

Time Series Econometrics, 2ST111

Lecture 1. Introduction and Overview

Yukai Yang

Department of Statistics, Uppsala University

Yukai Yang

(YY) Floor 3, B322
yukai.yang@statistik.uu.se

Starting Point

Prerequisites

- knowledge about the probability theory and statistics
- knowledge of linear regression
- one programming language: R and RStudio

Broad Outline of the Course

- We use Hamilton's book "Time Series Analysis".
- For the outline, see the schedule.

Theory and Practice

The ultimate goal of the course is:

We want to develop the tools necessary
for analyzing relevant problems in real time series data.

Lectures and classes are:

- **partly theoretical**

Deal with the mathematical structure of the models and explore the properties.

Necessary for understanding the tools (and for being critical towards them!).

- **partly empirical**

Feeling for real data. Hands-on experience.

Promote an interest for doing empirical analyses.

Introduce practical tools to perform analyses for e.g. MA theses.

The Exam

The ultimate goal of the exam is

to test whether you understand.

- the main results and the underlying intuition
- the tools and how they should be applied
- the details of specific econometric models

Today's lecture is about

- what time series econometrics is?
- overview
- readings, software and some resources

What is a time series?

A time series is a set of observations ordered by time.

Time series can be found in

- Economics (price indices, unemployment measurements, ...)
- Finance (stock prices, stock market indices, exchange rates, ...)
- Meteorology (temperature and precipitation records, ...)

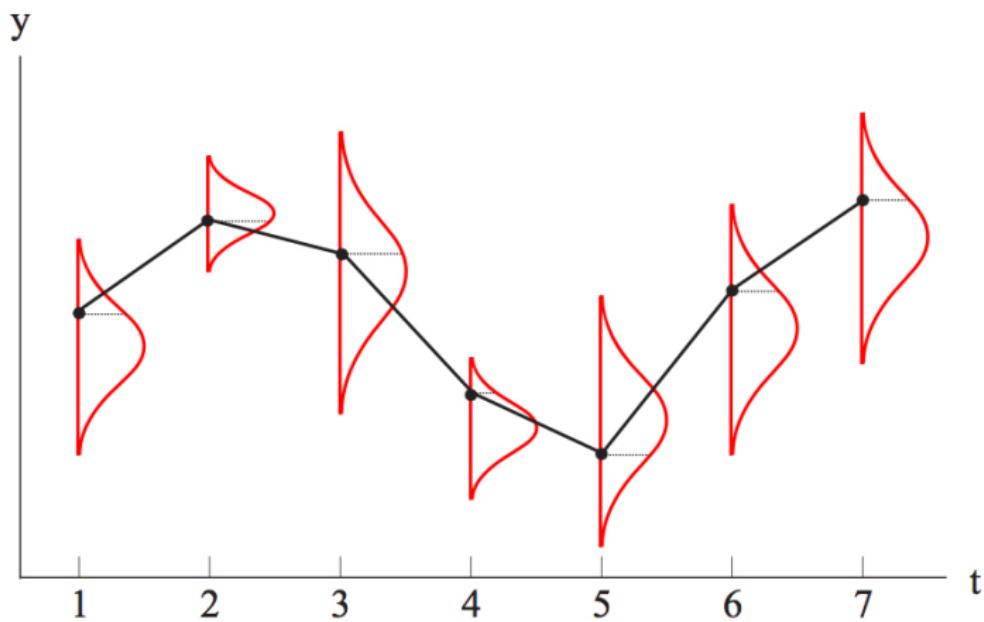
and in many other fields.

These observations are presumed equidistant over the time interval in most cases, but it is not always the case.

Time series data is a realization of a stochastic processes.

What is a time series?

Observation y_t is a realization of a random variable y_t .
Only one observation per random variable!



Example: S&P 500 stock market index

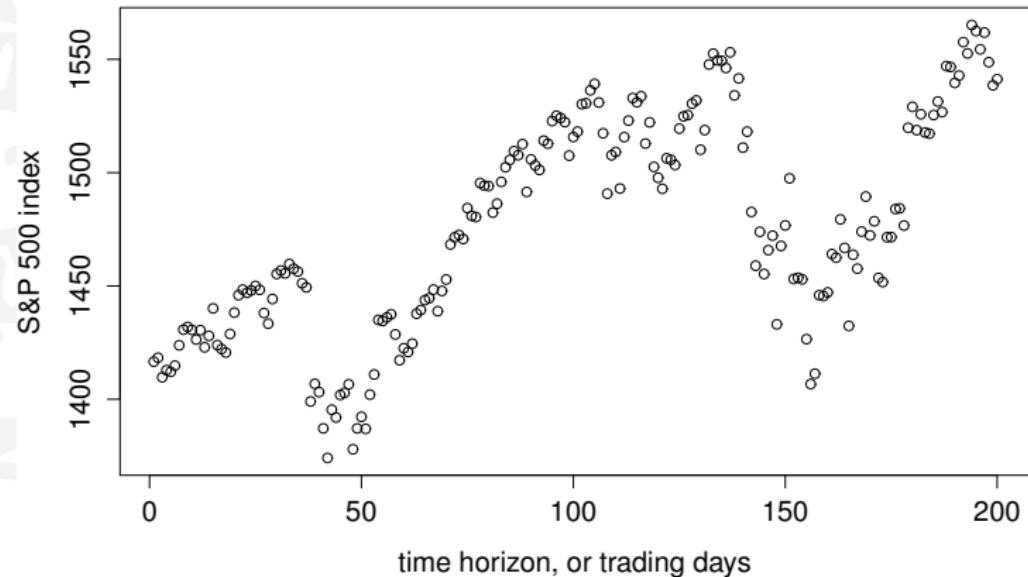
The sequence of the S&P 500 stock market index is a typical time series.
It is

- a portfolio of 500 stocks in the US stock market.
- chosen by some committee based on the market capitalization, liquidity, industry grouping and some other factors.
- supposed to be a leading indicator for the US market equities, and one of the most commonly used benchmarks for it. These companies (500 stocks) form a representative of the industries in the US economy.

Data source: finance.yahoo.com

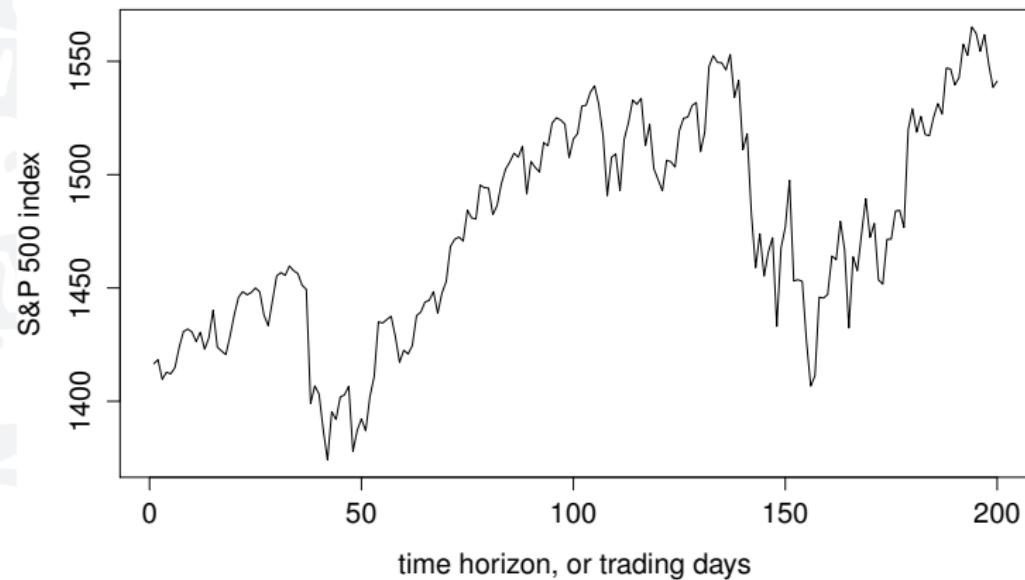
Example: S&P 500 stock market index

S&P daily closing price indices in circles, from 1 Jan to 17 Oct in 2007
not equidistant, 200 trading days



Example: S&P 500 stock market index

S&P daily closing price indices (time series plot), from 1 Jan to 17 Oct in 2007



Example: RW Model

- An example is the random walk (RW) model given by

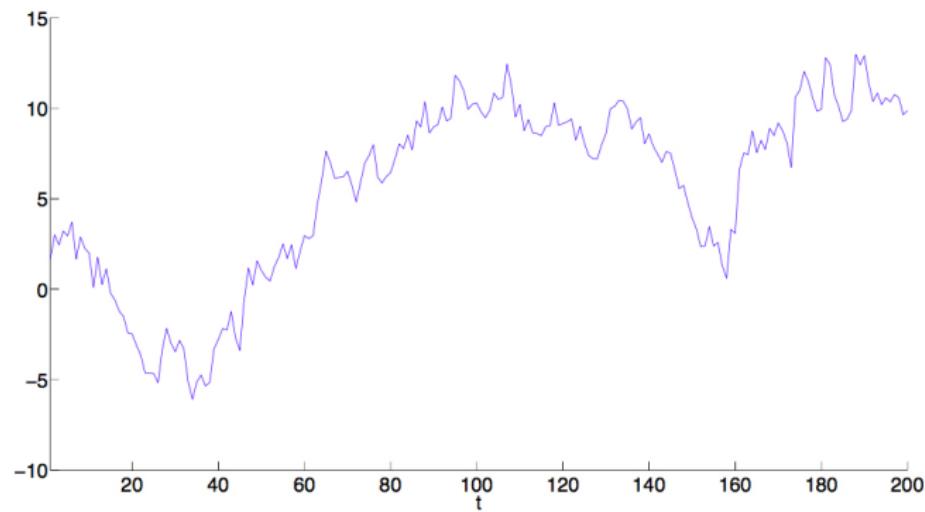
$$y_t = y_{t-1} + u_t \quad (1)$$

for $t = 1, 2, 3, \dots$

- The error term u_t are random variables assumed to be (mutually) independent, identically distributed (*i.i.d.*)
- u_t is independent of $y_{t-1}, y_{t-2}, \dots, y_0$.
- u_t is normally assumed to be normally (Gaussian) distributed with mean zero.

RW Model (one sample path)

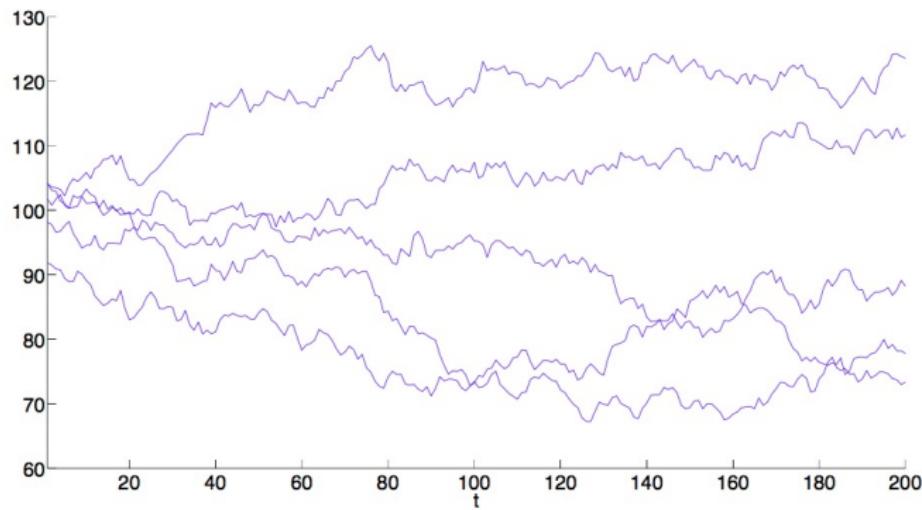
$$y_t = y_{t-1} + u_t$$



Sample path of a RW model with $u_t \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ and $y_0 = 0$.

RW Model (5 sample path)

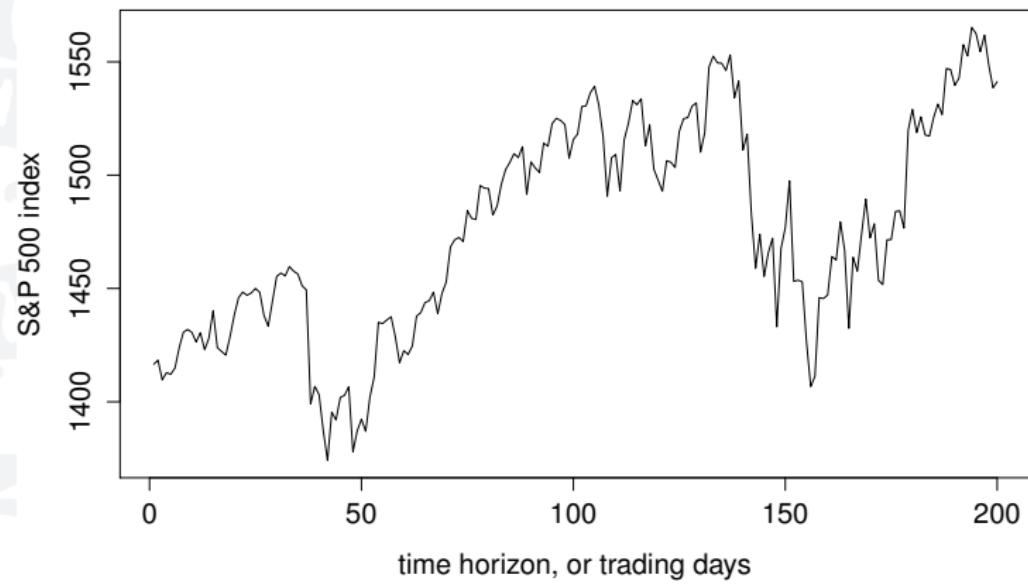
$$y_{i,t} = y_{i,t-1} + u_{i,t}, \quad i = 1, \dots, 5$$



Sample paths of a RW model with $u_{i,t} \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ and $y_{i,0} \stackrel{iid}{\sim} \mathcal{N}(100, 25)$.

Example: S&P 500 stock market index

Is this similar?



Time Series Analysis

- Time series analysis is important because it concerns the extraction of the useful information from the historical data.
- How to conduct the time series analysis given a particular time series?
 - 1 set up a hypothetical statistical model to represent the series in order to obtain insights into the mechanism that generates the data.
 - 2 once a satisfactory model has been formulated, to extrapolate from the model in order to anticipate (forecast) the future values of the time series.
 - 3 control future events via intervention

Example: RW Model

- Now we propose the RW model as the suitable one for daily S&P data.
- The best forecast of tomorrow's value is the current value. That is

$$E_t(y_{t+1}) = E_t(y_t) + E_t(u_{t+1}) = y_t \quad (2)$$

as $E_t(u_{t+1}) = 0$, where $E_t(\cdot)$ is the conditional expectation given the information available at time t .

- A time series econometrician faces the task to construct models capable of forecasting, interpreting, and testing hypothesis concerning economic data.

Model Diagnostics

- Having selected a time series model, the parameters of the model need to be estimated and its goodness of fit to the data can be checked.
- Some fundamental assumptions have to be checked as well, for example, no autocorrelation in residuals, no heteroskedasticity (implied by *i.i.d.* errors), and etc.
- If there are several suitable models for the data, we need to choose the best one based on some information criterion or criteria.
- If the model is satisfactory it may be used for forecasting.

Forecast Evaluation

- Once a time series has been analyzed and its future values have been forecast, it is reasonable to question how good the forecasts are. Typically, there will be several plausible models to extrapolate from in order to forecast the series.
- With forecasts from several models it is inevitable that the sample will show differences in forecast accuracy between the models.
- Because of this it is important to investigate how likely this outcome is due to pure chance, that is, whether the observed difference is statistically significant or not.

"Modelling Economic Series" by C.W.J. Granger

The basic objective is to affect the beliefs - and hence the behaviour - of other research workers and possibly other economic agents. These beliefs may be about the size or signs of certain coefficients (relating to a hypothesis from an economic theory), the quality of a forecasting technique or the relevance of some policy strategy, for example. The degree of belief, about the correctness of some hypothesis say, may be measured as a probability and can be affected by the outcome of an empirical investigation.

For example an economist may say that his or her belief about the statement 'inflation can be controlled by using a policy of money supply being attached to pre-announced slowly growing monetary targets is 0.4'. A carefully conducted empirical study may obtain results that change this economists' belief to 0.6 and consequentially change his or her behaviour. Of course, this propose is not limited to econometric modelling and can also apply to a new economic theory at all levels of sophistication.

Nonlinear/Nonstationary Time Series Models

- Nonlinear and nonstationary time series models have gained more and more attention in the last two decades. The fact is that there are empirical evidences that many realistic time series are non-Gaussian and have a structure that change over time.
- For example, many economic time series are known to show a large number of nonlinear features such as cycles, asymmetries, jumps, thresholds, heteroskedasticity and combinations thereof, that additionally need to be taken into account.

Characteristics of Economic Time Series

Many economic time series do not have a constant mean, and most exhibit phases of relative tranquility followed by periods of high volatility.

Stylized facts:

- a clear trend
- a high degree of persistence, especially from the shocks
- varying volatility
- co-movements between series

The Window

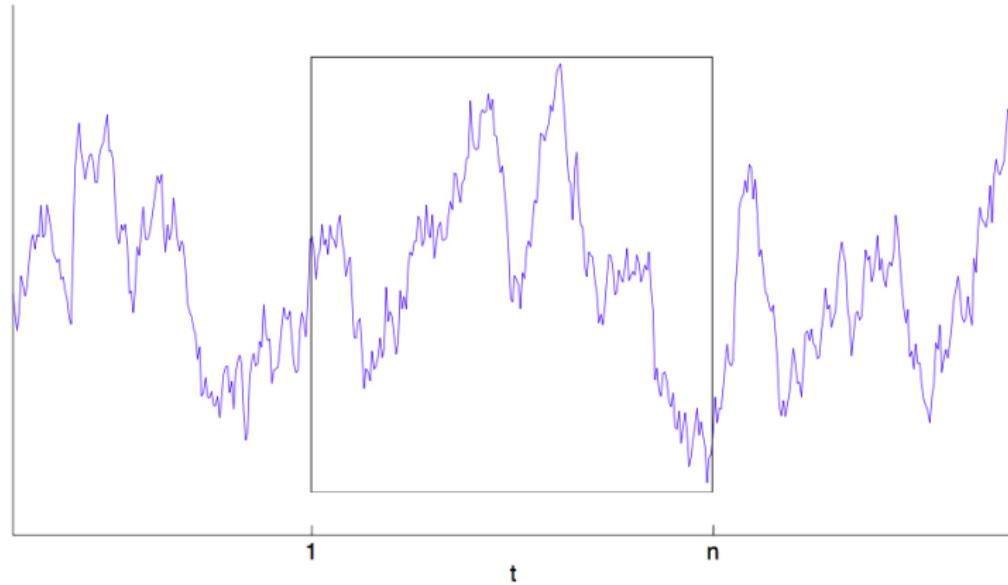
We think of the sample

$$\{y_t\}_{t=1}^n = y_1, y_2, \dots, y_n$$

as a 'window' out of an infinite past and infinite future:

$$\{y_t\}_{t=-\infty}^{\infty} = \dots, y_{-1}, y_0, \underbrace{y_1, y_2, \dots, y_n}_{\text{sample}}, y_{n+1}, y_{n+2}, \dots$$

The Window



Sample paths of a stochastic process $\{y_t\}_{t=-\infty}^{\infty}$.

LLN and CLT (iid case)

Let y_1, \dots, y_n be a sequence of *iid* random variables with finite mean and variance μ and σ^2 , respectively, and let $\bar{y}_n = \frac{1}{n} \sum_{t=1}^n y_t$.

- Law of Large Numbers (LLN)

$$\bar{y}_n \xrightarrow{P} \mu \quad \text{as } n \rightarrow \infty$$

- Central Limit Theorem (CLT)

$$\sqrt{n} \frac{\bar{y}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty$$

The assumption of independence tends to be made rather casually, even though it is often inappropriate.

LLN and CLT for dependent random variables

Question:

Does the LLN and CLT apply when y_1, \dots, y_n are dependent?

Answer:

Yes, but under certain conditions.

Example: MA(1) Model

A simple time series model with temporal dependence is the moving average model of order 1 (MA(1)).

$$y_t = \theta u_{t-1} + u_t,$$

where u_0, \dots, u_n is a sequence of *iid* random variables with finite mean and variance μ_u and σ_u^2 , respectively.

Thus, we have $\mu = E(y_t) = (\theta + 1)\mu_u$ and $Cov(y_t, y_{t-1}) = \theta\sigma_u^2$ for all integers t .

y_1, \dots, y_n are dependent random variables if $\theta \neq 0$.

Example: MA(1) Model

Moreover,

$$\bar{y}_n = \frac{1}{n} \sum_{t=1}^n (\theta u_{t-1} + u_t) = \theta \times \underbrace{\frac{1}{n} \sum_{t=1}^n u_{t-1}}_{\xrightarrow{P} \mu_u} + \underbrace{\frac{1}{n} \sum_{t=1}^n u_t}_{\xrightarrow{P} \mu_u} \xrightarrow{P} \mu$$

as $n \rightarrow \infty$. The LLN still applies.

It can be shown that

$$\sqrt{n} \frac{\bar{y}_n - \mu}{(\theta + 1)\sigma_u} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty.$$

The CLT applies.

Example: MA(1) Model, Proof for the CLT

By defining $\bar{u}_n = \frac{1}{n} \sum_{t=1}^n u_t$, we have $\bar{y}_n = \theta(\bar{u}_n + \frac{1}{n}u_0 - \frac{1}{n}u_n) + \bar{u}_n$.

By rearranging, we have

$$\bar{y}_n = (\theta + 1)\bar{u}_n + \underbrace{\frac{\theta}{n}(u_0 - u_n)}_{\xrightarrow{P} 0}.$$

Remember that $\sqrt{n}(\bar{u}_n - \mu_u)/\sigma_u \xrightarrow{d} \mathcal{N}(0, 1)$.

It follows that

$$\sqrt{n} \frac{\bar{y}_n - \mu}{(\theta + 1)\sigma_u} = \sqrt{n} \frac{\bar{u}_n - \mu_u}{\sigma_u} + \underbrace{\frac{\theta(u_0 - u_n)}{\sqrt{n}(\theta + 1)\sigma_u}}_{\xrightarrow{P} 0} \xrightarrow{d} \mathcal{N}(0, 1).$$

Central Question

How far can we relax the independence assumption of the classical theory and still preserve the validity of the LLN and CLT?

Recommended Readings

- Hayashi (2000), "Econometrics", Princeton University Press.
- Fuller (1995), "Introduction to Statistical Time Series", Wiley.
- Brockwell & Davis (2009), "Time Series: Theory and Methods", Springer.
- Lütkepohl (2010), "New Introduction to Multiple Time Series Analysis", Springer.
- White (2000), "Asymptotic Theory for Econometricians", Academic Press.

Web Resources

- MIT video courses in mathematics,
<http://ocw.mit.edu/courses/most-visited-courses/>



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 2. Difference Equations and Lag Operators

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Difference Equations (Hamilton, pp. 1-24)
 - first-order equations
 - p th-order equations
- Lag Operators (Hamilton, pp. 25-42)
 - first-order equations
 - p th-order equations
 - initial conditions

First-Order Difference Equations

Denote y_t the value of a variable at time t .

A linear first-order difference equation

$$y_t = \phi y_{t-1} + w_t \quad (1)$$

is an expression relating the variable y_t to its previous values.

- y_t as a linear function of y_{t-1} and w_t
- first-order due to that only y_{t-1} enters
- affine transformation

Example: Goldfeld's Model

Goldfeld's model (1973), estimated money demand function for US:

$$\begin{aligned}m_t &= 0.72m_{t-1} + w_t \\w_t &= 0.27 + 0.19I_t - 0.045r_{bt} - 0.019r_{ct}\end{aligned}\quad (2)$$

- m_t the log of the real money holdings of the public
- I_t the log of aggregate real income
- r_{bt} the log of the interest rate on bank accounts
- r_{ct} the log of the interest rate on commercial paper

First-Order Difference Equations

We have seen that w_t is **deterministic**, which means that y_t is perfectly predictable.

Question:

If a dynamic system is described by $y_t = \phi y_{t-1} + w_t$, what are the effects on y of changes in the value of w ?

Recursive Substitution

The answer is given by **Recursive Substitution**.

Let us expand y_2 in the following way:

$$\begin{aligned}y_2 &= \phi y_1 + w_2 = \phi(\phi y_0 + w_1) + w_2 \\&= \phi^2 y_0 + \phi w_1 + w_2.\end{aligned}\tag{3}$$

Likewise, for y_3 we have

$$\begin{aligned}y_3 &= \phi y_2 + w_3 = \phi(\phi^2 y_0 + \phi w_1 + w_2) + w_3 \\&= \phi^3 y_0 + \phi^2 w_1 + \phi w_2 + w_3.\end{aligned}\tag{4}$$

Recursive Substitution

By Recursive Substitution,

$$\begin{aligned}y_t &= \phi^t y_0 + \phi^{t-1} w_1 + \phi^{t-2} w_2 + \dots + w_t \\&= \phi^t y_0 + \sum_{i=1}^t \phi^{t-i} w_i.\end{aligned}\tag{5}$$

The effect on y_t of changing the value of w_i is, *ceteris paribus*,

$$\frac{\partial y_t}{\partial w_i} = \phi^{t-i},\tag{6}$$

where $\partial y_t / \partial w_i$ denotes the partial derivative of y_t w.r.t. w_i .

Dynamic Multipliers

Let us expand $y_{t+\tau}$ instead of y_t recursively up to y_{-k} :

$$\begin{aligned}y_{t+\tau} &= \phi^{t+\tau+k} y_{-k} + \phi^{t+\tau+k-1} w_{-k+1} + \dots + w_{t+\tau} \\&= \phi^{t+\tau+k} y_{-k} + \sum_{i=-k+1}^{t+\tau} \phi^{t+\tau-i} w_i,\end{aligned}\tag{7}$$

with

$$\frac{\partial y_{t+\tau}}{\partial w_i} = \phi^{t+\tau-i}.\tag{8}$$

Note that k is **not** involved.

By setting $i = t$, we have the **Dynamic Multiplier**

$$\frac{\partial y_{t+\tau}}{\partial w_t} = \phi^\tau,\tag{9}$$

only depending on τ .

Dynamic Multipliers

Remarks for the Dynamic Multiplier

$$\frac{\partial y_{t+\tau}}{\partial w_t} = \phi^\tau$$

- $0 < \phi < 1$, ϕ^τ decays geometrically.
- $-1 < \phi < 0$, ϕ^τ alternates in sign, $|\phi^\tau|$ decays geometrically.
- $1 < \phi$, ϕ^τ increases exponentially.
- $\phi < -1$, ϕ^τ alternates in sign, $|\phi^\tau|$ increases exponentially.

See Figure 1.1 on pp.4 in Hamilton.

Dynamic Multipliers

- The dynamic system is called **stable** if $|\phi| < 1$ and **explosive** if $|\phi| > 1$.
- The τ th dynamic multiplier is the response of y τ -step ahead to a single impulse in w . It is also referred to as the **impulse-response function**.
- Think about what if $|\phi| = 1$.

Long Run Effect

Sometimes we are interested in the effect of a **permanent change** in w , i.e. the effect when $w_t, w_{t+1}, \dots, w_{t+\tau}$ all increase by one unit. Consider again

$$y_{t+\tau} = \phi^{t+\tau+k} y_{-k} + \sum_{i=-k+1}^{t+\tau} \phi^{t+\tau-i} w_i.$$

Let $k = 1 - t$, we have

$$y_{t+\tau} = \phi^{\tau+1} y_{t-1} + \sum_{i=t}^{t+\tau} \phi^{t+\tau-i} w_i.$$

Thus, if $w_i = 1$ for $i = t, \dots, t + \tau$ (one unit), the **Long-Run Effect**

$$\sum_{i=t}^{t+\tau} \frac{\partial y_{t+\tau}}{\partial w_i} = \sum_{i=t}^{t+\tau} \phi^{t+\tau-i} = \phi^\tau + \phi^{\tau-1} + \dots + 1.$$

When $\tau \rightarrow \infty$, it converges to $1/(1 - \phi)$, if $|\phi| < 1$.

Cumulative Effect

We may be also interested in the **Cumulative Effect** of a one unit increase in w_t , that is

$$\sum_{\tau=0}^{\infty} \frac{\partial y_{t+\tau}}{\partial w_t}. \quad (10)$$

Provided that $|\phi| < 1$, it is the same as the long-run effect $1/(1 - \phi)$.

p th-Order Difference Equations

The linear first-order difference equation

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

is a special case ($p = 1$) of the linear p th-Order Difference Equation

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + w_t \quad (11)$$

in which the value of y at time t depends on p of its own lags
 $(y_{t-1}, \dots, y_{t-p})$ and the current value of w .

First-Order Vector Difference Equations

It is often convenient to rewrite the p th-order scalar difference equation as a **First-Order Vector Difference Equation**. Denote

$$\xi_t = \begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{pmatrix}_p, \quad \mathbf{F} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}_{p \times p}, \quad \mathbf{v}_t = \begin{pmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_p$$

Consider the following first-order vector difference equation:

$$\xi_t = \mathbf{F}\xi_{t-1} + \mathbf{v}_t. \tag{12}$$

In particular, when $p = 1$, $\xi_t = y_t$, $\mathbf{F} = \phi_1$, and $\mathbf{v}_t = w_t$ (first-order difference equation).

First-Order Vector Difference Equations

More clearly, the system of equations are

$$\begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (13)$$

Remarks

- The first-order vector system is equivalent to the p th-order scalar system (11).
- The advantage of rewriting the p th-order scalar system into a first-order vector system is that the latter one is often easier to handle.

First-Order Vector Difference Equations

Given the first-order vector difference equation (12), we expand $\xi_{t+\tau}$ up to $t - 1$ by recursive substitution as follows:

$$\xi_{t+\tau} = \mathbf{F}^{\tau+1} \xi_{t-1} + \mathbf{F}^\tau \mathbf{v}_t + \mathbf{F}^{\tau-1} \mathbf{v}_{t+1} + \dots + \mathbf{v}_{t+\tau}. \quad (14)$$

The system of equations are

$$\begin{pmatrix} y_{t+\tau} \\ y_{t+\tau-1} \\ y_{t+\tau-2} \\ \vdots \\ y_{t+\tau-p+1} \end{pmatrix} = \mathbf{F}^{\tau+1} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{pmatrix} + \mathbf{F}^\tau \begin{pmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \dots + \begin{pmatrix} w_{t+\tau} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (15)$$

First-Order Vector Difference Equations

Denote

$$\mathbf{F}^s = \begin{pmatrix} f_{11}^{(s)} & f_{12}^{(s)} & \dots & f_{1p}^{(s)} \\ f_{21}^{(s)} & f_{22}^{(s)} & \dots & f_{2p}^{(s)} \\ \vdots & \vdots & \ddots & \vdots \\ f_{p1}^{(s)} & f_{p2}^{(s)} & \dots & f_{pp}^{(s)} \end{pmatrix} \quad (16)$$

For the first equation, we have

$$\begin{aligned} y_{t+\tau} &= f_{11}^{(\tau+1)} y_{t-1} + f_{12}^{(\tau+1)} y_{t-2} + \dots + f_{1p}^{(\tau+1)} y_{t-p} + \\ &\quad \color{blue}{f_{11}^{(\tau)} w_t} + f_{11}^{(\tau-1)} w_{t+1} + \dots + w_{t+\tau}. \end{aligned} \quad (17)$$

Thus, the dynamic multiplier (at time t for τ -step ahead) is given by

$$\frac{\partial y_{t+\tau}}{\partial w_t} = \color{blue}{f_{11}^{(\tau)}} \quad (18)$$

Dynamic Multiplier

- For $p = 1$, $f_{11}^{(\tau)} = \phi_1^\tau$.
- More generally, for any positive integer p ,

$$\frac{\partial y_{t+1}}{\partial w_t} = f_{11}^{(1)} = \phi_1, \quad \frac{\partial y_{t+2}}{\partial w_t} = f_{11}^{(2)} = \phi_1^2 + \phi_2. \quad (19)$$

- Recall the impulse-response function.

Dynamic Multiplier

- For larger values of τ , Hamilton suggests to compute $f_{11}^{(\tau)}$ by numerical simulation, see Hamilton pp.10.
- A simple analytical characterization of the dynamic multiplier (18) can be obtained in terms of the eigenvalues of the matrix \mathbf{F} .
- The reason: it is related to the power of matrix \mathbf{F} .
- Recall that the eigenvalues of matrix \mathbf{F} are those (complex) numbers λ who satisfy $|\mathbf{F} - \lambda \mathbf{I}_p| = 0$.
- For a general p th-order system, this determinant is a p th-order polynomial in λ whose p solutions are the eigenvalues of \mathbf{F} . See Proposition 1.1 on pp.10 and its proof in Appendix 1.A on pp.21 in Hamilton.

General Solution of a p th-Order Difference Equation

Distinct Eigenvalues

The matrix \mathbf{F} with distinct eigenvalues can be decomposed (eigenvalue decomposition) as follows

$$\mathbf{F} = \mathbf{T}\Lambda\mathbf{T}^{-1}.$$

Remarks:

- The columns of the $p \times p$ matrix \mathbf{T} are the eigenvectors of \mathbf{F} .
- The elements on the main diagonal of the $p \times p$ diagonal matrix Λ are the eigenvalues.
- The decomposition is not unique. Different columns of \mathbf{T} can be switched, but certain eigenvalue corresponds to its eigenvector at certain position.
- Most software functions keep the eigenvalues in decreasing order.

General Solution of a p th-Order Difference Equation

Distinct Eigenvalues

It is related to the power of the matrix. To see this, check for example
 $\tau = 2$

$$\begin{aligned}\mathbf{F}^2 &= \mathbf{T} \boldsymbol{\Lambda} \mathbf{T}^{-1} \cdot \mathbf{T} \boldsymbol{\Lambda} \mathbf{T}^{-1} = \mathbf{T} \boldsymbol{\Lambda} (\mathbf{T}^{-1} \mathbf{T}) \boldsymbol{\Lambda} \mathbf{T}^{-1} \\ &= \mathbf{T} \boldsymbol{\Lambda} \boldsymbol{\Lambda} \mathbf{T}^{-1} = \mathbf{T} \boldsymbol{\Lambda}^2 \mathbf{T}^{-1}.\end{aligned}$$

By induction, we have the general result

$$\mathbf{F}^\tau = \mathbf{T} \boldsymbol{\Lambda}^\tau \mathbf{T}^{-1}. \quad (20)$$

where

$$\boldsymbol{\Lambda}^\tau = \begin{pmatrix} \lambda_1^\tau & 0 & \cdots & 0 \\ 0 & \lambda_2^\tau & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p^\tau \end{pmatrix}$$

General Solution of a p th-Order Difference Equation

Distinct Eigenvalues

Proposition 1.2 on pp.12 in Hamilton says that the dynamic multiplier has the close form

$$\frac{\partial y_{t+\tau}}{\partial w_t} = f_{11}^{(\tau)} = c_1 \lambda_1^\tau + c_2 \lambda_2^\tau + \dots + c_p \lambda_p^\tau \quad (21)$$

where

$$c_i = \frac{\lambda_i^{p-1}}{\prod_{k \neq i} (\lambda_i - \lambda_k)}.$$

Remarks:

- It can be shown that $\sum_{i=1}^p c_i = 1$. The dynamic multiplier is a **weighted average** of λ_i^τ .
- Some of the eigenvalues may be complex. They will appear as complex conjugates.

General Solution of a Second-Order Difference Equation

Distinct Eigenvalues

A summary of the dynamics for a Second-Order Difference Equation with a nice graph are given on pp.17-18 in Hamilton.

General Solution of a p th-Order Difference Equation

Repeated Eigenvalues

What if \mathbf{F} has repeated eigenvalues? Note that some c_i does not exist.

Solution: The previous result for the dynamic multiplier can be generalized using the Jordan decomposition.

$$\mathbf{F} = \mathbf{M}\mathbf{J}\mathbf{M}^{-1}. \quad (22)$$

See pp.18-19 in Hamilton for details.

Infinite History

If the modulus (absolute value) of the eigenvalues of \mathbf{F} are all less than one, that is, $|\lambda_i| < 1$, \mathbf{F}^τ goes to zero as $\tau \rightarrow \infty$.

If all values of w and y are taken to be bounded, we can think of a 'solution' of y_t in terms of the infinite history of w

$$y_t = w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \psi_3 w_{t-3} + \dots \quad (23)$$

where, likewise, $\psi_\tau = \partial y_{t+\tau} / \partial w_t = f_{11}^{(\tau)}$ is the row 1 column 1 element of \mathbf{F}^τ .

Cumulative Effect

Again, if all the eigenvalues of \mathbf{F} are less than one in modulus, it can be shown that the cumulative effect of a one-time change in w on y is

$$\sum_{\tau=0}^{\infty} \frac{\partial y_{t+\tau}}{w_t} = \frac{1}{1 - \phi_1 - \phi_2 - \dots - \phi_p} \quad (24)$$

Sample, Window and Time Series

We think of the sample

$$\{y_t\}_{t=1}^n = y_1, y_2, \dots, y_n$$

as a 'window' out of an infinite past and infinite future

$$\{y_t\}_{t=-\infty}^{\infty} = \dots, y_{-1}, y_0, \underbrace{y_1, y_2, \dots, y_n}_{\text{sample}}, y_{n+1}, y_{n+2}, \dots$$

The time series $\{y_t\}_{t=-\infty}^{\infty}$ is typically identified by its t th element.

Time Series Operators

A time series operator transforms one or more time series into a new time series.

Example (multiplication operator)

$$y_t = \beta x_t$$

Example (addition operator)

$$y_t = x_t + w_t$$

Note that they are transformations from $\{x_t\}_{t=-\infty}^{\infty}$ and $\{w_t\}_{t=-\infty}^{\infty}$ to $\{y_t\}_{t=-\infty}^{\infty}$, not just one observation at t .

Time Series Operators

Since the multiplication or addition operators amount to element-by-element multiplication or addition, they obey the fundamental laws of algebra (the commutative, associate and distributive laws).

For example (distributive),

$$\beta x_t + \beta w_t = \beta(x_t + w_t)$$

The Lag Operator

A highly useful time series operator is the **Lag Operator**, L .

By definition,

$$L x_t = x_{t-1}. \quad (25)$$

Furthermore,

$$L(L x_t) = L x_{t-1} = x_{t-2}.$$

The associate law holds, and then we have $L(L x_t) = (LL)x_t$. And we define the power of the lag operator $L^2 = LL$.

By induction, we have the general form

$$L^k x_t = x_{t-k}, \quad \text{for } k = 0, 1, 2, \dots \quad (26)$$

and the special case $L^0 x_t = x_t$.

The inverse of the lag operator can also be defined, $L^{-k} x_t = x_{t+k}$, and in general we have $L^{-j} L^k = L^{k-j}$.

The Lag Operator

Remarks:

- The lag operator is a **unary operator**, which only requires one operand. So it belongs to the family of the minus sign ($-$) or the factorial ($!$), but totally different from the multiplication (\times) and the addition ($+$) operators who are binary and require two operands.
- The lag operator is commutative **with** some other operators, and therefore, the lag operator is distributive **over** those operators.

$$L(x_t + w_t) = Lx_t + Lw_t, \quad (\text{distributive over } +)$$

Applying $+$ first (LHS) or L first (RHS) produces the same result (it commutes $+$).

$$L(x_t \cdot w_t) = Lx_t \cdot Lw_t, \quad (\text{distributive over } \cdot)$$

- The special case, the lag of a constant

$$L\beta = \beta$$

For better understanding, consider the lag operator L implies a function $\text{lag}(x_t) = x_{t-1}$ with only one argument (unary), the addition operator implies a function $\text{add}(x, y) = x + y$ with two arguments (binary).

The lag operator commutes the addition operator implies that

$$\text{lag}(\text{add}(x_t, y)) = \text{add}(\text{lag}(x_t), \text{lag}(y_t)). \quad (27)$$

The same result holds for the multiplication operator, and division, but not all (because you can define any kind of operator as you wish).

The Lag Operator

We think of the lag operator as a third operator in addition to the addition and the multiplication, and then we apply the fundamental laws of algebra carefully.

For example, you can do

$$y_t = (\alpha + \beta L)Lx_t \iff y_t = (\alpha L + \beta L^2)x_t$$

or

$$y_t = (1 - \lambda_1 L)(1 - \lambda_2 L)x_t \iff y_t = (1 - \lambda_2 L - \lambda_1 L - \lambda_1 \lambda_2 L^2)x_t$$

The expressions such as $\alpha L + \beta L^2$ and $1 - \lambda_2 L - \lambda_1 L - \lambda_1 \lambda_2 L^2$ without time varying terms x_t are referred to as **polynomials in the lag operator** or simply **lag polynomials**.

