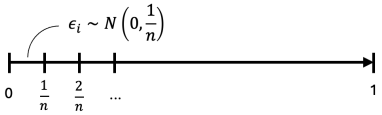
To derive the asymptotic distribution for the unit root process, we need further knowledge of Brownian motions. The idea is derive a continuous version of the unit root process, where each innovation is infinitesimally small. This is the topic of the next section.

21 Brownian Motion

21.1 Continuous random walk

Brownian motion, or **Wiener process**, is the continuous-time extension of a discrete-time random walk. To define the continuous version of a random walk, we cannot simply sum up infinite number of white noises, which will certainly explode. Instead, we chop up a finite interval into infinitely many small intervals, each one corresponding to a tiny Gaussian white noise. The following table shows how a discrete random walk extends to a continuous function.

Table 21.1: Random Walk and Brownian Motion

	Random Walk	Brownian Motion
		
Innovation	$\epsilon_i \sim \text{WN}(0,1)$	
Stochastic process	$y_t = \epsilon_1 + \cdots + \epsilon_t$	$W(t)_{t\in[0,1]} = \lim_{n\to\infty} \sum_{i=1}^{nt} \epsilon_i$
Expectation	$\mathbb{E}[y_t] = 0$	$\mathbb{E}[W(t)] = 0$
Variance	$\text{Var}[y_t] = t\text{Var}[\epsilon_i] = t$	$\text{Var}[W(t)] = nt\text{Var}[\epsilon_i] = t$
Quadratic variation	$\mathbb{E}\sum_{i=1}^{t}(y_i - y_{i-1})^2 = t$	$\int_0^t (dW)^2 = t$

Brownian motion $W(t)$ is a stochastic function. Its realized path is different each time we take a draw from it. But every piece of it follows a tiny Gaussian process. **The function is continuous, but nowhere differentiable.** It is hard to imagine such a function at first glance. But as we proceed, we will appreciate its amazing properties. Despite its path is random, the area under the curve integrates to a well-defined probability distribution. The squared changes (quadratic variation) even sum up to a deterministic constant.

21.2 Some properties

The **quadratic variation** is one of the most important properties of Brownian motions. Intuitively, it says the squared tiny changes sum up to a constant with probability 1 no

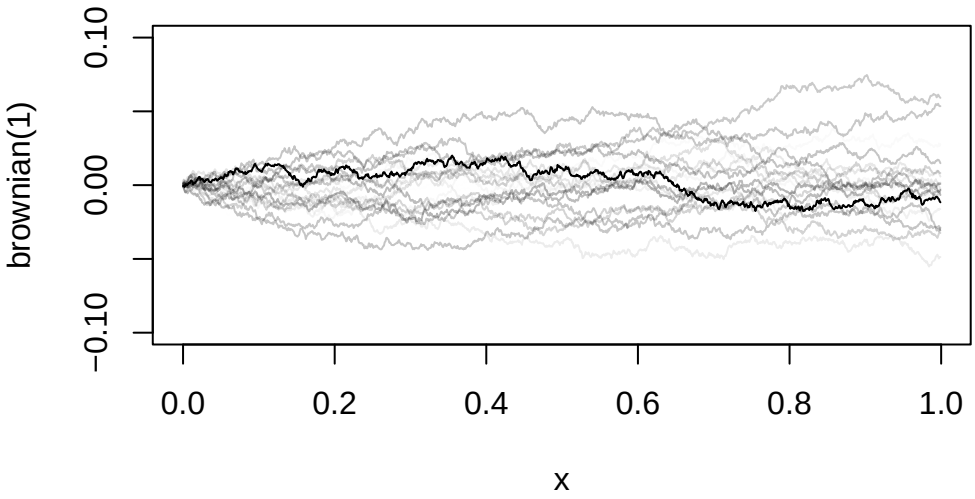


Figure 21.1: Multiple Realizations of Brownian Motion

matter which realized path it takes. Smooth functions will not have such property, because $(dX)^2$ diminishes much faster than dX, surely we get $\int_0^t (dX)^2 = 0$. Because Brownian motion fluctuations too much, the squared changes do not diminish away. To see this, imagine summing up the squares of infinitely many small increments up to t:

$$\lim_{n\to\infty} \sum_{i=1}^{nt} \left[X\left(\frac{i}{n}\right) - X\left(\frac{i-1}{n}\right) \right]^2 = \lim_{n\to\infty} \sum_{i=1}^{nt} \epsilon_i^2,$$

with $\epsilon_i \sim N\left(0, \frac{1}{n}\right)$. Therefore, $\mathbb{E}(\epsilon_i^2) = \frac{1}{n}$. Let $z_i = \epsilon_i^2$. The sum above can be rewritten as

$$\sum_{i=1}^{nt} z_i = nt \left(\frac{1}{nt} \sum_{i=1}^{nt} z_i \right) \overset{\text{LLN}}{\to} nt\mathbb{E}(z_i) = t.$$

The integral version is the quadratic variation

$$\int_0^t (dW)^2 = t, \tag{21.1}$$

or written in differential form

$$(dW)^2 = dt. \tag{21.2}$$

It should be stressed, W is not differentiable. We use the notation dW, but it is not the same as conventional differentials. The calculus for Brownian motions, the Itô calculus, which

will be introduced below, is a different class of calculus specifically designed for stochastic functions. Before we get to that, let's first have a look at some additional properties of Brownian motions.