First-Order Difference Equations (revisited)

The first-order difference equation can be written in terms of the lag operators

$$y_t = \phi L y_t + w_t \iff (1 - \phi L) y_t = w_t. \quad (28)$$

Consider 'multiplying' both sides of (28) by the lag polynomial

$$1 + \phi L + \phi^2 L^2 + \dots + \phi^{t-1} L^{t-1}.$$

This yields

$$(1 - \phi^t L^t) y_t = (1 + \phi L + \phi^2 L^2 + \dots + \phi^{t-1} L^{t-1}) w_t \quad (29)$$

or equivalently (same as that from recursive substitution),

$$y_t = \phi^t y_0 + w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \dots + \phi^{t-1} w_1 \quad (30)$$

First-Order Difference Equations (revisited)

Since $(1 + \phi L + \phi^2 L^2 + \dots + \phi^{t-1} L^{t-1})(1 - \phi L) = 1 - \phi^t L^t$, we have

$$1 + \phi L + \phi^2 L^2 + \dots + \phi^{t-1} L^{t-1} = \frac{1 - \phi^t L^t}{1 - \phi L}. \quad (31)$$

If $|\phi| < 1$, ϕ^t converges to zero as $t \rightarrow \infty$, and

$$1 + \phi L + \phi^2 L^2 + \dots = \lim_{t \rightarrow \infty} \frac{1 - \phi^t L^t}{1 - \phi L} = (1 - \phi L)^{-1}. \quad (32)$$

We find the inverse of $1 - \phi L$, such that $(1 - \phi L)^{-1}(1 - \phi L) = 1$.

First-Order Difference Equations (revisited)

Suppose that $|\phi| < 1$. We divide both sides of $(1 - \phi L)y_t = w_t$ by $1 - \phi L$:

$$(1 - \phi L)^{-1}(1 - \phi L)y_t = (1 - \phi L)^{-1}w_t.$$

Then we obtain

$$y_t = w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \phi^3 w_{t-3} + \dots \quad (33)$$

p th-Order Difference Equations (revisited)

The general p th-order difference equation

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + w_t \quad (34)$$

can be written in terms of the lag operator as well

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p) y_t = w_t, \quad (35)$$

where the lag polynomial in (35) can be factorized as

$$1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p = (1 - \lambda_1 L)(1 - \lambda_2 L) \times \dots \times (1 - \lambda_p L) \quad (36)$$

Why the lag polynomial can be factorized and how?

Consider the equation with complex number z

$$1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p = (1 - \lambda_1 z)(1 - \lambda_2 z) \times \dots \times (1 - \lambda_p z). \quad (37)$$

Is that possible to find $\lambda_1, \dots, \lambda_p$ such that, for any value of z , the equation holds? The answer is yes!

Immediately we find that the equation holds when $z = 0$. For $z \neq 0$, turn to the next page.

Why the lag polynomial can be factorized and how?

When $z \neq 0$, first define $\lambda = 1/z$, then divide both sides of the equation by z^p , and we obtain:

$$\lambda^p - \phi_1 \lambda^{p-1} - \phi_2 \lambda^{p-2} - \dots - \phi_p = (\lambda - \lambda_1)(\lambda - \lambda_2) \times \dots \times (\lambda - \lambda_p). \quad (38)$$

Now looks familiar? If so, you get it!

$\lambda_1, \dots, \lambda_p$ are actually the roots of the equation

$$\lambda^p - \phi_1 \lambda^{p-1} - \phi_2 \lambda^{p-2} - \dots - \phi_p = 0. \quad (39)$$

There must be p complex roots which can be repeated. If complex, then conjugates.

Why the lag polynomial can be factorized and how?

Recall the matrix \mathbf{F} in the corresponding first-order vector difference equation. The eigenvalue problem $|\mathbf{F} - \lambda \mathbf{I}_p| = 0$ or $|\lambda \mathbf{I}_p - \mathbf{F}| = 0$ is actually equivalent to the root-finding problem in (41).

To see this, consider the eigenvalue decomposition $\mathbf{F} = \mathbf{T}\Lambda\mathbf{T}^{-1}$. We have $|\lambda \mathbf{I}_p - \mathbf{F}| = |\lambda \mathbf{I}_p - \Lambda| = (\lambda - \lambda_1)(\lambda - \lambda_2) \times \dots \times (\lambda - \lambda_p) = 0$.

If you think that it is beautiful, you get it!

If all the p roots are found, the polynomial

$\lambda^p - \phi_1 \lambda^{p-1} - \phi_2 \lambda^{p-2} - \dots - \phi_p$ can be factorized like the RHS of (38).

Thus, we have (37).

$$\lambda \in \mathbb{C}, \quad F = \begin{bmatrix} \phi_1 & \cdots & \phi_p \\ \vdots & \ddots & 0 \\ 0 & \cdots & 1 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}, \quad T = \begin{bmatrix}$$

$$F = T \Lambda T^{-1}$$

$$|\lambda I_p - F| = |\lambda I_p - T \Lambda T^{-1}|$$

$$= |\lambda I_p -$$

p th-Order Difference Equations (revisited)

Given $\lambda = 1/z$ and $z \neq 0$, we have two equivalent root-finding problems

$$1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p = 0,$$

and

$$\lambda^p - \phi_1 \lambda^{p-1} - \phi_2 \lambda^{p-2} - \dots - \phi_p = 0.$$

The latter one is equivalent to the eigenvalue problem $|\lambda \mathbf{I}_p - \mathbf{F}| = 0$.

'Traditionally', or in the literature,

- we call the roots of the former one '**the roots of the lag polynomial**',
- and we call the roots of the latter one '**the eigenvalues of the companion matrix**', as \mathbf{F} is termed the **companion matrix** of the p th-order difference equation.

p th-Order Difference Equations (revisited)

Provided the factorization

$$1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p = (1 - \lambda_1 L)(1 - \lambda_2 L) \times \dots \times (1 - \lambda_p L),$$

if

- all the roots of the corresponding lag polynomial are greater than one in modulus (lie outside the unit circle or unit disk), or equivalently,
- all the eigenvalues of the corresponding companion matrix are less than one in modulus (lie inside the unit circle or unit disk),

then we call the p th-order difference equation **stable**, and each $1 - \lambda_i L$, $i = 1, \dots, p$, can be inverted, that is

$$(1 - \lambda_i L)^{-1} = 1 + \lambda_i L + \lambda_i^2 L^2 + \lambda_i^3 L^3 + \dots \quad (40)$$

p th-Order Difference Equations (revisited)

Thus, the p th-order difference equation for y_t

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p) y_t = w_t$$

can be transformed into

$$y_t = (1 - \lambda_1 L)^{-1} (1 - \lambda_2 L)^{-1} \dots (1 - \lambda_p L)^{-1} w_t, \quad (41)$$

by multiplying $(1 - \lambda_i L)^{-1}$, $i = 1, \dots, p$, on both sides.

p th-Order Difference Equations (revisited)

Recall the dynamic multiplier in (21). Likewise, if the eigenvalues λ_i are distinct, we have first

$$(1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1} \dots (1 - \lambda_p L)^{-1} = \sum_{i=1}^p \frac{c_i}{1 - \lambda_i L}. \quad (42)$$

See [2.4.8] on pp.34 in Hamilton. c_i s are defined in (21). Combined with (40), we have

$$\begin{aligned} y_t &= \sum_{i=1}^p \frac{c_i}{1 - \lambda_i L} \cdot w_t = \sum_{i=1}^p c_i (1 + \lambda_i L + \lambda_i^2 L^2 + \lambda_i^3 L^3 + \dots) \cdot w_t \\ &= w_t \sum_{i=1}^p c_i + w_{t-1} \sum_{i=1}^p c_i \lambda_i + w_{t-2} \sum_{i=1}^p c_i \lambda_i^2 + \dots \\ &= \sum_{j=0}^{\infty} \left(w_{t-j} \sum_{i=1}^p c_i \lambda_i^j \right) \quad (\text{get used to it}) \end{aligned} \quad (43)$$

p th-Order Difference Equations (revisited)

From (43), we can obtain the dynamic multiplier

$$\frac{\partial y_{t+\tau}}{\partial w_t} = \sum_{i=1}^p c_i \lambda_i^\tau$$

p th-Order Difference Equations (revisited)

(41) is often written as

$$\begin{aligned}y_t &= (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1} \dots (1 - \lambda_p L)^{-1} w_t, \\&= \prod_{i=1}^p (1 + \lambda_i L + \lambda_i^2 L^2 + \lambda_i^3 L^3 + \dots) w_t \\&= \psi_0 w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \dots \\&= \psi(L) w_t\end{aligned}\tag{44}$$

where $\psi(L) = \psi_0 + \psi_1 L + \dots$ represents the lag polynomial, when the difference equation is stable.

Whether the eigenvalues are distinct is irrelevant for this form.

However, when they are distinct, then $\psi_j = \frac{\partial y_{t+j}}{\partial w_t} = \sum_{i=1}^p c_i \lambda_i^j$, as c_i exit.

Initial Conditions

Given the p th-order difference equation

$$y_t = \sum_{i=1}^p \phi_i y_{t-i} + w_t,$$

p initial values of y

$$y_0, y_{-1}, \dots, y_{1-p}, ,$$

and a sequence of w

$$w_1, w_2, \dots, w_t, ,$$

we can calculate the sequence of y from time 1 to t

$$y_1, y_2, \dots, y_t, ,$$

Initial Conditions

However, there are many examples in economics and finance in which a theory does not specify the initial values $y_0, y_{-1}, \dots, y_{1-p}$. See the example and discussion on pp.36-42 in Hamilton.



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 3. Stationary ARMA Processes and Forecasting

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Stationary ARMA Processes (pp.43-71 in Hamilton)
 - Expectations, Stationarity & Ergodicity
 - MA, AR & ARMA Processes
 - Invertibility
- Forecasting (pp.72-116 in Hamilton)
 - Based on Conditional Expectation
 - Based on Linear Projection
 - Based on an Infinite Number of Observations
 - Based on a Finite Number of Observations

Stochastic Processes

- Consider the elements of an observed time series as being realizations (outcomes) of a stochastic (random) process. (Recall the graph in Lecture 1)
- In modeling such a process, we attempt to capture the characteristics.
- The univariate ARMA processes provide a very useful class of models for describing the dynamics of an individual time series.

Stochastic Processes

Suppose that we have observed a sample

$$\{y_t\}_{t=1}^T = \{y_1, y_2, \dots, y_T\}$$

of size T of some random variables $\{Y_t\}_{t=1}^T$.

If we could observe, which is not possible, the process for an infinite period, then the full sample is

$$\{y_t\}_{t=-\infty}^{\infty} = \{\dots, y_{-1}, y_0, y_1, y_2, \dots, y_T, y_{T+1}, \dots\},$$

from the random variables $\{Y_t\}_{t=-\infty}^{\infty}$.

They are both one realization of the underlying data generating process but with different sample sizes!

Stochastic Processes

If we could independently repeat the data generating process at time t for I times, then we can collect

$$\{y_t^{(i)}\}_{i=1}^I = \{y_t^{(1)}, y_t^{(2)}, \dots, y_t^{(I)}\}.$$

We obtain the **cross-sectional data**. Very often, this is **impossible** in time series.

Denote $f_{Y_t}(y)$ the unconditional density function of the random number Y_t at time t . We have

$$f_{Y_t}(y) \geq 0, \quad \text{and} \quad \int_{-\infty}^{\infty} f_{Y_t}(y) dy = 1$$

Then $f_{Y_t}(y_t)$ is the value of the density function when the argument y equals the observation y_t , to be precise.

Expectation

The expectation of the t th observation of a time series refers to the following integral, provided it exists:

$$E(Y_t) = \int_{-\infty}^{\infty} y f_{Y_t}(y) dy. \quad (1)$$

You will see or have seen the notation in the literature like $y \cdot f_{Y_t}$, which implies the integral above (most probably, it is $y \cdot \mu$ where $\mu = E(Y_t)$).

This existence is also termed **integrable**.

The **ensemble average** or **ensemble mean** of the observations at time t

$$I^{-1} \sum_{i=1}^I y_t^{(i)} \xrightarrow{P} E(Y_t) \quad \text{Strong LLN.} \quad (2)$$

So far, Y_t is the implied random variable.

Expectation

Some expectations

- $Y_t = \mu + \varepsilon_t$ implies $E(Y_t) = \mu$.
- $Y_t = \beta t + \varepsilon_t$ implies $E(Y_t) = \beta t$.
- If the expectation is time-varying, for example, a function of the date like above, we denote $E(Y_t) = \mu_t$.

Variance

The unconditional **variance** of the random variable Y_t is defined as follows

$$\gamma_{0t} = \text{Var}(Y_t) = E(Y_t - \mu_t)^2 = \int_{-\infty}^{\infty} (y - \mu_t)^2 f_{Y_t}(y) dy \quad (3)$$

Note that $E(Y_t - \mu_t)^2 = E((Y_t - \mu_t)^2)$, which differs from $E^2(Y_t - \mu_t) = (E(Y_t - \mu_t))^2$.

If $Y_t = \beta t + \varepsilon_t$, and $\varepsilon_t \sim (0, \sigma^2)$, then $\gamma_{0t} = E(Y_t - \beta t)^2 = E(\varepsilon_t^2) = \sigma^2$.

Autocovariance

The j th **autocovariance**

$$\begin{aligned}\gamma_{jt} &= \text{Cov}(Y_t, Y_{t-j}) = E(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j}) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \mu_t)(x - \mu_{t-j}) f_{Y_t, Y_{t-j}}(y, x) dy dx\end{aligned}\quad (4)$$

where $f_{Y_t, Y_{t-j}}(y, x)$ is the joint density function of the random variables Y_t and Y_{t-j} .

Note that

$$\int_{-\infty}^{\infty} f_{Y_t, Y_{t-j}}(y, x) dx = f_{Y_t}(y) \text{ and } \int_{-\infty}^{\infty} f_{Y_t, Y_{t-j}}(y, x) dy = f_{Y_{t-j}}(x).$$

This is telling the same story as [3.1.10] on pp.45 in Hamilton, but they look so different. Do you know why?

Autocovariance

- The autocovariance is the covariance between Y_t and its own lag.
- The 0th autocovariance, γ_{0t} , is the variance of Y_t .
- We have the ensemble average, if the pair of the observations $(y_t^{(i)}, y_{t-j}^{(i)})$ at time t and $t - j$ can be repeatedly independently sampled:

$$I^{-1} \sum_{i=1}^I (y_t^{(i)} - \mu_t)(y_{t-j}^{(i)} - \mu_{t-j}) \xrightarrow{P} \gamma_{jt}.$$

Stationarity

If neither the expectation μ_t nor the autocovariances γ_{jt} depend on the time t , the process for Y_t is said to be **covariance-stationary** or **weakly stationary**.

$$\begin{aligned} E(Y_t) &= \mu, && \text{for all } t; \\ E(Y_t - \mu)(Y_{t-j} - \mu) &= \gamma_j < \infty, && \text{for all } t \text{ and any } j. \end{aligned} \quad (5)$$

Example:

$Y_t = \mu + \varepsilon_t$, where $\varepsilon_t \stackrel{iid}{\sim} (0, \sigma^2)$,
with $E(Y_t) = \mu$, $\gamma_{0t} = \sigma^2$ and $\gamma_{jt} = 0$ for $j \neq 0$.

If a process is covariance-stationary, then it follows that

$$\gamma_j = E(Y_t - \mu)(Y_{t-j} - \mu) = E(Y_{t-j} - \mu)(Y_t - \mu) = \gamma_{-j}. \quad (6)$$

Stationarity

A process is said to be **strictly stationary** if, for any (integer) values of j_1, j_2, \dots, j_n , the joint distribution of $(Y_t, Y_{t+j_1}, Y_{t+j_2}, \dots, Y_{t+j_n})$ depends not on the time t , but only on j_1, j_2, \dots, j_n .

Remarks:

- If a strictly stationary process has finite autocovariances, then it is covariance-stationary.
- A covariance-stationary process may not be strictly stationary, as some higher moments (> 2) can be time dependent.
- The assumption of strict stationarity is too strong to verify in most cases in practice.
- By default, "stationary" means "covariance-stationary".

Gaussian Process

A process $\{Y_t\}$ is said to be **Gaussian**, if the joint density

$$f_{Y_t, Y_{t+j_1}, Y_{t+j_2}, \dots, Y_{t+j_n}}(y_0, y_1, y_2, \dots, y_n)$$

is multivariate Gaussian for any j_1, j_2, \dots, j_n .

A covariance-stationary Gaussian process is strictly stationary.

Defn: $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \Leftrightarrow \forall V \neq 0, V^T X \sim N(\dots).$

Ergodicity

Motivation:

Since we are not dealing with cross-sectional data, it is not realistic in practice to have $y_t^{(1)}, y_t^{(2)}, \dots, y_t^{(I)}$ at time t . We only have one single realization y_t . Can we still infer something from the following time average?

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

Whether the time average as such eventually converge to the ensemble $E(Y_t)$ for a stationary process has to do with **ergodicity**.

Ergodicity

A stationary process is said to be **ergodic for the mean**, if the time average converges in probability to $E(Y_t)$ as $T \rightarrow \infty$.

$$T^{-1} \sum_{t=1}^T y_t \xrightarrow{p} E(Y_t) \quad (7)$$

Remarks

- A process is ergodic for the mean provided that the autocovariance γ_j goes to zero sufficiently fast as $j \rightarrow \infty$.
- We will see (chapter 7 in Hamilton) that if $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ (absolute summability) holds for a stationary process Y_t , then Y_t is ergodic for the mean.

Ergodicity

A stationary process is said to be **ergodic for second moments**, if

$$(T-j)^{-1} \sum_{t=j+1}^T (y_t - \mu)(y_{t-j} - \mu) \xrightarrow{P} \gamma_j \quad (8)$$

for all j .

Remarks

- If Y_t is a stationary Gaussian process, the absolute summability $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ is sufficient for ergodicity for all moments.
- The Gaussian assumption offers great convenience.
- Sufficient conditions for more general cases can be found in chapter 7 in Hamilton.

Example: Stationary but Not Ergodic

Suppose that

$$Y_t = U_t + Z$$

where $U_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$, $Z \stackrel{iid}{\sim} \mathcal{N}(0, 1)$, and U_t and Z are independent to each other.

We have $E(Y_t) = 0$, and

$$\begin{aligned}\text{Cov}(Y_t, Y_{t-j}) &= E(U_t + Z)(U_{t-j} + Z) \\ &= \text{Var}(Z) = 1.\end{aligned}$$

Thus, the process Y_t is stationary with $\gamma_j = 1$ for all j .

Example: Stationary but Not Ergodic

However, when you observe the sample y_t (note that you cannot see u_t and z), the time average

$$\bar{y} = T^{-1} \sum_{t=1}^T y_t = T^{-1} \sum_{t=1}^T u_t + z \xrightarrow{P} z,$$

and $z \neq 0$ almost surely.

Even worse, when you resample it for I times, you will find that $\bar{y}^{(i)}$ are distinct, for $i = 1, \dots, I$, as $\bar{y}^{(i)} \xrightarrow{P} z^{(i)}$ and $z^{(i)}$ are distinct.

In reality, normally the data generating cannot be repeated. Most probably, you will regard z as a constant, due to the conditioning like $E(Y_t|z) = z$. This process, therefore, becomes ergodic for the mean.

White Noise

A **white noise process** is a sequence $\{\varepsilon_t\}_{-\infty}^{\infty}$ whose elements satisfy

$$E(\varepsilon_t) = 0 \tag{9}$$

$$E(\varepsilon_t^2) = \sigma^2 \tag{10}$$

$$E(\varepsilon_t \varepsilon_{t-j}) = 0 \quad \text{if } j \neq 0 \tag{11}$$

for all integers t and j .

A stronger version of the white noise process is to replace (11) by

$$\varepsilon_t \text{ and } \varepsilon_{t-j} \text{ are independent if } j \neq 0, \tag{12}$$

which is said to be the **independent white noise process**.

White Noise

Remarks:

- The white noise process is the basic building block for the ARMA processes.
- The white noise process, by construction, is stationary.
- The white noise process is called **Gaussian white noise process** if any joint distribution of $\varepsilon_t, \varepsilon_{t+j_1}, \dots, \varepsilon_{t+j_n}$ is Gaussian distributed.
- A Gaussian white noise process is strictly stationary.

Autocorrelation

The j th autocorrelation of a stationary process is defined as

$$\rho_j = \gamma_j / \gamma_0. \quad (13)$$

Remarks

- Autocorrelation comes from the correlation between Y_t and Y_{t-j}

$$\text{Corr}(Y_t, Y_{t-j}) = \frac{\text{Cov}(Y_t, Y_{t-j})}{\sqrt{\text{Var}(Y_t)\text{Var}(Y_{t-j})}} = \frac{\gamma_j}{\gamma_0} = \rho_j \quad (14)$$

- By the Cauchy-Schwarz inequality, $|\rho_j| \leq 1$ for all j .
- $\rho_0 = 1$.

$$\begin{aligned} \mathbf{a} \cdot \mathbf{b} &= \|\mathbf{a}\| \cdot \|\mathbf{b}\| \cdot \cos \theta \\ &\leq \|\mathbf{a}\| \cdot \|\mathbf{b}\| \end{aligned}$$



Moving Average Processes

Let $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ be a white noise process. The *qth-order moving average process* or $\text{MA}(q)$ is given by

$$Y_t = \mu + \sum_{i=0}^q \theta_i \varepsilon_{t-i} \quad (15)$$

where $\theta_0 = 1$ and $\theta_i \in \mathbb{R}$. It can be shown that

- The expectation $E(Y_t) = \mu$.
- $\{Y_t\}_{t=-\infty}^{\infty}$ is stationary for all $\theta_i \in \mathbb{R}$, with

$$\gamma_j = \begin{cases} 0 & \text{for } j > q \\ \sigma^2 \sum_{i=0}^{q-j} \theta_i \theta_{i+j} & \text{for } j = 0, \dots, q \\ \gamma_{-j} & \text{for } j < 0 \end{cases} \quad (16)$$

hence, $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ (absolutely summable) and $\rho_j = 0$ for $j > q$.

- If $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is a Gaussian white noise process, then $\{Y_t\}_{t=-\infty}^{\infty}$ is ergodic for all moments.

Moving Average Processes

Likewise, the infinite-order moving average process or $\text{MA}(\infty)$ is given by

$$Y_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} \quad (17)$$

where $\psi_0 = 1$ and $\psi_i \in \mathbb{R}$. $E(Y_t) = \mu$.

Recall the lag operator.

- The $\text{MA}(q)$ can be written as

$$Y_t = \mu + \theta(L) \varepsilon_t \quad (18)$$

where $\theta(L) = 1 + \theta_1 L + \dots + \theta_q L^q$.

- The $\text{MA}(\infty)$ can be written as

$$Y_t = \mu + \psi(L) \varepsilon_t \quad (19)$$

where $\psi(L) = 1 + \psi_1 L + \psi_2 L^2 + \dots$

Moving Average Processes

Appendix 3.A on pp.69-70 in Hamilton shows that $\text{MA}(\infty)$ is a well defined stationary process provided that

$$\sum_{i=0}^{\infty} \psi_i^2 < \infty \quad \text{square-summability}, \quad (20)$$

or, stronger and more often used,

$$\sum_{i=0}^{\infty} |\psi_i| < \infty \quad \text{absolute summability}, \quad (21)$$

We have $\sum_{i=0}^{\infty} |\psi_i| < \infty \implies \sum_{i=0}^{\infty} \psi_i^2 < \infty$.

Moving Average Processes

Remarks for MA(∞)

- The variance is $\text{Var}(Y_t) = \sigma^2 \sum_{i=0}^{\infty} \psi_i^2 < \infty$.
- The autocovariance is

$$\gamma_j = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+j} < \infty, \quad j = 0, 1, 2, \dots \quad (22)$$

- If the coefficients ψ_i are absolutely summable, the corresponding autocovariance is absolutely summable

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \quad (23)$$

See pp.70 in Hamilton.

- Recall (pp.15 in the slides) that if the autocovariance is absolutely summable, the MA(∞) is ergodic for the mean.
- If in addition ε_t is Gaussian, then you know... ergodic for all moments.

Autoregressive Processes

Let $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ be a white noise process. The *p*th-order autoregressive process or AR(*p*) is given by

$$Y_t = c + \sum_{i=1}^p \phi_i Y_{t-i} + \varepsilon_t, \quad (24)$$

where $c, \phi_i \in \mathbb{R}$. Alternatively we can write

$$\phi(L)Y_t = c + \varepsilon_t, \quad (25)$$

where $\phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$.

Let $w_t = c + \varepsilon_t$. From Lecturer 2, we know that this difference equation is **stable** when the roots of $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p = 0$ lie outside the unit disk, or equivalent by denoting $\lambda = 1/z$, the roots (eigenvalues of the companion matrix) of $\lambda^p - \phi_1 \lambda^{p-1} - \dots - \phi_p = 0$ lie inside the unit disk.

Autoregressive Processes

Proposition: The AR(p) process is stationary, if the corresponding difference equation is stable.

If the difference equation is stable, then Y_t has the MA(∞) representation:

$$Y_t = \mu + \psi(L)\varepsilon_t \quad (26)$$

where

$$\mu = c\phi^{-1}(L) = c\phi^{-1}(1) = \frac{c}{1 - \phi_1 - \phi_2 - \dots - \phi_p} \quad (27)$$

$$\begin{aligned}\psi(L) &= \phi^{-1}(L) = (1 - \phi_1 L - \dots - \phi_p L^p)^{-1} \\ &= (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1} \dots (1 - \lambda_p L)^{-1} \quad (\text{fact: } |\lambda_i| < 1) \\ &= \left(\sum_{i=0}^{\infty} \lambda_1^i L^i \right) \left(\sum_{i=0}^{\infty} \lambda_2^i L^i \right) \dots \left(\sum_{i=0}^{\infty} \lambda_p^i L^i \right) \\ &= 1 + \psi_1 L + \psi_2 L^2 + \dots\end{aligned} \quad (28)$$

Stability and Absolute Summability

- A Cauchy sequence $\alpha_j, j = 1, \dots$ is a sequence satisfying that, for any small positive number ϵ , there exists a N such that $|\alpha_n - \alpha_m| < \epsilon$ for any $n, m > N$.
- A sequence is convergent iff it is a Cauchy sequence.
- Given $|\lambda| < 1$, $1 - \lambda L$ is stable and has the inverse $\sum_{i=0}^{\infty} \lambda^i L^i$ which has absolutely summable coefficients $\sum_{i=0}^{\infty} |\lambda|^i < \infty$.

To see this, define

$$\alpha_j = \sum_{i=0}^j |\lambda|^i.$$

Assuming $n > m$ without loss of generality,

$|\alpha_n - \alpha_m| = \sum_{i=m+1}^n |\lambda|^i = |\lambda|^{m+1}(1 - |\lambda|^{n-m})/(1 - |\lambda|)$ goes to zero. Thus, α_j is Cauchy and then it is convergent (absolute summability).

Stability and Absolute Summability

- If two lag polynomials are both absolutely summable, its product is absolutely summable as well.

$$\sum_{i=0}^{\infty} \phi_i L^i \quad \text{with} \quad \sum_{i=0}^{\infty} |\phi_i| < \infty$$
$$\sum_{i=0}^{\infty} \psi_i L^i \quad \text{with} \quad \sum_{i=0}^{\infty} |\psi_i| < \infty$$

We need to check whether the lag polynomial
 $(\sum_{i=0}^{\infty} \phi_i L^i) (\sum_{i=0}^{\infty} \psi_i L^i)$ has absolutely summable coefficients.

Stability and Absolute Summability

The product $(\sum_{i=0}^{\infty} \phi_i L^i) (\sum_{i=0}^{\infty} \psi_i L^i)$ has the terms

	L^0	L^1	L^2	L^3	\dots
$\phi_0 \cdot$	ψ_0	ψ_1	ψ_2	ψ_3	\dots
$\phi_1 \cdot$		ψ_0	ψ_1	ψ_2	\dots
$\phi_2 \cdot$			ψ_0	ψ_1	\dots
\vdots			\vdots	\vdots	

They are

$$\begin{aligned} & \phi_0\psi_0 + (\phi_0\psi_1 + \phi_1\psi_0)L + (\phi_0\psi_2 + \phi_1\psi_1 + \phi_2\psi_0)L^2 + \dots \\ & = \sum_{k=0}^{\infty} \left(\sum_{i+j=k} \phi_i\psi_j \right) L^k. \end{aligned}$$

Then we need to check $\sum_{k=0}^{\infty} \left| \sum_{i+j=k} \phi_i\psi_j \right| < \infty$.

Stability and Absolute Summability

We have the inequality

$$\sum_{k=0}^{\infty} \left| \sum_{i+j=k} \phi_i \psi_j \right| \leq \sum_{k=0}^{\infty} \sum_{i+j=k} |\phi_i| |\psi_j|.$$

And

$$\sum_{k=0}^{\infty} \sum_{i+j=k} |\phi_i| |\psi_j| = \left(\sum_{i=0}^{\infty} |\phi_i| \right) \left(\sum_{j=0}^{\infty} |\psi_j| \right) < \infty \quad Q.E.D.$$

Conclusion : If a p -order lag polynomial is stable, then its inverse polynomial has absolutely summable coefficients.

Autoregressive Processes

$$Y_t = c + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} - \varepsilon_t$$
$$\phi(L) Y_t = c + \varepsilon_t$$

Now let us go back to AP(p).

Proposition: The AR(p) process is stationary, if the corresponding difference equation is stable.

$$Y_t = \phi'(L)(c + \varepsilon_t)$$

If the difference equation is stable, then Y_t has the MA(∞)

$$Y_t = \mu + \psi(L)\varepsilon_t$$
$$= \psi(L)(c + \varepsilon_t)$$
$$= \psi(L)c + \psi(L)\varepsilon_t$$
$$= \underbrace{\psi(L)c}_{\mu} + \underbrace{\psi(L)\varepsilon_t}_{\varepsilon_t - \mu}$$
$$\varepsilon_t - \mu = \psi(L)\varepsilon_t$$
$$\psi(L) = 1 + \psi_1 L + \psi_2 L^2 + \dots \Leftrightarrow \phi(L)(\underbrace{Y_t - \mu}_{\varepsilon_t}) = \varepsilon_t$$

The coefficients ψ_i are absolutely summable, definitely.

$$Y_t - \mu = \varepsilon_t = \phi_1 Z_{t-1} + \dots + \phi_p Z_{t-p}$$

Autoregressive Processes

The autocovariances of an AR(p) process are given by

$$\gamma_j = \begin{cases} \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \dots + \phi_p\gamma_{j-p} & \text{for } j = 1, 2, \dots \\ \phi_1\gamma_1 + \phi_2\gamma_2 + \dots + \phi_p\gamma_p + \sigma^2 & \text{for } j = 0 \\ \gamma_{-j} & \text{for } j < 0 \end{cases} \quad (29)$$

Remarks

- Actually the system of equations (29) for $j = 0, 1, \dots, p$ can be solved for $\gamma_0, \gamma_1, \dots, \gamma_p$, by using $\gamma_j = \gamma_{-j}$, as functions of $\sigma^2, \phi_1, \dots, \phi_p$.
- Recall the autocovariances of the stationary MA(∞) process. The same result (absolutely summable autocovariances) applies here if the AR(p) is stable.

Autoregressive Processes

The autocorrelations of an AR(p) process are given by

$$\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} + \dots + \phi_p \rho_{j-p}, \quad j = 1, 2, \dots \quad (30)$$

the so-called **Yule-Walker equations**.

Note that the autocovariances and the autocorrelations follow the same p th-order difference equation as the AR(p) process itself.

$$Y_t = C + \sum_{i=1}^p \phi_i Y_{t-i} + \varepsilon_t \Leftrightarrow {}^* \phi(L) Z_t = \varepsilon_t$$

$$Z_t = Y_t - M, \quad M = \frac{C}{\phi(1)}$$

$$(*) \quad Z_t = \sum_{i=1}^p \phi_i Z_{t-i} + \varepsilon_t$$

$$E(Z_t Z_{t+1}) = E(\phi_1 Z_{t-1}^2 + \phi_2 Z_{t-2} Z_{t-1} + \dots + \varepsilon_t \varepsilon_{t+1})$$

$$E(Y_t M)(Y_{t+1} M) = \gamma_1$$

$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1 + \dots + \phi_p \gamma_p$$

$$\gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0 + \phi_3 \gamma_1 + \dots$$

$$\gamma_3 = \phi_1 \gamma_2 + \phi_2 \gamma_1 + \phi_3 \gamma_0 + \dots$$

Autoregressive Moving Average Processes

Let $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ be a white noise process. The ARMA(p, q) is given by

$$Y_t = c + \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{i=0}^q \theta_i \varepsilon_{t-i}, \quad (31)$$

where $c, \phi_i, \theta_i \in \mathbb{R}$ and $\theta_0 = 1$. Alternatively we can write

$$\phi(L)Y_t = c + \theta(L)\varepsilon_t, \quad (32)$$

where $\phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$, and $\theta(L) = 1 + \theta_1 L + \dots + \theta_p L^p$.

Autoregressive Moving Average Processes

Assuming that the roots of $\phi(z) = 0$ lie outside the unit disk, both sides of (32) can be divided by $\phi(L)$ to obtain

$$Y_t = \mu + \psi(L)\varepsilon_t \quad (33)$$

where where

$$\mu = \frac{c}{1 - \phi_1 - \phi_2 - \dots - \phi_p}$$

$$\psi(L) = \phi^{-1}(L)\theta(L) = 1 + \psi_1 L + \psi_2 L^2 + \dots$$

The coefficients ψ_i are absolutely summable. Note that $\theta(L)$ has finite number of coefficients, and hence it is absolutely summable.

Autoregressive Moving Average Processes

The autocovariances of an ARMA(p, q) process can be computed using standard methods, for $j > q$ they are given by

$$\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \dots + \phi_p\gamma_{j-p}, \quad j = q+1, q+2, \dots$$

Remarks:

- The stationarity of the ARMA(p, q) process, where p, q are finite, depends entirely on the stability of $\phi(L)$, not on $\theta(L)$.
- An ARMA(p, q) process will have more complicated autocovariances γ_j for $j = 1, \dots, q$ than would the corresponding AR(p) process.
- There is a potential for redundant parameterization with ARMA processes, see pp.60-61 in Hamilton.

Invertibility

Consider the MA(1) process

$$Y_t - \mu = (1 + \theta L)\varepsilon_t.$$

Provided that $|\theta| < 1$, both sides of the equation can be multiplied by $(1 + \theta L)^{-1}$, where

$$(1 + \theta L)^{-1} = (1 - (-\theta)L)^{-1} = 1 + (-\theta)L + (-\theta)^2L^2 + (-\theta)^3L^3 + \dots$$

Then we have

$$(1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + \dots)(Y_t - \mu) = \varepsilon_t$$

which could be viewed as an AR(∞) representation.

Invertibility

Remarks:

- If an MA(1) representation can be rewritten as an AR(∞) representation by inverting $(1 + \theta L)$, then the MA(1) representation is said to be **invertible**.
- For an MA(1) process invertibility requires $|\theta| < 1$.
- For any invertible MA(1) representation, there is a noninvertible MA(1) representation with the same first and second moments as the invertible representation. See pp.65 in Hamilton for details.

Invertibility

Remarks:

- Either representation could be used as an equally valid description of any given MA(1) process.
- For estimation and forecasting purposes, we prefer to work with the invertible representation.
- The **innovation** (noise term, error term) associated with the invertible representation is sometimes called the **fundamental innovation**.
- The concept of invertibility can be extended to the general MA(q) process.

Forecasts Based on Conditional Expectation

Suppose we are interested in forecasting the random variable Y_{t+s} based on a set of variable \mathbf{x}_t available at time t .

$$\mathbf{x}_t = (1, y_t, y_{t-1}, \dots, y_{t-m+1})' \quad (34)$$

Let $Y_{t+s|t}^*$ denote such a s -step ahead forecast ($s = 1, 2, \dots$). Actually it is a function of \mathbf{x}_t .

The performance of the forecast $Y_{t+s|t}^*$ is evaluated in terms of some loss function $g : \mathbb{R} \rightarrow \mathbb{R}$.

Consider the quadratic loss function $g(x) = (Y_{t+s} - x)^2$. We choose the forecast $Y_{t+s|t}^*$ to minimize

$$E_t g(x).$$

Forecasts Based on Conditional Expectation

The mean squared error (MSE) associated with $Y_{t+s|t}^*$ is given by

$$E(Y_{t+s} - Y_{t+s|t}^*)^2 \quad (35)$$

We are actually finding a functional form for $Y_{t+s|t}^*$ with the argument \mathbf{x}_t in order to minimize the expected loss function Eg .

Suppose that $Y_{t+s|t}^* = h(\mathbf{x}_t)$ for some function $h(\cdot)$, then the forecast that minimizes

$$E(Y_{t+s} - Y_{t+s|t}^*)^2 = E(Y_{t+s} - h(\mathbf{x}_t))^2 \quad (36)$$

is given by $h(\mathbf{x}_t) = E(Y_{t+s} | \mathbf{x}_t)$.

Forecasts Based on Linear Projection

Let $\mathbf{h}'\mathbf{x}_t$ denote any arbitrary linear forecasting rule, then the forecast that minimizes

$$E(Y_{t+s} - \mathbf{h}'\mathbf{x}_t)^2 \quad (37)$$

is given by $Y_{t+s|t}^* = \hat{\mathbf{h}}'\mathbf{x}_t$, where $\hat{\mathbf{h}}'\mathbf{x}_t$ satisfies

$$E(Y_{t+s} - \hat{\mathbf{h}}'\mathbf{x}_t)\mathbf{x}_t' = \mathbf{0}' \quad (38)$$

Forecasts Based on Linear Projection

Remarks:

- $\hat{\mathbf{h}}' \mathbf{x}_t$ is called the **linear projection** of $Y + t + s$ on \mathbf{x}_t and is the optimal linear forecast.
- Since $E(Y_{t+s}|\mathbf{x}_t)$ offers the best possible forecast (in terms of MSE), we have that

$$E(Y_{t+s} - \hat{\mathbf{h}}' \mathbf{x}_t)^2 \geq E(Y_{t+s} - E(Y_{t+s}|\mathbf{x}_t))^2 \quad (39)$$

- By (38),

$$\hat{\mathbf{h}} = [E(\mathbf{x}_t \mathbf{x}'_t)]^{-1} E(\mathbf{x}_t Y_{t+s}) \quad (40)$$

- Hamilton uses the symbol \hat{E} to indicate a linear projection on a vector of random variables along with a constant term.
- Linear projection is closely related to OLS regression.

Forecasts Based on $\varepsilon_t, \varepsilon_{t-1}, \dots$

Consider a process with MA(∞) representation

$$Y_t - \mu = \psi(L)\varepsilon_t \quad (41)$$

where $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is a white noise process, and $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ with $\psi_0 = 1$ and is absolutely summable.

In addition, assume that, for simplicity, $\varepsilon_t, \varepsilon_{t-1}, \dots$ are observed and the parameters μ and ψ_1, ψ_2, \dots are known.

We are going to forecast Y_{t+s}

$$Y_{t+s} = \mu + \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \dots + \psi_{s-1} \varepsilon_{t+1} + \psi_s \varepsilon_t + \dots$$

The optimal linear forecast is

$$\hat{E}(Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \dots) = \mu + \psi_s \varepsilon_t + \dots$$

Forecasts Based on $\varepsilon_t, \varepsilon_{t-1}, \dots$

The accompanying forecast error

$$Y_{t+s} - \hat{E}(Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \dots) = \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \dots + \psi_{s-1} \varepsilon_{t+1}$$

And MSE

$$\text{E}(Y_{t+s} - \hat{E}(Y_{t+s} | \varepsilon_t, \varepsilon_{t-1}, \dots))^2 = (1 + \psi_1^2 + \dots + \psi_{s-1}^2)\sigma^2$$

In particular, if Y_t follows an MA(q) process with $\psi(L) = 1 + \theta_1 L + \dots + \theta_q L^q$, then the MSE increases with the increasing of s until $s = q$.

The forecast for $s > q$ is just μ and the MSE is always $(1 + \theta_1^2 + \dots + \theta_q^2)\sigma^2$.

Forecasts Based on $\varepsilon_t, \varepsilon_{t-1}, \dots$

It is convenient to introduce the compact lag operator expression of the s -step ahead forecast $\hat{E}(Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \dots)$.

First consider dividing $\psi(L)$ by L^s

$$\frac{\psi(L)}{L^s} = L^{-s} + \psi_1 L^{1-s} + \psi_2 L^{2-s} + \dots + \psi_{s-1} L^{-1} + \psi_s L^0 + \psi_{s+1} L^1 + \dots$$

and let $[\cdot]_+$ denote the annihilation operator, which replaces negative powers of L by zero,

$$\left[\frac{\psi(L)}{L^s} \right]_+ = \psi_s L^0 + \psi_{s+1} L^1 + \psi_{s+2} L^2 + \dots$$

Hence, we have the compact form

$$\hat{E}(Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \dots) = \mu + \left[\frac{\psi(L)}{L^s} \right]_+ \varepsilon_t \quad (42)$$

Forecasts Based on y_t, y_{t-1}, \dots

In practice, we observe y_t, y_{t-1}, \dots , but not $\varepsilon_t, \varepsilon_{t-1}, \dots$

Suppose that $Y_t - \mu = \psi(L)\varepsilon_t$ has an AR(∞) representation given by $\eta(L)(Y_t - \mu) = \varepsilon_t$, where $\eta(L) = \psi^{-1}(L)$ with $\eta(L) = \sum_{i=0}^{\infty} \eta_i L^i$, $\eta_0 = 1$ and the absolute summability.

We can construct $\varepsilon_t, \varepsilon_{t-1}, \dots$ based on y_t, y_{t-1}, \dots

Examples:

- AR(1) with $\eta(L) = 1 - \phi L$, then $(1 - \phi L)(y_t - \mu) = \varepsilon_t$, or

$$\varepsilon_t = (y_t - \mu) - \phi(y_{t-1} - \mu)$$

- MA(1) with $\eta(L) = (1 + \theta L)^{-1}$, then $(1 + \theta L)^{-1}(y_t - \mu) = \varepsilon_t$, or

$$\varepsilon_t = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - \theta^3(y_{t-3} - \mu) + \dots$$

Forecasts Based on y_t, y_{t-1}, \dots

The $\varepsilon_t, \varepsilon_{t-1}, \dots$ constructed from y_t, y_{t-1}, \dots can be plugged into the compact form (42)

$$\hat{E}(Y_{t+s}|y_t, y_{t-1}, \dots) = \mu + \left[\frac{\psi(L)}{L^s} \right]_+ \eta(L)(y_t - \mu). \quad (43)$$

This is called the [Wiener-Kolmogorov prediction formula](#).

Consider once again the AR(1) process with $\eta(L) = 1 - \phi L$ and $|\phi| < 1$. We have

$$\left[\frac{\psi(L)}{L^s} \right]_+ = \phi^s + \phi^{s+1}L + \phi^{s+2}L^2 + \dots = \frac{\phi^s}{1 - \phi L}.$$

Therefore, by the Wiener-Kolmogorov prediction formula, the optimal linear s -step ahead forecast is

$$\hat{E}(Y_{t+s}|y_t, y_{t-1}, \dots) = \mu + \phi^s(y_t - \mu).$$

Forecasts Based on a Finite Number of Observations

Consider forecasting a stationary AR(p) process with known parameters μ and $\phi_1, \phi_2, \dots, \phi_p$. From Lecture 2 we know that

$$\begin{aligned} Y_{t+s} - \mu &= f_{11}^{(s)}(Y_t - \mu) + f_{12}^{(s)}(Y_{t-1} - \mu) + \dots + f_{1p}^{(s)}(Y_{t-p+1} - \mu) \\ &\quad + \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s} + \psi_2 \varepsilon_{t+s} + \dots + \psi_{s-1} \varepsilon_{t+s} \end{aligned}$$

where $\psi_i = f_{11}^{(i)}$.

The optimal s -step ahead forecast is

$$\hat{E}(Y_{t+s}|y_t, y_{t-1}, \dots) = \mu + f_{11}^{(s)}(y_t - \mu) + f_{12}^{(s)}(y_{t-1} - \mu) + \dots + f_{1p}^{(s)}(y_{t-p+1} - \mu)$$

Forecasts Based on a Finite Number of Observations

Remarks:

- For forecasting the AR(p) process, we only need its p most recent observations, $y_t, y_{t-1}, \dots, y_{t-p+1}$.
- However, for MA or ARMA, we generally need infinite observations, y_t, y_{t-1}, \dots

Approximations to Optimal Forecasts

One approach to forecasting based on a finite number of values $y_t, y_{t-1}, \dots, y_{t-m+1}$ is to replace all presample ε 's with zero.

Precisely speaking, the idea is to replace $\hat{E}(Y_{t+s}|y_t, y_{t-1}, \dots)$ by

$$\hat{E}(y_t, y_{t-1}, \dots, y_{t-m+1}, \varepsilon_{t-m} = 0, \varepsilon_{t-m-1} = 0, \dots). \quad (44)$$

Exact Finite Sample Forecasts

An alternative approach is to calculate the linear projection of $Y_{t+s} - \mu$ on its m most recent values. To this end, let

$$\mathbf{x}_t = ((y_t - \mu), (y_{t-1} - \mu), \dots, (y_{t-m+1} - \mu))'$$

Then we look for a linear forecast of the form

$$\begin{aligned} Y_{t+s|t}^* - \mu &= \boldsymbol{\alpha}' \mathbf{x}_t \\ &= \alpha_1(y_t - \mu) + \alpha_2(Y_{t-1} - \mu) + \dots + \alpha_m(y_{t-m+1} - \mu) \end{aligned}$$

Exact Finite Sample Forecasts

Under the assumption of stationarity, the coefficients α_i can be calculated directly from (40)

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_s \\ \gamma_{s+1} \\ \vdots \\ \gamma_{s+m-1} \end{pmatrix}$$



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 4. Forecasting and Maximum Likelihood Estimation

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Forecasting (pp.108-116 in Hamilton)
- Maximum Likelihood Estimation
- Asymptotic Distribution Theory

Forecasting

- Wold's Decomposition Theorem
- The Box-Jenkins Modelling Philosophy
- Model Selection

Wold's Decomposition

Recall that any (covariance) stationary ARMA(p, q) process can be written as

$$Y_t = \mu + \psi(L)\varepsilon_t = \mu + (\psi_0 + \psi_1 L + \psi_2 L^2 + \dots)\varepsilon_t \quad (1)$$

where ε_t is interpreted as the, white noise, forecast error

$$Y_t - E(Y_t | y_{t-1}, y_{t-2}, \dots) \quad (2)$$

and where $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ with $\psi_0 = 1$.

It turns out that the above representation is **fundamental** for any stationary time series.

Wold's Decomposition

Wold's Decomposition

Any zero-mean covariance stationary process $\{Y_t\}_{t=-\infty}^{\infty}$ can be represented as

$$Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} + \kappa_t \quad (3)$$

where $\psi_0 = 1$, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ and

$$\varepsilon_t = Y_t - \hat{E}(Y_t | y_{t-1}, y_{t-2}, \dots) \quad (4)$$

is white noise. The linearly deterministic component

$$\kappa_t = \hat{E}(\kappa_t | y_{t-1}, y_{t-2}, \dots) \quad (5)$$

is uncorrelated with ε_{t-i} for any i .

Wold's Decomposition

Remarks:

- $\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$ is called the linearly indeterministic component of Y_t .
- If κ_t is zero for all t , then $\{Y_t\}_{t=-\infty}^{\infty}$ is called purely linearly indeterministic.
- The proposition essentially says that any stationary process can be expressed as the sum of two uncorrelated processes.
- The Wold's representation is named after Herman Wold, who was a professor in Statistics at Uppsala University.
- In practice, we seek a model that gives an adequate approximation to the data with as few parameters as possible. A typical assumption is that $\psi(L)$ can be expressed as

$$\sum_{i=0}^{\infty} \psi_i L^i = \frac{1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{1 + \phi_1 L + \phi_2 L^2 + \dots + \phi_p L^p} \quad (6)$$

The Box-Jenkins Modelling Philosophy

Box & Jenkins procedure can be broken down into four steps:

- Transform the data, if necessary, so that the assumption of covariance stationarity appears to be satisfied.
- Examine the transformed data to see which ARMA(p, q) process appears to be most appropriate.
- Estimate the parameters in $\phi(L)$ and $\theta(L)$ accompanying the chosen model.
- Assess the chosen models adequacy by checking whether the model assumptions are satisfied. If not, repeat the procedure.

The Box-Jenkins Modelling Philosophy

Remarks:

- They argued that, in practice, the more parameters to estimate, the more room there is to go wrong.
- Transformation: natural logarithm, square root, differencing, and etc.
- Model Identification or Specification: the ARMA form, plausible values of p and q selected by for example sample autocorrelations and partial autocorrelations.
- Estimation: OLS, ML, GMM and etc.
- Diagnostic Checking or Model evaluation: serial correlation test, heteroskedasticity test, Gaussianity test, and etc.

Model Selection

In practice, a time series analyst often ends up with two or more seemingly adequate parsimonious models.

In this case, some model selection [information criteria](#), such as AIC or BIC, can be used to help in choosing an appropriate model.

Maximum Likelihood Estimation

- The Method of Maximum Likelihood
- MLE for a Gaussian AR(1)
- Conditional MLE for a Gaussian AR(1)
- Conditional MLE for a Gaussian MA(1)
- Conditional MLE for a Gaussian ARMA(p, q)
- Statistical Inference with ML Estimation

Maximum Likelihood Estimation

Let us start with the general ARMA(p, q) model

$$Y_t = c + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad (7)$$

where ε_t is white noise with the variance σ^2 .

Now we start to consider the **reality**! What if all the values in the parameter vector

$$\boldsymbol{\theta} = (c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2)' \quad (8)$$

are unknown?

We have to estimate all of them based on the observations y_1, \dots, y_T .

Maximum Likelihood Estimation

The random numbers $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ are unobservable. But why not assume that they follow some distribution, for example Gaussian? This Gaussian assumption can be checked later.

Then we have the joint probability density function (value) of the sample (at the point y_1, \dots, y_T)

$$f_{Y_1, \dots, Y_T}(y_1, y_2, \dots, y_T; \theta). \quad (9)$$

Let us suppress the subscript Y_1, \dots, Y_T in the density function. We say $f(y_t; \theta) = f_{Y_t}(y_t; \theta)$ for random number Y_t and the observation y_t .

Similarly we have the conditional density

$$f(y_t|y_\tau, \theta) = f(y_t|y_\tau; \theta) = f_{Y_t|Y_\tau}(y_t|y_\tau; \theta).$$

Maximum Likelihood Estimation

When the sample $\{y_t\}_{t=1}^T$ is *iid*, the joint density can be decomposed as

$$f(y_1, y_2, \dots, y_T; \theta) = f(y_1; \theta)f(y_2; \theta)\dots f(y_T; \theta). \quad (10)$$

However, in time series, it is normally the case that y_t are **dependent**.

$f_{Y_1, \dots, Y_T}(\dots)$ is the joint density function, whose argument or input is the observations. However, if you regard all the observations as given and the parameters θ as the argument, it becomes a function of θ . This function is called the **likelihood function**.

$$L(\theta) = f(y_1, y_2, \dots, y_T; \theta) \quad (11)$$

f is a function of y , and L is a function of θ .

Maximum Likelihood Estimation

For a given sample of data with observed values y_1, \dots, y_T , the **maximum likelihood (ML) estimator** of θ is obtained by maximizing the likelihood function (11):

$$\hat{\theta} = \max_{\theta} L(\theta) \quad (12)$$

or, equivalently, by maximizing the **log-likelihood function**:

$$\hat{\theta} = \max_{\theta} \log L(\theta) \quad (13)$$

Estimator $\hat{\theta}$ means that it is a function of the observations (the observations can be changed).

Estimate $\hat{\theta}$ means that it is the function value of the corresponding estimator at certain observations.

The Likelihood for a Gaussian AR(1) Process

Consider the Gaussian AR(1) process

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots \quad (14)$$

with $\varepsilon_t \stackrel{iid}{\sim} N(0, \sigma^2)$. We have $\boldsymbol{\theta} = (c, \phi, \sigma^2)$.

By assuming stationarity, for Y_1 , we have

$$\begin{aligned} E(Y_1) &= \mu &= c/(1 - \phi) \\ E(Y_1 - \mu)^2 &= \sigma^2/(1 - \phi^2) \end{aligned}$$

The Likelihood for a Gaussian AR(1) Process

Since ε_t is Gaussian distributed, Y_1 is also Gaussian

$$Y_1 \sim N(c/(1-\phi), \sigma^2/(1-\phi^2))$$

And hence, the unconditional pdf of Y_1 is

$$f(y_1; \theta) = \left(\frac{2\pi\sigma^2}{1-\phi^2} \right)^{-\frac{1}{2}} \exp \left[-\frac{(y_1 - c/(1-\phi))^2}{2\sigma^2/(1-\phi^2)} \right]$$

The Likelihood for a Gaussian AR(1) Process

Since $Y_2 = c + \phi Y_1 + \varepsilon_2$, the distribution of Y_2 conditional on Y_1 is

$$Y_2 | Y_1 \sim N(c + \phi Y_1, \sigma^2)$$

Then

$$f(y_2 | y_1; \theta) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left[-\frac{(y_2 - (c + \phi y_1))^2}{2\sigma^2} \right]$$

The joint density of y_2 and y_1 is

$$f(y_2, y_1; \theta) = f(y_2 | y_1; \theta) \cdot f(y_1; \theta)$$

The Likelihood for a Gaussian AR(1) Process

Similarly, since $Y_3 = c + \phi Y_2 + \varepsilon_3$, the distribution of Y_3 conditional on Y_2 and Y_1 is

$$Y_3|Y_2, Y_1 \sim N(c + \phi Y_2, \sigma^2)$$

Then

$$f(y_3|y_2, y_1; \theta) = f(y_3|y_2; \theta) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(y_3 - (c + \phi y_2))^2}{2\sigma^2}\right]$$

The joint density of y_3 , y_2 and y_1 is

$$\begin{aligned} f(y_3, y_2, y_1; \theta) &= f(y_3|y_2, y_1; \theta) \cdot f(y_2, y_1; \theta) \\ &= f(y_3|y_2; \theta) \cdot f(y_2|y_1; \theta) \cdot f(y_1; \theta) \end{aligned}$$

The Likelihood for a Gaussian AR(1) Process

By induction, the distribution of Y_t conditional on Y_{t-1}, \dots, Y_1 is

$$Y_t | Y_{t-1}, \dots, Y_1 \sim N(c + \phi Y_{t-1}, \sigma^2)$$

Then

$$f(y_t | y_{t-1}, \dots, y_1; \theta) = f(y_t | y_{t-1}; \theta) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left[-\frac{(y_t - (c + \phi y_{t-1}))^2}{2\sigma^2} \right]$$

The joint density of y_t, \dots, y_1 is

$$f(y_t, \dots, y_1; \theta) = f(y_1; \theta) \prod_{i=2}^t f(y_i | y_{i-1}; \theta).$$

The likelihood function of θ given the observations y_T, \dots, y_1 is
 $f(y_T, \dots, y_1; \theta)$.

The Likelihood for a Gaussian AR(1) Process

The log-likelihood function for the stationary Gaussian AR(1) process is given by

$$\log L(\theta) = \log f(y_1; \theta) + \sum_{t=2}^T \log f(y_t | y_{t-1}; \theta) \quad (15)$$

Replace each f by the corresponding formula, rearrange...

Somewhat complex? There is another expression for that, in matrix form.

The Likelihood for a Gaussian AR(1) Process

Denote the column vector of observations $\mathbf{y} = (y_1, y_2, \dots, y_T)'$ and the random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_T)'$. \mathbf{Y} is multivariate Gaussian distributed, and \mathbf{y} is one realization of it.

Due to stationarity and Gaussian assumption, the unconditional distribution of each element in \mathbf{Y} is identical, but the elements are correlated.

We have $E(\mathbf{Y}) = \boldsymbol{\mu} = (\mu, \dots, \mu)'$, where $\mu = c/(1 - \phi)$.

The covariance matrix of \mathbf{Y} is defined as

$$E(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})' = \boldsymbol{\Omega}$$

The Likelihood for a Gaussian AR(1) Process

Note that Ω has elements

$$\begin{pmatrix} E(Y_1 - \mu)^2 & E(Y_1 - \mu)(Y_2 - \mu) & \cdots & E(Y_1 - \mu)(Y_T - \mu) \\ E(Y_2 - \mu)(Y_1 - \mu) & E(Y_2 - \mu)^2 & \cdots & E(Y_2 - \mu)(Y_T - \mu) \\ \vdots & \vdots & \ddots & \vdots \\ E(Y_T - \mu)(Y_1 - \mu) & E(Y_T - \mu)(Y_2 - \mu) & \cdots & E(Y_T - \mu)^2 \end{pmatrix}$$

In fact, they are autocovariances

$$\Omega = \begin{pmatrix} \gamma_0 & \gamma_{-1} & \cdots & \gamma_{1-T} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{2-T} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{T-1} & \gamma_{T-2} & \cdots & \gamma_0 \end{pmatrix}$$

The Likelihood for a Gaussian AR(1) Process

For a stationary AR(1) process, $\gamma_j = \sigma^2 \phi^j / (1 - \phi^2)$, which implies that

$$\Omega = \sigma^2 \mathbf{V},$$

where

$$\mathbf{V} = \frac{1}{1 - \phi^2} \begin{pmatrix} 1 & \phi & \dots & \phi^{T-1} \\ \phi & 1 & \dots & \phi^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \dots & 1 \end{pmatrix}$$

Thus, we can say that $\mathbf{Y} \sim N_T(\boldsymbol{\mu}, \Omega)$.

The Likelihood for a Gaussian AR(1) Process

Since $\mathbf{Y} \sim N_T(\boldsymbol{\mu}, \boldsymbol{\Omega})$, we have

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2} |\boldsymbol{\Omega}|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right] \quad (16)$$

with log-likelihood function

$$\log L(\boldsymbol{\theta}) = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu}). \quad (17)$$

Remember that $\boldsymbol{\Omega} = \sigma^2 \mathbf{V}$.

The Likelihood for a Gaussian AR(1) Process

Remarks:

- The two expressions are equivalent.
- However, neither of the two expressions have closed form for the ML estimator.
- The estimate has to be found through numerical optimization.
- Section 5.7 in Hamilton gives an introduction to the numerical optimization methods.

Conditional Likelihood for a Gaussian AR(1) Process

Question: What if we consider y_1 as deterministic and maximize the likelihood function conditional on y_1 ?

An advantage: Since y_1 is considered as non-stochastic, we do not need to assume that $|\phi| < 1$ (stationarity for y). Since $Y_t = c + \phi Y_{t-1} + \varepsilon_t$

$$Y_t | Y_{t-1}, \dots, Y_2, y_1 \sim N(c + \phi y_{t-1}, \sigma^2).$$

The conditional distribution of Y_t given all the available past is Gaussian.

Conditional Likelihood for a Gaussian AR(1) Process

The joint density function of y_T, \dots, y_2 conditional on y_1 is given by

$$f(y_T, \dots, y_2 | y_1, \theta) = \prod_{i=2}^T f(y_t | y_{t-1}; \theta), \quad (18)$$

where $f(y_t | y_{t-1})$ is the pdf of $N(c + \phi y_{t-1}, \sigma^2)$. We get rid of the annoying $f(y_1 | \theta)$.

The log-likelihood function is

$$\log L(\theta) = -\frac{T-1}{2} \log(2\pi) - \frac{T-1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^T (y_t - c - \phi y_{t-1})^2 \quad (19)$$

Conditional Likelihood for a Gaussian AR(1) Process

Denote $\beta = (c, \phi)'$ and $\mathbf{x}_t = (1, y_{t-1})'$. The first order condition (FOC) w.r.t. β implies that the ML estimator of β is

$$\hat{\beta} = \left[\sum_{t=2}^T \mathbf{x}_t \mathbf{x}'_t \right]^{-1} \left[\sum_{t=2}^T \mathbf{x}_t y_t \right], \quad (20)$$

which coincides with the OLS estimator for β .

The ML estimator of σ^2 is found in the similar way (FOC), by inputting $\hat{\beta}$.

$$\hat{\sigma}^2 = \frac{1}{T-1} \sum_{t=2}^T (y_t - \hat{c} - \hat{\phi} y_{t-1})^2 \quad (21)$$

which is the average squared residuals from the OLS regression. Note that $\hat{\sigma}^2$ is not unbiased.

Conditional Likelihood for a Gaussian AR(1) Process

Remarks:

- In contrast to the exact MLE, the conditional MLE offers a closed form solution.
- If the sample size T is large enough, the first observation y_1 makes a negligible contribution to the total likelihood.
- The exact and conditional MLE turn out to have the same asymptotic distribution, provided that $|\phi| < 1$.
- The conditional MLE in this case is consistent for both cases $|\phi| < 1$ and $|\phi| \leq 1$.
- The above results can be extended to AR(p) Gaussian processes.

Conditional Likelihood for a Gaussian MA(1) Process

Consider the Gaussian MA(1) process

$$Y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1}, \quad t = 0, \pm 1, \pm 2, \dots \quad (22)$$

where $\varepsilon_t \sim N(0, \sigma^2)$. We have the parameters $\boldsymbol{\theta} = (\mu, \theta, \sigma^2)'$.

Suppose that $\varepsilon_0 = 0$. Then

$$Y_1 | \varepsilon_0 \sim N(\mu, \sigma^2).$$

Moreover, given y_1 and $\varepsilon_0 = 0$, $\varepsilon_1 = y_1 - \mu$ is also known and

$$(Y_2 | y_1, \varepsilon_0 = 0) \sim N(\mu + \theta \varepsilon_1, \sigma^2)$$

Conditional Likelihood for a Gaussian MA(1) Process

Similarly, given y_1, y_2 and $\varepsilon_0 = 0, \varepsilon_2 = y_2 - \mu - \theta\varepsilon_1$ is also known and

$$(Y_3|y_2, y_1, \varepsilon_0 = 0) \sim N(\mu + \theta\varepsilon_2, \sigma^2)$$

Proceeding in this fashion, it follows that

$$\begin{aligned} f(y_t|y_{t-1}, \dots, y_1, \varepsilon_0 = 0, \boldsymbol{\theta}) &= f(y_t|\varepsilon_{t-1}, \boldsymbol{\theta}) \\ &= (2\pi\sigma^2)^{-1/2} \exp \left[-\frac{(y_t - \mu - \theta\varepsilon_{t-1})^2}{2\sigma^2} \right] \quad E[Y_t] \\ &= (2\pi\sigma^2)^{-1/2} \exp \left[-\frac{\varepsilon_t^2}{2\sigma^2} \right], \quad t = 2, 3, \dots \end{aligned} \tag{23}$$

Conditional Likelihood for a Gaussian MA(1) Process

The joint density function conditional on $\varepsilon_0 = 0$ is given by

$$f(y_T, \dots, y_1 | \varepsilon_0 = 0, \theta) = f(y_1 | \varepsilon_0, \theta) \prod_{t=2}^T f(y_t | \varepsilon_{t-1}, \theta)$$

The log-likelihood function is

$$\begin{aligned}\log L(\theta) &= \log f(y_T, \dots, y_1 | \varepsilon_0 = 0, \theta) \\ &= -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T \varepsilon_t^2\end{aligned}\quad (24)$$

where $\varepsilon_t = (y_t - \mu) - \theta\varepsilon_{t-1} = (y_t - \mu) - \theta[(y_{t-1} - \mu) - \theta\varepsilon_{t-2}] = \dots = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - \dots + (-1)^{t-1}\theta^{t-1}(y_1 - \mu) + (-1)^t\theta^t\varepsilon_0$.

Conditional Likelihood for a Gaussian MA(1) Process

Remarks:

- The conditional ML estimator for a Gaussian MA(1) process has no closed form.
- If $|\theta|$ is close enough to zero, the effect of imposing $\varepsilon_0 = 0$ will quickly die out, though it may not be a suitable assumption.
- If $|\theta| > 1$, the conditional approach is not reasonable.
- If the numerical optimization results in $|\hat{\theta}| > 1$, the results must be discarded.
- The above results can be extended to a Gaussian MA(q) process.
- The conditional log-likelihood function for a Gaussian MA(q) process is useful only if the process is invertible.

Conditional Likelihood for a Gaussian ARMA(p, q) Process

Consider the Gaussian ARMA(p, q) process

$$Y_t = c + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad t = 0, \pm 1, \pm 2, \dots \quad (25)$$

where $\varepsilon_t \sim N(0, \sigma^2)$. We have the parameters

$$\boldsymbol{\theta} = (c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2)'$$

Denote

$$\mathbf{y}_0 = (y_0, y_{-1}, \dots, y_{-p+1})'$$

and

$$\boldsymbol{\varepsilon}_0 = (\varepsilon_0, \varepsilon_{-1}, \dots, \varepsilon_{-q+1})'$$

Conditional on \mathbf{y}_0 and $\boldsymbol{\varepsilon}_0$, the sequence of ε_t can be calculated using the observations y_t by iterating

$E(\varepsilon_t)$

$$\varepsilon_t = y_t - \underbrace{c - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}}_{E(\varepsilon_t)}, \quad (26)$$

Conditional Likelihood for a Gaussian ARMA(p, q) Process

Likewise, we have the conditional log-likelihood function for a Gaussian ARMA(p, q) process

$$\log L(\theta) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T \varepsilon_t^2$$

where ε_t is given recursively by (26).

In order to maximize the log-likelihood, one approach is to assume, as before, that the initial value of y and ε are equal to their expected values.

$$\mathbf{y}_0 = c / (1 - \phi_1 - \dots - \phi_p) \mathbf{1}_p, \quad \boldsymbol{\varepsilon}_0 = \mathbf{0}_p.$$

Conditional Likelihood for a Gaussian ARMA(p, q) Process

Another approach, according to Box & Jenkins (1976, pp.211), is to set the initial ε equal to zero, but the initial y to their observed values. This means that first p y s become deterministic, and the sample size becomes $T - p$.

The corresponding log-likelihood function becomes

$$\log L(\theta) = -\frac{T-p}{2} \log(2\pi) - \frac{T-p}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=p+1}^T \varepsilon_t^2$$

Statistical Model, Likelihood, ML Estimation

- A statistical model for multivariate data x from a sample space S is given by a parametrized family of probability of joint densities $p(x|\theta)$, where the parameter θ is varying in a parameter set Θ . *concept*
- We define the **likelihood function** by fixing the data x , and consider the density as a function of the parameter: *[a, b]*.

$$L(\theta) = p(x|\theta), \quad \theta \in \Theta. \quad (27)$$

- We can also write $L(\theta|x)$.
- The Maximum Likelihood estimator is given as the value of θ that maximize the likelihood function

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta), \quad \text{or } \hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta) \quad (28)$$

- In this way, ML estimator becomes a function of the data, hence a random variable.

Hypothesis testing

- A hypothesis concerning the parameter θ is expressed as a restriction on the parameter set Θ : $H_0 : \theta \in \Theta_0$. The null hypothesis.
- The alternative hypothesis $H_1 : \theta \in \Theta / \Theta_0$.
- In most cases, it is more convenient to express the null hypothesis on the parameter as a restriction on θ : $H_0 : g(\theta) = 0$. In this case, $\Theta_0 = \{\theta : g(\theta) = 0\}$.
- The statistical model under the null becomes $\{p(x|\theta), \theta \in \Theta_0\}$, or $\{p(x|\theta), g(\theta) = 0\}$. The statistical model under the null is nested by the former statistical model $\theta \in \Theta$.

Hypothesis testing

- The ML estimator for the restricted θ is found by maximizing the likelihood function under the constraint specified by Θ_0 . Denote by $\tilde{\theta}$.
- We have the Likelihood Ratio test statistic: $Q(x) = L(\tilde{\theta})/L(\hat{\theta})$.
- p -value is defined to be: $pv = \sup_{\theta \in \Theta_0} \text{Prob}\{Q(x) \leq Q(x_{obs})\}$.
- Sig. level and critical value c_α : $\sup_{\theta \in \Theta_0} \text{Prob}\{Q(x) \leq c_\alpha\} = \alpha$.
- Power of the test: $\text{Prob}(Q(x) \leq c_\alpha)$, given $\theta \in \Theta/\Theta_0$.

Likelihood, score and information

- Given the pair $x = (x_1, \dots, x_T)$ and θ , assuming that θ is a m -dimensional vector, we have the $m \times 1$ score vector:

$$S_T(\theta) = d \log p(x|\theta) / d\theta \quad (29)$$

and the $m \times m$ information matrix:

$$I_T(\theta) = -dS_T/d\theta = -d^2 \log p(x|\theta) / d\theta^2 \quad (30)$$

- We have:

$$E_\theta(S_T(\theta)) = 0 \quad (31)$$

$$E_\theta(I_T(\theta)) = E_\theta(S_T(\theta)S_T(\theta)') = \text{Var}(S_T(\theta)). \quad (32)$$

Proof

Obviously we have:

$$\int p(x|\theta)dx = 1 \quad (33)$$

$$\int \frac{\partial p(x|\theta)}{\partial \theta} dx = 0 \quad (34)$$

$$\int \frac{\partial^2 p(x|\theta)}{\partial \theta^2} dx = 0 \quad (35)$$

Note that $S_T(\theta) = \frac{\partial \log p(x|\theta)}{\partial \theta} = p(x|\theta)^{-1} \frac{\partial p(x|\theta)}{\partial \theta}$. Then

$$E_\theta(S_T(\theta)) = \int p(x|\theta)^{-1} \frac{\partial p(x|\theta)}{\partial \theta} p(x|\theta) dx = 0.$$

$$\text{And } I_T(\theta) = -\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} = -p(x|\theta)^{-1} \frac{\partial^2 p(x|\theta)}{\partial \theta^2} + \frac{\partial \log p(x|\theta)}{\partial \theta} \left(\frac{\partial \log p(x|\theta)}{\partial \theta} \right)'.$$

Thus,

$$E_\theta(I_T(\theta)) = E_\theta\left(\frac{\partial \log p(x|\theta)}{\partial \theta} \left(\frac{\partial \log p(x|\theta)}{\partial \theta} \right)'\right) = E_\theta(S_T(\theta)S_T(\theta)') = \text{Var}(S_T(\theta)).$$

The three tests

- Suppose that θ_0 is the true value of the parameter θ .
- The likelihood ratio test:

$$Q(x) = -2 \log \frac{L(\theta_0)}{L(\hat{\theta})} = 2(\log(L(\hat{\theta})) - \log(L(\theta_0))) \quad (36)$$

- The Wald test:

$$\hat{\theta} \sim N(\theta_0, I_T(\hat{\theta})^{-1}) \quad (37)$$

- The Lagrange-multiplier or score test:

$$S_T(\theta_0) \sim N(0, I_T(\theta_0)) \quad (38)$$

The relations between the three tests

Stochastic Taylor expansion:

$$\log L(\theta_0) = \log L(\hat{\theta}) + (\theta_0 - \hat{\theta})' S_T(\hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta}) I_T(\hat{\theta})(\theta_0 - \hat{\theta}) + \dots \quad (39)$$

Since $S_T(\hat{\theta}) = 0$, Taylor again $0 = S_T(\hat{\theta}) = S_T(\theta_0) - I_T(\theta_0)(\hat{\theta} - \theta_0) + \dots$,
and hence

$$S_T(\theta_0) = I_T(\theta_0)(\hat{\theta} - \theta_0) + \dots \quad (40)$$

We have the relations:

$$\begin{aligned} -2 \log \frac{L(\theta_0)}{L(\hat{\theta})} &= (\theta_0 - \hat{\theta}) I_T(\hat{\theta})(\theta_0 - \hat{\theta}) + \dots = S_T(\theta_0)' I_T(\theta_0)^{-1} S_T(\theta_0) + \dots \\ &\sim \chi^2(m) \end{aligned} \quad (41)$$



To be continued! Thank you!

Notes and Supplementary Readings

Time Series Econometrics 2ST111



UPPSALA
UNIVERSITET

Yukai Yang, PhD

Uppsala Universitet
Statistiska Institutionen
P.O. 513, 751 20 Uppsala
Sverige

yukai.yang@statistik.uu.se

Contents

1 Stationary ARMA processes	1
2 Maximum likelihood and the three tests	3
2.1 maximum likelihood	3
2.2 hypothesis testing, score and information	4
2.3 the three tests	6
3 Asymptotic distribution theory	9
3.1 convergence in probability	9
3.2 convergence in mean square	9
3.3 convergence in distribution	10
3.4 small o and big o	11
3.5 central limit theorem	11
4 Martingale difference sequence and central limit theorem	12

1 Stationary ARMA processes

A stationary autoregressive moving average process of order p, q , ARMA(p, q) is defined to be

$$\phi(L)(y_t - \mu) = \theta(L)\varepsilon_t, \quad (1.1)$$

where the lag polynomial $\phi(L) = 1 - \sum_{i=1}^p \phi_i L^i$ is stable and the lag polynomial $\theta(L) = \sum_{i=0}^q \theta_i L^i$ is absolutely summable. The error sequence $\{\varepsilon_t\}$ is a white noise process (zero mean, constant variance and no serial correlation, but not necessarily independent through time). In most cases, it is presumed to be Gaussian white noise, which implies the likelihood form, the ergodicity for every moment, and the independency through time. The random variable y_t in (1.1) has the invariant unconditional expectation μ .

A stable lag polynomial $\phi(L)$ implies that it can be factorized as follows:

$$\phi(L) = 1 - \sum_{i=1}^p \phi_i L^i = \prod_{i=1}^p (1 - \lambda_i L), \quad (1.2)$$

with $|\lambda_i| < 1$ for $i = 1, \dots, p$. Consider the special case when $p = 1$, $\phi(L) = 1 - \phi_1 L = 1 - \lambda_1 L$ with $\phi_1 = \lambda_1$, and the stability implies that $|\phi_1| < 1$.

An absolutely summable lag polynomial $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ satisfies $\sum_{i=0}^{\infty} |\psi_i| < \infty$. Apparently the lag polynomial $\theta(L)$ is absolutely summable as it is a finite series. Actually it can be regarded as $\psi_i = \theta_i$ when $i \leq q$ and $\psi_i = 0$ when $i > q$, a special case of the infinite series.

We have shown in the slides that a stable lag polynomial can be inverted to an absolutely summable one in the following way:

$$\phi(L)^{-1} = \left(1 - \sum_{i=1}^p \phi_i L^i \right)^{-1} = \prod_{i=1}^p (1 - \lambda_i L)^{-1}; \quad (1.3)$$

provided that $|\lambda_i| < 1$, we have

$$(1 - \lambda_i L)^{-1} = 1 + \sum_{j=1}^{\infty} \lambda_i^j L^j, \quad (1.4)$$

which satisfies $1 + \sum_{j=1}^{\infty} |\lambda_i^j| = (1 - |\lambda_i|)^{-1} < \infty$ (absolutely summable); and thus,

$$\phi(L)^{-1} = \prod_{i=1}^p \left(1 + \sum_{j=1}^{\infty} \lambda_i^j L^j \right) = \sum_{i=0}^{\infty} \tilde{\psi}_i L^i, \quad (1.5)$$

with $\tilde{\psi}_0 = 1$ and $\sum_{i=0}^{\infty} |\tilde{\psi}_i| < \infty$ due to that the product of absolutely summable lag polynomials are still absolutely summable.

A lag polynomial is invertible if its inverse exists. The stable lag polynomial is invertible, as its inverse is given by (1.5).

However, note that an absolutely summable lag polynomial may not be invertible. Consider the case when $q = 1$, $1 + \tilde{\theta}_1 L$ and $\tilde{\theta}_1 > 1$, see the example in Hamilton (1994) for more details.

The absolutely summable lag polynomial that we obtain by inverting the stable one is invertible, because, from (1.5), clearly we have

$$\left(\sum_{i=0}^{\infty} \tilde{\psi}_i L^i \right)^{-1} = \phi(L). \quad (1.6)$$

We presume that the absolutely summable lag polynomial $\theta(L)$ is invertible.

Hence, the following lag polynomial

$$\psi(L) = \sum_{i=1}^{\infty} \psi_i L^i = \phi(L)^{-1} \theta(L) \quad (1.7)$$

is invertible and absolutely summable.

Now let us consider the ARMA process (1.1) again. It follows that

$$\phi(L)(y_t - \mu) = \phi(L)y_t - \phi(L)\mu = \phi(L)y_t - c = \theta(L)\varepsilon_t,$$

where the intercept $c = \phi(1)\mu$, and $\phi(1) = 1 - \sum_{i=1}^p \phi_i$ (this is how the intercept comes). By rearranging and multiplying $\phi(L)^{-1}$ on both sides, we obtain

$$y_t = \phi(L)^{-1}c + \phi(L)^{-1}\theta(L)\varepsilon_t = \mu + \psi(L)\varepsilon_t, \quad (1.8)$$

which is exactly the Wold's decomposition. y_t has time-invariant expectation μ . The absolute summability of $\psi(L)$ ensures that y_t has finite unconditional variance, which is also time-invariant. The autocovariance structure γ_j can be obtained, see the slides or Hamilton (1994). It turns out that the autocovariances do not depend on time. Therefore, the ARMA process (1.1) is covariance-stationary.

From ARMA(p, q) to MA(∞) if we multiply the both sides of (1.1) by $\phi(L)^{-1}$, which is an absolutely summable lag polynomial, this yields a stationary moving average process of infinite order:

$$y_t - \mu = \phi(L)^{-1}\theta(L)\varepsilon_t = \psi(L)\varepsilon_t, \quad (1.9)$$

where $\psi(L)$ is an absolutely summable lag polynomial with infinite terms.

From ARMA(p, q) to AR(∞) if we multiply the both sides of (1.1) by $\theta(L)^{-1}$, this yields a stationary autoregressive process of infinite order:

$$\theta(L)^{-1}\phi(L)(y_t - \mu) = \varepsilon_t, \quad (1.10)$$

where $\theta(L)^{-1}\phi(L)$ is actually the inverse of $\psi(L)$.

Same rules hold for the conversions from AR(p) to MA(∞) and from MA(q) to AR(∞).

2 Maximum likelihood and the three tests

Consider the statistical model for the data X of sample size T from a sample space S . For example, in univariate case (the dependent variable is a scalar), the set X stacks the sequence of observations $\{y_t\}_{t=1}^T$, i.e., $X = (y_1, y_2, \dots, y_T)$. Each observation y_t is a realization of its random variable and takes a value from its support $y_t \in S_t$. Then the sample space $S = S_1 \otimes S_2 \otimes \dots \otimes S_T$.

The joint density function of X given the set of the parameters θ (any values of θ) is denoted by $p(X|\theta)$. The parameters take the values in the space Θ , i.e., $\theta \in \Theta$. Denote m the dimension of the parameter set (number of parameters).

Suppose that $\theta_0 \in \Theta$ is the set of the true parameters, from which the observations are drawn. The true data generating process has the joint density function $p(X|\theta_0)$. Unfortunately I do not know the value of θ_0 .

2.1 maximum likelihood

Given the observations X , the likelihood function is $L(\theta) = p(X|\theta)$ for $\theta \in \Theta$, which is a function of θ . The maximum likelihood (ML) estimator is defined as

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta), \quad (2.1)$$

or equivalently,

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log L(\theta), \quad (2.2)$$

where $\log L(\theta)$ is the logarithm of the likelihood, or simply log-likelihood function. Normally the log-likelihood function (2.2) is easier to handle as it takes the summation form, if the joint density belongs to the exponential family.

Consider the following AR(1) model

$$y_t = \phi y_{t-1} + \varepsilon_t, \quad (2.3)$$

for $t = 1, \dots, T$ with y_0 given, where $\varepsilon_t \stackrel{iid}{\sim} N(0, \sigma^2)$. The data set is $X = (y_1, y_2, \dots, y_T)$, and the parameter set is $\theta = (\phi, \sigma^2)$. The corresponding likelihood function is

$$L(\theta) = p(X|\theta) = \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_t - \phi y_{t-1})^2}{2\sigma^2} \right\}. \quad (2.4)$$

The log-likelihood function is

$$\log L(\theta) = -\frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \phi y_{t-1})^2. \quad (2.5)$$

We see that, if the first order condition is applied here to find the optimum, the log-likelihood (2.5) is much easier to handle than the likelihood (2.4).

It is worth noting that in practice people solve the following the maximization problem in order to find the ML estimate.

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log L(\theta)/T, \quad (2.6)$$

where $\log L(\theta)/T$ is usually termed the average log-likelihood function.

One advantage of the average log-likelihood is to control the magnitude of the log-likelihood function. Note that the magnitude of the average log-likelihood is rather stable no matter how large the sample size T is. Moreover, the consistency of the ML estimator relies on whether the average log-likelihood will converge to some function with a unique global maximum with the maximizer being the true parameter.

2.2 hypothesis testing, score and information

A hypothesis concerning the parameter θ is expressed as a restriction on the parameter space Θ , i.e., the null hypothesis $H_0 : \theta \in \Theta_0 \subset \Theta$.

Consider the AR(1) model (2.3) whose parameter space is $\Theta = \Phi \otimes \Sigma$ where $\Phi = \mathbb{R}$ is the parameter space for ϕ and $\Sigma = \mathbb{R}^+$ for σ^2 . A possible null hypothesis for the AR(1) model (2.3), for example, can be $\phi = 0$ or equivalently $\Theta_0 = \{0\} \otimes \Sigma$.

The ML estimator under the unrestricted model with $\theta \in \Theta$ is $\hat{\theta} = (\hat{\phi}, \hat{\sigma}^2)$. We denote $\tilde{\theta}$ the ML estimator under the restricted model with $\theta \in \Theta_0$.

$$\tilde{\theta} = \arg \max_{\theta \in \Theta_0} \log L(\theta), \quad (2.7)$$

For the above case, we have $\tilde{\theta} = (0, \hat{\sigma}^2)$.