By Central Limit Theorem, it holds that

$$\frac{1}{\sqrt{nt}}W(t) = \sqrt{nt}\frac{1}{nt}\sum_{i=1}^{nt}\epsilon_i \to N\left(0, \frac{1}{n}\right)$$

Therefore,

$$W(t) \sim \sqrt{nt}N\left(0, \frac{1}{n}\right) \sim N(0, t); \tag{21.3}$$

It follows that, for any $r < s$,

$$W(s) - W(t) \sim N(0, s - t); \tag{21.4}$$

As s and t become arbitrarily close, we have

$$dW \sim N(0, dt). \tag{21.5}$$

In essence, Brownian motion is the accumulation of tiny independent Gaussian innovations. We give the formal definition of Brownian motions below.

Brownian motion

A Brownian motion (Wiener process) is a stochastic function $W(t)$ such that

1. $W(0) = 0$;
2. For $0 \leq t \leq s \leq 1$, $W(s) - W(t) \sim N(0, s - t)$;
3. For any realization, $W(t)$ is continuous in t.

Brownian motions are frequently used to model stock returns. For a fixed horizon T, the returns are normally distributed, with volatility scaled by $\sqrt{T}$. And the returns over different periods are independent, which means no predictability from past returns to future returns.

21.3 Itô calculus

Lemma 21.1. *Let $F(W)$ be a "smooth" function of a Brownian motion $W(t)$. Then*

$$dF = F'dW + \frac{1}{2}F''dt.$$

Proof. For an informal proof, it immediately follows from the Taylor expansion

$$F(W(t+h)) - F(W(t)) = F'(W(t+h) - W(t)) + \frac{1}{2}F''(W(t+h) - W(t))^2 + \cdots$$

As $h \to 0$,

$$dF = F'dW + \frac{1}{2}F''(dW)^2.$$

By Equation 21.2, $(dW)^2 = dt$. Therefore,

$$dF = F'dW + \frac{1}{2}F''dt.$$

□

This formula is known as the **Itô's lemma**, which is the key equation of Itô calculus. Note that how this differs from the differential formula for normal functions: $dF = F'dW$. The *second-order* term does not disappear precisely because the quadratic variation does not go to zero.

Example 21.1. $F(W) = W^2 \implies dF = 2WdW + dt$

Example 21.2. Let's do a more involved example to familiar ourselves with the Itô's lemma, especially how the second-order differentiation of the Brownian motions plays out in computation.

We like to model the continuous-time stock price with Brownian motions. Let S_t be the stock price at time t. Assume the behavior of S_t follows a stochastic differential equation:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW$$

That is, the percentage of S_t is a continuous time random walk with drift μ and volatility σ. Let's compute the log-return of the stock over the horizon T:

$$R_T = \ln(S_T) - \ln(S_0) = f(S_T)$$

Apply second-order Taylor expansion:

$$\begin{aligned}
dR_T &= f' dS_T + \frac{1}{2} f''(dS_T)^2 \\
&= \frac{1}{S_T} dS_T - \frac{1}{2}\frac{1}{S_T^2}(dS_T)^2 \\
&= \frac{dS_T}{S_T} - \frac{1}{2}\left(\frac{dS_T}{S_T}\right)^2 \\
&= (\mu dt + \sigma dW) - \frac{1}{2}(\mu dt + \sigma dW)^2 \\
&= (\mu dt + \sigma dW) - \frac{1}{2}(\mu^2 dt^2 + 2\mu\sigma dt dW + \sigma^2 dW^2) \\
&\to (\mu dt + \sigma dW) - \frac{1}{2}\sigma^2 dt \\
&= \left(\mu - \frac{1}{2}\sigma^2\right) dt + \sigma dW
\end{aligned}$$

The second last step holds because, as $dt \to 0$, the terms dt^2 and $dtdW$ tend to zero faster than dW^2. The only term left is $dW^2 = dt$. If we define integral as the inverse of differentiation, we have

$$\begin{aligned}
R_T = \int_0^T dR_T &= \left(\mu - \frac{1}{2}\sigma^2\right)\int_0^T dt + \sigma \int_0^T dW \\
&= \left(\mu - \frac{1}{2}\sigma^2\right) T + \sigma\sqrt{T}\epsilon_T
\end{aligned}$$

where ϵ_T is a standard normal variable. The model tells us that the log-return of a stock over a fixed horizon of T is normally distributed with mean $(\mu - \sigma^2/2)T$ and standard deviation of

$\sigma\sqrt{T}$. Everything looks familiar, except the Ito's term, $\sigma^2/2$, which comes from the non-zero second-order differential dW^2. The famous Black-Scholes formula for option pricing is derived from this model.

Definition 21.1. Stochastic integrals as the reverse operation of the stochastic differentiation. Note that we change $W(t)$ to $W(s)$ when it enters as the integrand.

1. $\int_0^t dW = W(t)$;
2. $F(W(t)) = \int_0^t f(W(s))dW$, if $dF = f\ dW$;
3. $F(t, W(t)) = \int_0^t f(s, W(s))dW + \int_0^t g(s, W(s))ds$, if $dF = f\ dW + g\ dt$.

Example 21.3. Given $dW^2 = 2W dW + dt$, take integral on both sides:

$$W^2(t) = 2\int_0^t W(s)dW + \int_0^t ds$$

$$\implies \int_0^t W dW = \frac{1}{2}[W^2(t) - t].$$

Setting $t = 1$, it follows that

$$\int_0^1 W dW = \frac{1}{2}[W^2(1) - 1].$$

Note that $W(1) \sim N(0, 1)$. So $W^2(1) \sim \chi^2(1)$ with expectation 1. Thus $\int_0^1 W dW$ is centered at 0 but skewed.