The score vector is defined to be the first-order derivative of the log-likelihood function

$$S_T(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}. \quad (2.8)$$

It is a m -vector of functions of the parameter vector θ . People use very often the average score, which is defined as $\bar{S}_T(\theta) = S_T(\theta)/T$.

The information matrix is defined as follows:

$$I_T(\theta) = -\frac{\partial S_T(\theta)}{\partial \theta} = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}, \quad (2.9)$$

which is a $m \times m$ matrix. People use very often the average information, which is defined as $\bar{I}_T(\theta) = I_T(\theta)/T$.

Under the regularity conditions¹, the ML estimator $\hat{\theta}$ makes $S_T(\hat{\theta}) = 0$ which is exactly the first-order condition for maximization. And $I_T(\hat{\theta})$ must be positive definite as $\hat{\theta}$ is a maximizer.

The Fisher's information is defined as follows

$$\mathcal{I}_T(\theta) = E_\theta[I_T(\theta)]. \quad (2.10)$$

E_θ reads the expectation with respect to θ . This mean that, given any $\theta \in \Theta$, for the fixed sample size T , we resample the data from the data generating process $p(X|\theta)$ for infinite times (population). For each sample, we compute its $I_T(\theta)$ and then take the average $E_\theta[I_T(\theta)]$. Note

¹If you are interested in the regularity conditions, please refer to the corresponding articles and books. We will not discuss about them here.

that the given θ for the data generating is known for this procedure, but we can change it for another procedure. Thus, in a formal representation, it is

$$\mathcal{I}_T(\theta) = \int_S I_T(\theta)p(X|\theta)dX. \quad (2.11)$$

Still the Fisher's information is a function of θ . We input θ first, sample using the θ , and compute the function value $\mathcal{I}_T(\theta)$ at θ .

The information matrix $I_T(\theta)$ and the Fisher's information $\mathcal{I}_T(\theta)$ will both go to infinity when the sample size $T \rightarrow \infty$. However, the average information matrix and the average Fisher's information may converge to each other under certain conditions², i.e., $\bar{I}_\theta(\theta) \xrightarrow{P} \mathcal{I}(\theta)$ where $\mathcal{I}(\theta) = \lim_{T \rightarrow \infty} \mathcal{I}_T(\theta)/T$. This is the reason why sometimes one can replace $\mathcal{I}_T(\theta)$ by $I_T(\theta)$.

Theorem 1. *The score vector has the following properties:*

$$E_\theta[S_T(\theta)] = \mathbf{0}, \quad (2.12)$$

$$\text{Var}_\theta[S_T(\theta)] = \mathcal{I}_T(\theta). \quad (2.13)$$

Proof. Obviously we have:

$$\int_S p(X|\theta) dX = 1 \quad (2.14)$$

$$\int_S \frac{\partial p(X|\theta)}{\partial \theta} dX = 0 \quad (2.15)$$

$$\int_S \frac{\partial^2 p(X|\theta)}{\partial \theta^2} dX = 0 \quad (2.16)$$

Note that

$$S_T[\theta] = \frac{\partial \log p(X|\theta)}{\partial \theta} = p(X|\theta)^{-1} \frac{\partial p(X|\theta)}{\partial \theta}.$$

Then

$$E_\theta[S_T(\theta)] = \int_S p(X|\theta)^{-1} \frac{\partial p(X|\theta)}{\partial \theta} p(X|\theta) dX = \int_S \frac{\partial p(X|\theta)}{\partial \theta} dX = 0.$$

And

$$I_T[\theta] = -\frac{\partial^2 \log p(X|\theta)}{\partial \theta^2} = -p(X|\theta)^{-1} \frac{\partial^2 p(X|\theta)}{\partial \theta^2} + \frac{\partial \log p(X|\theta)}{\partial \theta} \left(\frac{\partial \log p(X|\theta)}{\partial \theta} \right)'.$$

Thus,

$$E_\theta[I_T(\theta)] = E_\theta \left[\frac{\partial \log p(x|\theta)}{\partial \theta} \left(\frac{\partial \log p(x|\theta)}{\partial \theta} \right)' \right] = E_\theta[S_T(\theta)S_T(\theta)'] = \text{Var}[S_T(\theta)].$$

□

The regularity conditions include the differentiability of the density function and the existence of the above integrals in the proof.

The theorem says that, if the observed data X with size T is a sample of draws from the model with the unknown true parameter θ_0 , then the score $S_T(\theta_0)$ should have expectation 0 (around zero) and the covariance $\mathcal{I}_T(\theta_0)$. We will see in the following section that this is the basic idea of the score test.

²Note that the certain conditions are not identical to the regularity conditions.

2.3 the three tests

In this section, our interest focuses on the hypothesis testing for the null hypothesis $H_0 : \theta \in \Theta_0$. We call the model without any restrictions on the parameters, i.e., $\theta \in \Theta$, the *unrestricted model*, and we call the model with restrictions on the parameters, i.e., $\theta \in \Theta_0$, the *restricted model*. As defined in the previous section, $\hat{\theta}$ is the ML estimator under the unrestricted model, and $\tilde{\theta}$ the ML estimator under the restricted model.

Let us say that the parameter set can be split into two parts $\theta = (\alpha, \beta)$, and the true values of the parameters are $\theta_0 = (\alpha_0, \beta_0)$. The ML estimators for them under the unrestricted model are then given by $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$. Actually we are testing $H_0 : \alpha = \check{\alpha}$ where the vector α has dimension k with $1 \leq k \leq m$, and hence the ML estimators under the restricted model are precisely $\tilde{\theta} = (\check{\alpha}, \tilde{\beta})$.

Based on the log-likelihood function, there are three tests: Wald test, Lagrange-multiplier or score test, and likelihood ratio test.

The Wald test statistic takes the form as follows:

$$Q_{\text{wald}} = (\hat{\alpha} - \check{\alpha})' \hat{\Sigma}_\alpha (\hat{\alpha} - \check{\alpha}), \quad (2.17)$$

where the $k \times k$ matrix $\hat{\Sigma}_\alpha$ is the α -corresponding main diagonal block in $I_T(\hat{\theta})$. Typically the Wald test compares Q_{wald} with the $\chi^2(k)$ distribution, yields the p -value and makes statistical inferences. This is the one-sided test version. One can make it more flexible in the sense that one just compares $\hat{\alpha}$ with the Gaussian distribution $N(\check{\alpha}, \hat{\Sigma}_\alpha^{-1})$, or even computes t -ratios for each element of $\hat{\alpha}$, and then the test becomes a two-sided one.

After all, the Wald test tells the story that, if the null hypothesis is true, then

$$\hat{\alpha} \sim N(\check{\alpha}, \hat{\Sigma}_\alpha^{-1}) \quad (2.18)$$

approximately, due to the fact that $\hat{\theta} \sim N(\theta_0, I_T(\theta_0)^{-1})$ approximately under certain conditions. Note that $I_T(\theta_0)^{-1}$ is the Cramér-Rao lower bound for $\text{Var}(\hat{\theta})$.

(2.18) is asymptotically valid for stationary ARMA models with white noise errors under regularity conditions. The central limit theorem ensures that $\sqrt{T}(\hat{\theta} - \theta_0)$ will converge to a zero-mean multivariate Gaussian distribution with covariance $I(\theta_0)^{-1}$. Note that $I(\theta_0)$ is the limit of the average Fisher's information. (2.18) also says that $\hat{\alpha} \xrightarrow{P} \alpha_0$, as $I_T(\theta_0)$ will explode.

Suppose that the null hypothesis is true, i.e., $\check{\alpha} = \alpha_0$. Since the ML estimator is consistent, $\hat{\theta} \approx \theta_0$, $I_T(\hat{\theta}) \approx I_T(\theta_0)$, and $I_T(\hat{\theta}) \approx I_T(\hat{\theta})$ as explained in the previous section. Then we believe that (2.18) is true and (2.17) is approximately $\chi^2(k)$ distributed.

However, in many cases, (2.18) is incorrect and (2.17) is not $\chi^2(k)$ distributed. For example, when the true model is a random walk or contains a unit root, the ML estimator is still consistent, but $\sqrt{T}(\hat{\alpha} - \alpha_0)$ does not converge to a Gaussian distribution. Then (2.18) does not hold, and (2.17) is not χ^2 distributed any more. This mistake results in a poor size property for the test. For a certain case in your analysis or research, you need to check carefully. In the literature, a common way to fix the Wald test is to suggest a better distribution for (2.17) to improve the size.

The Lagrange-multiplier (LM) test statistic or the score test statistic takes the form as follows:

$$Q_{lm} = S_T(\tilde{\theta})' I_T(\tilde{\theta})^{-1} S_T(\tilde{\theta}) = T \bar{S}_T(\tilde{\theta})' \bar{I}_T(\tilde{\theta})^{-1} \bar{S}_T(\tilde{\theta}). \quad (2.19)$$

Typically the LM test compares Q_{lm} with the $\chi^2(k)$ distribution, yields the p -value and makes the inferences. It is a one-sided test. It follows the fact that the score vector at the true parameters $S_T(\theta_0)$ should have zero mean and covariance $I_T(\theta_0)$, which is absolutely correct under the regularity conditions. But the distribution of the score is definitely not Gaussian in finite sample cases $T < \infty$.

The Gaussianity, again, comes from the central limit theorem under certain conditions. So in practice, you have to check carefully. If (2.19) is not asymptotically Gaussian distributed, then one needs to suggest another distribution for better size property.

Denote $S_\alpha(\theta) = \partial \log L(\theta)/\partial \alpha$ and $S_\beta(\theta) = \partial \log L(\theta)/\partial \beta$. If $S_\beta(\tilde{\theta}) = 0$ which is the first-order condition for the ML estimation of the restricted model, the LM test can be simplified to

$$Q_{lm} = S_\alpha(\tilde{\theta})' \tilde{\Sigma}_\alpha^{-1} S_\alpha(\tilde{\theta}) = T \bar{S}_\alpha(\tilde{\theta})' \bar{\Sigma}_\alpha^{-1} \bar{S}_\alpha(\tilde{\theta}), \quad (2.20)$$

where \bar{S}_α is the average score, and $\tilde{\Sigma}_\alpha$ and $\bar{\Sigma}_\alpha$ are the α -corresponding main diagonal block in $I_T(\tilde{\theta})$ and $\bar{I}_T(\tilde{\theta})$, respectively.

Note that (2.19) and (2.20) are identical if $S_\beta(\tilde{\theta}) = 0$, but the equality does not hold when some nonlinear parameter enter, e.g., σ^2 . In order to get rid of $S_{\sigma^2}(\tilde{\theta})$ in Q_{lm} , then you need to check the Fisher's information. You will find very soon that $I_T(\tilde{\theta})$ may be block diagonal with α - β off diagonal elements being zero, or that (2.19) and (2.20) are asymptotically equivalent if $I(\theta_0)$ is block diagonal with α - β off diagonal elements being zero. Don't forget that, in principle, $I_T(\theta_0)$ is the one that we should take for the LM test, which must be block diagonal.

The likelihood ratio (LR) test aims to compare how far the likelihood functions under the restricted and the unrestricted models are away to each other. If the ratio between them is close enough to one, then the null hypothesis is acceptable. The LR test takes the form as follows:

$$Q_{lr} = -2 \log \left(\frac{L(\hat{\theta})}{L(\tilde{\theta})} \right) = 2 \log L(\hat{\theta}) - 2 \log L(\tilde{\theta}). \quad (2.21)$$

As $\Theta_0 \subset \Theta$, $L(\tilde{\theta}) \leq L(\hat{\theta})$, where the equality holds when $\tilde{\theta} = \hat{\theta}$ given that $\hat{\theta}$ is the unique global maximum of the likelihood function $L(\theta)$. This implies that $L(\tilde{\theta})/L(\hat{\theta}) \leq 1$ and $Q_{lr} \geq 0$. If the two likelihood functions are close enough, we can accept the null hypothesis, which means that this is a one-sided test.

Under certain conditions (you need to check each time), the LR test statistic can be compared with the $\chi^2(k)$ distribution. Typically people compute the p -value of the LR test for the hypothesis testing based on the χ^2 distribution, but confess that there may be size-distortion.

In the slides, we have shown that the three tests are asymptotically equivalent. We discuss it in more details but not rigorously in the following.

First let us apply the Taylor expansion on the log-likelihood function at the estimator under the restricted model $\log L(\tilde{\theta})$. We expand it around the estimator under the unrestricted model. If the null hypothesis is true, then the two estimators should be asymptotically identical, which means that the expansion is valid and the approximation is accurate.

$$\log L(\tilde{\theta}) = \log L(\hat{\theta}) + (\tilde{\theta} - \hat{\theta})' S_T(\hat{\theta}) - \frac{1}{2} (\tilde{\theta} - \hat{\theta})' I_T(\hat{\theta})(\tilde{\theta} - \hat{\theta}) + o(||\tilde{\theta} - \hat{\theta}||^2) \quad (2.22)$$

where $o(||\tilde{\theta} - \hat{\theta}||^2)$ is the term that converges to zero faster than $||\tilde{\theta} - \hat{\theta}||^2$, i.e., $o(||\tilde{\theta} - \hat{\theta}||^2)/||\tilde{\theta} - \hat{\theta}||^2 \rightarrow 0$ as $||\tilde{\theta} - \hat{\theta}|| \rightarrow 0$, where $|| \cdot ||$ is a vector norm.

Note that $S_T(\hat{\theta}) = 0$ (for σ^2 , consider the concentrated log-likelihood). Then we have

$$-2(\log L(\tilde{\theta}) - \log L(\hat{\theta})) = (\hat{\theta} - \tilde{\theta})' I_T(\hat{\theta})(\hat{\theta} - \tilde{\theta}) + o(||\hat{\theta} - \tilde{\theta}||^2), \quad (2.23)$$

where the left-hand side is the LR test, and the right-hand side is the Wald test. Note that the right-hand side is not the exact form of the Wald test that we introduced before. The Wald test only use $\hat{\alpha}$ and $\check{\alpha}$ in the formula. Consider that if you are testing $H_0 : \theta = \check{\theta}$ with $k = m$, then the right-hand side becomes exactly the Wald test.

Let us again expand $S_T(\hat{\theta})$ around $\tilde{\theta}$, and we obtain

$$S_T(\hat{\theta}) = 0 = S_T(\tilde{\theta}) - I_T(\tilde{\theta})(\hat{\theta} - \tilde{\theta}) + o(||\hat{\theta} - \tilde{\theta}||), \quad (2.24)$$

and hence

$$S_T(\tilde{\theta}) = I_T(\tilde{\theta})(\hat{\theta} - \tilde{\theta}) + o(||\hat{\theta} - \tilde{\theta}||), \quad (2.25)$$

We have the relations:

$$S_T(\tilde{\theta})' I_T(\tilde{\theta})^{-1} S_T(\tilde{\theta}) = (\tilde{\theta} - \hat{\theta})' I_T(\hat{\theta})(\tilde{\theta} - \hat{\theta}) + o(||\hat{\theta} - \tilde{\theta}||). \quad (2.26)$$

where the left-hand side is the LM test, and the right-hand side is the Wald test.

Remember that, if the null hypothesis is true, $\check{\alpha} = \alpha_0$ and both $\hat{\theta}$ and $\tilde{\theta}$ converge to θ_0 in probability. The distribution of all the three tests, under certain conditions, converges to $\chi^2(k)$.

This is not a rigorous proof for the asymptotic equivalence of the three tests, but it gives you the clue about how they are related to each other.

Though the three tests are asymptotically equivalent, they are totally different when you conduct them. In order to produce the Wald test statistic, you estimate the unrestricted model and get the estimates $\hat{\theta}$ and the information $I_T(\hat{\theta})$. For the LM test, instead of the unrestricted model, you estimate the restricted model and get the estimates $\tilde{\theta}$, the score $S_T(\tilde{\theta})$ and the information $I_T(\tilde{\theta})$. For the LR test, you have to estimate both the restricted model and the unrestricted model, compute the log-likelihood functions for both of them.

In practice, you may face the problem: "which test shall I choose?". It depends on many things. The convenience is one of the main reasons. If the unrestricted model is much more difficult to estimate than the restricted one, then you may probably consider the LM test. Sometimes the LR test may offer great convenience for derivation and the following statistical inference, though you have to estimate both models.

3 Asymptotic distribution theory

We make a summary of the asymptotic distribution theory here. First of all, in this section, we are talking about the sequence of random numbers, i.e., $\{X_t\}_{t=1}^{\infty}$. There are infinite random numbers in the sequence, from $t = 1$ to infinity. Each random number X_t follows its own distribution, may not be identical.

3.1 convergence in probability

We say that the sequence of the random variables X_t converges in probability to a constant c , if for any (small) positive number ε and any (small) positive δ , there exists a (big) integer N , such that for any $t \geq N$,

$$\text{Prob}\{|X_t - c| > \delta\} < \varepsilon, \quad \text{A.S. Convergence: } \mathbb{P}\left\{\lim_{t \rightarrow \infty} X_t = c\right\} = 1 \quad (3.1)$$

or, equivalently,

$$\lim_{t \rightarrow \infty} \text{Prob}\{|X_t - c| > \delta\} = 0, \quad (3.2)$$

Since each X_t has its own distribution, $\text{Prob}\{|X_t - c| > \delta\}$ implies a sequence of probabilities given δ . So the convergence in probability simply means that this sequence of probabilities converges to zero.

The convergence in probability can be represented in two ways:

$$\text{plim}_{t \rightarrow \infty} X_t = c, \quad X_t \sim o_p(1) \Leftrightarrow X_t \xrightarrow{P} 0 \quad (3.3)$$

in which you can simply skip $t \rightarrow \infty$, or equivalently,

$$X_t \xrightarrow{P} c. \quad Z_t = \frac{X_t}{t^n} \quad (3.4)$$

Consider if there are two sequences of random variables X_t and Y_t . Then $X_t \xrightarrow{P} Y_t$ simply implies that $X_t - Y_t \xrightarrow{P} 0$. Note that this does not mean that they converge in probability to the same constant. Instead they converge to the same random variable.

We say that an estimator $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{P} \theta_0$ where θ_0 is the true value of the parameter. In this case, we have hidden the subscript T (the sample size of the data) in $\hat{\theta}$.

3.2 convergence in mean square

The sequence of the random variables X_t is said to converge in mean square to a constant c , if for any (small) positive number ε , there exists a (big) integer N , such that for any $t \geq N$,

$$\mathbb{E}(X_t - c)^2 < \varepsilon, \quad (3.5)$$

or, equivalently,

$$\lim_{t \rightarrow \infty} \mathbb{E}(X_t - c)^2 = 0. \quad (3.6)$$

$E(X_t - c)^2$ implies a sequence of these mean squares (moments).

The convergence in mean square can be represented as follows:

$$X_t \xrightarrow{m.s.} c. \quad (3.7)$$

It can also be called convergence in quadratic mean. It can be shown (Proposition 7.2 in Hamilton) that

$$X_t \xrightarrow{m.s.} c \implies X_t \xrightarrow{p} c, \quad (3.8)$$

while the other way around does not hold. And you need to know that, for a sequence of *i.i.d.* random numbers, its sample mean converges in mean square to its expectation.

3.3 convergence in distribution

Consider a sequence of random variables X_t for $t = 1, \dots, \infty$, with the cumulative distribution functions (cdf) $F_t(x)$ for $t = 1, \dots, \infty$. Suppose that there exists a random variable X with the cdf $F(x)$ such that

$$\begin{aligned} X_t &\xrightarrow{P} X \Rightarrow X_t \xrightarrow{d} X \\ \lim_{t \rightarrow \infty} F_t(x) &= F(x) \quad X_t(\omega) \xrightarrow{P} X(\omega) \quad X_t(\omega) \xrightarrow{d} X(\omega). \end{aligned} \quad (3.9)$$

at any value x where $F(\cdot)$ is continuous. Then X_t is said to converge in distribution to X ,

$$X_t \xrightarrow{d} X. \quad (3.10)$$

If you regard a constant c as a random variable with zero variance, you can say $X_t \xrightarrow{P} c \iff X_t \xrightarrow{d} c$.

In Proposition 7.3 (a) in Hamilton, it is true that

$$Y_t \xrightarrow{d} Y \text{ and } X_t - Y_t \xrightarrow{P} 0 \implies X_t \xrightarrow{d} Y. \quad (3.11)$$

However, it is NOT true that

$$Y_t \xrightarrow{d} Y \text{ and } X_t \xrightarrow{d} Y \implies X_t - Y_t \xrightarrow{P} 0. \quad (3.12)$$

The reason is that the convergence in distribution \xrightarrow{d} only ensures that the limit distributions of X_t and Y_t have the same cdf, but they can be independent.

Proposition 7.3 (b) is in fact part of the Slutsky theorem. The complete version is as follows:

If $X_t \xrightarrow{P} c$ and $Y_t \xrightarrow{d} Y$, then

$$(1) \quad X_t + Y_t \xrightarrow{d} c + Y; \quad (3.13)$$

$$(2) \quad X_t Y_t \xrightarrow{d} c Y; \quad (3.14)$$

$$(3) \quad Y_t/X_t \xrightarrow{d} Y/c. \quad (3.15)$$

The third one tells the story that if $X_t \xrightarrow{P} c$ then $1/X_t \xrightarrow{P} 1/c$.

3.4 small o and big o

In this section, we introduce the so-called "stochastic orders". The stochastic orders involves two notations: $o_p(\cdot)$ the small o and $O_p(\cdot)$ the big O.

For a sequence of random variables X_t , we say that $X_t = o_p(t^n)$ for some $n \in \mathbb{R}$ if $X_t/t^n \xrightarrow{P} 0$ as $t \rightarrow \infty$. For example, if $X_t \xrightarrow{P} 0$, then we say $X_t = o_p(1)$. Note that, in the literature, sometime people write $o_p(n)$ instead of $o_p(t^n)$.

For a sequence of random variables X_t , we say that $X_t = O_p(t^n)$ for some $n \in \mathbb{R}$ if for any (small) positive number ε , there exists a positive number (can be big) c and a positive integer N such that for any $t > N$

$$\text{Prob}\{|X_t|/t^n > c\} < \varepsilon \quad (3.16)$$

as $t \rightarrow \infty$.

Note that this definition is written in a different way in contrast to the definition of convergence in distribution. If $X_t = O_p(1)$, then X_t sequence is often referred to as "tight". Similarly, if $X_t = O_p(t^n)$, then $t^{-n}X_t$ is tight. Intuitively, the tightness means that the corresponding limiting distribution is somewhat "regular" in the sense that you can always find finite lower and upper bounds such that most of the probability "mass" is inside the bounded area, or "boundedness in probability".

We have some properties for the small o and big O:

- $X_t \xrightarrow{P} c \implies X_t = O_p(1)$
- $X_t \xrightarrow{d} X \implies X_t = O_p(1)$
- $X_t = op(t^n) \implies X_t = O_p(t^n)$
- $X_t = O_p(t^n) \implies X_t = op(t^m)$ if $m > n$
- if $X_t = O_p(t^n)$ and $Y_t = O_p(t^m)$, then $X_t + Y_t = O_p(t^{\max(n,m)})$ and $X_t Y_t = O_p(t^{n+m})$, and the same results hold for $o_p(\cdot)$
- if $X_t = O_p(t^n)$ and $Y_t = o_p(t^m)$, then $X_t Y_t = op(t^{n+m})$

3.5 central limit theorem

We have shown the central limit theorem for *i.i.d.* sequence in the previous lecture.

Proposition 7.4 in Hamilton has another name: the "delta method". We give the following theorem based on the small o and big O theory for free.

Theorem 2 (Stochastic Taylor Expansion). *Let X_t be a sequence of random variable in \mathbb{R}^k with $X_t = c + O_p(t^n)$, where $c \in \mathbb{R}^k$ and $n < 0$, such that $t^n \rightarrow 0$ as $t \rightarrow \infty$. Then if f is continuously differentiable at c , we have*

$$f(X_t) = f(c) + f'(c)(X_t - c) + o_p(t^n) \quad (3.17)$$

Now we see that the well known "delta method" or Proposition 7.4 is a special case of it. But this theorem is more general.

4 Martingale difference sequence and central limit theorem

Definition 1 (Martingale). *A sequence of stochastic variables Y_t is called a martingale with respect to the information \mathcal{F}_{t-1} available at time t , if Y_t has finite expectation and is measurable with respect to \mathcal{F}_t , and it holds that*

$$\mathbb{E}(Y_t | \mathcal{F}_{t-1}) = Y_{t-1}. \quad (4.1)$$

Intuitively, a martingale sequence has the property that the expected value of tomorrow's random variable is just today's observation, given all the information available today. Note that there is no requirements for any other moments, and even there is no need to be *i.i.d.* for martingale sequence. This offers great convenience for empirical data modelling and great robustness for the statistical inference. So does the martingale difference sequence as follows.

Definition 2 (Martingale difference sequence). *A sequence of stochastic variables X_t is called a martingale difference sequence with respect to the information \mathcal{F}_t available at time t , if X_t has finite expectation and is measurable with respect to \mathcal{F}_t , and it holds that*

$$\mathbb{E}(X_t | \mathcal{F}_{t-1}) = 0. \quad (4.2)$$

You may ask why it has the name martingale difference sequence (MDS). Define $X_t = Y_t - Y_{t-1}$ where Y_t is martingale. Then clearly $\mathbb{E}(X_t | \mathcal{F}_{t-1}) = 0$. The MDS is the first order difference of a martingale.

Each element in the sequence of a MDS has unconditional zero mean due to

$$\mathbb{E}(X_t) = \mathbb{E}[\mathbb{E}(X_t | \mathcal{F}_{t-1})] = 0 \quad (4.3)$$

You can skip the " L^1 -Mixingales" which is more general on pp.190, but please read the law of large number and the central limit theorem for the MDS.

In the following, we give another version of the vector MDS CLT, which may be used in your futher research.

Theorem 3 (Brown (1971)). *Let \mathbf{X}_t , with finite variance, be a d -dimensional martingale difference sequence with respect to the information \mathcal{F}_{t-1} available at time t . Assume that, as $T \rightarrow \infty$,*

$$T^{-1} \sum_{t=1}^T \mathbb{E}(\mathbf{X}_t \mathbf{X}'_t | \mathcal{F}_{t-1}) \xrightarrow{P} \boldsymbol{\Sigma}, \quad (4.4)$$

where $\boldsymbol{\Sigma}$ is positive definite. Assume further that, as $T \rightarrow \infty$, either

$$T^{-1} \sum_{t=1}^T \mathbb{E} \left[\|\mathbf{X}_t\|^2 \mathbf{1}\{\|\mathbf{X}_t\| > \delta \sqrt{T}\} | \mathcal{F}_t \right] \xrightarrow{P} 0, \quad \text{or} \quad (4.5)$$

$$T^{-1} \sum_{t=1}^T \mathbb{E} \left[\|\mathbf{X}_t\|^2 \mathbf{1}\{\|\mathbf{X}_t\| > \delta \sqrt{T}\} \right] \xrightarrow{P} 0 \quad (4.6)$$

hold for all $\delta > 0$. Then it holds that

$$T^{-1/2} \sum_{t=1}^T X_t \xrightarrow{d} N_d(\mathbf{0}, \boldsymbol{\Sigma}). \quad (4.7)$$

This theorem can also be found in Hall and Heyde (1980). Note that $\|\cdot\|$ is a matrix norm, and that $\mathbf{1}\{\cdot\}$ is an indicator function such that $\mathbf{1}\{\text{True}\} = 1$ and $\mathbf{1}\{\text{False}\} = 0$.

For better understanding, consider the example

$$X_t = \varepsilon_t Z_{t-1}, \quad (4.8)$$

where X_t is a scalar, the sequence ε_t is i.i.d. $(0, \omega)$, the sequence of scalars Z_t is stationary with zero mean and is ergodic in variance, i.e. $T^{-1} \sum_{t=1}^T Z_t^2 \xrightarrow{P} E(Z_t^2) = \sigma^2$, and ε_t is independent of $(Z_{t-1}, \varepsilon_{t-1}, Z_{t-2}, \dots)$. With

$$\mathcal{F}_t = \sigma(\varepsilon_i, Z_i), \quad i = 1, \dots, t,$$

X_t is a MDS with respect to \mathcal{F}_t . Moreover,

$$T^{-1} \sum_{t=1}^T E[X_t^2 | \mathcal{F}_{t-1}] = \omega T^{-1} \sum_{t=1}^T Z_{t-1}^2 \xrightarrow{P} \omega \sigma^2. \quad (4.9)$$

Next, by the stationarity of X_t

$$T^{-1} \sum_{t=1}^T E[X_t^2 \mathbf{1}\{|X_t| > \delta \sqrt{T}\}] = E[X_t^2 \mathbf{1}\{|X_t| > \delta \sqrt{T}\}] \rightarrow 0 \quad (4.10)$$

as $T \rightarrow \infty$. The latter convergence holds by dominated convergence as $X_t^2 \mathbf{1}\{|X_t| > \delta \sqrt{T}\} \leq X_t^2$ and $E(X_t^2) < \infty$. This shows that (4.4) and (4.6) hold and hence Theorem 3 gives

$$T^{-1/2} \sum_{t=1}^T X_t \xrightarrow{d} N(0, \omega \sigma^2). \quad (4.11)$$

References

- Brown, B. M.: 1971, Martingale central limit theorems, *The Annals of Mathematical Statistics* **42**, 59–66.
- Hall, P. and Heyde, C. C.: 1980, *Martingale Limit Theory and its Applications*, Academic Press, New York.
- Hamilton, J. D.: 1994, *Time Series Econometrics*, Princeton University Press.

Time Series Econometrics, 2ST111

Lecture 6. Linear Regression Models
Covariance Stationary Vector Processes

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

Outline:

- Linear Regression Models
- Covariance Stationary Vector Processes (not 10.2-10.3)

The Content

- Repetition of econometrics content
- Introduce results we will rely on later: covariance matrices of estimators under heteroscedasticity, in AR models, etc
- Traditional assumptions not fulfilled, but what are the consequences?
- We focus on the linear regression model:

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + u_t \quad (1)$$

- Note that much of the head-scratching in practice comes from the fact that the data is fundamentally different from what we usually deal with in econometrics courses
- By varying the assumptions for \mathbf{x}_t and u_t , several different cases arise

The Basics of Linear Regression

- Given some data, (y_1, \dots, y_T) , the OLS estimator of the $k \times 1$ parameter vector β in (1) is the minimizer of the residual sum of squares:

$$RSS = \sum_{t=1}^T (y_t - \mathbf{x}'_t \beta)^2 \quad (2)$$

- We write the estimator for β as

$$\mathbf{b} = \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t y_t \right) \quad (3)$$

given that $\left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1}$ exists.

- The existence of the inverse is equivalent to a full rank assumption, which in turn requires $T > k$

The Basics of Linear Regression

- We may write (1) with matrix notation:

$$\begin{aligned}y_1 &= \mathbf{x}'_1 \boldsymbol{\beta} + u_1 \\y_2 &= \mathbf{x}'_2 \boldsymbol{\beta} + u_2 \\&\vdots \\y_T &= \mathbf{x}'_T \boldsymbol{\beta} + u_T\end{aligned}\tag{i}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_1 \boldsymbol{\beta} + u_1 \\ \mathbf{x}'_2 \boldsymbol{\beta} + u_2 \\ \vdots \\ \mathbf{x}'_T \boldsymbol{\beta} + u_T \end{pmatrix} \tag{ii}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_T \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{pmatrix} \tag{iii}$$

$$\underbrace{\mathbf{y}}_{T \times 1} = \underbrace{\mathbf{X}}_{T \times k} \underbrace{\boldsymbol{\beta}}_{k \times 1} + \underbrace{\mathbf{u}}_{T \times 1} \tag{iv}$$

The Basics of Linear Regression

- The OLS estimator can thus also be written as

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ &= \left((\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_T) \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_T' \end{pmatrix} \right)^{-1} (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_T) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} \\ &= \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t y_t \right)\end{aligned}$$

- The latter form is convenient from an asymptotic point of view, as we can divide by T in both the first term (i.e. the inverse) and the second term and using various LLN/CLT arguments

The Basics of Linear Regression

- The residual is $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b}$, which Hamilton refers to as the *sample* residual (as opposed to the *population* residual $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$)
- I will use residual to refer to the sample residual and error term to refer to the population residual
- The residuals are by construction orthogonal to \mathbf{X} :

$$\begin{aligned}\hat{\mathbf{u}}'\mathbf{X} &= (\mathbf{y}' - \mathbf{b}'\mathbf{X}')\mathbf{X} \\ &= (\mathbf{y}' - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} \\ &= \mathbf{y}'\mathbf{X} - \mathbf{y}'\mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{=\mathbf{I}_k} \\ &= \mathbf{0}\end{aligned}$$

- A useful representation of the OLS estimator is:

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{4}$$

Case 1: Classical Regression Assumptions

- The “classical” regression assumptions
- Assumptions:
 - a) \mathbf{x}_t is a vector of deterministic variables
 - b) u_t is i.i.d. with mean 0 and variance σ^2
 - c) u_t is Gaussian
- b) and c) is the same as saying u_t is iid $\sim N(0, \sigma^2)$

Case 1: Classical Regression Assumptions

- By using the useful form of \mathbf{b} from (4), the expectation is easily seen to be

$$\begin{aligned}E(\mathbf{b}) &= E\left(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right) \\&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}) \\&= \beta\end{aligned}$$

The variance-covariance matrix is

$$\begin{aligned}E[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'] &= E\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right)\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right)'\right] \\&= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_T\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Case 1: Classical Regression Assumptions

- Without the normality assumption (but assuming a) and b)), the Gauss-Markov theorem establishes optimality of \mathbf{b} within the class of unbiased and linear estimators
- If we also assume normality, \mathbf{b} is *exactly* Gaussian
- In general, if $\mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}')$
- So $\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- Important:** t and F tests are exact due to this

Case 2 and 3

- Case 2: a) \mathbf{x}_t stochastic and independent of u_s , all t, s ; b)
 $u_t \sim i.i.d. N(0, \sigma^2)$
 - The results are the same (except that $\mathbf{b}|\mathbf{X}$ is Gaussian and not \mathbf{b})
- Case 3: a) \mathbf{x}_t stochastic and independent of u_s for all t, s ; b) u_t non-Gaussian but i.i.d. with mean zero, variance σ^2 , and
 $E(u_t^4) = \mu_4 < \infty$
 - There are additional technical assumptions in c)-e), which assume \mathbf{x}_t to be ‘well-behaved’, allowing for the use of previous results for martingale difference sequences
 - Still unbiased, but only asymptotically Gaussian
 - t and F tests are inexact (asymptotically valid)

Case 4: Autoregression

Case 4: Stationary autoregression with independent errors

Assumption (8.4)

The regression model is

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t \quad (5)$$

with

- roots of $(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = 0$ outside of the unit circle,
- $\{\epsilon_t\}$ an i.i.d. sequence with $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = \sigma^2$, and $E(\epsilon_t^4) = \mu_4 < \infty$.

Case 4: Autoregression

- The autoregression can be written as a typical regression model:

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + u_t \quad (6)$$

where $\mathbf{x}'_t = (1 \ y_{t-1} \ \cdots \ y_{t-p})$ and $u_t = \epsilon_t$.

- However**, it cannot satisfy the independence assumption found in the previous cases:

Assumption (8.2(a), 8.3(a))

\mathbf{x}_t stochastic and independent of u_s for all t, s .

Case 4: Autoregression

- The assumption states that \mathbf{x}_t and u_{t-1} are independent. But:

$$\mathbf{x}_t = \begin{pmatrix} 1 \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{pmatrix}$$

$$E(\mathbf{x}_t u_{t-1}) = E \begin{pmatrix} u_{t-1} \\ y_{t-1} u_{t-1} \\ y_{t-2} u_{t-1} \\ \vdots \\ y_{t-p} u_{t-1} \end{pmatrix}$$

- Note that $E(y_{t-1} u_{t-1}) = E(u_{t-1}^2) = \sigma^2$
- Since $E(\mathbf{x}_t u_{t-1}) \neq \mathbf{0}$, \mathbf{x}_t and u_{t-1} are dependent and Assumption 8.2(a) and 8.3(a) cannot hold
- Thus, the previous cases considered are insufficient in the case of an autoregression

Case 4: Autoregression

- Consequence: \mathbf{b}_T is not unbiased

$$E(\mathbf{b}) = \beta + E \left[\left(T^{-1} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} T^{-1} \sum_{t=1}^T \mathbf{x}_t u_t \right] \quad (7)$$

- What about asymptotics? Consider the first part:

$$\begin{aligned} \sum_{t=1}^T \frac{\mathbf{x}_t \mathbf{x}'_t}{T} &= \sum_{t=1}^T \frac{1}{T} \begin{pmatrix} 1 \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{pmatrix} (1 \quad y_{t-1} \quad \cdots \quad y_{t-p}) \\ &= \begin{pmatrix} 1 & T^{-1} \sum y_{t-1} & \cdots & T^{-1} \sum y_{t-p} \\ T^{-1} \sum y_{t-1} & T^{-1} \sum y_{t-1}^2 & \cdots & T^{-1} \sum y_{t-1} y_{t-p} \\ \vdots & \vdots & \ddots & \vdots \\ T^{-1} \sum y_{t-p} & T^{-1} \sum y_{t-p} y_{t-1} & \cdots & T^{-1} \sum y_{t-p}^2 \end{pmatrix} \end{aligned}$$

Case 4: Autoregression

- From Prop 7.5, we know that $T^{-1} \sum_{t=1}^T y_{t-j} \xrightarrow{P} E(y_t) = \mu$ and that

$$T^{-1} \sum_{t=1}^T y_{t-i} y_{t-j} \xrightarrow{P} E(y_{t-i} y_{t-j}) = \gamma_{|i-j|} + \mu^2.$$

- This gives us:

$$\mathbf{Q} = \begin{pmatrix} 1 & \mu & \mu & \cdots & \mu \\ \mu & \gamma_0 + \mu^2 & \gamma_1 + \mu^2 & \cdots & \gamma_{p-1} + \mu^2 \\ \mu & \gamma_1 + \mu^2 & \gamma_0 + \mu^2 & \cdots & \gamma_{p-2} + \mu^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu & \gamma_{p-1} + \mu^2 & \gamma_{p-2} + \mu^2 & \cdots & \gamma_0 + \mu^2 \end{pmatrix}$$

- The second term is an MDS and separately the limits are:

$$\left(\frac{\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t'}{T} \right)^{-1} \xrightarrow{P} \mathbf{Q}^{-1}, \quad \left(\frac{\sum_{t=1}^T \mathbf{x}_t u_t}{\sqrt{T}} \right) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q})$$

Case 4: Autoregression

- Thus,

$$\mathbf{b}_T - \boldsymbol{\beta} \xrightarrow{P} \mathbf{0} \quad (8)$$

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}) \quad (9)$$

- In an autoregression, the OLS estimator is biased but consistent
- t and F tests are asymptotically valid
- Example: AR(1)

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

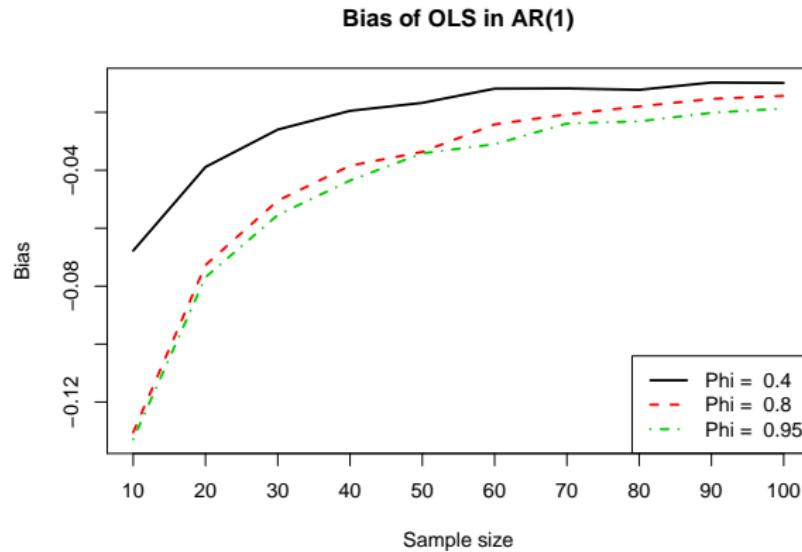
where the process is stationary ($|\phi| < 1$).