Example 21.4. Given $d(tW) = W dt + t\ dW$ (verify this with Ito's lemma), we have

$$tW(t) = \int_0^t W(s)ds + \int_0^t s\ dW$$

$$\begin{aligned}\implies \int_0^t W(s)ds &= tW(t) - \int_0^t s\ dW \\ &= t\int_0^t dW - \int_0^t s\ dW \\ &= \int_0^t (t-s)dW.\end{aligned}$$

Making Sense of Itô Calculus

Let $F(t)$ be a continuous-time trading strategy that holds an amount of $F(t)$ of a stock at time t. The stock price is a Brownian motion $W(t)$. dW represents the movement of stock price. Consider the Itô's integral

$$Y_t = \int_0^t F\ dW$$

The stochastic integral represents the payoff of the trading strategy up to time t. Note that the integral always evaluates *at the left*. So the strategy can only make decision based on available information.

Proposition 21.1. *Let $W(t)$ be a Brownian motion. Let $f(t)$ be a nonrandom function of time. Then*

1. $\mathbb{E}\left[\int_0^t f(s)dW\right] = 0$;
2. $\mathbb{E}\left[\left(\int_0^t f(s)dW\right)^2\right] = \mathbb{E}\left[\int_0^t f^2 ds\right]$ *(Itô isometry);*
3. $\int_0^t f(s)dW \sim N\left(0, \int_0^t f^2 ds\right)$.

If $f(t)$ represents a trading strategy and the stock price follows a Brownian motion, the theorem tells us the expected payoff of this strategy is zero; more precisely, the payoff follows a Gaussian distribution.

Example 21.5. Following the last example,

$$\int_0^t W ds = \int_0^t (t-s)dW$$

$f(s) = t - s$ is a nonrandom function, apply the theorem above

$$\text{var}\left[\int_0^t W ds\right] = \int_0^t (t-s)^2 ds = \frac{1}{3}t^3$$

Therefore,

$$\int_0^t W(s) ds \sim N\left(0, \frac{1}{3}t^3\right).$$

Thus, the integral, the area under the curve of a Brownian motion, follows a Gaussian distribution.

21.4 Unit root process

Consider a unit root process,

$$y_t = y_{t-1} + \epsilon_t = \sum_{j=1}^{t} \epsilon_j,$$

where $\epsilon_t \sim \text{WN}(0,1)$, $t = 1, 2, \ldots, T$. It is not surprising to see the unit root process converges to Brownian motion if stabilizing it by $T^{-1/2}$:

$$\begin{aligned}
\frac{1}{\sqrt{T}} y_t &= \frac{1}{\sqrt{T}} \sum_{j=1}^{t} \epsilon_j \\
&= \frac{1}{\sqrt{T}} \sum_{j=1}^{Tr} \epsilon_j \quad (r = t/T) \\
&= \sqrt{r} \left(\frac{1}{\sqrt{Tr}} \sum_{j=1}^{Tr} \epsilon_j \right) \\
&\to \sqrt{r} N(0,1) \sim N(0, r) \quad \text{(by CLT)} \\
&\to W(r) \quad \text{(by definition).}
\end{aligned}$$

If ϵ_t has variance σ^2, we would have $T^{-1/2} y_t \to \sigma W(t/T)$. Note if y_t is stationary, y_t will not deviate too far from $\mathbb{E}(y_t)$, we would have $T^{-1/2} y_t \to 0$.

Now let's consider the behaviour of the mean: $\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$. Define $\xi_T(r) = T^{-1/2} y_t$, where $r = t/T$. We have

$$\begin{aligned}
\frac{1}{\sqrt{T}}\bar{y} &= \frac{1}{\sqrt{T}}\left(\frac{1}{T}\sum_{t=1}^{T} y_t\right) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{T}} y_t\right) = \frac{1}{T}\sum_{t=1}^{T}\xi\left(\frac{t}{T}\right) \\
&= \Delta r \sum_{r=0}^{1} \xi(r) \quad (r = t/T, \Delta r = 1/T) \\
&\to \int_0^1 \sigma W(r) dr \quad (\Delta r \to 0)
\end{aligned}$$

Remember for stationary process, we would have $\bar{y} \to \mathbb{E}(y_t)$. With unit root process, the mean no longer converges to a constant, but to a distribution $\int_0^1 W(r)dr \sim N(0, \frac{1}{3})$.

For higher orders of y_t, we have

$$\begin{aligned}
\frac{1}{T^2}\sum_{t=1}^{T} y_t^2 &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{T}} y_t\right)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\xi\left(\frac{t}{T}\right)\right]^2 \\
&= \sum_{r=0}^{1}[\xi(r)]^2 \Delta r \\
&\to \int_0^1 \sigma^2 W^2(r) dr
\end{aligned}$$

By continuous mapping theorem, it can be shown, in general

$$\frac{1}{T^{1+k/2}}\sum_{i=1}^{T} y_t^k \to \sigma^k \int_0^1 W^k(r) dr.$$

Consider the numerator of the OLS estimator of the unit root process,

$$\begin{aligned}
\sum_{t=1}^{T} y_{t-1}\epsilon_t &= \sum_{i=1}^{T}(\epsilon_1 + \cdots + \epsilon_{t-1})\epsilon_t = \sum_{s<t}^{T} \epsilon_s \epsilon_t \\
&= \frac{1}{2}\left[(\epsilon_1 + \cdots + \epsilon_T)^2 - \sum_{t=1}^{T}\epsilon_t^2\right] \\
&= \frac{1}{2} y_T^2 - \frac{1}{2}\sum_{t=1}^{T}\epsilon_t^2
\end{aligned}$$

Divide it by T, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} y_{t-1}\epsilon_t &= \frac{1}{2}\left(\frac{1}{\sqrt{T}}y_T\right)^2 - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_t^2 \\
&= \frac{1}{2}[\xi^2(1) - \hat{\sigma}^2] \\
&\to \frac{1}{2}[\sigma^2 W^2(1) - \sigma^2] \\
&= \frac{1}{2}\sigma^2[W^2(1) - 1] \\
&= \sigma^2 \int_0^1 W dW.
\end{aligned}$$

We summarize the important results below.