- The matrix \mathbf{Q} is just a scalar, $E(y_{t-1}^2) = \gamma_0 = \sigma^2 / (1 - \phi^2)$
- The asymptotic distribution result above thus implies that

$$\sqrt{T}(\hat{\phi}_T - \phi) \xrightarrow{d} N\left(0, \sigma^2 \frac{1 - \phi^2}{\sigma^2}\right) = N(0, 1 - \phi^2)$$

Case 4: Autoregression

- How large is the bias?
- Model: $y_t = \phi y_{t-1} + \epsilon_t$, with $T = 10, 20, \dots, 100$ and $\phi = 0.4, 0.8, 0.95$
- Bias: $\hat{\phi} - \phi$



Case 5: Errors Gaussian with Known Var-Cov Matrix

- Further relaxation of assumptions:
 - (a) x_t stochastic
 - (b) $u|X \sim N(\mathbf{0}, \sigma^2 V)$
 - (c) V is a known positive definite matrix
- Heteroskedasticity: V diagonal, but $V \neq \sigma^2 I_T$
- Autocorrelation: V non-diagonal
- OLS still unbiased:

$$E((\mathbf{b} - \boldsymbol{\beta})|X) = (X'X)^{-1}X'E(u|X) = \mathbf{0}$$

and since $E(Y) = E_X(E(Y|X))$:

$$E((\mathbf{b} - \boldsymbol{\beta})) = E_X\{E[(\mathbf{b} - \boldsymbol{\beta})|X]\} = E_X(\mathbf{0}) = \mathbf{0}$$

- For the estimator:
$$\mathbf{b}|X \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}X'VX(X'X)^{-1})$$
- OLS inefficient, use GLS (Section 8.3)

Case 6: Errors Uncorr but with General Heteroskedasticity

- Unknown and general heteroskedasticity:

Assumption (8.6)

a) \mathbf{x}_t stochastic, including perhaps lags of y

b) $\mathbf{x}_t u_t$ is an MDS

c) $E(u_t^2 \mathbf{x}_t \mathbf{x}_t') = \boldsymbol{\Omega}_t$ (positive definite) and

$$(i) \sum_{t=1}^T \frac{\Omega_t}{T} \rightarrow \boldsymbol{\Omega}$$

$$(ii) \sum_{t=1}^T \frac{u_t^2 \mathbf{x}_t \mathbf{x}_t'}{T} \xrightarrow{p} \boldsymbol{\Omega}$$

d)-e) \mathbf{x}_t and u_t well-behaved such that certain asymptotic results apply

- Example: let $\mathbf{x}_t = x_t$ and suppose that $E(x_t^2) = \mu_2$ and $E(x_t^4) = \mu_4$. Suppose the heteroskedasticity is of the form

$$E(u_t^2 | x_t) = a + bx_t^2$$

Case 6: Errors Uncorr but with General Heteroskedasticity

- In this case,

$$\begin{aligned}\Omega_t &= E(u_t^2 x_t^2) = E_x[E(u_t^2 | x_t^2)x_t^2] = E_x [(a + bx_t^2)x_t^2] \\ &= a\mu_2 + b\mu_4\end{aligned}$$

so $\Omega_t = \Omega$ for all t and (i) is satisfied.

- By the LLN we have (ii)

$$\sum_{t=1}^T \frac{u_t^2 x_t^2}{T} \xrightarrow{P} E(u_t^2 x_t^2) = \Omega$$

- Case 6 thus allows for fairly general types of conditional heteroskedasticity

Case 6: Errors Uncorr but with General Heteroskedasticity

- By Assumption 8.6 (and Proposition 7.9):

$$\left(\sum_{t=1}^T \frac{\mathbf{x}_t \mathbf{x}'_t}{T} \right)^{-1} \xrightarrow{p} \mathbf{Q}^{-1}$$
$$\sum_{t=1}^T \frac{\mathbf{x}_t u_t}{\sqrt{T}} \xrightarrow{d} N(\mathbf{0}, \Omega)$$

- Hence

$$\sqrt{T}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q}^{-1} \Omega \mathbf{Q}^{-1})$$

- White (1980) proposed: use $\hat{\mathbf{Q}}_T = T^{-1} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t$ and $\hat{\Omega}_T = T^{-1} \sum_{t=1}^T \hat{u}_t^2 \mathbf{x}_t \mathbf{x}'_t$
- Then (Proposition 8.3):

$$\hat{\mathbf{Q}}_T^{-1} \hat{\Omega}_T \hat{\mathbf{Q}}_T^{-1} \xrightarrow{p} \mathbf{Q}^{-1} \Omega \mathbf{Q}^{-1}$$

Case 6: Errors Uncorr but with General Heteroskedasticity

- We can treat \mathbf{b}_T as if

$$\mathbf{b}_T \approx N\left(\boldsymbol{\beta}, \frac{\hat{\mathbf{Q}}_T^{-1} \hat{\Omega}_T \hat{\mathbf{Q}}_T^{-1}}{T}\right)$$

- As it turns out, the expression for the variance is quite nice:

$$\frac{\hat{\mathbf{Q}}_T^{-1} \hat{\Omega}_T \hat{\mathbf{Q}}_T^{-1}}{T} = (\mathbf{X}'_T \mathbf{X}_T)^{-1} \left(\sum_{t=1}^T \hat{u}_t^2 \mathbf{x}_t \mathbf{x}'_t \right) (\mathbf{X}'_T \mathbf{X}_T)^{-1}$$

- Consistent, even when an unknown form of heteroskedasticity is present
- Also known as the sandwich estimator

Case 6: Errors Uncorr but with General Heteroskedasticity

- Note here that a model with autocorrelation may have it either in the variable itself, or in the error term:

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

$$\epsilon_t = u_t$$

$$u_t \sim iid(0, \sigma^2)$$

$$y_t = \epsilon_t$$

$$\phi(L)\epsilon_t = u_t$$

$$u_t \sim iid(0, \sigma^2)$$

- However, for zero-mean processes the two models are identical:

$$y_t = [\phi(L)]^{-1}u_t$$

$$u_t \sim iid(0, \sigma^2)$$

- With regressors in the equations as well, things get more complicated. You may find the details in last part of the chapter

Introduction to Vector Autoregressions

- Stochastic vector processes is a straight-forward generalization of univariate processes
- Most of the previous results are (principally) the same, only need to adapt them to a multivariate context
- Chapter 10 is quite technical, with more of a mathematical focus on vector time series
- Chapter 11 is also technical (it's still Hamilton), but focuses more on empirical issues and interpretations

Introduction to Vector Autoregressions

- Previously, we considered univariate stochastic processes such as an autoregression:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (10)$$

where we assumed

$$E(\varepsilon) = 0, \quad E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \sigma^2, & \text{for } t = \tau, \\ 0, & \text{otherwise.} \end{cases}$$

- The generalization to a vector process is made by replacing the scalar y_t by the $n \times 1$ vector

$$\mathbf{y}_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{n,t} \end{pmatrix}$$

Introduction to Vector Autoregressions

- The vector equivalent of (10) is thus:

$$\mathbf{y}_t = \mathbf{c} + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \cdots + \Phi_p \mathbf{y}_{t-p} + \varepsilon_t \quad (11)$$

where \mathbf{y}_t , \mathbf{c} and ε are all $n \times 1$ and each Φ_j is $n \times n$.

- Alternatively, the process can be formulated using a lag polynomial:

$$(\mathbf{I}_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p) \mathbf{y}_t = \mathbf{c} + \varepsilon_t$$

where

$$E(\varepsilon_t) = \mathbf{0},$$

$$E(\varepsilon_t \varepsilon'_\tau) = \begin{cases} \Omega, & \text{for } t = \tau, \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

Introduction to Vector Autoregressions

- Stationarity applies in the same way to vector processes
- Covariance stationary if $E(\mathbf{y}_t)$ and $E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-j} - \boldsymbol{\mu})']$ are independent of t
- Take expectations of (11):

$$E(\mathbf{y}_t) = \mathbf{c} + \Phi_1 E(\mathbf{y}_{t-1}) + \Phi_2 E(\mathbf{y}_{t-2}) + \cdots + \Phi_p E(\mathbf{y}_{t-p}) + E(\varepsilon_t)$$

- If \mathbf{y}_t is covariance stationary, $E(\mathbf{y}_t) = E(\mathbf{y}_{t-1}) = \cdots = \boldsymbol{\mu}$ for all t :

$$\boldsymbol{\mu} = \mathbf{c} + (\Phi_1 + \Phi_2 + \cdots + \Phi_p)\boldsymbol{\mu}$$

$$\boldsymbol{\mu} = (\mathbf{I} - \Phi_1 - \Phi_2 - \cdots - \Phi_p)^{-1}\mathbf{c}$$

- Thus, this is the obvious multivariate extension of the unconditional mean for an AR(p) process:

$$E(y_t) = (1 - \phi_1 - \phi_2 - \cdots - \phi_p)^{-1}c$$

$$E(\mathbf{y}_t) = (\mathbf{I} - \Phi_1 - \Phi_2 - \cdots - \Phi_p)^{-1}\mathbf{c}$$

Introduction to Vector Autoregressions

- It is sometimes useful to write the process in deviations from the mean. Note that from the previous slide:

$$\mathbf{c} = \boldsymbol{\mu} - \Phi_1\boldsymbol{\mu} - \Phi_2\boldsymbol{\mu} - \cdots - \Phi_p\boldsymbol{\mu}$$

- So (11) is

$$\mathbf{y}_t = \boldsymbol{\mu} - \Phi_1\boldsymbol{\mu} - \Phi_2\boldsymbol{\mu} - \cdots - \Phi_p\boldsymbol{\mu}$$

$$+ \Phi_1\mathbf{y}_{t-1} + \Phi_2\mathbf{y}_{t-2} + \cdots + \Phi_p\mathbf{y}_{t-p} + \varepsilon_t$$

$$(\mathbf{y}_t - \boldsymbol{\mu}) = \Phi_1(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \Phi_2(\mathbf{y}_{t-2} - \boldsymbol{\mu}) + \cdots + \Phi_p(\mathbf{y}_{t-p} - \boldsymbol{\mu}) + \varepsilon_t$$

Introduction to Vector Autoregressions

- Just as for the AR process, we can write this as a process of order one (companion form):

$$\begin{pmatrix} \mathbf{y}_t - \mu \\ \mathbf{y}_{t-1} - \mu \\ \vdots \\ \mathbf{y}_{t-p} - \mu \end{pmatrix} \vec{\mathbf{Y}}_t = \begin{pmatrix} \mathbf{y}_{1,t} \\ \mathbf{y}_{2,t} \end{pmatrix} = \begin{pmatrix} \phi_{11}' \mathbf{y}_{1,t-1} + \phi_{21}' \mathbf{y}_{2,t-1} \\ \phi_{12}' \mathbf{y}_{1,t-1} + \phi_{22}' \mathbf{y}_{2,t-1} \end{pmatrix} + \vec{\varepsilon}_t$$
$$= \phi \begin{pmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{2,t-1} \end{pmatrix}$$
$$= \begin{pmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-1} - \mu \\ \mathbf{y}_{t-2} - \mu \\ \vdots \\ \mathbf{y}_{t-p-1} - \mu \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

- Or in short:

$$\xi_t = \mathbf{F}\xi_{t-1} + \mathbf{v}_t$$

Introduction to Vector Autoregressions

- Just as for the AR process, the effects of previous shocks must eventually die out

$$\begin{aligned}\boldsymbol{\xi}_t &= \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t \\ &= \mathbf{F}^2\boldsymbol{\xi}_{t-2} + \mathbf{F}\mathbf{v}_{t-1} + \mathbf{v}_t \\ &= \mathbf{F}^3\boldsymbol{\xi}_{t-3} + \mathbf{F}^2\mathbf{v}_{t-2} + \mathbf{F}\mathbf{v}_{t-1} + \mathbf{v}_t \\ &\vdots \\ &= \mathbf{F}^s\boldsymbol{\xi}_{t-s} + \mathbf{F}^{s-1}\mathbf{v}_{t-s+1} + \cdots + \mathbf{F}\mathbf{v}_{t-1} + \mathbf{v}_t\end{aligned}\tag{12}$$

- By Proposition 10.1, the eigenvalues of \mathbf{F} satisfy

$$|\mathbf{I}_n\lambda^p - \Phi_1\lambda^{p-1} - \Phi_2\lambda^{p-2} - \cdots - \Phi_p| = 0$$

- Hence, covariance stationary if and only if all solutions λ satisfy $|\lambda| < 1$, i.e. *inside* the unit circle.

Introduction to Vector Autoregressions

- But, similarly, we have that

$$\begin{aligned} & |\mathbf{I}_n \lambda^p - \Phi_1 \lambda^{p-1} - \Phi_2 \lambda^{p-2} - \cdots - \Phi_p| \\ &= \lambda^{np} |\mathbf{I}_n - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p| \end{aligned}$$

where $z = \lambda^{-1}$.

- So, this is equivalent to

$$|\mathbf{I}_n - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p| = 0,$$

so the process is stationary if all solutions z satisfy $|z| > 1$, i.e. *outside* of the unit circle.

Introduction to Vector Autoregressions

- MA processes also exist in vector form, called VMA(q):

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \boldsymbol{\Theta}_1 \boldsymbol{\varepsilon}_{t-1} + \cdots + \boldsymbol{\Theta}_q \boldsymbol{\varepsilon}_{t-q}$$

- As in the univariate case, VMA(q) processes:

- are always stationary
- have $\boldsymbol{\Gamma}_j = \mathbf{0}$ if $|j| > q$

- Compare the autocovariances for q -th order processes:

$$\boldsymbol{\Gamma}_j = \begin{cases} \boldsymbol{\Theta}_j \boldsymbol{\Omega} + \boldsymbol{\Theta}_{j+1} \boldsymbol{\Omega} \boldsymbol{\Theta}'_1 + \cdots + \boldsymbol{\Theta}_q \boldsymbol{\Omega} \boldsymbol{\Theta}'_{q-j}, & j = 1, \dots, q \\ \boldsymbol{\Omega} \boldsymbol{\Theta}'_{-j} + \boldsymbol{\Theta}_1 \boldsymbol{\Omega} \boldsymbol{\Theta}'_{-j+1} + \cdots + \boldsymbol{\Theta}_{q-j} \boldsymbol{\Omega} \boldsymbol{\Theta}'_q, & j = -1, \dots, -q \\ \mathbf{0}, & |j| > q \end{cases}$$

$$\gamma_j = \begin{cases} (\theta_j + \theta_{j+1}\theta_1 + \cdots + \theta_q\theta_{q-j})\sigma^2, & j = 1, \dots, q \\ 0, & |j| > q \end{cases}$$

- For the VMA(∞) process, we change notation and write the model as:

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\varepsilon}_{t-2} + \cdots$$

Introduction to Vector Autoregressions

- If the process

$$(\mathbf{I}_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p)(\mathbf{y}_t - \boldsymbol{\mu}) = \boldsymbol{\varepsilon}_t$$

is stationary, it admits a moving average representation:

$$\begin{aligned}\mathbf{y}_t - \boldsymbol{\mu} &= \boldsymbol{\varepsilon}_t + \Psi_1 \boldsymbol{\varepsilon}_{t-1} + \Psi_2 \boldsymbol{\varepsilon}_{t-2} + \cdots \\ &= (\mathbf{I}_n + \Psi_1 L + \Psi_2 L^2 + \cdots) \boldsymbol{\varepsilon}_t\end{aligned}$$

- Hence,

$$(\mathbf{I}_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p)(\mathbf{I}_n + \Psi_1 L + \Psi_2 L^2 + \cdots) = \mathbf{I}_n$$

and restrictions on L^s coefficients yield

$$\Psi_1 - \Phi_1 = \mathbf{0}$$

$$\Psi_2 - \Phi_1 \Psi_1 - \Phi_2 = \mathbf{0}$$

$$\Psi_3 - \Phi_2 \Psi_1 - \Phi_1 \Psi_2 - \Phi_3 = \mathbf{0}$$

and so on.

Introduction to Vector Autoregressions

- From the previous slide:

$$\Psi_1 = \Phi_1$$

$$\Psi_2 = \Phi_1 \Psi_1 + \Phi_2$$

$$\Psi_3 = \Phi_2 \Psi_1 + \Phi_1 \Psi_2 + \Phi_3$$

- Two examples:

Example: VAR(1) ($p = 1$)

$$\Psi_1 = \Phi_1 = \Phi_1$$

$$\Psi_2 = \Phi_1 \Psi_1 = \Phi_1^2$$

$$\Psi_3 = \Phi_1 \Psi_2 = \Phi_1^3$$

Example: VAR(2) ($p = 2$)

$$\Psi_1 = \Phi_1 = \Phi_1$$

$$\Psi_2 = \Phi_1 \Psi_1 + \Phi_2 = \Phi_1^2 + \Phi_2$$

$$\Psi_3 = \Phi_1 \Psi_2 + \Phi_2 \Psi_1 = \Phi_1^3 + \Phi_1 \Phi_2 + \Phi_2 \Phi_1$$

Introduction to Vector Autoregressions

- For the VMA(q) we saw that the autocovariances were different for negative and positive j . Why is this?
- The autocovariance is

$$\Gamma_j = E [(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-j} - \boldsymbol{\mu})']$$

- In the univariate case $\gamma_j = \gamma_{-j}$, but $\Gamma_j \neq \Gamma_{-j}$.
- Instead, $\Gamma_j = \Gamma'_{-j}$
- The reason is that generally:

$$\text{Cov}(y_{k,t+j}, y_{l,t}) \neq \text{Cov}(y_{k,t}, y_{l,t+j})$$

- We have before noted that e.g. in an AR(1)

$$\text{Cov}(y_{t+1}, \epsilon_t) = \phi \neq \text{Cov}(y_t, \epsilon_{t+1}) = 0$$

Forecasting

- For forecasting, consider again (12) but for $t + s$:

$$\xi_{t+s} = \mathbf{F}^s \xi_t + \mathbf{F}^{s-1} \mathbf{v}_{t+1} + \cdots + \mathbf{F} \mathbf{v}_{t+s-1} + \mathbf{v}_{t+s} \quad (13)$$

- If we view $t + 1, t + 2, \dots, t + s$ as the future, then this expresses the future ξ_{t+s} as a function of past (known) $(\mathbf{y} - \boldsymbol{\mu})$ and future errors $\boldsymbol{\varepsilon}$
- If we want to forecast \mathbf{y}_{t+s} , then the first $n \times 1$ elements of ξ_{t+s} are what we want
- Recall equation [4.2.20] for Y_{t+s} :

$$Y_{t+s} - \mu = f_{11}^{(s)}(Y_t - \mu) + f_{12}^{(s)}(Y_{t-1} - \mu) + \cdots + f_{1p}^{(s)}(Y_{t-p+1} - \mu) \\ + \epsilon_{t+s} + \psi_1 \epsilon_{t+s-1} + \psi_2 \epsilon_{t+s-2} + \cdots + \psi_{s-1} \epsilon_{t+1}$$

Forecasting

- Difference between the \mathbf{F} matrices:

$$\mathbf{F}_{AR}^s = \underbrace{\begin{pmatrix} f_{11}^{(s)} & f_{12}^{(s)} & \dots & f_{1p}^{(s)} \\ f_{21}^{(s)} & f_{22}^{(s)} & \dots & f_{2p}^{(s)} \\ \vdots & \vdots & \ddots & \vdots \\ f_{p1}^{(s)} & f_{p2}^{(s)} & \dots & f_{pp}^{(s)} \end{pmatrix}}_{p \times p}, \quad \mathbf{F}_{VAR}^s = \underbrace{\begin{pmatrix} \mathbf{F}_{11}^{(s)} & \mathbf{F}_{12}^{(s)} & \dots & \mathbf{F}_{1p}^{(s)} \\ \mathbf{F}_{21}^{(s)} & \mathbf{F}_{22}^{(s)} & \dots & \mathbf{F}_{2p}^{(s)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{F}_{p1}^{(s)} & \mathbf{F}_{p2}^{(s)} & \dots & \mathbf{F}_{pp}^{(s)} \end{pmatrix}}_{np \times np}$$

- Hence, straight-forward to get an expression of only the first n rows of ξ_{t+s} :

$$\mathbf{y}_{t+s} - \boldsymbol{\mu} = \mathbf{F}_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + \mathbf{F}_{12}^{(s)}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \dots + \mathbf{F}_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}) \\ + \varepsilon_{t+s} + \Psi_1 \varepsilon_{t+s-1} + \Psi_2 \varepsilon_{t+s-2} + \dots + \Psi_{s-1} \varepsilon_{t+1} \quad (14)$$

- The forecast of \mathbf{y}_{t+s} is therefore:

$$\hat{\mathbf{y}}_{t+s|t} = \boldsymbol{\mu} + \mathbf{F}_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + \mathbf{F}_{12}^{(s)}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \dots + \mathbf{F}_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}) \quad (15)$$

$$H_0: \tilde{\mu} = 0$$

$$H_1: \tilde{\mu} \neq 0$$

$$T \cdot E(\bar{Y}_T - \mu)(\bar{Y}_T - \mu)' \rightarrow \sum \Gamma_i$$

$$T \cdot E(\bar{Y}_T - \mu)(\bar{Y}_T - \mu)' \approx \sum \Gamma_i$$

$$\Rightarrow \text{Var}(y) = E(I)(I) = \frac{1}{T} \sum \Gamma_i$$

$$\Rightarrow \bar{Y} - \mu \sim N(0, \frac{1}{T} \sum \Gamma_i)$$

$$\Rightarrow \frac{\bar{Y} - \tilde{\mu}}{\sqrt{\frac{1}{T} \sum \Gamma_i}} \sim N(0, 1)$$

$$\Rightarrow \left(\frac{\bar{Y} - \tilde{\mu}}{\sqrt{\frac{1}{T} \sum \Gamma_i}} \right)^2 \sim \chi^2$$

The Sample Mean of a Vector Process

- Proposition 10.5 says that for a covariance stationary process \mathbf{y}_t , with expectation μ and $E[(\mathbf{y}_t - \mu)(\mathbf{y}_{t-j} - \mu)'] = \Gamma_j$ absolutely summable, it follows that $\bar{\mathbf{y}}_T \xrightarrow{P} \mu$ and

$$\mathbf{S} = \lim_{T \rightarrow \infty} \left\{ T \cdot E[(\bar{\mathbf{y}}_T - \mu)(\bar{\mathbf{y}}_T - \mu)'] \right\} = \sum_{v=-\infty}^{\infty} \Gamma_v$$

- If we assume a VMA(q), $\Gamma_j = \mathbf{0}$ for all $|j| > q$. We can estimate Γ_v , where $v = 0, 1, \dots, q$, by

$$\hat{\Gamma}_v = T^{-1} \sum_{t=v+1}^T (\mathbf{y}_t - \bar{\mathbf{y}}_T)(\mathbf{y}_{t-v} - \bar{\mathbf{y}}_T)'$$

which is consistent as long as \mathbf{y}_t is ergodic for second moments.

- \mathbf{S} can then be consistently estimated by

$$\hat{\mathbf{S}} = \hat{\Gamma}_0 + \sum_{v=1}^q (\hat{\Gamma}_v + \hat{\Gamma}'_v)$$

The Sample Mean of a Vector Process

- $\hat{\mathbf{S}}$ is not guaranteed to be positive semi-definite
- An adjusted estimator was proposed by Newey and West, and it is known as the Newey and West estimator:

$$\tilde{\mathbf{S}} = \hat{\mathbf{r}}_0 + \sum_{v=1}^q \left(1 - \frac{v}{q+1}\right) (\hat{\mathbf{r}}_v + \hat{\mathbf{r}}'_v)$$

- Note: with the VMA(q), we say that the autocovariances are zero for $v > q$, but even if $E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_s - \boldsymbol{\mu})']$ is non-zero for all t and s (e.g. VAR), $\tilde{\mathbf{S}}$ will still be consistent if q , the threshold for non-zero autocovariances, is allowed to increase alongside T
- In particular, if $q \rightarrow \infty$ and $T \rightarrow \infty$ such that

$$\frac{q}{T^{1/4}} \rightarrow 0,$$

then $\tilde{\mathbf{S}} \xrightarrow{P} \mathbf{S}$.

Final remarks

Some things to note:

- For an AR, we have only one dimension for its size; p , the number of lags
- For a VAR, we write $\text{VAR}(p)$ and the cross-sectional dimension n is usually omitted
- Three types of possible error-term covariance:
 - Between equations, same time point
 - Within equations, over time
 - Between equations, over time

$$E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \Omega, & t = \tau \\ 0, & t \neq \tau \end{cases}$$

- With intercepts, n variables and p lags, the number of parameters to be estimated is $n(np + 1)$. Thus, an 8-variable VAR with four lags needs to estimate 264 parameters.



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 7. Vector Autoregressions

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Chapter 11: Vector Autoregressions (not 11.6-11.7)
- *Vector Autoregressions* by Stock and Watson (2001) (unless you're interested, you may skip the parts about structural models)

Some history



- Christopher A. Sims, Princeton University
- Awarded the Nobel Prize in Economics in 2011 together with Thomas J. Sargent "for their empirical research on cause and effect in the macroeconomy"
- "Macroeconomics and Reality" (1980, in *Econometrica*) is a seminal paper in the field
- Many more important contributions

ADL (Autoregressive distributed lag)

$$Y_t = c + \beta_1 Y_{t-1} + \dots$$

ML Estimation

- Suppose that the model is

$$\mathbf{y}_t = \mathbf{c} + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \cdots + \Phi_p \mathbf{y}_{t-p} + \varepsilon_t$$

where $\varepsilon_t \sim \text{i.i.d. } N(\mathbf{0}, \Omega)$.

- Assume we have a sample of length $T + p$, i.e.

$\mathbf{y}_{-p+1}, \dots, \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_T$. The simplest method of estimation is then to condition upon $(\mathbf{y}_{-p+1}, \dots, \mathbf{y}_0)$ and maximize the conditional likelihood

$$f_{\mathbf{Y}_T, \mathbf{Y}_{T-1}, \dots, \mathbf{Y}_1 | \mathbf{Y}_0, \dots, \mathbf{Y}_{-p+1}}(\mathbf{y}_T, \mathbf{y}_{T-1}, \dots, \mathbf{y}_1 | \mathbf{y}_0, \dots, \mathbf{y}_{-p+1}; \theta)$$

where θ contains all the unknowns $\mathbf{c}, \Phi_1, \dots, \Phi_p$ and Ω .

ML Estimation

- For notational convenience, let

$$\mathbf{x}_t = \begin{pmatrix} 1 \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{pmatrix}, \quad \boldsymbol{\Pi} = \begin{pmatrix} \mathbf{c}' \\ \boldsymbol{\Phi}'_1 \\ \boldsymbol{\Phi}'_2 \\ \vdots \\ \boldsymbol{\Phi}'_p \end{pmatrix} \implies \mathbf{y}_t = \boldsymbol{\Pi}' \mathbf{x}_t + \varepsilon_t$$

- It thus follows that

$$\mathbf{y}_t | \mathbf{y}_{t-1}, \dots, \mathbf{y}_{-p+1} \sim N(\boldsymbol{\Pi}' \mathbf{x}_t, \boldsymbol{\Omega})$$

and

$$f_t = (2\pi)^{-n/2} |\boldsymbol{\Omega}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t)' \boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t) \right\}$$

ML Estimation

- The joint density is the product of the individual conditional densities:

$$\begin{aligned} & f_{\mathbf{Y}_T, \mathbf{Y}_{T-1}, \dots, \mathbf{Y}_1 | \mathbf{Y}_0, \dots, \mathbf{Y}_{-p+1}}(\mathbf{y}_T, \mathbf{y}_{T-1}, \dots, \mathbf{y}_1 | \mathbf{y}_0, \dots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}) \\ &= \prod_{t=1}^T f_{\mathbf{Y}_t | \mathbf{Y}_{t-1}, \dots, \mathbf{Y}_{t-p}}(\mathbf{y}_t | \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p}; \boldsymbol{\theta}) \\ &= \prod_{t=1}^T f_t. \end{aligned}$$

- Eventually, we end up with the log likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = c + \frac{T}{2} \log(|\boldsymbol{\Omega}^{-1}|) - \frac{1}{2} \sum_{t=1}^T (\mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t)' \boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t)$$

ML Estimation

- To maximize, differentiate with respect to $\boldsymbol{\Pi}$, set to 0 and solve
- Useful derivative: For a symmetric matrix \mathbf{W} ,

$$\frac{\partial}{\partial \mathbf{A}} (\mathbf{x} - \mathbf{As})' \mathbf{W} (\mathbf{x} - \mathbf{As}) = -2\mathbf{W}(\mathbf{x} - \mathbf{As})\mathbf{s}'$$

- Hence,

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \boldsymbol{\Pi}'} &= \sum_{t=1}^T \boldsymbol{\Omega}^{-1} (\mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t) \mathbf{x}_t' = \mathbf{0} \\ &= \sum_{t=1}^T (\mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t) \mathbf{x}_t' = \mathbf{0}\end{aligned}$$

- The ML estimator is therefore the OLS estimator,

$$\hat{\boldsymbol{\Pi}} = \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{y}_t' \right)$$

ML Estimation

- Furthermore, maximum (conditional) likelihood estimation of a VAR is equivalent to equation-by-equation OLS
- By straight-forward matrix calculus, the ML estimator of the error covariance matrix can be shown to be

$$\hat{\Omega} = T^{-1} \sum_{t=1}^T \hat{\epsilon}_t \hat{\epsilon}'_t$$

where $\hat{\epsilon}_t = \mathbf{y}_t - \hat{\Pi}' \mathbf{x}_t$, i.e. evaluated at $\hat{\Pi}$.

Lag selection

- Likelihood ratio test for
 - H_0 : The model has p_0 lags
 - H_1 : The model has $p_1 > p_0$ lags
- The likelihood under H_0 ($i = 0$) and H_1 ($i = 1$)

$$\mathcal{L}(\hat{\Omega}_i, \hat{\Pi}_i) = c + \frac{T}{2} \log(|\hat{\Omega}_i^{-1}|) - \frac{1}{2} \sum_{t=1}^T \hat{\varepsilon}'_{t,i} \hat{\Omega}_i^{-1} \hat{\varepsilon}_{t,i}, \quad i = 0, 1$$

- The last term is for both $i = 0$ and $i = 1$ equal to $-Tn/2$. Put this into a new constant $c^* = c - Tn/2$. Thus, minus two times the log likelihood ratio is

$$\begin{aligned}\Lambda &= -2 \left(\mathcal{L}(\hat{\Omega}_0, \hat{\Pi}_0) - \mathcal{L}(\hat{\Omega}_1, \hat{\Pi}_1) \right) = \\ &= -2 \left(c^* + \frac{T}{2} \log(|\hat{\Omega}_0^{-1}|) - c^* - \frac{T}{2} \log(|\hat{\Omega}_1^{-1}|) \right) \\ &= T \left(\log(|\hat{\Omega}_1^{-1}|) - \log(|\hat{\Omega}_0^{-1}|) \right) \\ &= T \left(\log(|\hat{\Omega}_0|) - \log(|\hat{\Omega}_1|) \right)\end{aligned}$$

Lag selection

- How many restrictions are imposed under H_0 ?
 - Each equation has $p_1 - p_0$ fewer lags per variable, i.e. $n(p_1 - p_0)$ parameters are restricted to 0
 - n equations, so $n^2(p_1 - p_0)$ restrictions
- Under H_0 , $\Lambda \sim \chi^2(n^2(p_1 - p_0))$
- Finding the lag length using this procedure means sequential testing of hypotheses, quite complicated to control significance levels
- Sometimes prediction is the objective - the correct order of the model is uninteresting, a model suitable for forecasting is desired

Lag selection

- It is important to choose an appropriate lag length: too few will make the residuals correlated, too many make estimates imprecise and forecasts worse
- If forecasting is the objective, one can find the order which minimizes some forecast measure
- The usual model selection criteria are often used:

$$AIC(p) = \ln |\hat{\Omega}| + \frac{2}{T} n(np + 1)$$

$$BIC(p) = \ln |\hat{\Omega}| + \frac{\ln T}{T} n(np + 1)$$

$$HQ(p) = \ln |\hat{\Omega}| + \frac{2 \ln(\ln T)}{T} n(np + 1)$$

- It is common to use these criteria together with residual tests (e.g. for autocorrelation)

Lag selection

- Having selected the lag length, how do we summarize and present the results?
- There are often a huge number of parameters involved, so looking at the estimated coefficients individually is pointless
- Some main tools:
 - Impulse response analysis: summarizes the dynamics in the model
 - Granger causality: are certain variables important for the prediction of others?
 - Variance decomposition: how much of the unexplained variance in one variables can be traced back to unexplained shocks to other variables?

Impulse responses

- Impulse responses are often of great interest to researchers
- How do shocks transmit in the system?
- Consider a trivariate VAR and a shock in variable 1 at time $t = 0$
 - Because of the lag structure, the shock in variable 1 affects variables 1-3 at $t = 1$
 - Similarly, at $t = 2$, *all* variables are affected by *all* variables
- Example of a trivariate VAR:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \epsilon_{3,t} \end{pmatrix}$$

- So the shock affects i) y_1 directly, ii) y_2 at $t = 1$, and iii) y_3 at $t = 2$ (since the shock in y_1 goes through y_2 as there is no direct connection between y_1 and y_3)

Impulse responses

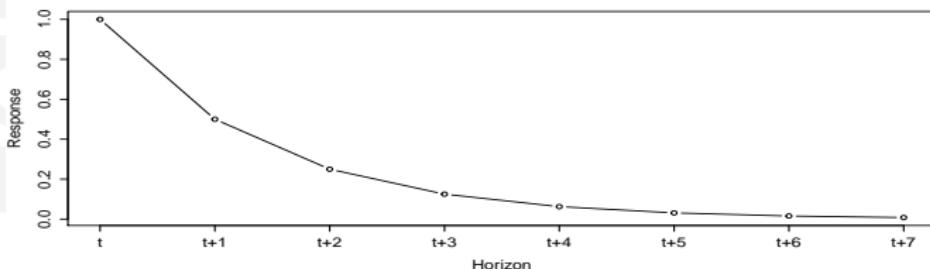
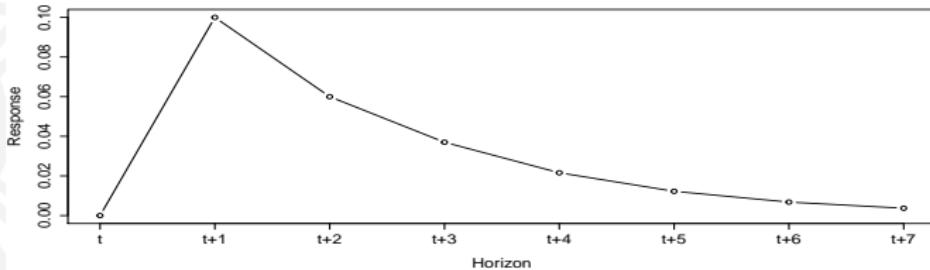
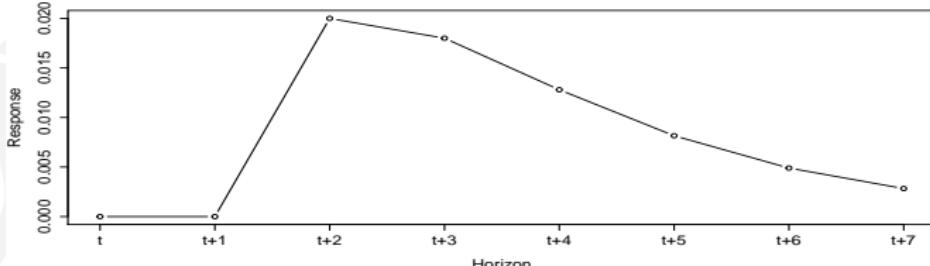
- The impulse response for a univariate process is $\frac{\partial y_{t+s}}{\partial \epsilon_t}$
- For a multivariate model, we might be interested in the effect on variable i at time $t + s$ of a shock to variable j at time t :

$$\frac{\partial y_{i,t+s}}{\partial \epsilon_{j,t}}$$

- Example: what is the effect on the inflation rate of a monetary policy shock?
- From the VMA(∞) form,

$$\frac{\partial \mathbf{y}_{t+s}}{\partial \epsilon'_t} = \frac{\partial}{\partial \epsilon'_t} (\boldsymbol{\mu} + \boldsymbol{\varepsilon}_{t+s} + \boldsymbol{\Psi}_1 \boldsymbol{\varepsilon}_{t+s-1} + \boldsymbol{\Psi}_2 \boldsymbol{\varepsilon}_{t+s-2} + \dots) = \boldsymbol{\Psi}_s$$

- Element (i,j) is $\frac{\partial y_{i,t+s}}{\partial \epsilon_{j,t}}$, so plotting this for $s = 0, 1, 2, \dots$ produces a plot of the impulse response function

Variable 1**Variable 2****Variable 3**

$$IRF: E(Y_{ith} | F_t, \varepsilon_{it}=6)$$

$$- E(Y_{i,2th} | F_t, \varepsilon_{it}=0)$$

Cholesky: $\Sigma = PP'$

$P = \Delta$ matrix

$$\varepsilon_t \sim (\omega, \Sigma)$$

$$u_t = P^{-1}\varepsilon_t \sim (0, I)$$

$$P^{-1}\varepsilon_t = \begin{bmatrix} P_{11} & 0 \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} = \begin{bmatrix} P_{11}\varepsilon_{t1} \\ \dots \end{bmatrix}$$

Impulse responses

- One serious problem: what is the meaning of this?
- Recall: $E(\varepsilon_t \varepsilon_t')$ = Ω , which is (usually) not a diagonal matrix
- Example for the trivariate VAR:

$$\Omega = \begin{pmatrix} 2.25 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 0.74 \end{pmatrix}$$

- If we again consider a shock in y_1 , this means that this shock is likely accompanied by a shock in y_2 as well
- The workaround is to orthogonalize the system

Impulse responses

- Cholesky decomposition: $\Omega = \mathbf{P}\mathbf{P}'$, and we let $\mathbf{v}_t = \mathbf{P}^{-1}\boldsymbol{\varepsilon}_t$:

$$E(\mathbf{v}_t\mathbf{v}_t') = \mathbf{P}^{-1}E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') (\mathbf{P}^{-1})' = \mathbf{P}^{-1}\mathbf{P}\mathbf{P}' (\mathbf{P}^{-1})' = \mathbf{I}$$

- VMA(∞) form again:

$$\begin{aligned}\mathbf{y}_t &= \mu + \sum_{s=0}^{\infty} \Psi_s \boldsymbol{\varepsilon}_{t-s} \\ &= \mu + \sum_{s=0}^{\infty} \Psi_s \mathbf{P}\mathbf{P}^{-1} \boldsymbol{\varepsilon}_{t-s} \\ &= \mu + \sum_{s=0}^{\infty} \Psi_s^* \mathbf{v}_{t-s}\end{aligned}$$

Impulse responses

- With orthogonal errors, the derivative $\frac{\partial y_{i,t+s}}{\partial v_{j,t}}$ makes sense as an isolated change
- However, new problems arise: the order of the variables matter because of the decomposition

$$\frac{\partial \mathbf{y}_{t+s}}{\partial v_{j,t}} = \boldsymbol{\Psi}_s \mathbf{p}_j$$

where \mathbf{p}_j is column j of \mathbf{P} , a lower triangular matrix.

- Example: three variables

$$\frac{\partial \mathbf{y}_{t+s}}{\partial v_{1,t}} = \boldsymbol{\Psi}_s \begin{pmatrix} p_{11} \\ p_{21} \\ p_{31} \end{pmatrix}, \quad \frac{\partial \mathbf{y}_{t+s}}{\partial v_{2,t}} = \boldsymbol{\Psi}_s \begin{pmatrix} 0 \\ p_{22} \\ p_{32} \end{pmatrix}, \quad \frac{\partial \mathbf{y}_{t+s}}{\partial v_{3,t}} = \boldsymbol{\Psi}_s \begin{pmatrix} 0 \\ 0 \\ p_{33} \end{pmatrix}$$

- Order matters, and it cannot be determined by statistical procedures but must be chosen

Impulse responses

- Consider a simple bivariate VAR(1):

$$\begin{aligned}y_t &= 0.5y_{t-1} + 0.2x_{t-1} + \epsilon_{y,t} \\x_t &= 0.3y_{t-1} - 0.1x_{t-1} + \epsilon_{x,t}\end{aligned}$$

which we write as

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} 0.5 & 0.2 \\ 0.3 & -0.1 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{y,t} \\ \epsilon_{x,t} \end{pmatrix}.$$

- With covariance

$$\Omega = \begin{pmatrix} 2 & 0.2 \\ 0.2 & 3 \end{pmatrix} = \overbrace{\begin{pmatrix} 1.41 & 0 \\ 0.14 & 1.73 \end{pmatrix}}^P \overbrace{\begin{pmatrix} 1.41 & 0.14 \\ 0 & 1.73 \end{pmatrix}}^{P'}$$

Impulse responses

- Thus, for an orthogonal shock in y :

$$\frac{\partial}{\partial v_{y,t}} \begin{pmatrix} y_{t+1} \\ x_{t+1} \end{pmatrix} = \Psi_1 \begin{pmatrix} p_{11} \\ p_{21} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.2 \\ 0.3 & -0.1 \end{pmatrix} \begin{pmatrix} 1.41 \\ 0.14 \end{pmatrix} = \begin{pmatrix} 0.75 \\ 0.27 \end{pmatrix}$$

- What if we instead had ordered x before y ?