Key Rules Summary

1. $\int_0^T W dt = N(0, \frac{T^3}{3})$
2. $\int_0^T W dW = \frac{1}{2}[W^2(T) - T]$
3. $\frac{1}{T^{3/2}}\sum_{t=1}^{T} y_t \to \sigma \int_0^1 W(r)dr$
4. $\frac{1}{T^{1+k/2}}\sum_{i=1}^{T} y_t^k \to \sigma^k \int_0^1 W^k(r)dr$
5. $\frac{1}{T}\sum_{t=1}^{T} y_{t-1}\epsilon_t \to \sigma^2 \int_0^1 W dW$
6. $\frac{1}{T^2}\sum_{t=1}^{T} y_{1t}y_{2t} \to \sigma_1\sigma_2 \int_0^1 W_1(r)W_2(r)dr$

The last rule was given without proof, as we will need it in the following chapters.

22 Unit Root Process (contd)

22.1 Univariate case

We now have all the ingredient to further analyse the unit root process

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

where $\phi = 1$, and its OLS estimator

$$T(\hat{\phi} - 1) = \frac{T^{-1} \sum_t y_{t-1}\epsilon_t}{T^{-2} \sum_t y_{t-1}^2}.$$

We have shown that

$$T^{-1} \sum_t y_{t-1}\epsilon_t \to \sigma^2 \int_0^1 W dW = \frac{\sigma^2}{2}(W^2(1) - 1)$$
$$T^{-2} \sum_t y_{t-1}^2 \to \sigma^2 \int_0^1 W^2 ds$$

Therefore,

$$T(\hat{\phi} - 1) \to \frac{\int_0^1 W dW}{\int_0^1 W^2 ds}.$$

$\int W dW$ is centered around 0, meaning $\hat{\phi}$ is consistent for large samples. But it is biased in small smaples. Moreover, the distribution is not Gaussian, rending all conventional t-test or F-test meaningless. We contrast the properties of stationary processes and unit root processes below.

Table 22.1: Stationary AR(1) process vs unit root process

	Stationary	Unit Root
Model	$y_t = \phi y_{t-1} + \epsilon_t$	$y_t = y_{t-1} + \epsilon_t$

	Stationary	Unit Root
Asymptotic distribution of $\hat{\phi}$	$\sqrt{T}(\hat{\phi}-\phi) \to N(0, 1-\phi^2)$	$\sqrt{T}(\hat{\phi}-1) \to \frac{\int W dW}{\int W^2 dt}$
Asymptotic distribution of t-statistics	$t \to N(0,1)$	$t \to \frac{\int W dW}{\sqrt{\int W^2 dt}}$

22.2 Spurious regression

We now dive deeper into the nature of spurious regression problem presented at the beginning of the chapter. We formulate the problem as below. Suppose

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

where y_t and x_t are unit root processes and there does not exist (α, β) such that the residual u_t is stationary. In this case, OLS is likely to produce spurious result: even if y_t is completely unrelated to x_t, the estimated value of $\hat{\beta}$ is likely to appear to be statistically significantly different from zero.

Spurious regression happens when

1. Dependent/independent variables are non-stationary;
2. The residual is non-stationary for all possible values of the coefficient vector.

To understand why this happens, consider the OLS estimator:

$$\hat{b} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} T & \sum x_t \\ \sum x_t & \sum x_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_t \\ \sum x_t y_t \end{bmatrix}$$

To account for different convergent speed, similar to the trend-stationary case, we multiply the estimators by a matrix,

$$\begin{aligned} \begin{bmatrix} \sqrt{T}^{-1} & \\ & 1 \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} &= \begin{bmatrix} \sqrt{T}^{-1} & \\ & 1 \end{bmatrix} \begin{bmatrix} T & \sum x_t \\ \sum x_t & \sum x_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sqrt{T}^{-1} & \\ & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sqrt{T}^{-1} & \\ & 1 \end{bmatrix} \begin{bmatrix} \sum y_t \\ \sum x_t y_t \end{bmatrix} \\ &= \begin{bmatrix} 1 & T^{-3/2} \sum x_t \\ T^{-3/2} \sum x_t & T^{-2} \sum x_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum T^{-3/2} y_t \\ T^{-2} \sum x_t y_t \end{bmatrix} \\ &\to \begin{bmatrix} 1 & \int W_X dt \\ \int W_X dt & \int W_X^2 dt \end{bmatrix}^{-1} \begin{bmatrix} \int W_Y dt \\ \int W_X W_Y dt \end{bmatrix} \end{aligned}$$

This means, $\hat{\alpha}$ actually diverges. Because it needs to be divided by $\sqrt{T}$ to be able to converge to a stable distribution, rather than being multiplied by a stabilizing factor. $\hat{\beta}$ converges, but it is not consistent. If there is no α, we would have

$$\hat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} \to \frac{\int W_X W_Y dt}{\int W_X^2 dt},$$

which is inconsistent. So we won't get zero even in very large samples.

The OLS estimate of the variance of u_t also diverges. It needs to be divided by T to converge:

$$\begin{aligned}
\frac{1}{T}\hat{\sigma}^2 &= \frac{1}{T^2}\sum (y_t - \hat{\beta} x_t)^2 \\
&= \frac{1}{T^2}\sum y_t^2 - 2\hat{\beta}\frac{1}{T^2}\sum x_t y_t + \hat{\beta}^2 \frac{1}{T^2}\sum x_t^2 \\
&\to \int W_Y^2 dt - 2\hat{\beta}\int W_X W_Y dt + \hat{\beta}^2 \int W_X^2 dt \\
&\to \int W_Y^2 dt - \frac{(\int W_X W_Y dt)^2}{\int W_X^2 dt}.
\end{aligned}$$

The t or F statistics also diverge. t-stat has to be divided by $\sqrt{T}$ to converge; F-stat needs to be divided by T to converge.

$$t = \frac{\hat{\beta}}{\hat{\sigma}}\sqrt{\sum x_t^2} = \sqrt{T}\frac{\hat{\beta}}{\sqrt{T^{-1}\hat{\sigma}^2}}\sqrt{T^{-2}\sum x_t^2} \to \sqrt{T}\cdot C$$

Thus, as sample size T grows, t-test will appear very large and significant, despite y_t and x_t are completely independent.

22.3 Cures for spurious regression

1. Include lagged values of both dependent and independent variables in the regression:

   $$y_t = \alpha + \phi y_{t-1} + \beta x_t + \gamma x_{t-1} + u_t$$

   Now there exists a coefficient vector $[\phi, \beta, \gamma] = [1, 0, 0]$ such that u_t is $I(0)$ stationary. In this case, OLS yields consistent estimates for all the coefficients. t-test converges to Gaussian, though F-test of joint hypotheses has non-standard asymptotic distribution. We will come back to this point later.

2. Difference the data to stationary:

$$\Delta y_t = \alpha + \beta \Delta x_t + u_t$$

   Because Δy_t and Δx_t are all $I(0)$. Standard OLS is valid.

3. Estimate with Cochrane-Orcutt adjustment for first-order correlations in the residuals. This method is asymptotically equivalent to the second method.

22.4 Summary

! Key Point Summary

1. The OLS estimator for unit root coefficient converges to non-standard distributions involving Brownian motions. Thus, standard statistical inferences are meaningless.
2. Regressing unit root processes lead to spurious results, because the diverging behavior of t-stats makes artificially significant values.
3. Include lagged values or difference the data to stationary when working with non-stationary time series.

23 Unit Root Test

A presence of unit root necessitates special treatment in empirical applications. Therefore it is of vital importance to pre-test the existence of unit root.

However, there is no clear cut between stationary and unit root processes for finite samples. Consider an AR(1) process with $\phi = 0.999$, which is a stationary process that behaves very close to a unit root process. In other words, unit root and stationary processes differ in their implications at *infinite* time horizons, but for any *finite* number of observations, there is always a representation from either class of models that could account for all the observed features of the data. So it is *not* possible to tell whether the DGP is stationary or not. We can formulate testable hypothesis only if we were willing to restrict the class of models being considered. Suppose we were committed to an AR(1) model: $y_t = \phi y_{t-1} + \epsilon_t$. The hypothesis $\phi = 1$ is definitely testable.

23.1 Dickey-Fuller Test

Consider an AR(1) process

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

assuming no series correlation in the innovations $u_t \sim IID(0, \sigma^2)$. We have shown that under the hypothesis $\phi = 1$,

$$T(\hat{\phi} - 1) \to \frac{\int W dW}{\int W^2 dt}.$$

We can the hypothesis $H_0 : \ \phi = 1$ utilizing this distribution. The critical values can be obtained by Monte Carlo simulations. The test was proposed by Fuller (1976).

The Dickey-Fuller tests involve three sets of equations depending whether a drift or trend is included, assuming *iid* innovations.

$$\begin{aligned}
\Delta y_t &= \gamma y_{t-1} + u_t \\
\Delta y_t &= \alpha_0 + \gamma y_{t-1} + u_t \\
\Delta y_t &= \alpha_0 + \gamma y_{t-1} + \alpha_2 t + u_t
\end{aligned}$$

Testing $\phi = 1$ is equivalent to testing $\gamma = 0$. The critical values depends on the form of the regression and the sample size (including a drift or trend results in different limiting distributions for γ).

Table 23.1: Critical values of Dickey-Fuller tests

Model	Hypothesis	95%	99%
Default	$\gamma = 0$	-1.95	-2.60
With drift	$\gamma = 0$	-2.89	-3.51
	$\alpha_0 = \gamma = 0$	4.71	6.70
With drift and trend	$\gamma = 0$	-3.45	-4.04
	$\gamma = \alpha_2 = 0$	6.49	8.73
	$\alpha_0 = \gamma = \alpha_2 = 0$	4.88	6.50

23.2 Augmented Dickey-Fuller Test

The assumption that ϵ_t being uncorrelated is too strong for empirical applications. Suppose the data is generated by an AR(p) process with an unit root,

$$\begin{aligned} a(L)y_t &= \epsilon_t \\ (1-L)y_t &= \underbrace{b^{-1}(L)\epsilon_t}_{u_t} \end{aligned}$$

where $a(L) = (1-L)b(L)$ in which $b(L)$ is stationary. In this case, u_t will be autocorrelated. In empirical works, it is more reasonable to assume u_t being serially correlated.