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} -0.1 & 0.3 \\ 0.2 & 0.5 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{x,t} \\ \epsilon_{y,t} \end{pmatrix}$$

$$\Omega = \begin{pmatrix} 3 & 0.2 \\ 0.2 & 2 \end{pmatrix} = \overbrace{\begin{pmatrix} 1.73 & 0 \\ 0.12 & 1.41 \end{pmatrix}}^{\mathbf{P}_*} \overbrace{\begin{pmatrix} 1.73 & 0.12 \\ 0 & 1.41 \end{pmatrix}}^{\mathbf{P}'_*}$$

- Another orthogonal shock, still in y :

$$\frac{\partial}{\partial v_{y,t}} \begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \Psi_1 \begin{pmatrix} 0 \\ p_{22}^* \end{pmatrix} = \begin{pmatrix} -0.1 & 0.3 \\ 0.2 & 0.5 \end{pmatrix} \begin{pmatrix} 0 \\ 1.41 \end{pmatrix} = \begin{pmatrix} 0.28 \\ 0.70 \end{pmatrix}$$

Forecast error variance decomposition

- A closely related concept is forecast error variance decomposition
- How much of the variance of the forecast error of $y_{i,t+s}$ is due to an exogenous shock to $y_{j,t}$?
- Recall two of our previous expressions:

$$\mathbf{y}_{t+s} = \boldsymbol{\mu} + \mathbf{F}_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + \mathbf{F}_{12}^{(s)}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \cdots + \mathbf{F}_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}) \\ + \boldsymbol{\varepsilon}_{t+s} + \boldsymbol{\Psi}_1 \boldsymbol{\varepsilon}_{t+s-1} + \boldsymbol{\Psi}_2 \boldsymbol{\varepsilon}_{t+s-2} + \cdots + \boldsymbol{\Psi}_{s-1} \boldsymbol{\varepsilon}_{t+1}$$

$$\hat{\mathbf{y}}_{t+s} = \boldsymbol{\mu} + \mathbf{F}_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + \mathbf{F}_{12}^{(s)}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \cdots + \mathbf{F}_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu})$$

- The forecast error is therefore:

$$\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s} = \boldsymbol{\varepsilon}_{t+s} + \boldsymbol{\Psi}_1 \boldsymbol{\varepsilon}_{t+s-1} + \boldsymbol{\Psi}_2 \boldsymbol{\varepsilon}_{t+s-2} + \cdots + \boldsymbol{\Psi}_{s-1} \boldsymbol{\varepsilon}_{t+1}$$

- This means that the forecast error is due to exogenous innovations

Forecast error variance decomposition

- The MSE of the forecast is:

$$\begin{aligned}MSE(\hat{\mathbf{y}}_{t+s|t}) &= E[(\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t})(\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t})'] \\&= \Omega + \Psi_1 \Omega \Psi_1' + \Psi_2 \Omega \Psi_2' + \cdots + \Psi_{s-1} \Omega \Psi_{s-1}'\end{aligned}\quad (1)$$

since $E(\varepsilon_t \varepsilon_\tau') = \mathbf{0}$ if $t \neq \tau$

- Key idea:** how much does each of the *orthogonal* disturbances contribute to this MSE?
- To orthogonalize, let $\Omega = \mathbf{P} \mathbf{P}'$ and

$$\varepsilon_t = \mathbf{P} \mathbf{v}_t = \mathbf{p}_1 v_{1,t} + \mathbf{p}_2 v_{2,t} + \cdots + \mathbf{p}_n v_{n,t}$$

- The v_{it} and v_{jt} terms are orthogonal and have unit variance, so

$$\begin{aligned}\Omega &= E(\varepsilon_t \varepsilon_t') \\&= \mathbf{p}_1 \mathbf{p}_1' V(v_{1,t}) + \mathbf{p}_2 \mathbf{p}_2' V(v_{2,t}) + \cdots + \mathbf{p}_n \mathbf{p}_n' V(v_{n,t}) \\&= \mathbf{p}_1 \mathbf{p}_1' + \mathbf{p}_2 \mathbf{p}_2' + \cdots + \mathbf{p}_n \mathbf{p}_n'\end{aligned}\quad (2)$$

Forecast error variance decomposition

- Now: take the MSE expression in (1) and replace Ω with (2)

$$MSE(\hat{\mathbf{y}}_{t+s|t}) =$$

$$\sum_{j=1}^n (\mathbf{p}_j \mathbf{p}'_j + \boldsymbol{\Psi}_1 \mathbf{p}_j \mathbf{p}'_j \boldsymbol{\Psi}'_1 + \boldsymbol{\Psi}_2 \mathbf{p}_j \mathbf{p}'_j \boldsymbol{\Psi}'_2 + \cdots + \boldsymbol{\Psi}_{s-1} \mathbf{p}_j \mathbf{p}'_j \boldsymbol{\Psi}'_{s-1})$$

- Each j in the sum is the contribution by each variable to the MSE at horizon s
- Notation: call the term in brackets $\Xi_{j,s}$ and $\sum_{j=1}^n \Xi_{j,s} = \Xi_s$
- The proportion of forecast error variance of variable m attributable to variable j at horizon s is then

$$\frac{\Xi_{j,s}(m, m)}{\sum_{j=1}^n \Xi_{j,s}(m, m)} = \frac{\Xi_{j,s}(m, m)}{\Xi_s(m, m)}$$

- Numerator: the diagonal of $\Xi_{j,s}$ gives the contribution of variable j to MSE
- Denominator: the diagonal contains the variables' total MSEs

Granger causality

- Granger causality has little to do with causality; it is used to see if lags of one variable are useful in forecasting another
- A variable y Granger-causes x if lags of y improve forecasts of x
- More specifically, if the MSE of a forecast of x_{t+s} based on (x_t, x_{t-1}, \dots) is the same as a forecast based on both (x_t, x_{t-1}, \dots) and (y_t, y_{t-1}, \dots) , then y does *not* Granger-cause x
- In a VAR model, this is simply a joint test of certain coefficients being zero

Granger causality

- A bivariate VAR(p):

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \phi_{11}^{(1)} & \phi_{12}^{(1)} \\ \phi_{21}^{(1)} & \phi_{22}^{(1)} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \dots + \begin{pmatrix} \phi_{11}^{(p)} & \phi_{12}^{(p)} \\ \phi_{21}^{(p)} & \phi_{22}^{(p)} \end{pmatrix} \begin{pmatrix} x_{t-p} \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}$$

- The optimal forecast for x_{t+1} is:

$$\begin{aligned} \hat{E}(x_{t+1}|x_t, x_{t-1}, \dots, y_t, y_{t-1}, \dots) \\ = \phi_{11}^{(1)} x_t + \dots + \phi_{11}^{(p)} x_{t-p+1} + \phi_{12}^{(1)} y_t + \dots + \phi_{12}^{(p)} y_{t-p+1} \end{aligned}$$

- Thus, if $\phi_{12}^{(1)} = \dots = \phi_{12}^{(p)} = 0$, the forecast depends only on lagged values of x itself and we get:

$$\hat{E}(x_{t+1}|x_t, x_{t-1}, \dots, y_t, y_{t-1}, \dots) = \hat{E}(x_{t+1}|x_t, x_{t-1}, \dots)$$

Granger causality

- To test for Granger causality, we regress x_t on lags of x and y :

$$x_t = \sum_{i=1}^p \alpha_i x_{t-i} + \sum_{i=1}^p \beta_i y_{t-i} + u_t$$

- Conduct an F -test with the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Extensions

There are many extensions

- Threshold VARs
- Smooth Transition VARs
- Markov-switching VARs
- Time-varying parameters VARs
- VARs with stochastic volatility
- Factor-augmented VARs
- etcetera...

Stock and Watson (2001)

- Stock and Watson: two of the leading macroeconometricians in the world
- This paper discusses how well VAR models handle what they're frequently used to do
 - Data description
 - Forecasting
 - (Structural inference)
 - (Policy analysis)

- VAR models come in one of three forms: reduced, recursive or structural
- **Reduced** form is what we have discussed so far, where each variable is a (linear) function of past values of itself and the other variables
- **Recursive** form we used when we had orthogonalized impulse responses, since adding contemporaneous lags is equivalent to doing a Cholesky decomposition
- **Structural** VARs are based on economic theory and make use of identifying assumptions therein
- Their data ranges from 1960Q1-2000Q4 and includes π (inflation rate), u (unemployment rate) and R (the federal funds rate, i.e. an interest rate)

Recursive:

$$P^{-1} \begin{pmatrix} Y_{1t} \\ Y_{2t} \end{pmatrix}, P^{-1} \varepsilon_t = V_t \sim (0, I)$$

$$\begin{pmatrix} P_{11} & 0 \\ P_{21} & P_{22} \end{pmatrix} \begin{pmatrix} Y_{1t} \\ Y_{2t} \end{pmatrix} = (\cdot)$$

$$P_{11} Y_{1t} = \dots$$

$$Y_{1t} = \frac{1}{P_{11}} (\dots)$$

$$P_{22} Y_{2t} = P_{21} Y_{1t} + \dots$$

$$= \left(\frac{P_{21}}{P_{11}} \right) (\dots)$$

Data description: Granger-Causality

- Granger-Causality Test: what variables help predict others?
 - H_0 : the regressor does not Granger-cause the dependent variable
-

A. Granger-Causality Tests

Dependent Variable in Regression

Regressors	π	u	R
π	0.00	0.31	0.00
u	0.02	0.00	0.00
R	0.27	0.01	0.00

Figure: p -values of Granger-causality tests (Table 1, Panel A)

Data description: Variance decomposition

- The (forecast error) variance decomposition tells us the percentage of the error in forecasting a variable (e.g. inflation) that is due to specific shocks in another variable (such as unemployment) at a specific horizon (like 4 quarters)

B.i. Variance Decomposition of π

Forecast Horizon	Forecast Standard Error	Variance Decomposition (Percentage Points)		
		π	u	R
1	0.96	100	0	0
4	1.34	88	10	2
8	1.75	82	17	1
12	1.97	82	16	2

Figure: Variance decomposition (Table 1, Panel B.i)

Data description: Variance decomposition

B.i. Variance Decomposition of π

Forecast Horizon	Forecast Standard Error	Variance Decomposition (Percentage Points)		
		π	u	R
1	$\sqrt{\Xi_1(1, 1)}$	0.96	100	0
4	$\sqrt{\Xi_4(1, 1)}$	1.34	88	10
8	$\sqrt{\Xi_8(1, 1)}$	1.75	82	17
12	$\sqrt{\Xi_{12}(1, 1)}$	1.97	82	16

- The numbers in the u column are given by

$$\frac{\Xi_{2,1}(1, 1)}{\Xi_1(1, 1)} = 0, \quad \frac{\Xi_{2,4}(1, 1)}{\Xi_4(1, 1)} = 0.10 \quad \begin{array}{l} \text{■ Error for variable 1} \\ \text{■ at horizon 1, 4, 8, 12} \\ \text{■ due to variable 2} \end{array}$$
$$\frac{\Xi_{2,8}(1, 1)}{\Xi_8(1, 1)} = 0.17, \quad \frac{\Xi_{2,12}(1, 1)}{\Xi_{12}(1, 1)} = 0.16$$

Data description: Variance decomposition

B.ii. Variance Decomposition of u

Forecast Horizon	Forecast Standard Error	Variance Decomposition (Percentage Points)			
		π	u	R	
1	$\sqrt{\Xi_1(2, 2)}$	0.23	1	99	0
4	$\sqrt{\Xi_4(2, 2)}$	0.64	0	98	2
8	$\sqrt{\Xi_8(2, 2)}$	0.79	7	82	11
12	$\sqrt{\Xi_{12}(2, 2)}$	0.92	16	66	18

- The numbers in the u column are given by

$$\frac{\Xi_{2,1}(2, 2)}{\Xi_1(2, 2)} = 0.99, \quad \frac{\Xi_{2,4}(2, 2)}{\Xi_4(2, 2)} = 0.98$$

$$\frac{\Xi_{2,8}(2, 2)}{\Xi_8(2, 2)} = 0.82, \quad \frac{\Xi_{2,12}(2, 2)}{\Xi_{12}(2, 2)} = 0.66$$

Data description: Variance decomposition

B.iii. Variance Decomposition of R

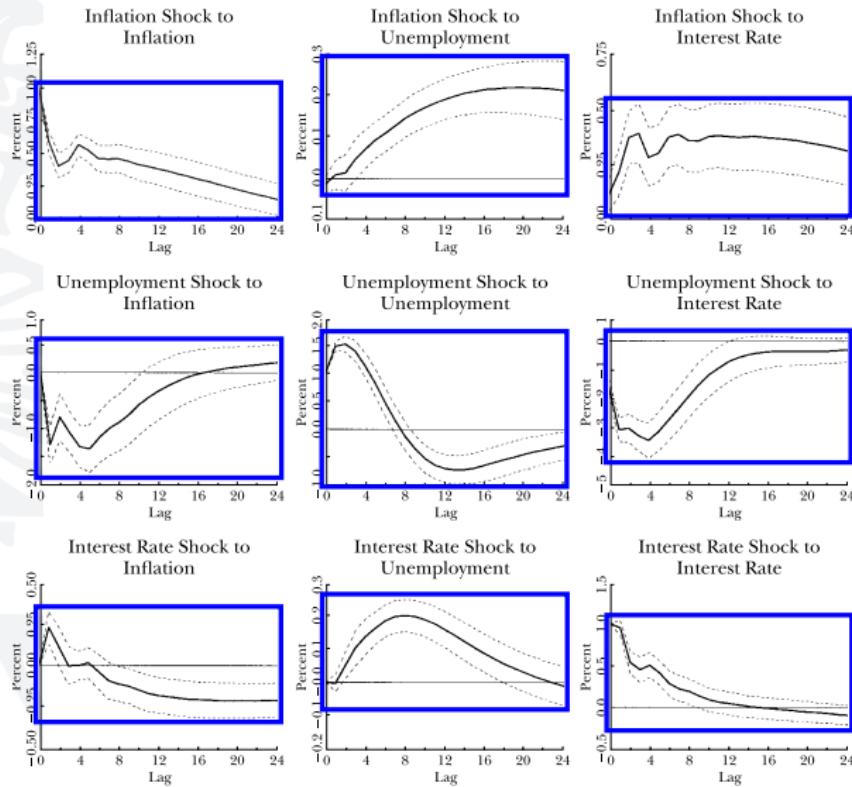
Forecast Horizon	Forecast Standard Error	Variance Decomposition (Percentage Points)		
		π	u	R
1	$\sqrt{\Xi_1(3, 3)}$	0.85	2	19
4	$\sqrt{\Xi_4(3, 3)}$	1.84	9	50
8	$\sqrt{\Xi_8(3, 3)}$	2.44	12	60
12	$\sqrt{\Xi_{12}(3, 3)}$	2.63	16	59

- The numbers in the u column are given by

$$\frac{\Xi_{2,1}(3, 3)}{\Xi_1(3, 3)} = 0.19, \quad \frac{\Xi_{2,4}(3, 3)}{\Xi_4(3, 3)} = 0.50$$

$$\frac{\Xi_{2,8}(3, 3)}{\Xi_8(3, 3)} = 0.60, \quad \frac{\Xi_{2,12}(3, 3)}{\Xi_{12}(3, 3)} = 0.59$$

Data description: Impulse responses



Data description: Forecasting

- VARs are often used for forecasting
- Pseudo out-of-sample forecasting exercise for the period 1985Q1-2000Q4 using a rolling forecast window:
 - Estimate model on data through 1984Q4 and predict h steps ahead
 - Add one more data point: estimate model on data through 1985Q1 and predict h steps ahead
 - Continue until the sample ends, repeat for $h = 2, 4, 8$
- Call the forecasts $\hat{\pi}_t^{(h)}$, $\hat{u}_t^{(h)}$ and $\hat{R}_t^{(h)}$
- Stock and Watson evaluate the forecasts using the standard measure RMSE:

$$RMSE_h(\pi) = \sqrt{\sum_{t=1984Q4+h}^{2000Q4} \frac{(\pi_t - \hat{\pi}_t^{(h)})^2}{\text{\# of forecasts}}}$$

Data description: Forecasting

- It is common practice to include AR(1) and random walks as benchmark models

Forecast Horizon	Inflation Rate			Unemployment Rate			Interest Rate		
	RW	AR	VAR	RW	AR	VAR	RW	AR	VAR
2 quarters	0.82	0.70	0.68	0.34	0.28	0.29	0.79	0.77	0.68
4 quarters	0.73	0.65	0.63	0.62	0.52	0.53	1.36	1.25	1.07
8 quarters	0.75	0.75	0.75	1.12	0.95	0.78	2.18	1.92	1.70

Figure: RMSE of pseudo out-of-sample forecasts

- It is often quite difficult to beat simple AR models, but here the VAR is most of the time slightly better

Conclusions

- VAR models have been very useful tools for macroeconometricians for almost four decades
- There are limitations, but competing models are usually much more complicated for little or no gain
- Much of the recent research is focused on fixing its limitations: dealing with overparametrization, allowing for nonlinearities in various ways, using larger data sets



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 8. Nonstationarity & Deterministic Trends

Yukai Yang

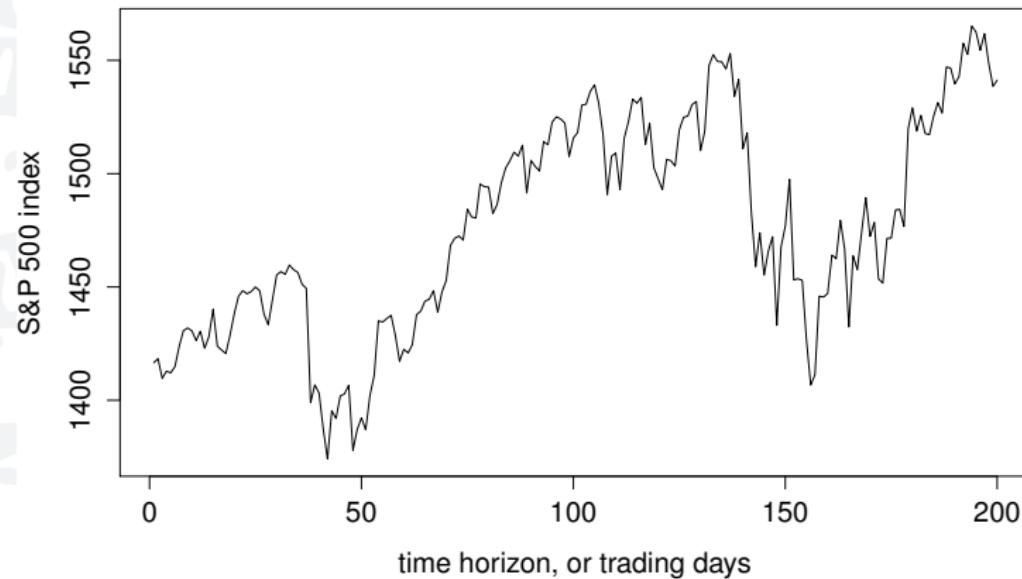
Department of Statistics, Uppsala University

Outline of Today's Lecture

- Models of Nonstationary Time Series (pp.435-453 in Hamilton)
 - Deterministic Time Trend & Unit Root Approaches
 - Unit Root Process
 - ARIMA(p, d, q) Process
 - Linear vs. Exponential Time Trends
 - Comparison of Trend-Stationary & Unit Root Processes
- Processes with Deterministic Time Trends (pp.454-474 in Hamilton)
 - Asymptotic Results for OLS Estimators for the Simple Trend-Stationary Process
 - Order in Probability
 - Asymptotic Results for OLS Estimators for the Trend-Stationary AR(p) Process

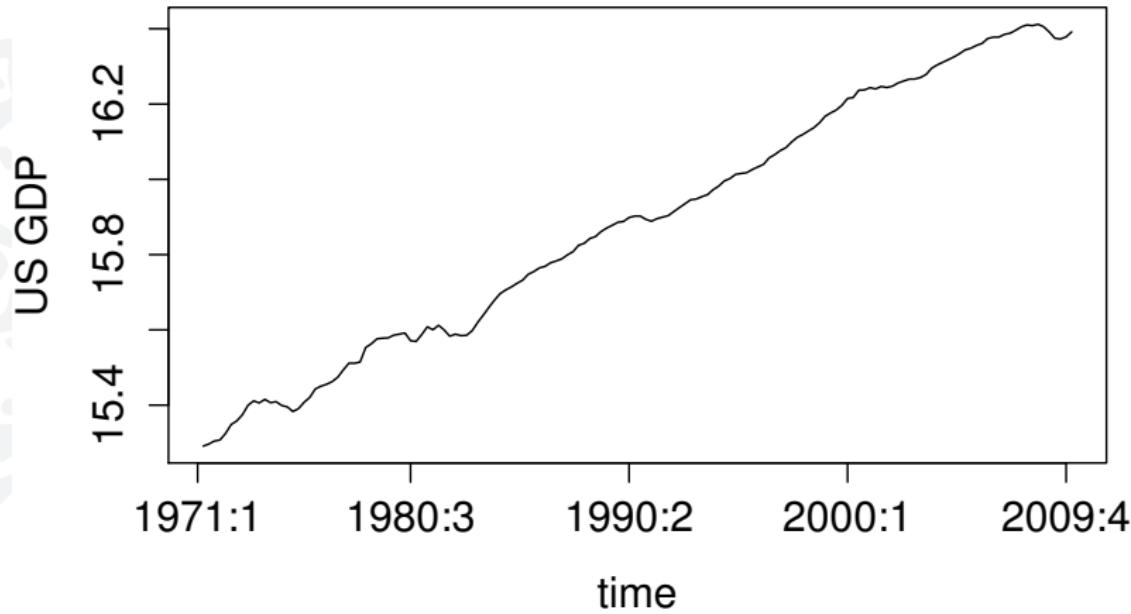
Why Nonstationarity instead of Stationary ARMA?

S&P daily closing price indices (time series plot), from 1 Jan to 17 Oct in 2007



Why Nonstationarity instead of Stationary ARMA?

Monthly log US GDP from Jan 1971 to April 2009



Models for Nonstationary Time Series

We consider two different approaches for modeling nonstationary time series

- 1 A **deterministic time trend** approach

$$y_t = \alpha + \delta t + \psi(L)\varepsilon_t, \quad (1)$$

where $\alpha, \delta \in \mathbb{R}$ and $\alpha + \delta t$ is a deterministic time trend.

- 2 A **unit root** approach:

$$\Delta y_t = \delta + \psi(L)\varepsilon_t, \quad (2)$$

where $\Delta = 1 - L$ and $\psi(1) \neq 0$.

From now on, the small letter y_t will be used for both random variables and the observations.

Models for Nonstationary Time Series

Remarks

- The stochastic process given by (1) is sometimes said to be **trend-stationary**, because if one subtracts the trend δt from it, the result is a stationary process.
- The condition that $\psi(1) \neq 0$ for the unit root process (2) ensures that y_t is nonstationary. $\Sigma_i \text{iid } (0, \sigma^2)$
- The prototypical example of a unit root process (2) is obtained when $\psi(L) = 1$

$$y_t = y_{t-1} + \delta + \varepsilon_t, \quad (3)$$

which is known as a **random walk** with drift δ .

- There are several other approaches. For example, fractionally integrated processes and processes with occasional discrete shifts in trend. See pp.447-451 in Hamilton.

Unit Root Process

To see that the condition $\psi(1) \neq 0$ ensures that y_t is nonstationary, suppose that y_t is stationary with $MA(\infty)$ representation

$$y_t = \mu + \chi(L)\varepsilon_t. \quad (4)$$

By taking the first-order difference, we have

$$(1 - L)y_t = \underbrace{(1 - L)\mu}_{=0} + \underbrace{(1 - L)\chi(L)}_{=\psi(L)}\varepsilon_t, \quad (5)$$

where $\psi(1) = (1 - 1)\chi(1) = 0$.

Claim: Let $y_t = \mu + X(L) \varepsilon_t$.

Then $\psi(1) \neq 0 \Rightarrow Y_t$ not stationary.

Proof: Can show Y_t stationary $\Rightarrow \psi(1) = 0$.

Let $(1-L)Y_t$

Unit Root Process

It is sometimes convenient to work with a slightly different representation of the unit root process in (2). Let

$$y_t = \alpha + \delta t + u_t, \quad (6)$$

where u_t is a zero-mean ARMA(p, q) process

$$\phi(L)u_t = \theta(L)\varepsilon_t. \quad (7)$$

Assume that ε_t is white noise, the MA lag polynomial $\theta(L)$ is invertible, and the AR lag polynomial is stable.

Unit Root Process

If the lag polynomial in the AR part is stable, (7) can be written as

$$u_t = \frac{1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{(1 - \lambda_1 L)(1 - \lambda_2 L)\dots(1 - \lambda_p L)} \varepsilon_t = \psi(L) \varepsilon_t \quad (8)$$

where $\sum_{i=0}^{\infty} |\psi_i| < \infty$. It is exactly the form in (1) (trend stationary)!

Unit Root Process

Now suppose that one $\lambda_i = 1$ and $|\lambda_j| < 1$ for $j \neq i$. Without loss of generality, let $\lambda_1 = 1$

$$(1 - L)u_t = \frac{1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{(1 - \lambda_2 L)(1 - \lambda_3 L)\dots(1 - \lambda_p L)} \varepsilon_t = \psi^*(L) \varepsilon_t \quad (9)$$

with $\sum_{i=0}^{\infty} |\psi_i^*| < \infty$.

By taking the first-order difference of y_t , we obtain

$$(1 - L)y_t = \underbrace{(1 - L)\alpha}_{=0} + \underbrace{(1 - L)\delta t}_{=\delta} + (1 - L)u_t = \delta + \psi^*(L) \varepsilon_t, \quad (10)$$

which is exactly the unit root process in (2).

Unit Root Process

$$u_t = L(1) \Rightarrow \Delta u_t = L(0)$$
$$\varepsilon_t = L(0) \Rightarrow \sum \varepsilon_t = L(1)$$

Remarks:

- (6) together with (9) and (10) explain why (2) is called a **unit root process**. One of the roots of the lag polynomial in the AR part of u_t equals one, and all other roots lie outside the unit disk.
- The unit root process (2) with **only one** unit root is also called **integrated of order 1**, or simply **I(1)**.
- $\psi^*(L)\varepsilon_t$ in (9) and (10) with $\psi^*(1) \neq 0$ is called **I(0)** process.
- If, unfortunately, **two roots** equal one, and the other roots lie outside the unit disk, then y_t has to be **differenced twice** to reach I(0).

$$\Delta^2 y_t = \kappa + \psi^*(L)\varepsilon_t. \quad (11)$$

y_t in this case is called **I(2)**.

- Think about in which case $\kappa \neq 0$ (quadratic trend).

ARIMA(p, d, q)

A general stochastic process is called an **autoregressive integrated moving average process**, or simply an **ARIMA(p, d, q)** process. It takes the form as follows

$$\begin{aligned} y_t &= \alpha + \delta t + u_t &\sim I(d) \\ \Delta^d \phi(L) u_t &= \theta(L) \varepsilon_t &\sim I(0) \end{aligned} \tag{12}$$

with $\phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$ stable, and $\theta(L) = 1 + \theta_1 L + \dots + \theta_q L^q$ invertible. We only consider the case when the integration order d is a **non-negative integer**, or it's gonna be fractionally integrated.

Taking d th difference produces a stationary ARMA(p, q) process $\Delta^d y_t$.

If $d = 0$ but $\delta \neq 0$, it is trend stationary.

ARIMA(p, d, q)

The ARIMA(p, d, q) can also be rewritten in the following form:

- First, take the d th difference on y_t process

$$\Delta^d y_t = \Delta^d \alpha + \Delta^d \delta t + \Delta^d u_t.$$

- Since $\Delta^d \phi(L)u_t = \theta(L)\varepsilon_t$, then $\Delta^d u_t = \phi(L)^{-1}\theta(L)\varepsilon_t = \psi(L)\varepsilon_t$

$$\Delta^d y_t = \Delta^d \alpha + \Delta^d \delta t + \psi(L)\varepsilon_t.$$

- Multiply both sides by $\phi(L)$

$$\Delta^d \phi(L)y_t = \Delta^d \phi(1)\alpha + \Delta^d \phi(L)\delta t + \theta(L)\varepsilon_t.$$

Note that $\phi(L)\delta t = \gamma t + \eta$ and $\phi(1)\alpha = \zeta$. Then

$$\Delta^d \phi(L)y_t = \Delta^d \tilde{\alpha} + \Delta^d \tilde{\delta}t + \theta(L)\varepsilon_t$$

where $\tilde{\alpha} = \eta + \zeta$ and $\tilde{\delta} = \gamma$. This ARIMA form is more often employed.

Linear vs. Exponential Time Trends

- In practice, many economic time series exhibit an exponential trend rather than a linear trend
- Because of this, it is common to take logs of economic time series before attempting to model them with the trend-stationary or unit root process, respectively.
- Note that

$$\begin{aligned}\Delta \log y_t &= \log y_t - \log y_{t-1} = \log \frac{y_t}{y_{t-1}} \\ &= \log \frac{y_{t-1} + y_t - y_{t-1}}{y_{t-1}} = \log \left(1 + \frac{y_t - y_{t-1}}{y_{t-1}} \right) \\ &\approx \frac{y_t - y_{t-1}}{y_{t-1}} \quad \text{provided that } \left| \frac{y_t - y_{t-1}}{y_{t-1}} \right| \text{ very small.}\end{aligned}$$

This is referred to as the growth rate in discrete time.

Linear vs. Exponential Time Trends

- In finance, the continuous compound growth rate is widely used.
- Suppose that the value P goes to F after one year. The annual growth rate R_y is computed from the identity

$$F = P(1 + R_y).$$

- If the corresponding interest is compounded every $1/n$ period, then the compound annual growth rate R_n is computed from

$$F = P \left(1 + \frac{R_n}{n}\right)^n.$$

- Suppose that the compound time period can be infinitely small, then the continuous compound annual growth rate r is obtained from

$$\lim_{n \rightarrow \infty} \left(1 + \frac{r}{n}\right)^n = \exp r = \frac{F}{P}, \quad \text{or} \quad r = \log F - \log P.$$

Linear vs. Exponential Time Trends

- The continuous compound growth rate has the very nice feature that

$$\exp(r(t_1 - t_0)) \exp(r(t_2 - t_1)) = \exp(r(t_2 - t_0))$$

where t_0, t_1, t_2 are time points.

- If the growth rate r_t is a time-varying from time point t_0 to point t_1 , the discrete growth rate for that period is

$$R = \exp \int_{t_0}^{t_1} r_t dt,$$

- It is reasonable to assume that some growth rates are $I(0)$.

Comparison of Trend-Stationary & Unit Root Processes

Let us compare the forecasts of a trend-stationary and unit root process.

To forecast a trend-stationary process, the known deterministic component

$$\alpha + \delta t$$

is simply added to the forecast of the stationary component

$$\psi(L)\varepsilon_t$$

Comparison of Trend-Stationary & Unit Root Processes

Hence, for the trend-stationary process

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

the *s*-step ahead forecast of y_{t+s} at time t is

$$\begin{aligned}\hat{y}_{t+s|t} &= \hat{E}(y_{t+s}|y_t, y_{t-1}, \dots) \\ &= \alpha + \delta(t+s) + \psi_s \varepsilon_t + \psi_{s+1} \varepsilon_{t-1} + \psi_{s+2} \varepsilon_{t-2} + \dots\end{aligned}\quad (13)$$

The absolute summability of $\{\psi_i\}_{i=0}^{\infty}$ implies that this *s*-step ahead forecast converges in mean square to the **time trend**. That is

$$\lim_{s \rightarrow \infty} E[\hat{y}_{t+s|t} - \alpha - \delta(t+s)]^2 = 0$$

Comparison of Trend-Stationary & Unit Root Processes

By contrast, it can be shown that the s -step-ahead forecast of y_{t+s} at time t for the unit process

$$\Delta y_t = \delta + \psi(L)\varepsilon_t$$

is given by

$$\hat{y}_{t+s|t} = s\delta + y_t + (\psi_s + \psi_{s-1} + \dots + \psi_1)\varepsilon_t + (\psi_{s+1} + \psi_s + \dots + \psi_2)\varepsilon_{t-1} + \dots \quad (14)$$

In particular, if $\psi(L)\varepsilon_t = \varepsilon_t$, then

$$\hat{y}_{t+s|t} = s\delta + y_t \quad (15)$$

Comparison of Trend-Stationary & Unit Root Processes

It can be shown that the mean squared error (MSE)

$$\text{MSE} = E(y_{t+s} - \hat{y}_{t+s|t})^2 \quad (16)$$

for the trend-stationary process **converges to a constant** as $s \rightarrow \infty$.

By contrast, the MSE for the unit root process **diverges** as $s \rightarrow \infty$.

Comparison of Trend-Stationary & Unit Root Processes

It can also be shown that, for the trend-stationary process, the dynamic multiplier is $\partial y_{t+s}/\partial \varepsilon_t = \psi_s$, and hence

$$\lim_{s \rightarrow \infty} \frac{\partial y_{t+s}}{\partial \varepsilon_t} = 0. \quad (17)$$

By contrast, for the unit root process, $\partial y_{t+s}/\partial \varepsilon_t = \sum_{i=0}^s \psi_i$, and hence

$$\lim_{s \rightarrow \infty} \frac{\partial y_{t+s}}{\partial \varepsilon_t} = \sum_{i=0}^{\infty} \psi_i. \quad (18)$$

Thus, the effect of (or a shock occurring to) ε_t dies out eventually in trend stationary process, but is permanent in unit root process.

Trend-Stationary Process

The trend-stationary process

$$y_t = \alpha + \delta t + \psi(L)\varepsilon_t \quad (19)$$

with $\sum_{i=0}^{\infty} |\psi_i| < \infty$ and ε_t is white noise.

Consider the special case when $\psi(L) = 1$, simply

$$y_t = \alpha + \delta t + \varepsilon_t \quad (20)$$

where α and β are unknown. Alternatively, we write it as

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t \quad (21)$$

where $\mathbf{x}_t = (1, t)'$ and $\boldsymbol{\beta} = (\alpha, \delta)'$.

OLS for the Simple Trend-Stationary Process

Denote $\hat{\beta}_T$ the OLS estimator for the parameter vector β , given the sample y_1, \dots, y_T of size T

$$\hat{\beta}_T = \begin{pmatrix} \hat{\alpha}_T \\ \hat{\delta}_T \end{pmatrix} = \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t y_t \right). \quad (22)$$

It can be readily shown that

$$\hat{\beta}_T = \beta + \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right). \quad (23)$$

OLS for the Simple Trend-Stationary Process

Hence

$$\hat{\beta}_T - \beta = \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right),$$

or equivalently

$$\begin{aligned} \begin{pmatrix} \hat{\alpha}_T - \alpha \\ \hat{\delta}_T - \delta \end{pmatrix} &= \left[\sum_{t=1}^T \begin{pmatrix} 1 & t \\ t & t^2 \end{pmatrix} \right]^{-1} \left[\sum_{t=1}^T \begin{pmatrix} \varepsilon_t \\ t\varepsilon_t \end{pmatrix} \right] \\ &= \begin{pmatrix} \sum 1 & \sum t \\ \sum t & \sum t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum \varepsilon_t \\ \sum t\varepsilon_t \end{pmatrix}, \end{aligned}$$

where \sum denotes $\sum_{t=1}^T$.

OLS for the Simple Trend-Stationary Process

In order to find a **non-degenerate** limiting distribution (chapter 8 in Hamilton), typically we consider the statistic

$$\sqrt{T}(\hat{\beta}_T - \beta) = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right).$$

Recall in maximum likelihood that $\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$ under certain conditions.

OLS for the Simple Trend-Stationary Process

Usually one assumes that

$$\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \xrightarrow{P} \mathbf{Q} \quad (24)$$

for some nonsingular matrix \mathbf{Q} , and that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q})$$

which implies that

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}).$$

OLS for the Simple Trend-Stationary Process

However, this procedure does **not** work for the trend stationary process.

To see this, let us check the assumptions. First,

$$\begin{aligned}\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t &= \begin{pmatrix} \sum 1 & \sum t \\ \sum t & \sum t^2 \end{pmatrix} \\ &= \begin{pmatrix} T & T(T+1)/2 \\ T(T+1)/2 & T(T+1)(2T+1)/6 \end{pmatrix} \quad (25)\end{aligned}$$

Then, $T^{-1} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t$ definitely **diverges** as $T \rightarrow \infty$.

What about ... try $T^{-3} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t$ instead of $T^{-1} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t$?

$$\lim_{T \rightarrow \infty} \frac{1}{T^3} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t = \begin{pmatrix} 0 & 0 \\ 0 & 1/3 \end{pmatrix}$$

which is singular! Usch!

OLS for the Simple Trend-Stationary Process

- It turns out that the OLS estimators $\hat{\alpha}_T$ and $\hat{\delta}_T$ have different asymptotic rates of convergence.
- In order to arrive at a non-degenerate limiting distribution, $\hat{\alpha}_T$ must be multiplied by $T^{1/2}$, while $\hat{\delta}_T$ by $T^{3/2}$.
- Then define the matrix below

$$\mathbf{S}_T = \begin{pmatrix} T^{1/2} & 0 \\ 0 & T^{3/2} \end{pmatrix}.$$

If we left-multiply \mathbf{S}_T to $\hat{\beta}_T - \beta$, what will happen?

OLS for the Simple Trend-Stationary Process

$$\begin{aligned}\mathbf{S}_T(\hat{\beta}_T - \beta) &= \mathbf{S}_T \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right) \\ &= \mathbf{S}_T \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \mathbf{S}_T \mathbf{S}_T^{-1} \left(\sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right) \\ &= \left[\mathbf{S}_T^{-1} \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right) \mathbf{S}_T^{-1} \right]^{-1} \left(\mathbf{S}_T^{-1} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right) \\ &= \mathbf{Q}_T^{-1} \mathbf{u}_T\end{aligned}$$

where

$$\mathbf{Q}_T = \mathbf{S}_T^{-1} \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right) \mathbf{S}_T^{-1} \quad \text{and} \quad \mathbf{u}_T = \mathbf{S}_T^{-1} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t.$$

OLS for the Simple Trend-Stationary Process

- First, we see that $\mathbf{Q}_T \rightarrow \mathbf{Q}$ as $T \rightarrow \infty$, where, from (25),

$$\mathbf{Q} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}$$

is nonsingular.

- Second,

$$\mathbf{u}_T = \begin{pmatrix} T^{-1/2} & 0 \\ 0 & T^{-3/2} \end{pmatrix} \begin{pmatrix} \sum \varepsilon_t \\ \sum t \varepsilon_t \end{pmatrix}.$$

The first element of \mathbf{u}_T

$$u_{1T} = \sqrt{T} \times \frac{1}{T} \sum_{t=1}^T \varepsilon_t \quad (\text{Familiar? Yes! CLT}),$$

and the second element

$$u_{2T} = \frac{1}{T^{3/2}} \sum_{t=1}^T t \varepsilon_t = \sqrt{T} \times \frac{1}{T} \sum_{t=1}^T \left(\frac{t}{T} \right) \varepsilon_t,$$

OLS for the Simple Trend-Stationary Process

- Suppose that sequence ε_t is independently identical distributed with zero mean, finite constant variance σ^2 , and $E(\varepsilon_t^4) < \infty$.
- Consider first the limiting distribution of u_{1T} .

Theorem (Classical CLT)

Let $\bar{y}_T = T^{-1} \sum_{t=1}^T y_t$, where y_1, \dots, y_T is a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2 . Then

$$\sqrt{T}(\bar{y}_T - \mu) \xrightarrow{d} N(0, \sigma^2)$$

-
- Thus, it follows that

$$u_{1T} \xrightarrow{d} N(0, \sigma^2) \quad (26)$$

OLS for the Simple Trend-Stationary Process

- Next we consider the limiting distribution of u_{2T} .
- Define $v_t = (\frac{t}{T})\varepsilon_t$. Since

$$E(v_t) = E(v_t | v_{t-1}, v_{t-2}, \dots) = \left(\frac{t}{T}\right) E(\varepsilon_t) = 0,$$

for all t , v_t is a martingale difference sequence (MDS).

Theorem (MDS CLT, Proposition 7.8 in Hamilton)

Let $\{y_t\}_{t=1}^{\infty}$ be a MDS with $\bar{y}_T = T^{-1} \sum_{t=1}^T y_t$. Suppose that

- 1 $E(y_t^2) = \sigma_t^2$ with $\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \sigma_t^2 = \sigma^2 < \infty$,
- 2 $E|y_t|^r < \infty$ for some $r > 2$ and all t , and
- 3 $T^{-1} \sum_{t=1}^T y_t^2 \xrightarrow{P} \sigma^2$.

Then

$$\sqrt{T} \bar{y}_T \xrightarrow{d} N(0, \sigma^2)$$

OLS for the Simple Trend-Stationary Process

- Let us verify that conditions 1-3 of the MDS CLT are satisfied for v_t .
- Condition 1,

$$\sigma_t^2 = E(v_t^2) = \left(\frac{t}{T}\right)^2 E(\varepsilon_t^2) = \left(\frac{t}{T}\right)^2 \sigma^2$$

with

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \sigma_t^2 &= \frac{1}{T} \sum_{t=1}^T \left(\frac{t}{T}\right)^2 \sigma^2 = \frac{\sigma^2}{T^3} \sum_{t=1}^T t^2 \\ &= \frac{\sigma^2}{T^3} \times \frac{T(T+1)(2T+1)}{6} \rightarrow \frac{\sigma^2}{3} \end{aligned}$$

as $T \rightarrow \infty$. Therefore, condition 1 is satisfied.

OLS for the Simple Trend-Stationary Process

- Condition 2, check $r = 4$,

$$E|v_t^4| = E(v_t^4) = \left(\frac{t}{T}\right)^4 E(\varepsilon_t^4) < \infty,$$

which is true as it has been presumed.

- Condition 3,

$$\frac{1}{T} \sum_{t=1}^T v_t^2 = \frac{1}{T} \sum_{t=1}^T \left[\left(\frac{t}{T} \right) \varepsilon_t \right]^2 \xrightarrow{p} \frac{\sigma^2}{3}$$

is verified by checking $T^{-1} \sum_{t=1}^T v_t^2 - \sigma^2/3 \xrightarrow{m.s.} 0$ on pp.459 in Hamilton.

- Thus, it follows that

$$u_{2T} \xrightarrow{d} N(0, \sigma^2/3) \quad (27)$$

OLS for the Simple Trend-Stationary Process

- Finally, the vector version is shown on pp.459 in Hamilton

$$\begin{pmatrix} T^{-1/2} \sum \varepsilon_t \\ T^{-1/2} \sum (\frac{t}{T}) \varepsilon_t \end{pmatrix} \xrightarrow{d} N_2(\mathbf{0}, \sigma^2 \mathbf{Q}) \quad (28)$$

where

$$\mathbf{Q} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}$$

- Then we have

$$\begin{pmatrix} T^{1/2}(\hat{\alpha}_T - \alpha) \\ T^{3/2}(\hat{\delta}_T - \delta) \end{pmatrix} \xrightarrow{d} N_2(\mathbf{0}, \mathbf{Q}^{-1} \times \sigma^2 \mathbf{Q} \times \mathbf{Q}^{-1}) = N_2(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}) \quad (29)$$

OLS for the Simple Trend-Stationary Process

- Look at the model again

$$y_t = \alpha + \delta t + \varepsilon_t$$

- We see that the estimators $\hat{\alpha}_T$ and $\hat{\delta}_T$ are both consistent. In particular, $\hat{\delta}$ is called **super consistent**, as $\hat{\delta}_T$ is consistent and $\hat{\delta}_T = \delta + O_p(T^{-3/2})$ with $-3/2 < -1/2$.
- The super consistency means that the estimator converges faster in terms of the **order in probability** than the **square root convergence**.
- Square root convergence $\sqrt{T}X_T \xrightarrow{d} X$.
- Super consistency: $\hat{\delta}_T - \delta \xrightarrow{P} 0$, and $\sqrt{T}(\hat{\delta}_T - \delta) \xrightarrow{P} 0$ and even $T(\hat{\delta}_T - \delta) \xrightarrow{P} 0$, while $\hat{\alpha}_T$ is not so super.

OLS for the Trend-Stationary AR(p) Process

- Chapter 16.3 in Hamilton considers a more general trend stationary process generated by

$$y_t = \alpha + \delta t + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t,$$

where ε_t is an independent white noise process with $E(\varepsilon_t^4) < \infty$, and $\phi(L)$ is stable.

- The same approach used to establish the limiting distribution of the OLS estimators $\hat{\alpha}_T$ and $\hat{\delta}_T$ for the simple trend-stationary process are used to establish the asymptotic distribution of the OLS estimators

$$\hat{\alpha}_T, \hat{\delta}_T, \hat{\phi}_{1,T}, \dots, \hat{\phi}_{p,T}$$

for the trend-stationary AR(p) process.



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 9. Univariate and Multivariate Processes with Unit Roots

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Univariate Processes with Unit Roots (pp.475-543 in Hamilton)
 - Introduction
 - Brownian Motion
 - The Functional CLT
 - The Continuous Mapping Theorem
 - Inference for the Simple Random Walk

Introduction

Consider OLS estimation of the parameter ρ in the simple Gaussian AR(1) process

$$y_t = \rho y_{t-1} + u_t, \quad t = 1, 2, \dots \quad (1)$$

where $u_t \stackrel{iid}{\sim} N(0, \sigma^2)$ and the initial value $y_0 = 0$.

The OLS estimator for ρ is given by

$$\hat{\rho}_T = \left(\sum_{t=1}^T y_{t-1}^2 \right)^{-1} \left(\sum_{t=1}^T y_{t-1} y_t \right) = \frac{\sum_{t=1}^T y_{t-1} y_t}{\sum_{t=1}^T y_{t-1}^2} \quad (2)$$

and we have

$$\hat{\rho}_T - \rho = \left(\sum_{t=1}^T y_{t-1}^2 \right)^{-1} \left(\sum_{t=1}^T y_{t-1} u_t \right) = \frac{\sum_{t=1}^T y_{t-1} u_t}{\sum_{t=1}^T y_{t-1}^2} \quad (3)$$

Introduction

If $|\rho| < 1$, then definitely

$$\sqrt{T}(\hat{\rho}_T - \rho) \xrightarrow{d} N(0, 1 - \rho) \quad (4)$$

What if $\rho = 1$? that is,

$$(1 - L)y_t = u_t \quad (5)$$

Check (4) and you will find immediately that the variance is $1 - \rho = 0$.
The limiting distribution is degenerate and collapses to a point mass.

$$\sqrt{T}(\hat{\rho}_T - \rho) \xrightarrow{P} 0 \quad (6)$$

Introduction

To obtain a non-degenerate limiting distribution for $\hat{\rho}_T$ in the unit root case, it turns out that we have to multiply (or scale) $\hat{\rho}_T$ by T rather than \sqrt{T} .

To get a better sense of why scaling by T is necessary when $\rho = 1$, note that $T(\hat{\rho}_T - 1)$ can be written as

$$T(\hat{\rho}_T - 1) = T \times \frac{\sum_{t=1}^T y_{t-1} u_t}{\sum_{t=1}^T y_{t-1}^2} = \frac{T^{-1} \sum_{t=1}^T y_{t-1} u_t}{T^{-2} \sum_{t=1}^T y_{t-1}^2} \quad (7)$$

Introduction

First, consider the numerator of (7). It can be shown that

$$\frac{1}{\sigma^2} \times \frac{1}{T} \sum_{t=1}^T y_{t-1} u_t \xrightarrow{d} \frac{1}{2}(X - 1), \quad (8)$$

where $X \sim \chi^2(1)$.

Second, consider the denominator of (7):

$$T^{-2} \sum_{t=1}^T y_{t-1}^2. \quad (9)$$

Introduction

Let us consider the expectation $E\left(\sum_{t=1}^T y_{t-1}^2\right)$. Since

$$y_{t-1} \sim N(0, \sigma^2(t-1)),$$

we get

$$E\left(\sum_{t=1}^T y_{t-1}^2\right) = \sum_{t=1}^T E(y_{t-1}^2) = \sigma^2 \sum_{t=1}^T (t-1) = \frac{\sigma^2(T-1)T}{2} \quad (10)$$

Now we see that, if we divide $\sum_{t=1}^T y_{t-1}^2$ by T^2 , its expectation will converge to $\sigma^2/2$ without T .

Introduction

Summary:

- If $\rho = 1$ (random walk process), $\hat{\rho}_T - 1$ should be multiplied by T instead of \sqrt{T} to obtain a non-degenerate limiting distribution.
- This limiting distribution is not the usual Gaussian distribution. It is a ratio involving a $\chi^2(1)$ distribution in the numerator and another **nonstandard** distribution in the denominator.
- We would like to describe this limiting distribution. This can be done in terms of functionals of **Brownian motion**.

Brownian Motion

Suppose that

$$y_t = y_{t-1} + \varepsilon_t, \quad t = 1, 2, \dots \quad (11)$$

where $\varepsilon_t \stackrel{iid}{\sim} N(0, 1)$ and $y_0 = 0$.

Then if we expand all the y_{t-1}, \dots , we get $y_t = y_0 + \varepsilon_1 + \dots + \varepsilon_t \sim N(0, t)$.

Letting $s > t$, we have

$$\begin{aligned} y_s - y_t &= (y_0 + \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_t + \varepsilon_{t+1} + \dots + \varepsilon_s) - (y_0 + \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_t) \\ &= \varepsilon_{t+1} + \dots + \varepsilon_s \end{aligned}$$

implying that $y_s - y_t \sim N(0, s - t)$.

For any $t < s < r < q$, the two random variables $y_s - y_t$ and $y_q - y_r$ are independent.

Brownian Motion

In particular, consider the change between y_{t-1} and y_t

$$y_t - y_{t-1} = \varepsilon_t$$

Suppose that we view ε_t as the sum of two independent Gaussian random variables

$$\varepsilon_t = e_{1t} + e_{2t}$$

where $e_{it} \stackrel{iid}{\sim} N(0, 1/2)$.

We might associate e_{1t} with the change between y_{t-1} and $y_{t-1/2}$

$$y_{t-1/2} - y_{t-1} = e_{1t}, \tag{12}$$

and e_{2t} with the change between $y_{t-1/2}$ and y_t

$$y_t - y_{t-1/2} = e_{2t}. \tag{13}$$

Brownian Motion

Note that (12) added to (13) implies (11)

$$y_t - y_{t-1} = e_{1t} + e_{2t},$$

where $e_{1t} + e_{2t} \stackrel{iid}{\sim} N(0, 1)$.

That is, sampled at $t = 1, 2, \dots$ the stochastic process defined by (12) and (13) are equivalent to (11) except the frequency.

In addition, (12) and (13) describe a stochastic process defined not only for $t = 1, 2, \dots$ but also for $t = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots$

For both integer and noninteger values of s and t , ($s > t$)

$$y_s - y_t \sim N(0, s - t).$$

$y_s - y_t$ and $y_q - y_r$ are independent for any $t < s < r < q$.

Brownian Motion

Similarly, we could view the change between y_{t-1} and y_t

$$y_t - y_{t-1} = \varepsilon_t$$

as the sum of n independent Gaussian variables

$$\varepsilon_t = e_{1t} + e_{2t} + \dots + e_{nt},$$

where $e_{it} \stackrel{iid}{\sim} N(0, 1/n)$ and

$$y_{t-(n-1)/n} - y_{t-1} = e_{1t}$$

$$y_{t-(n-2)/n} - y_{t-(n-1)/n} = e_{2t}$$

⋮

$$y_{t-1/n} - y_{t-2/n} = e_{(n-1)t}$$

$$y_t - y_{t-1/n} = e_{nt}$$

Brownian Motion

Remarks:

- The result would be a stochastic process with the same properties as the process in (11), defined at a finer and finer grid of time points as we increase the value of n .
- The limit as n tends to infinity is a continuous-time stochastic process known as **standard Brownian motion**. The value of this process at time t is denoted by $W(t)$.
- Brownian motion is named after the biologist Robert Brown whose research dates to the 1820s. A standard Brownian motion is also known as a Wiener process.
- A continuous-time process is a random variable that takes on a value for any nonnegative real number t , this is in contrast to a discrete-time process, which is only defined for integer values of t .

Brownian Motion

Definition (Standard Brownian Motion)

Standard Brownian motion $W(\cdot)$ is a continuous time process, associating each $0 \leq t \leq 1$ with the scalar random variable $W(t)$ such that:

- 1 $W(0) = 0$
- 2 For any $0 \leq t_1 < t_2 < \dots < t_k \leq 1$ then random variables

$$W(t_2) - W(t_1), W(t_3) - W(t_2), \dots, W(t_k) - W(t_{k-1})$$

are independent and jointly multivariate Gaussian distributed, with

$$W(s) - W(t) \sim N(0, s - t)$$

- for any $0 \leq t < s \leq 1$.
- 3 $W(t)$ has continuous sample paths.

Brownian Motion

Remarks:

- Though $W(t)$ has continuous sample paths, it can be shown that its sample paths are nowhere differentiable.
- Other continuous time processes can be generated from standard Brownian motion.
- Since, by definition, $W(t) \sim N(0, t)$, it follows that

$$W_\sigma(t) = \sigma W(t) \sim N(0, \sigma^2 t).$$

- Similarly, it is readily seen that $W^2(t)$ is $t \times \chi^2(1)$ distributed. In particular, $W^2(1) \sim \chi^2(1)$.

Functional Central Limit Theorem

One of the uses of Brownian motion is to allow for more general statements of the CLT.

Recall the classical CLT.

Theorem

Let $\bar{y}_T = T^{-1} \sum_{t=1}^T y_t$, where y_1, \dots, y_T is a sequence of i.i.d. random variables with finite mean μ and variance σ^2 . Then

$$\sqrt{T}(\bar{y}_T - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Functional Central Limit Theorem

Suppose that u_1, \dots, u_T is an *i.i.d.* sequence with mean zero and variance σ^2 , and consider the estimator

$$\bar{u}_{\lfloor T/2 \rfloor} = \frac{1}{\lfloor T/2 \rfloor} \sum_{t=1}^{\lfloor T/2 \rfloor} u_t, \quad (14)$$

where $\lfloor \cdot \rfloor$ denotes the floor function (integer part for positive numbers).

Given a sample of size T , for an even T , this estimator uses only the first half of the sample and discards the other half.

Clearly, this estimator also satisfies the classical CLT

$$\sqrt{\lfloor T/2 \rfloor} \times \bar{u}_{\lfloor T/2 \rfloor} \xrightarrow{d} N(0, \sigma^2), \quad \text{as } T \rightarrow \infty. \quad (15)$$

Moreover, $\bar{u}_{\lfloor T/2 \rfloor}$ would be **independent** of an estimator that uses only the second half of the sample.

Functional Central Limit Theorem

More generally, we can construct a random variable $X_T(r)$ that uses only the first r th fraction of the sample u_1, \dots, u_T

$$X_T(r) = \frac{1}{T} \sum_{t=1}^{\lfloor rT \rfloor} u_t, \quad (16)$$

where $0 \leq r \leq 1$.

Thus, by construction

$$X_T(r) = \begin{cases} 0 & \text{for } 0 \leq r < 1/T \\ u_1/T & \text{for } 1/T \leq r < 2/T \\ (u_1 + u_2)/T & \text{for } 2/T \leq r < 3/T \\ \vdots & \\ (u_1 + u_2 + \dots + u_n)/T & \text{for } r = 1 \end{cases}$$

Functional Central Limit Theorem

For $r > 0$, it can be shown that

$$\sqrt{T} X_T(r) \xrightarrow{d} N(0, r\sigma^2).$$

Hence

$$\frac{1}{\sigma} \sqrt{T} X_T(r) \xrightarrow{d} N(0, r).$$

This implies that

$$\frac{1}{\sigma} \sqrt{T} [X_T(r_2) - X_T(r_1)] \xrightarrow{d} N(0, r_2 - r_1)$$

for any $0 \leq r_1 \leq r_2 \leq 1$.

Functional Central Limit Theorem

In addition, note that the random variable

$$X_T(r_2) - X_T(r_1) = \frac{1}{T} \sum_{t=1}^{\lfloor r_2 T \rfloor} u_t - \frac{1}{T} \sum_{t=1}^{\lfloor r_1 T \rfloor} u_t = \frac{1}{T} \sum_{t=\lfloor r_1 T \rfloor + 1}^{\lfloor r_2 T \rfloor} u_t$$

is independent of $X_T(r)$ for any $0 \leq r \leq r_1$.

Functional Central Limit Theorem

So it should not be surprising that

$$\frac{1}{\sigma} \sqrt{T} X_T(\cdot) \xrightarrow{d} W(\cdot) \quad (17)$$

for $0 \leq r \leq 1$. This is the **functional central limit theorem**.

For example, when the functions in (17) are evaluated at $r = 1$, we have

$$\frac{1}{\sigma} \sqrt{T} X_T(1) \xrightarrow{d} W(1)$$

where $X_T(1) = T^{-1} \sum_{t=1}^T u_t$ and $W(1)$ is the standard normal distribution.

The classical CLT is a special case of the functional CLT.

Functional Central Limit Theorem

Remarks:

- The expression $X_T(\cdot)$ denotes a function, while $X_T(r)$ denotes the value that function assumes at time r (a random variable).
- In previous lectures, we defined the convergence in distribution for (a sequence of) random variables. Now this definition can be extended to (a sequence of) random functions.

FCLT Again

- Suppose that ε_t is i.i.d $(0, \sigma^2)$, but not necessarily normally distributed!
- The functional central limit theorem (FCLT) (or the invariance principle) tells us that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{\textcolor{red}{s}} \varepsilon_t \xrightarrow{d} \sigma W(\textcolor{red}{r}) \sim N(0, r\sigma^2)$$

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{\textcolor{red}{T}} \varepsilon_t \xrightarrow{d} \sigma W(\textcolor{red}{1}) \sim N(0, \sigma^2)$$

as both T and s go to infinity and $s/T \rightarrow r$. Remember this!

Continuous Mapping Theorem

Proposition 7.3 (c) on pp.184 in Hamilton says that

- if the sequence of random variables $\{x_t\}_{t=1}^{\infty}$ converges in distribution to x , i.e., $x_t \xrightarrow{d} x$, and
- if the function $g : \mathbb{R} \mapsto \mathbb{R}$ is a continuous function, then

$$g(x_t) \xrightarrow{d} g(x)$$

A similar result holds for a sequence of random functions. The analog to the function $g(\cdot)$ is a continuous **functional**.

Continuous Mapping Theorem

The continuous mapping theorem states that if

$$S_T(\cdot) \xrightarrow{d} S(\cdot)$$

and $g(\cdot)$ is continuous functional, then

$$g(S_T(\cdot)) \xrightarrow{d} g(S(\cdot)) \quad (18)$$

For example,

$$\sqrt{T}X_T(\cdot) \xrightarrow{d} \sigma W(\cdot) \quad \text{and} \quad [\sqrt{T}X_T(\cdot)]^2 \xrightarrow{d} \sigma^2 [W(\cdot)]^2$$

and even

$$\int_0^1 \sqrt{T}X_T(x)dx \xrightarrow{d} \int_0^1 \sigma W(x)dx \quad \text{and...}$$

Inference for the Simple Random Walk

Consider the simple random walk

$$y_t = y_{t-1} + u_t, \quad t = 1, 2, \dots$$

where u_1, \dots, u_T is an *i.i.d.* with mean zero and variance σ^2 and $y_0 = 0$.

By recursion,

$$y_t = u_1 + u_2 + \dots + u_t$$

Note that

$$X_T(r) = \begin{cases} 0 = y_0/T & \text{for } 0 \leq r < 1/T \\ u_1/T = y_1/T & \text{for } 1/T \leq r < 2/T \\ (u_1 + u_2)/T = y_2/T & \text{for } 2/T \leq r < 3/T \\ \vdots & \\ (u_1 + u_2 + \dots + u_n)/T = y_T/T & \text{for } r = 1 \end{cases}$$

Inference for the Simple Random Walk

Please see Figure 17.1 on pp.484 in Hamilton!

A simple geometrical argument shows that

$$\int_0^1 X_T(r) dr = \frac{y_1}{T^2} + \frac{y_2}{T^2} + \dots + \frac{y_{T-1}}{T^2}$$

or

$$\int_0^1 \sqrt{T} X_T(r) dr = T^{-3/2} \sum_{t=1}^T y_{t-1}$$

Inference for the Simple Random Walk

Recall that

$$\sqrt{T}X_T(\cdot) \xrightarrow{d} \sigma W(\cdot).$$

Therefore, by the continuous mapping theorem

$$\int_0^1 \sqrt{T}X_T(r)dr \xrightarrow{d} \int_0^1 \sigma W(r)dr$$

which implies that

$$T^{-3/2} \sum_{t=1}^T y_{t-1} \xrightarrow{d} \sigma \int_0^1 W(r)dr,$$

as the sample size T tends to infinity.

Inference for the Simple Random Walk

A similar approach can be used to describe the limiting distribution of

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2$$

Let

$$S_T(r) = T[X_T(r)]^2$$

Then $S_T(r)$ can be written as

$$S_T(r) = \begin{cases} 0 & \text{for } 0 \leq r < 1/T \\ y_1^2/T & \text{for } 1/T \leq r < 2/T \\ y_2^2/T & \text{for } 2/T \leq r < 3/T \\ \vdots & \\ y_T^2/T & \text{for } r = 1 \end{cases}$$

Inference for the Simple Random Walk

A simple geometrical argument shows that

$$\int_0^1 S_T(r) dr = \frac{y_1^2}{T^2} + \frac{y_2^2}{T^2} + \dots + \frac{y_{T-1}^2}{T^2}$$

or equivalently

$$\int_0^1 S_T(r) dr = \frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2$$

Inference for the Simple Random Walk

Recall that

$$S_T(\cdot) \xrightarrow{d} \sigma^2[W(\cdot)]^2.$$

Therefore, by the continuous mapping theorem

$$\int_0^1 S_T(r) dr \xrightarrow{d} \int_0^1 \sigma^2[W(r)]^2 dr$$

which implies that

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2 \xrightarrow{d} \sigma^2 \int_0^1 W(r)^2 dr, \quad (19)$$

as the sample size T tends to infinity.

Inference for the Simple Random Walk

Consider now the limiting distribution of the test statistic

$$T(\hat{\rho}_T - 1) = \frac{T^{-1} \sum_{t=1}^T y_{t-1} u_t}{T^{-2} \sum_{t=1}^T y_{t-1}^2}. \quad (20)$$

In (7) and (8), we see that the numerator of (20) converges in distribution to $\frac{\sigma^2}{2}(X - 1)$, where $X \sim \chi^2(1)$ as $T \rightarrow \infty$.

Recall that the random variable $W(1)^2$ is $\chi^2(1)$ distributed.

Hence, another way to describe the limiting distribution of the numerator of (20) is using a functional of Brownian motion

$$\frac{1}{T} \sum_{t=1}^T y_{t-1} u_t \xrightarrow{d} \frac{\sigma^2}{2} (W(1)^2 - 1) \quad (21)$$

Inference for the Simple Random Walk

Since (20) is a continuous function of the LHSs of (21) and (19), it follows that, under the null hypothesis that $\rho = 1$, the OLS estimator $\hat{\rho}_T$ is characterized by

$$T(\hat{\rho}_T - 1) \xrightarrow{d} \frac{\frac{1}{2}(W(1)^2 - 1)}{\int_0^1 W(r)^2 dr} \quad (22)$$

Remark:

In practice, critical values for the test statistic in (22) are found by calculating the exact finite-sample distribution of $T(\hat{\rho}_T - 1)$ for a given sample size T , under the assumption that u_t are Gaussian distributed. This can be done either by Monte Carlo simulation, or by using exact numerical procedures. Read pp.488 in Hamilton.

You can use the results in Proposition 17.1 on pp.486 in Hamilton. Note that $\xi_t = y_t$.

Dickey-Fuller Tests

- Now consider the somewhat general model

$$y_t = \rho y_{t-1} + \alpha + \delta t + \varepsilon_t \quad (23)$$

where ε_t is *i.i.d.* with zero mean and finite variance σ^2 .

- We are interested in whether $\rho = 1$ (unit root), and we test it based on the observations y_t .
- Dickey-Fuller tests are several unit root tests for different situations (different assumptions), but they all assume that there is not autocorrelation in the errors ε_t .

Dickey-Fuller Test for Case 1

- The regression model

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (24)$$

- Assumptions: $\alpha = 0$ and $\delta = 0$
- Null hypothesis $H_0 : \rho = 1$
- The alternative $H_1 : |\rho| < 1$
- The test has been given in (22)
- There are two versions for the test, ρ version in (22), and t -ratio version

$$t_T \xrightarrow{d} \frac{\frac{1}{2}(W(1)^2 - 1)}{\sqrt{\int_0^1 W(r)^2 dr}} \quad (25)$$

Dickey-Fuller Test for Case 2

- The regression model

$$y_t = \rho y_{t-1} + \alpha + \varepsilon_t \quad (26)$$

- Assumptions: $\delta = 0$
- Null hypothesis $H_0 : \rho = 1$ and $\alpha = 0$
- The alternative $H_1 : |\rho| < 1$ or $\alpha \neq 0$
- The joint test for the null hypothesis is in [17.4.25] on pp.492.
- The tests for $\rho = 1$ are given in [17.4.28] on pp.492 (ρ) and [17.4.36] on pp.494 (t -ratio).
- If the null is true, the model is simply a random walk.

Dickey-Fuller Test for Case 3

- The regression model

$$y_t = \rho y_{t-1} + \alpha + \varepsilon_t \quad (27)$$

- Assumptions: $\delta = 0$ and $\alpha \neq 0$
- Null hypothesis $H_0 : \rho = 1$
- The alternative $H_1 : |\rho| < 1$
- The test is given in [17.4.46] on pp.492 (ρ). Note that it is the marginal distribution of $\hat{\rho}_T$
- Both $\hat{\alpha}_T$ and $\hat{\rho}_T$ converge to Gaussian, but with different rates of convergence.
- If the null is true, the model is $y_t = y_0 + \alpha t + \sum_{s=1}^t \varepsilon_s$. Random walk with drift αt .
- We see that, from cases 2 and 3, the asymptotic distributions of ρ are different based on different beliefs about the true value of α .

Dickey-Fuller Test for Case 4

- The regression model

$$y_t = \rho y_{t-1} + \alpha + \delta t + \varepsilon_t \quad (28)$$

- Assumptions: α can be anything
- Null hypothesis $H_0 : \rho = 1, \delta = 0$ and $\alpha = \alpha_0$
- The alternative $H_1 : |\rho| < 1$ or $\delta \neq 0$ or $\alpha \neq \alpha_0$
- The model can be reparameterized as follows

$$y_t = \alpha^* + \rho^* \xi_{t-1} + \delta^* t + \varepsilon_t \quad (29)$$

where $\alpha^* = (1 - \rho)\alpha$, $\rho^* = \rho$, $\delta^* = \delta + \rho\alpha$ and $\xi_t = y_{t-1} - \alpha(t - 1)$. Moreover, $\xi_t = y_0 + \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_t$.

- The new and equivalent null hypothesis is $H_0 : \rho^* = 1, \alpha^* = 0$ and $\delta^* = \alpha_0$.
- The joint test is given in [17.4.53] on pp.499 (ρ). The t -ratio of ρ is given in [17.4.55].

Remarks

If $|\rho| < 1$,

- $\alpha \neq 0$ is simply an intercept
- $\delta \neq 0$ is a linear trend.

If $\rho = 1$,

- $\alpha \neq 0$ will become a drift term (linear trend)
- $\delta \neq 0$ will become a quadratic trend.

In practice, you may assume that either $|\rho| < 1$ or $\rho = 1$,

- take 1st and 2nd order differences, see whether they look stationary
- if the 2nd order difference shows strong stationarity, you may skip case 4.
- you have to plot the data, see whether there is a clear linear trend; the linear trend may come from either ($|\rho| < 1$ and $\delta \neq 0$) or ($\rho = 1$ and $\alpha \neq 0$).
- choose one or several DF tests and analyse.

Serial correlation

- In all cases, it was assumed that the error term is independent (hence not serially correlated). We can test for serial correlation by using, for example, the Breusch-Godfrey autocorrelation test.
- If we find serial correlation, we should take it into account. An easy strategy for this is to respecify the estimation equation by adding lagged first differences. The corresponding unit root tests are called **augmented Dickey-Fuller (ADF) tests**.
- Another strategy is to estimate the autocovariances (nuisance parameters) γ_i of the errors and construct new tests similar to the DF tests. The resulting unit root tests are called **Phillips-Perron tests** but their finite-sample performance are poor in contrast to the ADF tests.



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 10. Unit Roots in Multivariate Time Series and Cointegration

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Unit Roots in Multivariate Time Series
 - Asymptotic Results for Nonstationary Vector Processes
 - Vector Autoregressions Containing Unit Roots
 - Spurious Regressions
- Cointegration
 - Introduction

Multivariate Standard Brownian Motion

We introduce the **multivariate standard Brownian Motion**, the definition on pp.544 in Hamilton,

Definition

n -dimensional standard Brownian motion $\mathbf{W}(\cdot)$ is a continuous-time process associating each data $r \in [0, 1]$ with the $(n \times 1)$ vector $\mathbf{W}(r)$ satisfying the following:

- 1 $\mathbf{W}(0) = 0$;
- 2 For any dates $0 \leq r_1 < r_2 < \dots < r_k \leq 1$, the changes $[\mathbf{W}(r_2) - \mathbf{W}(r_1)], [\mathbf{W}(r_3) - \mathbf{W}(r_2)], \dots, [\mathbf{W}(r_k) - \mathbf{W}(r_{k-1})]$ are independent multivariate Gaussian with $\mathbf{W}(s) - \mathbf{W}(r) \sim N_n(\mathbf{0}, (s - r)\mathbf{I}_n)$;
- 3 For any given realization, $\mathbf{W}(r)$ is continuous in r with probability 1.

Multivariate Standard Brownian Motion

Remarks:

- The univariate standard Brownian motion (BM) is a special case of the multivariate standard Brownian motion (MBM or VBM), $n = 1$.
- The univariate BM can be easily extended to the VBM by adding more independent univariate BMs. Note that the covariance for $\mathbf{W}(s) - \mathbf{W}(r)$ is simply $(s - r)\mathbf{I}_n$ implying that they are independent.
- You can simply change all the scalars in the previous lecture to vectors, y_t to \mathbf{y}_t , ε_t to $\boldsymbol{\varepsilon}_t$, the same results hold for the multivariate Brownian motion.
- The matrix $\mathbf{W}(r)\mathbf{W}(r)'$ is Wishart distributed.

Functional Central Limit Theorem

Suppose that $\varepsilon_1, \dots, \varepsilon_T$ is an *i.i.d.* n -vector sequence with mean zero and covariance \mathbf{I}_n .

Then we can construct a random vector $\mathbf{X}_T(r)$ that uses only the first r th fraction of the sample $\varepsilon_1, \dots, \varepsilon_T$

$$\mathbf{X}_T(r) = \frac{1}{T} \sum_{t=1}^{\lfloor rT \rfloor} \varepsilon_t, \quad (1)$$

where $0 \leq r \leq 1$.

The corresponding functional central limit theorem

$$\sqrt{T} \mathbf{X}_T(\cdot) \xrightarrow{d} \mathbf{W}(\cdot) \quad (2)$$

for $0 \leq r \leq 1$. Note that $\mathbf{W}(r) \sim N_n(\mathbf{0}, r\mathbf{I}_n)$.

Functional Central Limit Theorem

Suppose that there is another vector sequence \mathbf{v}_t such that $\mathbf{v}_t = \mathbf{P}\boldsymbol{\varepsilon}_t$, and that $\mathbf{P}\mathbf{P}' = \Omega$ which is positive definite. The sequence of \mathbf{v}_t is *i.i.d.* with mean zero and covariance Ω .

Let $\mathbf{X}_T^*(r)$ be

$$\mathbf{X}_T^*(r) = \frac{1}{T} \sum_{t=1}^{\lfloor rT \rfloor} \mathbf{v}_t, \quad (3)$$

where $0 \leq r \leq 1$.

Then the functional central limit theorem

$$\sqrt{T}\mathbf{X}_T^*(\cdot) \xrightarrow{d} \mathbf{P} \cdot \mathbf{W}(\cdot) \quad (4)$$

for $0 \leq r \leq 1$. Note that $\mathbf{P}\mathbf{W}(r) \sim N_n(\mathbf{0}, r\Omega)$.

Vector I(0) process

Recall that

- A linear zero-mean vector I(0) process is a vector MA(∞) process

$$\mathbf{u}_t = \Psi(L)\mathbf{v}_t, \quad \Psi(L) = \Psi_0 + \Psi_1 L + \Psi_2 L^2 + \dots, \quad (5)$$

satisfying 2 conditions:

- 1 $\Psi(1) \neq \mathbf{0}$ (but not necessarily of full rank), (ensures I(0)) and
 - 2 the matrix $\Psi_s = (\psi_{ij}^{(s)})$ is one-summable, meaning $\sum_{s=0}^{\infty} s|\psi_{ij}^{(s)}| < \infty$ for all $i, j = 1, 2, \dots, n$ (we need it later).
- The Long-Run covariance (LRV) matrix of \mathbf{u}_t is

$$\text{LRV}(\mathbf{u}_t) \equiv \lim_{T \rightarrow \infty} \text{Var}[\sqrt{T}\bar{\mathbf{u}}_T] = \Psi(1)\Omega\Psi(1)'. \quad (6)$$

Vector I(1) process

- A **vector I(1) process** is defined as

$$\Delta \mathbf{y}_t = \boldsymbol{\delta} + \mathbf{u}_t \quad (7)$$

where $\mathbf{u}_t = \Psi(L)\mathbf{v}_t$ and $\boldsymbol{\delta} = E(\Delta \mathbf{y}_t)$ is a vector of constants. Hence,

$$\Delta \mathbf{y}_t = \boldsymbol{\delta} + \Psi(L)\mathbf{v}_t \quad (8)$$

is the vector moving average (VMA) representation of a vector I(1) process.

- In levels, \mathbf{y}_t can be written as

$$\mathbf{y}_t = \mathbf{y}_0 + \boldsymbol{\delta} t + \sum_{s=1}^t \mathbf{u}_s \quad (9)$$

Beveridge-Nelson decomposition

- Using $\Psi(L) = \Psi(1) + \Delta\alpha(L)$ where $\alpha(L) = \sum_{j=0}^{\infty} \alpha_j L^j$, with $\alpha_j = -(\Psi_{j+1} + \Psi_{j+2} + \dots)$ for $j = 0, 1, \dots$, we can write

$$\mathbf{u}_t = \Psi(1)\mathbf{v}_t + \boldsymbol{\eta}_t - \boldsymbol{\eta}_{t-1}, \quad (10)$$

where $\boldsymbol{\eta}_t = \alpha(L)\mathbf{v}_t$ is a zero-mean I(0) process, and $\alpha(L)$ is absolutely summable (ensured by the one-summability).

- Substitution of (10) in (9) gives

$$\mathbf{y}_t = \underbrace{\delta t}_{\text{linear trend}} + \underbrace{\Psi(1) \sum_{s=1}^t \mathbf{v}_s}_{\substack{\text{stochastic trend}}} + \underbrace{\boldsymbol{\eta}_t}_{\text{cycle}} + \underbrace{\mathbf{y}_0 - \boldsymbol{\eta}_0}_{\text{initial condition}} \quad (11)$$

FCLT for Serially Dependent Vector Processes

Now suppose that $\delta = \mathbf{0}$ and $\mathbf{y}_0 = \mathbf{0}$.

$$\mathbf{y}_t = \sum_{s=1}^t \mathbf{u}_s = \Psi(1) \sum_{s=1}^t \mathbf{v}_s + \boldsymbol{\eta}_t - \boldsymbol{\eta}_0 \quad (12)$$

Let $\mathbf{X}_T^{**}(r)$ be

$$\mathbf{X}_T^{**}(r) = \frac{1}{T} \sum_{t=1}^{\lfloor rT \rfloor} \mathbf{u}_t = \mathbf{y}_{\lfloor rT \rfloor} / T, \quad (13)$$

where $0 \leq r \leq 1$.

Then the functional central limit theorem

$$\sqrt{T} \mathbf{X}_T^{**}(\cdot) \xrightarrow{d} \Psi(1) \cdot \mathbf{P} \cdot \mathbf{W}(\cdot) \quad (14)$$

for $0 \leq r \leq 1$. Note that $\Psi(1)\mathbf{P}\mathbf{W}(r) \sim N_n(\mathbf{0}, r\Psi(1)\Omega\Psi(1)')$.

Asymptotic Results for Nonstationary Vector Processes

- Proposition 18.1 on pp.547 in Hamilton summarizes the results.
- Note that we use different notations compared to the ones in Hamilton.
- The differences are mainly, in Hamilton,
 - \mathbf{v}_t for *i.i.d.* error vectors with covariance \mathbf{I}_n while we use $\boldsymbol{\varepsilon}_t$;
 - $\boldsymbol{\varepsilon}_t$ for $\mathbf{P}\mathbf{v}_t$ while we use \mathbf{v}_t ;
 - $\xi_t = \sum_{s=1}^t \mathbf{u}_s$ while we use \mathbf{y}_t .

An Alternative Representation of a VAR(p) Process

Consider the following VAR(p) process

$$\Phi(L)\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{v}_t, \quad (15)$$

where $\Phi(L) = \mathbf{I}_n - \Phi_1 L - \dots - \Phi_p L^p$, and \mathbf{v}_t is defined as before.

The lag polynomial can be rewritten as

$$\begin{aligned}\Phi(L) &= \mathbf{I}_n - \Phi_1 L - \dots - \Phi_p L^p \\ &= (\mathbf{I}_n - \rho L) - (\zeta_1 L + \zeta_2 L^2 + \dots + \zeta_{p-1} L^{p-1})(1 - L)\end{aligned}$$

where $\rho = \sum_{s=1}^p \Phi_s$ and $\zeta_s = -(\Phi_{s+1} + \Phi_{s+2} + \dots + \Phi_p)$.

Then it follows that

$$\mathbf{y}_t = \rho \mathbf{y}_{t-1} + \zeta_1 \Delta \mathbf{y}_{t-1} + \zeta_2 \Delta \mathbf{y}_{t-2} + \dots + \zeta_{p-1} \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\alpha} + \mathbf{v}_t. \quad (16)$$

A Very Strong Assumption

- If $\mathbf{I}_n = \rho$, we can equivalently analyze the first-order difference of \mathbf{y}_t , which is a VAR($p - 1$) process.

$$\Delta \mathbf{y}_t = \zeta_1 \Delta \mathbf{y}_{t-1} + \zeta_2 \Delta \mathbf{y}_{t-2} + \dots + \zeta_{p-1} \Delta \mathbf{y}_{t-p+1} + \alpha + \mathbf{v}_t. \quad (17)$$

- By assuming that $\zeta(L)$ is stable, the model can be

$$\Delta \mathbf{y}_t = \zeta(1)^{-1} \alpha + \zeta(L)^{-1} \mathbf{v}_t = \delta + \mathbf{u}_t, \quad (18)$$

where $\zeta(L)^{-1} = \Psi(L)$, $\delta = \Psi(1)\alpha$ and $\mathbf{u}_t = \Psi(L)\mathbf{v}_t$.

- The assumption $\mathbf{I}_n = \rho$ is so strong that we could rarely find it in reality.
- Note that $\mathbf{I}_n = \rho$ implies $|\mathbf{I}_n - \rho| = 0$, but the other way around does not hold.
- First we consider the testing for the case $\mathbf{I}_n = \rho$, and then the more interesting case $|\mathbf{I}_n - \rho| = 0$ follows.

The Case with No Drift

- The regression model

$$\mathbf{y}_t = \rho \mathbf{y}_{t-1} + \zeta_1 \Delta \mathbf{y}_{t-1} + \zeta_2 \Delta \mathbf{y}_{t-2} + \dots + \zeta_{p-1} \Delta \mathbf{y}_{t-p+1} + \alpha + \mathbf{v}_t.$$

- Assumptions: the lag polynomial $\zeta(L)$ is stable, or equivalently the roots of the polynomial

$$|\mathbf{I}_n - \zeta_1 z - \zeta_2 z^2 - \dots - \zeta_{p-1} z^{p-1}| = 0 \quad (19)$$

are all outside the unit disk.

- Null hypothesis $H_0 : \rho = \mathbf{I}_n$ and $\alpha = \mathbf{0}$
- From [18.2.18] on pp.551, we see that the estimators have different rates of convergence. In particular, $\hat{\rho} - \mathbf{I}_n = O_p(T^{-1})$
- The test is given by [18.2.25] on pp.552. Note that the parameters are split into two parts, ζ s and (α, ρ) .

The Case with Drift

- The regression model

$$\mathbf{y}_t = \rho \mathbf{y}_{t-1} + \zeta_1 \Delta \mathbf{y}_{t-1} + \zeta_2 \Delta \mathbf{y}_{t-2} + \dots + \zeta_{p-1} \Delta \mathbf{y}_{t-p+1} + \alpha + \mathbf{v}_t.$$

- Assumptions: the lag polynomial $\zeta(L)$ is stable and $\alpha \neq 0$.
- Null hypothesis $H_0 : \rho = \mathbf{I}_n$
- Note that there is a reparametrization in [18.2.43] on pp.556!
- The rates of convergence of the estimators are shown in [18.2.45] on pp.556.
- The test for equation i is given by [18.2.49] on pp.557. Note that the parameters are split into two parts, ζ_i s and $(\alpha_i^*, \rho_i^*, \gamma_i)$.

Vector Autoregressions Containing Unit Roots

Remarks:

- We can test all the parameters including ζ s. The "null hypothesis" in previous pages stresses that they are somewhat "assumed" but still the zero parameters are put inside the regression.
- We see that the limiting distributions of these parameters depend closely on the assumptions or (better say) beliefs.
- Though these tests are not so often used in reality, but the tests for the case $n = 1$ is widely used, which are exactly the augmented Dickey-Fuller tests.