If we difference y_t once, we have

$$\begin{aligned} b(L)\Delta y_t &= \epsilon_t \\ y_t &= y_{t-1} + \sum_{j=1}^{p} \beta_j \Delta y_{t-j} + \epsilon_t \end{aligned}$$

This motivates Dickey-Fuller tests with lags $\{\Delta y_{t-j}\}$. This is called augmented Dickey-Fuller test. The set of equations change to

$$\Delta y_t = \gamma y_{t-1} + \sum_{j=1}^{p} \beta_j \Delta y_{t-j} + u_t$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \sum_{j=1}^{p} \beta_j \Delta y_{t-j} + u_t$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \alpha_2 t + \sum_{j=1}^{p} \beta_j \Delta y_{t-j} + u_t$$

The coefficients on Δy_{t-j} converge to Gaussian. The coefficient on y_{t-1} converges to non-standard distribution. The critical values are unchanged with lags are included.

23.3 Phillips-Perron Test

Another approach to test unit root is proposed by Phillips and Perron (1988), which also assumes autocorrelated errors. Our Brownian motion theories derived from *iid* innovations can be extended to autocorrelated innovations:

$$\xi_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{Tr} u_t \to \omega W(r)$$

where $\omega^2 = \sum_{-\infty}^{\infty} \gamma_j$ is the long-run variance for the autocorrelated process $\{u_t\}$.

But, with autocorrelated errors, the limiting distribution of $\hat{\phi}$ is slightly different, since

$$\begin{aligned} \frac{1}{T} \sum_t y_{t-1} u_t &= \frac{1}{2T} y_T^2 - \frac{1}{2T} \sum_t u_t^2 \\ &\to \frac{1}{2} [\omega^2 W^2(1) - \sigma_u^2] \\ &\to \omega^2 \int W dW + \frac{\omega^2 - \sigma_u^2}{2} \end{aligned}$$

where $\frac{\omega^2 - \sigma_u^2}{2}$ is a "nuisance" parameter. The Phillips and Perron proposed a test statistics correcting the nuisance parameter:

$$T(\hat{\phi} - 1) + \frac{\frac{1}{2}\hat{\omega}^2 - \hat{\sigma}_u^2}{\frac{1}{T^2} \sum y_t^2} \to \frac{\int W dW}{\int W^2 dt}$$

where $\hat{\omega}^2$ can be estimated by Newey-West, $\hat{\sigma}_u^2$ is the estimated variance of the residuals. After the correction, the unit root test can be applied to processes with autocorrelated errors.

24 Cointegration

We have shown that regressing on non-stationary time series might lead to spurious regressions that produce nonsensible large t-values. The conventional asymptotic theorems no longer hold for series with very high persistence such as unit root processes. But unit roots are not always an enemy, sometimes it can be a friend, as in the case of cointegration. In this case, the OLS estimates are super-consistent. They are consistent even when there is an endogeneity issue (which is amazing!)

24.1 Cointegration and super-consistency

Consider a regression with two random walks x_t and y_t:

$$y_t = \beta x_t + e_t$$

where e_t is stationary. In this case, y_t and x_t are **cointegrated**, because the linear combination of the two $I(1)$ processes becomes $I(0)$. If this is the case, we no longer have a spurious regression. In fact, $\hat{\beta}$ is not only consistent, but **super-consistent**. Consider the OLS estimator

$$T(\hat{\beta} - \beta) = \frac{T^{-1} \sum x_t e_t}{T^{-2} \sum x_t^2} \to \frac{\int W_1 dW_2}{\int W_1^2 dt}.$$

Note that $\int W_1 dW_2$ is centered at zero. Hence $\hat{\beta}$ is consistent, despite the distribution is non-Gaussian. It converges at rate T, faster than the standard case $\sqrt{T}$. So it is called super-consistency.

The argument extends to general regressions with multiple regressors. As long as there exists a vector of coefficients that makes the non-stationary variables cointegrated, the OLS estimator is super-consistent. For example,

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

If y_t is $I(1)$, then $y_t - y_{t-1}$ is $I(0)$ (cointegrated with itself). Therefore, $[0, 1]$ is the cointegration vector that makes e_t stationary even if x_t is not cointegrated with y_t. In this case, OLS estimates for $\hat{\rho}$ and $\hat{\beta}$ will be super-consistent.

The intuition of super-consistency is that OLS minimizes squared residuals. If the coefficients deviate from the cointegration vector, $\hat{e}_t^2$ would diverge. $\hat{e}_t^2$ is minimized only when the coefficients coincide with the cointegration vector. That makes OLS converges even faster.

The super-consistency is so strong, that the OLS estimators for cointegrated variables are consistent even when there is endogeneity problem. We demonstrate this with an example. Suppose x_t follows a random walk, and y_t cointegrates with x_t:

$$\begin{aligned} x_t &= x_{t-1} + u_t \\ y_t &= \beta x_t + e_t \end{aligned}$$

Assume u_t and e_t are correlated with $\mathbb{E}(e_t^2) = \mathbb{E}(u_t^2) = 1$ and $\text{cov}(e_t, u_t) = 1$. For simplicity, also assume $e_t = \phi u_t + \sqrt{1-\phi^2}\eta_t$ where η_t is iid standard normal. As e_t is correlated with x_t through u_t, there is clearly an endogeneity problem. The OLS estimator is given by

$$T(\hat{\beta} - \beta) = \frac{\frac{1}{T}\sum x_t e_t}{\frac{1}{T^2}\sum x_t^2} = \frac{\frac{1}{T}\sum x_{t-1} e_t + \frac{1}{T}\sum u_t e_t}{\frac{1}{T^2}\sum x_t^2}$$

where

$$\begin{aligned} \frac{1}{T}\sum x_{t-1} e_t &= \frac{\phi}{T}\sum x_{t-1} u_t + \frac{\sqrt{1-\phi^2}}{T}\sum x_{t-1}\eta_t \\ &\to \phi \int W_1 dW_1 + \sqrt{1-\phi^2}\int W_1 dW_2 \end{aligned}$$

Therefore,

$$T(\hat{\beta} - \beta) \to \frac{\phi \int W_1 dW_1 + \sqrt{1-\phi^2}\int W_1 dW_2 + \phi}{\int W_1^2 dt}.$$

When $\phi \neq 0$, that is endogneity exists, the limiting distribution is shifted. As a result $\hat{\beta}$ has a finite sample bias of order $\frac{1}{T}$. But as $T \to \infty$, the estimator is consistent. Figure 24.1 demonstrates the convergence of $\hat{\beta}$ as sample size increases (the true $\beta = 0.5$).

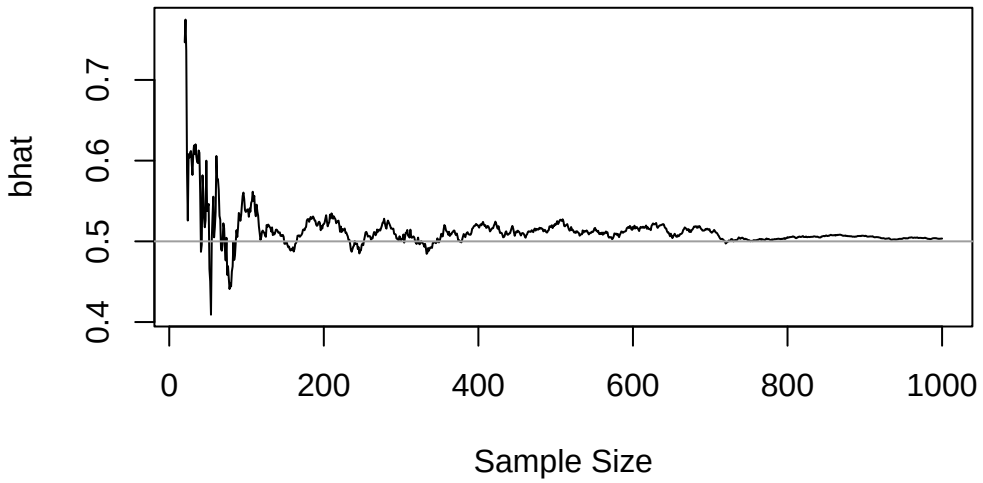


Figure 24.1: Superconsistentcy with cointegrated variables (even with endogeneity)

24.2 Inference under cointegration

We have seen the limiting distributions of persistent regressors could be non-standard. But luckily, in many cases, we can still get Gaussian distributions. This is hard to believe. But it can be shown. Consider a regression of the **canonical form** — that is a regression with four types of regressors: stationary $I(0)$, non-stationary $I(1)$, constant and trend. It can be shown any regression can be rewritten in the canonical form.

$$y_t = \boldsymbol{\gamma} \boldsymbol{z_t} + e_t$$

where

$$\boldsymbol{z}_t = \begin{bmatrix} F_1(L) & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ F_2(L) & G & H & 0 \\ F_3(L) & T & K & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\epsilon}_t \\ 1 \\ \boldsymbol{\eta}_t \\ t \end{bmatrix}$$

and e_t is stationary. Consider the OLS estimator: $\hat{\gamma} = \gamma + (Z'Z)'Z'e$ with a scaling matrix

$$Q = \begin{bmatrix} \sqrt{T}\boldsymbol{I}_{k_1} & & & \\ & \sqrt{T} & & \\ & & T\boldsymbol{I}_{k_3} & \\ & & & T^{3/2} \end{bmatrix}$$

Multiply them together,

$$
\begin{aligned}
Q(\hat{\gamma}-\gamma) &= (Q^{-1}Z'ZQ^{-1})^{-1}Q^{-1}Z'e \\
&= \begin{bmatrix} \underset{k_1\times k_1}{\text{const}} & 0 & 0 & 0 \\ 0 & \multicolumn{3}{c}{\text{Functions of}} \\ 0 & \multicolumn{3}{c}{\text{Brownian}} \\ 0 & \multicolumn{3}{c}{\text{motions}} \end{bmatrix}^{-1} \begin{bmatrix} N(0,?) \\ N(0,?) \\ ?\int W dW? \\ N(0,?) \end{bmatrix}
\end{aligned}
$$

We don't care the specific functional forms of the converged distribution. What matters is the speed of convergence. Note that

$$
\begin{aligned}
\sqrt{T}(\hat{\gamma}_1-\gamma_1) &\to N(0,?) \\
\sqrt{T}(\hat{\gamma}_2-\gamma_2) &\to \text{Something 2} \\
T(\hat{\gamma}_3-\gamma_3) &\to \text{Something 3} \\
T^{3/2}(\hat{\gamma}_4-\gamma_4) &\to \text{Something 4}
\end{aligned}
$$

Constant and stationary regressors have the slowest converging speed $\sqrt{T}$. Only stationary regressors converge to Gaussian distribution.