Spurious Regressions

- Consider the $I(1)$ vector sequence \mathbf{y}_t whose difference is simply

$$\Delta \mathbf{y}_t = \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\varepsilon}_t$ has covariance \mathbf{I}_n .

- For simplicity, assume $n = 2$. Let us regress the following model

$$y_{1t} = \alpha + \gamma y_{2t} + \epsilon_t \quad (20)$$

- We know that $\alpha = 0$ and $\gamma = 0$.
- However,

$$\begin{pmatrix} T^{-1/2} \hat{\alpha}_T \\ \hat{\gamma}_T \end{pmatrix} \xrightarrow{d} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \quad (21)$$

where h_1 and h_2 are given by [18.3.9] on pp.559.

- We see that neither of them are consistent because h_1 and h_2 have random variables with non-zero expectation and non-zero variances.
- Even worse, $\hat{\alpha}_T$ diverges.

Cointegration

Engel and Granger's cointegration:

- The $I(1)$ process \mathbf{y}_t defined in (8) is cointegrated with cointegrating vector $\mathbf{a} \neq 0$ (of dimension $n \times 1$), if $\mathbf{a}'\mathbf{y}_t$ is trend-stationary.
- Multiplying (11) on both sides by \mathbf{a}' , we obtain

$$\mathbf{a}'\mathbf{y}_t = \mathbf{a}'\delta t + \mathbf{a}'\Psi(1) \sum_{s=1}^t \mathbf{v}_s + \mathbf{a}'\eta_t + \mathbf{a}'(\mathbf{y}_0 - \eta_0) \quad (22)$$

and $\mathbf{a}'\mathbf{y}_t$ is trend-stationary if $\mathbf{a}'\Psi(1) = \mathbf{0}'$.

- The example on pp.572 about the purchasing power parity (PPP).
- The sufficient condition for $\mathbf{a}'\Psi(1) = \mathbf{0}'$ holds for a non-zero vector \mathbf{a} is that $\Psi(1)$ has reduced rank.
- Suppose that the null space of $\Psi(1)$ has dimension h . There exist h \mathbf{a} vectors who are linearly independent such that $\mathbf{A}'\Psi(1) = \mathbf{0}'$, where $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_h)$. \mathbf{A} is not unique!

Related Concepts

- The **cointegrating (CI) rank** is the number of linearly independent cointegrating vectors. If the CI rank is equal to h , then $\text{rank}(\Psi(1)) = n - h$.
- The **CI space** is the space spanned by the cointegrating vectors.
- It may also happen that $\mathbf{a}'\boldsymbol{\delta} = 0$. Then $\mathbf{a}'\mathbf{y}_t$ is stationary, rather than trend stationary.
- In that case $\text{rank}([\boldsymbol{\delta} : \Psi(1)]) = \text{rank}(\Psi(1))$: since the left matrix has $1 + n$ columns and n rows, its rank can't exceed $\text{rank}(\Psi(1))$. This implies that $\boldsymbol{\delta}$ is a linear combination of the columns of $\Psi(1)$.

Implications

- If $h = n$, $\text{rank}(\Psi(1)) = n - h = 0 \Rightarrow \Psi(1) = \mathbf{0}$, which is ruled out since \mathbf{u}_t is $I(0)$.
- $\text{LRV}(\Delta \mathbf{y}_t) = \text{LRV}(\mathbf{u}_t) = \Psi(1)\Omega\Psi(1)'$ is positive definite if and only if $\Psi(1)$ is of full rank.
- Therefore, \mathbf{y}_t cannot be cointegrated if $\text{LRV}(\Delta \mathbf{y}_t)$ is positive definite. In this case, each element of $\Delta \mathbf{y}_t$ has its $\text{LRV} > 0$ and is a univariate $I(1)$ process.
- Let $\mathbf{y}'_t = (y_{1t}, y'_{2t})$ and $\mathbf{a}' = (a_1, a'_2)$. If $h = 1$ and $a_1 \neq 0$, then y_{1t} is cointegrated with some elements in y_{2t} (i.e. $\mathbf{a}_2 \neq 0$). But y_{2t} is not cointegrated (itself) without y_{1t} , i.e. there is no CI vector such as $(0, \mathbf{b}')$.
If $h > 1$ (more than 1 CI vector), then \mathbf{y}_{2t} is also cointegrated (itself).

The Stock-Watson Common Trend Representation

- Fact 1: if $\text{rank}(\Psi(1)) = n - h$, there exists a non-singular $n \times n$ matrix \mathbf{G} , and a $n \times (n - h)$ matrix \mathbf{F} of full column rank, such that $\Psi(1)\mathbf{G} = [\mathbf{F} : \mathbf{0}_{n \times h}]$.
- Then the stochastic trend component of y_t in (11) can be written

$$\begin{aligned}\Psi(1) \sum_{s=1}^t \mathbf{v}_s &= \Psi(1) \mathbf{G} \mathbf{G}^{-1} \sum_{s=1}^t \mathbf{v}_s \\ &= [\mathbf{F} : \mathbf{0}_{n \times h}] \begin{pmatrix} \tau_t \\ \dots \\ \mathbf{v}_t \end{pmatrix} = \mathbf{F} \tau_t\end{aligned}$$

Therefore, an I(1) system with a CI rank equal to h has $h - h$ "common stochastic trends", which are the elements of τ_t .

Cointegrated VAR

- We don't use the VMA form to model cointegration. Normally we use the VAR model.
- We need to model \mathbf{y}_t , not just $\Delta \mathbf{y}_t$!
- We transform the VAR into its VECM form.

Stationary VAR

- A VAR(p) model with stable lag polynomial, implies that y_t is stationary, therefore not cointegrated. What conditions must be imposed if we want the VAR to allow for cointegration of \mathbf{y}_t ?
- We write the VAR(p) as

$$\mathbf{y}_t - \mathbf{a} - \mathbf{d}t = \mathbf{w}_t \quad (23)$$

$$\Phi(L)\mathbf{w}_t = \mathbf{v}_t \quad (24)$$

which is equivalent to

$$\Phi(L)\mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\gamma}t + \mathbf{v}_t \quad (25)$$

for $\boldsymbol{\alpha} = \Phi(1)\mathbf{a} - (\sum_{j=1}^p j\Phi_j)\mathbf{d}$ and $\boldsymbol{\gamma} = \Phi(1)\mathbf{d}$.

I(1) VAR and Reduced Rank Condition

- $\mathbf{w}_t \sim I(1)$ and $\mathbf{v}_t \sim I(0)$.
- Multiply both sides of (24) by $\Delta = 1 - L$:
 $\Phi(L)\Delta\mathbf{w}_t = (1 - L)\mathbf{v}_t$, and substitute $\Psi(L)\mathbf{v}_t$ (Wold representation) for $\Delta\mathbf{w}_t$: $\Phi(L)\Psi(L)\mathbf{v}_t = (1 - L)\mathbf{v}_t$.
This must be true for any \mathbf{v}_t , hence

$$\Phi(L)\Psi(L) = (1 - L)\mathbf{I}_n \quad (26)$$

- Let $L = 1$, we see that $\Phi(1)\Psi(1) = 0$.
- For cointegration, we need $\Psi(L)$ to be one-summable and $\text{rank}(\Psi(1)) = n - h$.
- The essential condition for this is that $\text{rank}(\Phi(1)) = h < n$ (reduced rank).
- Denote $\Pi = -\Phi(1)$ hereafter. $\text{rank}(\Pi) = \text{rank}(\Phi(1))$.

- If $\text{rank}(\boldsymbol{\Pi}) = h$, there exist two $n \times h$ matrices $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta}$, each of rank h , such that

$$\boldsymbol{\Pi} = \tilde{\boldsymbol{\alpha}}\boldsymbol{\beta}' \quad (27)$$

Hence, $\tilde{\boldsymbol{\alpha}}\boldsymbol{\beta}'\boldsymbol{\Psi}(1) = \mathbf{0} \Rightarrow \boldsymbol{\beta}'\boldsymbol{\Psi}(1) = \mathbf{0}$, which shows that the rows of $\boldsymbol{\beta}'$ are cointegrating vectors.

- The matrices $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta}$ are not uniquely defined, since $\tilde{\boldsymbol{\alpha}}\boldsymbol{\beta}' = \tilde{\boldsymbol{\alpha}}\mathbf{H}\mathbf{H}^{-1}\boldsymbol{\beta}'$ for any non-singular matrix \mathbf{H} (of dimension $h \times h$).
- For estimation, h^2 identification restrictions need to be imposed. For example, if $h = 1$, $\boldsymbol{\beta}' = (\beta_1, \beta_2)$ must be normalized to e.g. $(1, -b)$ where $b = -\beta_2/\beta_1$.

VECM Representation

- We have

$$\Delta \mathbf{y}_t = \tilde{\alpha} \beta' y_{t-1} + \alpha + \gamma t + \zeta_1 \Delta \mathbf{y}_{t-1} + \dots + \zeta_{p-1} \Delta \mathbf{y}_{t-p+1} + \mathbf{v}_t \quad (28)$$

which is the vector error-correction model (VECM).

- The variables $\beta' y_{t-1}$ are the "cointegrating errors" (or "disequilibrium terms") which are corrected for in each equation of the system through the "loading coefficients" in the matrix $\tilde{\alpha}$.
- If $\beta' y_t$ has no trend, then $\beta' \mathbf{d} = 0$ and $\gamma = -\tilde{\alpha} \beta' \mathbf{d} = 0$. In this case, the VECM does not include the linear trend term although it may be present in some elements of y_t as we see in equation (23).

A Cointegrated VAR(1)

- Consider the simple VAR(1) process:

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{v}_t, \quad (29)$$

with $\mathbf{A}(z) = \mathbf{I} - \mathbf{A}z$ has at least one root at $z = 1$ if $|\mathbf{A}(1)| = 0$. Equivalently, the corresponding $\boldsymbol{\Pi}$ has reduced rank where $\boldsymbol{\Pi} = \mathbf{A} - \mathbf{I}$.

- We assume that $\mathbf{y}_t \sim I(1)$. Very important!
- The corresponding VECM is

$$\Delta \mathbf{y}_t = \boldsymbol{\Pi} \mathbf{y}_{t-1} + \mathbf{v}_t \quad (30)$$

with $\boldsymbol{\Pi} = \tilde{\boldsymbol{\alpha}}\boldsymbol{\beta}'$, h is the CI rank.

- Since $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta}$ are both $n \times h$ full rank matrices, there exist $\tilde{\boldsymbol{\alpha}}_{\perp}$ and $\boldsymbol{\beta}_{\perp}$, which are both $n \times (n-h)$ matrices, s.t. $\tilde{\boldsymbol{\alpha}}'_{\perp}\tilde{\boldsymbol{\alpha}} = 0$ and $\boldsymbol{\beta}'_{\perp}\boldsymbol{\beta} = 0$.

Roots and Eigenvalues

- If (29) is stable (the lag polynomial $\mathbf{A}(L)$ is stable), the roots of $|\mathbf{A}(z)| = |\mathbf{I} - \mathbf{A}z| = 0$ are all outside the unit circle, i.e. $|z| > 1$.
- This is equivalent to the eigenvalue problem $|\lambda\mathbf{I} - \mathbf{A}| = 0$, which implies that the modulus of all the eigenvalues of \mathbf{A} matrix are all smaller than 1, i.e. $|\lambda| < 1$.
- If the system contains unit roots, we say, some of the roots or eigenvalues are equal to one.

Long-run relations

- Multiplying both sides of (30) by β' yields

$$\beta'(\mathbf{y}_t - \mathbf{y}_{t-1}) = \beta'\tilde{\alpha}\beta'\mathbf{y}_{t-1} + \beta'\mathbf{v}_t$$

$$\beta'\mathbf{y}_t = (\beta'\tilde{\alpha} + \mathbf{I})\beta'\mathbf{y}_{t-1} + \beta'\mathbf{v}_t$$

$$\rightarrow \mathbf{s}_t = \mathbf{B}\mathbf{s}_{t-1} + \boldsymbol{\eta}_t = \mathbf{B}^t\mathbf{s}_0 + \sum_{i=0}^{t-1} \mathbf{B}^i \boldsymbol{\eta}_{t-i} \quad (31)$$

$$= \sum_{i=0}^{\infty} \mathbf{B}^i \boldsymbol{\eta}_{t-i}, \quad (32)$$

where $\mathbf{s}_t = \beta'\mathbf{y}_t \sim \mathbf{I}(0)$, $\mathbf{B} = \beta'\tilde{\alpha} + \mathbf{I}$ and $\boldsymbol{\eta}_t = \beta'\mathbf{v}_t \sim \mathbf{I}(0)$.

- This process contains the linear combinations of \mathbf{y}_t , which are stationary or asymptotically stable process over time.
- β consists of h linearly independent vectors, and it is called long-run relations or cointegrating relations if $\beta'\mathbf{y}_t \sim \mathbf{I}(0)$.
- These linear combinations are not unique. For any $K \neq 0$, $K\beta'\mathbf{y}_t$ is also stationary.

Why $\beta'y_t$ is asymptotically stable?

- Due to the important assumption: $|\mathbf{A}(z) = 0|$ has $n - h$ unit roots and the other roots are outside the unit circle.
- $\Pi = \mathbf{A} - \mathbf{I} = -\mathbf{A}(1)$ has reduced rank and can be decomposed by $\tilde{\alpha}\beta'$. Cl rank is h . And $\mathbf{A} = \mathbf{I} + \tilde{\alpha}\beta'$.
- Check the following derivation carefully

$$\begin{aligned} |\mathbf{A}(z)| = 0 &\implies |(\beta, \beta_{\perp})' \mathbf{A}(z) (\beta, \beta_{\perp})| = 0 \\ &\implies \begin{vmatrix} \beta'\beta - \beta'\mathbf{A}\beta z & -\beta'\mathbf{A}\beta_{\perp}z \\ -\beta'_{\perp}\mathbf{A}\beta z & \beta'_{\perp}\beta_{\perp} - \beta'_{\perp}\mathbf{A}\beta_{\perp}z \end{vmatrix} = 0 \quad (33) \\ &\implies |\mathbf{I}_h - (\mathbf{I}_h + \beta'\tilde{\alpha})z| |\mathbf{I}_{n-h} - \mathbf{I}_{n-h}z| = 0 \quad (34) \end{aligned}$$

where $\mathbf{I}_h + \beta'\tilde{\alpha} = \mathbf{B}$

- The other roots ($|\mathbf{I}_h - (\mathbf{I}_h + \beta'\tilde{\alpha})z| = 0$) are outside the unit circle as assumed...

The Pushing Force

- Multiplying both sides of (30) by $\tilde{\alpha}'_{\perp}$ yields

$$\begin{aligned}\tilde{\alpha}'_{\perp} \Delta \mathbf{y}_t &= \tilde{\alpha}'_{\perp} \mathbf{v}_t \\ \tilde{\alpha}'_{\perp} \mathbf{y}_t &= \tilde{\alpha}'_{\perp} \sum_{i=0}^{t-1} \mathbf{v}_{t-i} + \tilde{\alpha}'_{\perp} \mathbf{y}_0.\end{aligned}\quad (35)$$

- $\tilde{\alpha}'_{\perp} \sum_{i=0}^{t-1} \mathbf{v}_{t-i}$ in (35) is the common stochastic trends of the I(1) VAR(1) process. We see that there are $n - h$ common stochastic trends, or unit roots in the vector system.
- The common stochastic trends are not unique. For any full rank $(n - h) \times (n - h)$ matrix K , $K \tilde{\alpha}'_{\perp} \sum_{i=0}^{t-1} \mathbf{v}_{t-i}$ common trends as well.
- $\tilde{\alpha}'_{\perp} \sum_{i=0}^{t-1} \mathbf{v}_{t-i}$ is also called the pushing force.

Granger's VMA Representation

- The beautiful identity:

$$\beta_{\perp}(\tilde{\alpha}'_{\perp}\beta_{\perp})^{-1}\tilde{\alpha}'_{\perp} + \tilde{\alpha}(\beta'\tilde{\alpha})^{-1}\beta' = \mathbf{I}. \quad (36)$$

Thus, we have

$$\begin{aligned}\mathbf{y}_t &= (\beta_{\perp}(\tilde{\alpha}'_{\perp}\beta_{\perp})^{-1}\tilde{\alpha}'_{\perp} + \tilde{\alpha}(\beta'\tilde{\alpha})^{-1}\beta')\mathbf{y}_t \\ &= (\beta_{\perp}(\tilde{\alpha}'_{\perp}\beta_{\perp})^{-1})\tilde{\alpha}'_{\perp}\mathbf{y}_t + (\tilde{\alpha}(\beta'\tilde{\alpha})^{-1})\beta'\mathbf{y}_t\end{aligned}$$

Replace the red parts by the common trends and the long-run relations:

$$y_t = \mathbf{C} \sum_{i=0}^{t-1} \mathbf{v}_{t-i} + \mathbf{C}\mathbf{y}_0 + \tilde{\alpha}(\beta'\tilde{\alpha})^{-1} \left(\sum_{i=0}^{\infty} \mathbf{B}^i \boldsymbol{\eta}_{t-i} \right) \quad (37)$$

where $\mathbf{C} = \beta_{\perp}(\tilde{\alpha}'_{\perp}\beta_{\perp})^{-1}\tilde{\alpha}'_{\perp}$.

Cointegrated VAR with Intercept and Trend

- We consider the following VAR(1) model

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\mu} + \boldsymbol{\delta}t + \mathbf{v}_t, \quad (38)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ may not be zero. $\mathbf{A}(z)$ contains r unit roots and the others roots are outside the unit circle.

- The corresponding VECM

$$\Delta\mathbf{y}_t = \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\mu} + \boldsymbol{\delta}t + \mathbf{v}_t, \quad (39)$$

where $\boldsymbol{\Pi} = \tilde{\boldsymbol{\alpha}}\boldsymbol{\beta}'$.

- $\boldsymbol{\beta}'\mathbf{y}_t$ is trend stationary. But the "one-summability" should be carefully checked.
- The pushing force may contain quadratic trend:

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}'_\perp \Delta\mathbf{y}_t &= \tilde{\boldsymbol{\alpha}}'_\perp \boldsymbol{\mu} + \tilde{\boldsymbol{\alpha}}'_\perp \boldsymbol{\delta}t + \tilde{\boldsymbol{\alpha}}'_\perp \mathbf{v}_t \\ \tilde{\boldsymbol{\alpha}}'_\perp \mathbf{y}_t &= \tilde{\boldsymbol{\alpha}}'_\perp \sum_{i=0}^{t-1} (\boldsymbol{\mu} + \boldsymbol{\delta}(t-i) + \mathbf{v}_{t-i}) + \tilde{\boldsymbol{\alpha}}'_\perp \mathbf{y}_0.\end{aligned} \quad (40)$$

The role of deterministic terms

- The Granger VMA representation is

$$\mathbf{y}_t = \mathbf{C}\mathbf{y}_0 + \mathbf{C} \sum_{i=0}^{t-1} (\boldsymbol{\mu} + \boldsymbol{\delta}(t-i) + \mathbf{v}_{t-i}) \quad (41)$$

$$+ \tilde{\alpha}(\beta' \tilde{\alpha})^{-1} \left(\sum_{i=0}^{\infty} \mathbf{B}^i \beta' (\boldsymbol{\mu} + \boldsymbol{\delta}(t-i) + \mathbf{v}_{t-i}) \right) \quad (42)$$

where $\mathbf{C} = \beta_{\perp} (\tilde{\alpha}' \beta_{\perp})^{-1} \tilde{\alpha}'_{\perp}$ and the last term is trend stationary.

- If $\boldsymbol{\delta} = \tilde{\alpha} \kappa$, where κ is an $h \times h$ matrix, there will be no quadratic trend in the system!
- Given $\boldsymbol{\delta} = 0$, if $\boldsymbol{\mu} = \tilde{\alpha} \gamma$, where γ is an $h \times h$ matrix, there will be no deterministic trend in the system!

Restricted intercept and trend

- Given CI h , the deterministic terms can be written as
$$\mathbf{d}_t = \boldsymbol{\mu} + \boldsymbol{\delta}t = \tilde{\boldsymbol{\alpha}}\kappa_0 + \tilde{\boldsymbol{\alpha}}_{\perp}\kappa_1 + (\tilde{\boldsymbol{\alpha}}\gamma_0 + \tilde{\boldsymbol{\alpha}}_{\perp}\gamma_1)t$$
- the following models (hypotheses) have nested relations:

$$H(h) : \mathbf{d}_t = \tilde{\boldsymbol{\alpha}}\kappa_0 + \tilde{\boldsymbol{\alpha}}_{\perp}\kappa_1 + (\tilde{\boldsymbol{\alpha}}\gamma_0 + \tilde{\boldsymbol{\alpha}}_{\perp}\gamma_1)t \quad (43)$$

$$H^*(h) : \mathbf{d}_t = \tilde{\boldsymbol{\alpha}}\kappa_0 + \tilde{\boldsymbol{\alpha}}_{\perp}\kappa_1 + \tilde{\boldsymbol{\alpha}}\gamma_0 t \text{ (no quadratic trend)} \quad (44)$$

$$H_1(h) : \mathbf{d}_t = \tilde{\boldsymbol{\alpha}}\kappa_0 + \tilde{\boldsymbol{\alpha}}_{\perp}\kappa_1 \text{ (no trend in } \boldsymbol{\beta}'y_t \text{)} \quad (45)$$

$$H_1^*(h) : \mathbf{d}_t = \tilde{\boldsymbol{\alpha}}\kappa_0 \text{ (no trend)} \quad (46)$$

$$H_2(h) : \mathbf{d}_t = 0 \text{ (no deterministic terms)} \quad (47)$$



To be continued! Thank you!

Time Series Econometrics, 2ST111

Lecture 11. Cointegration

Yukai Yang

Department of Statistics, Uppsala University

Outline of Today's Lecture

- Engle-Granger's Procedure for the Cointegration Analysis
- Johansen's Procedure for the Cointegration Analysis
- The Parameter Estimation and Reduced Rank Regression
- The Trace Test

Engle-Granger's Procedure

- Suppose that x_t and $y_t \sim I(1)$. This can be checked by for example the augmented DF test.
- If x_t and y_t can be cointegrated, then there exists a linear combination \mathbf{a} of the two variables, which can be normalized as $\mathbf{a} = (1, -a)'$, such that $\mathbf{a}'(x_t, y_t)' = x_t - ay_t \sim I(0)$.
- Therefore, if we regress the model as follows

$$x_t = c + by_t + u_t,$$

then $\hat{b} \xrightarrow{P} a$ in a super consistent way, and in particular the residuals should satisfy $\hat{u}_t \sim I(0)$, as u_t is $I(0)$.

- We test whether \hat{u}_t contains unit root, or $\hat{u}_t \sim I(1)$, and then make the conclusion whether x_t and y_t can be cointegrated.

Engle-Granger's Procedure

Remarks:

- When x_t and y_t are not cointegrated, \hat{u}_t is $I(1)$ almost surely. And \hat{b} will not converge to zero, though $b = 0$ (spurious regression).
- E-G procedure requires that the variables x_t , y_t and perhaps more, should be all $I(1)$. This makes sense because if $x_t \sim I(0)$ and $y_t \sim I(1)$, then there is naturally a linear combination $\mathbf{a} = (1, 0)$ such that $\mathbf{a}'(x_t, y_t)' \sim I(0)$, which is not cointegration.
- However, E-G procedure suffers from the problem that, if there are more than two variables x_t , y_t and z_t , there may be more than one linear combination, say, two linearly independent linear combinations, that makes the cointegration, but E-G can only find one of them.

Engle-Granger's Procedure

Remarks:

- Consider

$$x_t = \rho_1 u_t + \epsilon_{1t}; \quad y_t = \rho_2 u_t + \epsilon_{2t}; \quad z_t = \rho_3 u_t + \epsilon_{3t},$$

where $u_t \sim I(1)$, $\epsilon_{it} \sim I(0)$, and ρ_i are non-zero real numbers, $i = 1, 2, 3$.

- The three random variables share the same $I(1)$ process u_t . We can say that the vector system, or the system (x_t, y_t, z_t) , is pushed by the same underlying common stochastic trend. This gives normally nice economic insight.
- $\mathbf{a}_1 = (1, -\rho_1/\rho_2, 0)'$ is one cointegrating vector. $\mathbf{a}_2 = (1, 0, -\rho_1/\rho_3)'$ is another one. So is any non-zero linear combination of \mathbf{a}_1 and \mathbf{a}_2 .
- E-G procedure just finds the cointegrating vector which produces the smallest residual-sum-squares for the regression model

$$x_t = c + b_1 y_t + b_2 z_t + u_t$$

$$Y_t \sim Z(1)$$

$$\alpha'_1 Y_t \sim Z(0) \quad \alpha_1 \alpha_1 + \alpha_2 \alpha_2$$

$$\alpha'_2 Y_t \sim Z(0)$$

The Characteristic Polynomial

Before introducing the Johansen's procedure, we have to introduce

- The n -dimensional VAR(p)

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \dots + A_p y_{t-p} + \Phi D_t + \varepsilon_t \quad (1)$$

where D_t is vector containing the deterministic terms and any possible exogenous variables, ε_t is i.i.d. $\sim N_n(0, \Omega)$, $t = 1, \dots, T$ with initial variables y_{-p+1}, \dots, y_0 .

- It can be written as follows

$$A(L)y_t = \Phi D_t + \varepsilon_t \quad (2)$$

where $A(L) = I_n - A_1 L - A_2 L^2 - \dots - A_p L^p$ is the lag polynomial. $A(z)$ is termed the **characteristic polynomial** of the dynamic system.

The Basic Assumption

- Johansen's cointegration procedure presumes that the data vector y_t is no more than $I(1)$. More specifically, if $y_t = (y_{1t}, \dots, y_{nt})'$, then either $y_{it} \sim I(0)$ or $\Delta y_{it} \sim I(0)$ holds for $i = 1, \dots, n$.
- To ensure this, Johansen's procedure employs the following assumption explicitly

Assumption 1 in Johansen (1995)

The characteristic polynomial satisfies the condition that if $|A(z)| = 0$, then either $|z| > 1$ or $z = 1$.

- $|z| < 1$ is termed the explosive root. $|z| = 1$ is not necessarily a unit root. Instead, we call $|z| = 1$ but $z \neq 1$ a seasonal root which is ruled out.

The VECM Form

The corresponding VECM is:

$$\Delta y_t = \Pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \Phi D_t + \varepsilon_t \quad (3)$$

where $\Pi = \sum_{i=1}^p A_i - I_n$ and $\Gamma_i = -\sum_{j=i+1}^p A_j$.

Remarks:

- If $y_t \sim I(1)$ and can be cointegrated, Π has reduced rank $r < n$ and can be decomposed as $\Pi = \alpha\beta'$. Note that the CI rank is r now!
- In estimation, we assume that we know r , though you can try to estimate using every possible r .
- I'd like to stress again, we presume that the system is at most $I(1)$.

Remarks

- $r = 0$ implies that $\Pi = 0$, and hence $r = 0$, $I(1)$ but no cointegration! The number of stochastic trends are $n - r = n$, and hence there are n unit roots in the system.
- $r = n$ implies that Π is full rank, and hence the number of stochastic trends are $n - r = 0$, NO unit root! Thus, $I(0)$ asymptotically stable! (It is not $I(1)$ but at most $I(1)$, which means it can only be $I(0)$! If you check the VECM, you will get the same conclusion.)
- A cointegrated VAR and its VECM should have a Π whose rank is an integer satisfying $1 \leq r < n$.
- If the rank r is testable (yes), based on the Johansen's procedure, we can test $r = 0$ (no cointegration), $r = n$ ($I(0)$, a more general DF test), and $r = r_0$ (the number of the linearly independent long-run relations), compared to E-G procedure.

The nonlinear regression model

- Denoting $Z_{0t} = \Delta y_t$, $Z_{1t} = y_{t-1}$ and $Z_{2t} = (\Delta y'_{t-1}, \Delta y'_{t-2}, \dots, \Delta y'_{t-p+1}, D'_t)'$, (3) can be written as:

$$Z_{0t} = \alpha \beta' Z_{1t} + \underline{\Psi} Z_{2t} + \varepsilon_t, \quad (4)$$

where $\Psi = (\Gamma_1, \Gamma_2, \dots, \Gamma_{p-1}, \Phi)$. *nuisance parameter*

- (4) is the regression model we will consider in the estimation.
- It is nonlinear regression due to the reduced rank property of Π .
- α and Ψ are called "short-run adjustments".
- People are more interested in the long-run parameter β , the short-run adjustment ΨZ_{2t} is more or less "nuisance parameter".
- ΨZ_{2t} can be removed, so that it looks like a simple VAR(1) model without deterministic terms.

The log-likelihood function

- The log-likelihood function is given by

$$\log L(\alpha, \beta, \Psi, \Omega) = -\frac{1}{2} T \log |\Omega| -$$

$$\frac{1}{2} \sum_{t=1}^T (Z_{0t} - \alpha \beta' Z_{1t} - \Psi Z_{2t})' \Omega^{-1} (Z_{0t} - \alpha \beta' Z_{1t} - \Psi Z_{2t}) \quad (5)$$

- The Maximum Likelihood estimators of the parameters $(\alpha, \beta, \Psi, \Omega)$ are obtained by maximizing this function.
- When ε_t is not Gaussian distributed, it is called "Quasi-Maximum Likelihood".
- In addition, I introduce the notation for the product moment matrices: $M_{ij} = T^{-1} \sum_{t=1}^T Z_{it} Z_{jt}'$, $i, j = 0, 1, 2$. Note that $M_{ij} = M_{ji}'$.

Estimating Ψ given α and β

- The first order condition (FOC) is for estimating Ψ are given by

$$\sum_{t=1}^T (Z_{0t} - \alpha\beta' Z_{1t} - \hat{\Psi} Z_{2t}) Z_{2t}' = 0 \quad (6)$$

- It can be rewritten as

$$M_{02} = \alpha\beta' M_{12} + \hat{\Psi} M_{22}$$

such that

$$\hat{\Psi}(\alpha, \beta) = M_{02} M_{22}^{-1} - \alpha\beta' M_{12} M_{22}^{-1} \quad (7)$$

- Note that Z_{it} are from the data set, and hence M_{ij} are always known.

A tricky way to get rid of Ψ

- Replace the Ψ by $\hat{\Psi}(\alpha, \beta)$ in (5), and get concentrated LL function

$$\log L(\alpha, \beta, \Omega) = \frac{1}{2} \sum_{t=1}^T (Z_{0t} - \alpha\beta' Z_{1t} - (\textcolor{red}{M_{02}M_{22}^{-1}} - \alpha\beta' \textcolor{red}{M_{12}M_{22}^{-1}})Z_{2t})'$$

$$\Omega^{-1}(Z_{0t} - \alpha\beta' Z_{1t} - (\textcolor{red}{M_{02}M_{22}^{-1}} - \alpha\beta' \textcolor{red}{M_{12}M_{22}^{-1}})Z_{2t}) + \dots$$

$$= \frac{1}{2} \sum_{t=1}^T ((Z_{0t} - \textcolor{red}{M_{02}M_{22}^{-1}}Z_{2t}) - \alpha\beta'(Z_{1t} - \textcolor{red}{M_{12}M_{22}^{-1}}Z_{2t}))'$$

$$\Omega^{-1}((Z_{0t} - \textcolor{red}{M_{02}M_{22}^{-1}}Z_{2t}) - \alpha\beta'(Z_{1t} - \textcolor{red}{M_{12}M_{22}^{-1}}Z_{2t})) + \dots$$

- This implies two simple linear regressions! The residuals of the two regressions are given by

$$R_{0t} = Z_{0t} - \textcolor{red}{M_{02}M_{22}^{-1}}Z_{2t} \tag{8}$$

$$R_{1t} = Z_{1t} - \textcolor{red}{M_{12}M_{22}^{-1}}Z_{2t} \tag{9}$$

A tricky way to get rid of Ψ (cont.)

- The two regressions are just: 1. regress Z_{0t} on Z_{2t} ; 2. regress Z_{1t} on Z_{2t} ; and then collect the residuals R_{0t} and R_{1t} .
- The concentrated log-likelihood function becomes

$$\log L(\alpha, \beta, \Omega) = \frac{1}{2} \sum_{t=1}^T (R_{0t} - \alpha\beta'R_{1t})'\Omega^{-1}(R_{0t} - \alpha\beta'R_{1t}) + \dots$$

- This concentrated log-likelihood function implies a Gaussian model (a new regression)

$$R_{0t} = \alpha\beta'R_{1t} + \tilde{\varepsilon}_t, \quad (10)$$

with $\tilde{\varepsilon}_t \sim N(0, \Omega)$. This implies that Π can be estimated by regressing R_{0t} on R_{1t} if it is full rank.

- The key point is that we can forget about Ψ by using R_{0t} and R_{1t} .
- Recall that Z_{0t} contains Δy_t , Z_{1t} contains y_{t-1} and Z_{2t} the others.
- If no Z_{2t} , $R_{0t} = Z_{0t}$, $R_{1t} = Z_{1t}$! Simple VAR(1) without anything else.

Estimating α and Ω given β

- Due to the reduced rank restriction, (10) can not be estimated directly! We resort to the concentrated likelihood again.
- Denote $S_{ij} = T^{-1} \sum_{t=1}^T R_{it} R'_{jt} = M_{ij} - M_{i2} M_{22}^{-1} M_{2j}$, $i = 0, 1$. (This notation will be used very often!)
- Supposing that β is known, α and Ω can be estimated by simple linear regression:

$$\hat{\alpha}(\beta) = S_{01}\beta(\beta' S_{11}\beta)^{-1}, \quad (11)$$

$$\begin{aligned}\hat{\Omega}(\beta) &= S_{00} - \hat{\alpha}(\beta)(\beta' S_{11}\beta)^{-1}\hat{\alpha}(\beta)' \\ &= S_{00} - S_{01}\beta(\beta' S_{11}\beta)^{-1}\beta' S_{10}.\end{aligned} \quad (12)$$

Reduced Rank Regression

- Actually this is the method suggested by Anderson (1951), which is termed Reduced Rank Regression (RRR). It solves the ML problem when there is a reduced rank parameter matrix in the vector system.
- We see that,

$$\arg \max_{\beta} \log L(\beta) = \arg \min_{\beta} \log L(\beta)^{-\frac{2}{\pi}}, \quad (13)$$

and apart from the constant term in the log-likelihood function,

$$\log L(\beta)^{-\frac{2}{\pi}} = |\hat{\Omega}(\beta)| = |S_{00} - S_{01}\beta(\beta' S_{11} \beta)^{-1}\beta' S_{10}|. \quad (14)$$

Reduced Rank Regression

- We apply the following identity

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}| |\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}| = |\Sigma_{22}| |\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}|. \quad (15)$$

Thus,

$$\begin{aligned} |S_{00} - S_{01}\beta(\beta' S_{11}\beta)^{-1}\beta' S_{10}| &= \\ |\beta' S_{11}\beta|^{-1} |S_{00}| |\beta' S_{11}\beta - \beta' S_{10}(S_{00})^{-1} S_{01}\beta| & \\ \propto |\beta' S_{11}\beta|^{-1} |\beta' [S_{11} - S_{10}(S_{00})^{-1} S_{01}] \beta| & \quad (16) \end{aligned}$$

- Minimizing this ratio implies an eigenvalue problem!

$$|\lambda I - A| = 0, \quad |B\lambda - A| = 0$$

Reduced Rand Regression

- (16) is minimized by solving the eigenvalue problem:

$$|\rho S_{11} - [S_{11} - S_{10}(S_{00})^{-1}S_{01}]| = 0, \quad (17)$$

and choose $\hat{\beta} = (v_1, v_2, \dots, v_r)$ the eigenvectors corresponding to the **smallest** r eigenvalues.

- Or equivalently, for $\lambda = 1 - \rho$, by solving the eigenvalue problem:

$$|\lambda S_{11} - S_{10}S_{00}^{-1}S_{01}| = 0 \quad (18)$$

and choose $\hat{\beta} = (v_1, v_2, \dots, v_r)$ the eigenvectors corresponding to the **biggest** r eigenvalues.

- Therefore, we have

$$\begin{aligned} \max \log L^{-\frac{2}{\pi}}(r) &= |S_{00}| |\hat{\beta}' S_{11} \hat{\beta}|^{-1} |\hat{\beta}' [S_{11} - S_{10}(S_{00})^{-1}S_{01}] \hat{\beta}| \\ &= |S_{00}| \prod_{i=1}^r \hat{\rho}_i = |S_{00}| \prod_{i=1}^r (1 - \hat{\lambda}_i) \end{aligned} \quad (19)$$

The ML estimator and the LR test

- There is an identification problem for $\hat{\beta}$, since the value of the likelihood will not be changed for any $\tilde{\beta} = \hat{\beta}\kappa$, for any $r \times r$ full rank κ . Eigenvalue decomposition is not unique in the same manner. We can just use the eigenvectors the software gives, or choose some normalization scheme.
- Immediately, we have the Likelihood ratio test for $H(r) : \text{rank}(\Pi) = r$ against $H(n) : \text{rank}(\Pi) = n$ based on (19)

$$Q(H(r)|H(n))^{-\frac{2}{T}} = \frac{|S_{00}| \prod_{i=1}^r (1 - \hat{\lambda}_i)}{|S_{00}| \prod_{i=1}^M (1 - \hat{\lambda}_i)}. \quad (20)$$

This is the well-known Johansen's "trace" test statistic (it is an LR test):

$$-2 \log Q(H(r)|H(n)) = -T \sum_{i=r+1}^M (1 - \hat{\lambda}_i). \quad (21)$$

The Johansen's trace test

- One may ask why (24) is called "trace". The reason is that, as $T \rightarrow \infty$, it converges in distribution to a trace (DF test with $M - r$ degrees of freedom):

$$-2 \log Q(H(r)|H(n)) \xrightarrow{d} \text{tr} \left\{ \int_0^1 (\mathrm{d}W) F' \left(\int_0^1 FF' \mathrm{d}u \right)^{-1} \int_0^1 F(\mathrm{d}W)' \right\} \quad (22)$$

where W stands for the standard multivariate Wiener process with dimension $n - r$. F is a random function whose dimension depends on the deterministic term.

- The distribution of this trace is not necessarily standard, and it depends greatly on the functional form of F .
- The functional form of F is determined by the deterministic term.
- Recall the possible forms of the deterministic term in last section. There are five possible forms of F .
- Critical values are generated by means of simulation.

The Johansen's trace test (cont.)

- In particular, when $H_2(r) : d_t = 0, F = W$. A simple DF-type test, easy to simulate.
- The LR test does not depend on the choice of Γ_j , i.e. the lag length, which means there is no augmented DF!
- The test procedure for determining the rank is applied in a sequential way.
- Estimate using the RRR method only once, and find the $\lambda_i, i = 1, \dots, n$. Sort λ_i in decreasing order!
- $\Pi(r_1)$ is a special case of $\Pi(r_0)$ if $r_1 < r_0$, and of course is nested by the latter one. The test procedure is conducted by the sequence: test $r = n - 1$ against $r = n$, if accepted then test $r = n - 2$ against $r = n$... until rejection.
- This procedure can also be understood as: test $r \leq n - 1$ against $r > n - 1$, test $r \leq n - 2$ against $r > n - 2$... until rejection.

The Johansen's trace test (cont.)

- Different from the Granger's procedure, Johansen's procedure can find more than one cointegration relation.
- It is flexible in the sense that some variables can be $I(0)$ and the limit distribution is independent of the choice of the lag.
- Univariate unit root test finds one root, whereas the Johansen's test find how many common stochastic trends (several roots).
- It suffers from the size distortion problem in finite sample cases when the dimension grow up. The possible solutions are: 1. apply the possible Bartlett corrections; or 2. use the bootstrapping method.
- For details, see theorem 11.1 in Johansen (1995).

Hypothesis testing for long-run parameters

- The null hypothesis $H_0 : \beta = H\varphi$ is a very tricky way to represent the equivalent hypothesis $R'\beta = 0$, since $R'H = 0$ and they are actually orthogonal to each other.
- φ contains the freely varying parameters in β .
- We impose $n - s$ restrictions on the space spanned by β , $\implies R$ is $n \times (n - s)$, and hence, $\implies H$ is $n \times s$. H and R span the null space to each other.
- By replacing $\beta = H\varphi$, the model can be estimated by means of RRR and the null hypothesis can be tested.
- The new eigenvalue problem:

$$|\lambda^* H' S_{11} H - H' S_{10} S_{00}^{-1} S_{01} H| = 0 \quad (23)$$

- The LR test statistic for $\beta = H\varphi$:

$$-2 \log Q(H_0 | H(r)) = T \sum_{i=1}^r \log \{(1 - \lambda_i^*) / (1 - \hat{\lambda}_i)\}. \quad (24)$$

which is asymptotically χ^2 distributed with degrees of freedom



Thank you for taking the course!