Now consider a general regression,

$$y_t = \boldsymbol{\beta x_t} + e_t$$

where e_t is stationary. Sims (1990) shows that we can always find a linear combination $\boldsymbol{Dx_t = z_t}$ that transforms the regression into a canonical form

$$y_t = \boldsymbol{\gamma z_t} + e_t$$

where $\boldsymbol{\gamma = \beta D^{-1}}$. This means, if a component of $\hat{\beta}$ can be written as a linear combination of $\hat{\gamma}_1, \hat{\gamma}_3, \hat{\gamma}_4$, its distribution will be dominated by the behavior of $\hat{\gamma}_1$ due to its slower convergence speed. As such, it will behave like asymptotic normal and converge at speed $\sqrt{T}$.

In essence, the coefficients that can be represented as a linear combination involving stationary regressors will be asymptotically normal and converge at rate $\sqrt{T}$. Consider an example,

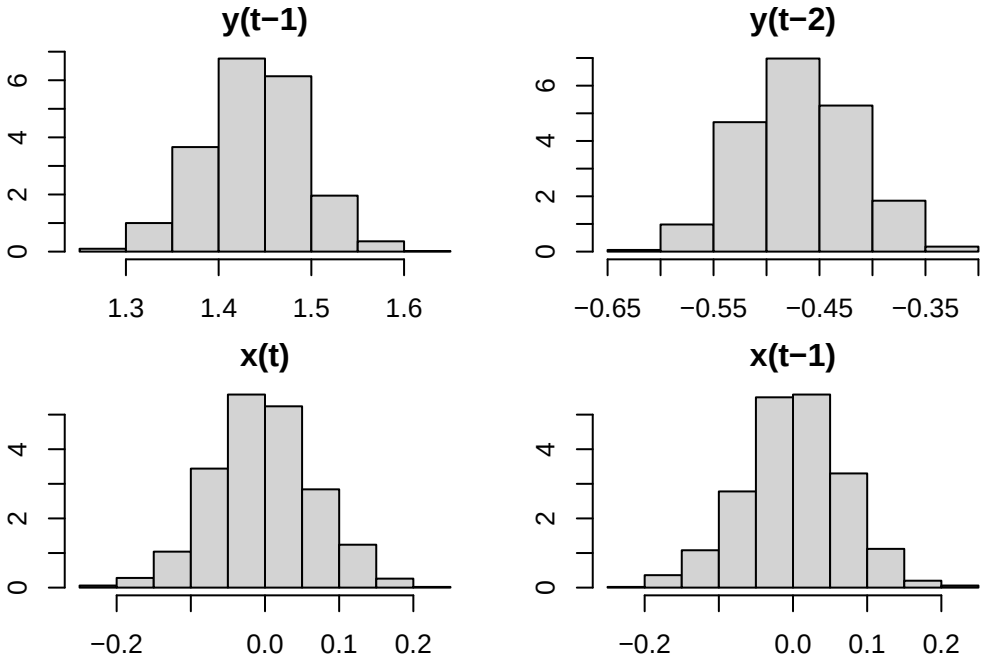
$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \beta_1 x_t + \beta_2 x_{t-1} + e_t$$

where y_t and x_t are $I(1)$. The regression can be rewritten as

$$
\begin{aligned}
y_t &= \alpha + \rho_1 \Delta y_{t-1} + (\rho_1+\rho_2) y_{t-2} + \beta_1 \Delta x_t + (\beta_1+\beta_2) x_{t-1} + e_t \\
&= \alpha + \rho_1 \Delta y_{t-1} + \lambda y_{t-2} + \beta_1 \Delta x_t + \delta x_{t-1} + e_t
\end{aligned}
$$

in which, ρ_1 and β_1 are coefficients on stationary regressors, therefore converging to Gaussian; $\rho_2 = \lambda - \rho_1$ and $\beta_2 = \delta - \beta_1$ are linear combinations involving coefficients on stationary regressors, whose distributions will be dominated by that of ρ_1 and β_1. Hence, all coefficients will have asymptotically normal distributions. Standard inference applies. We verify the claim with a Monte Carlo simulation.

```
library(dynlm)
beta = sapply(1:1000, function(i) {
  x = arima.sim(list(order=c(1,1,0),ar=.5), 200)
  y = arima.sim(list(order=c(0,1,1),ma=.7), 200)
  coef(dynlm(y ~ L(y,1:2) + L(x,0:1)))
}) |> t()
{
  par(mfrow=c(2,2), mar=c(2,2,2,2))
  hist(beta[,'L(y, 1:2)1'], freq=F, main="y(t-1)")
  hist(beta[,'L(y, 1:2)2'], freq=F, main="y(t-2)")
  hist(beta[,'L(x, 0:1)0'], freq=F, main="x(t)")
  hist(beta[,'L(x, 0:1)1'], freq=F, main="x(t-1)")
}
```



24.3 Conclusion

Key Takeaways

1. Regressions with non-stationary regressors are likely to be spurious regressions. Do not regress two non-stationary series unless they are cointegrated.
2. Including lags of dependent and independent variables protests you from spurious regressions.
3. Persistence in time series makes statistical inference complicated. Sometimes standard inference works, but not always.

Epilogue

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.

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

Stock, James H, and Mark W Watson. 2016. "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." In *Handbook of Macroeconomics*, 2:415–525.

———. 2020. *Introduction to Econometrics*. Pearson.

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