Lecture 1: Difference Equation

First Order Deterministic Difference Equation

1. Consider a linear first order deterministic difference equation $y_t = \phi y_{t-1} + \phi_0$.

2. Solving the difference equation amounts to expressing $y_t$ as a function of $t$ and $y_0$, the initial value.

3. In general, the solution (for any difference equation) is composed of two parts

$$y_t = y^c_t + y^e_t,$$

where $y^e$ measures the (intertemporal) equilibrium, $y^c_t$ measures the deviation from the equilibrium.

4. We obtain $y^e$ by letting it first equal a constant number. If this fails, then try a linear function of time, a quadratic function of time, and so on.

   (a) If $\phi \neq 1$, $y^e = \frac{\phi_0}{1-\phi}$. So the equilibrium is a fixed point.

   (b) If $\phi = 1$, $y^e = \phi_0 t$. This is a moving equilibrium.

5. The complementary (homogeneous) form of the difference equation is $y^c_t = \phi y^c_{t-1}$, and we get $y^c_t$ from this complementary form

   (a) We guess $y^c_t$ takes the form of $y^c_t = kx^t$, where both $k$ and $x$ are unknown.

   Substituting this guess solution to the complementary form we have

   $$x - \phi = 0.$$ 

   This is the characteristic equation of the difference equation. The solution of the characteristic equation, $x$, is called characteristic root. In this case,

   $$x = \phi$$

   is the characteristic root.

   (b) Now we know $y^c_t = k\phi^t$ and $y_t = y^c_t + y^e = k\phi^t + y^e$. 

2
6. We use the initial value \( y_0 \) to determine \( k \). Letting \( t = 0 \) we get

\[
k = y_0 - y^e
\]

7. Finally, we solve the difference equation, and the solution is

\[
y_t = \begin{cases} 
\left( y_0 - \frac{\phi_0}{1-\phi} \right) \phi^t + \frac{\phi_0}{1-\phi}, & (\phi \neq 1) \\
y_0 + \phi_0 t, & (\phi = 1) 
\end{cases}
\]

8. Remarks

(a) The solution depends on whether \( \phi \) equals one.

(b) If \( \phi = 1 \), the solution always diverges, i.e., \( \lim_{t \to \infty} y_t = \infty \)

(c) If \( \phi \neq 1 \), the solution can have different time paths.

i. If

\[ |\phi| = |x| < 1, \]

i.e., the characteristic root is less than one in absolute value, then \( \lim_{t \to \infty} y_t = y^e \). The solution converges, and converges to the equilibrium.

ii. If \( \phi < 0 \), the time path of \( y_t \) shows oscillation.

9. So if \(-1 < \phi < 0\), \( y_t \) converges to equilibrium in oscillation. This is consistent the observed “overshooting” behavior of some economic variables.

10. Example:

   - Consider the Cobweb Model, which is composed of three equations

   \[
   \begin{align*}
   Q^D_t &= a - bP_t, \quad (a > 0, b > 0) \\
   Q^S_t &= c + dP_{t-1}, \quad (c > 0, d > 0) \\
   Q^D_t &= Q^S_t
   \end{align*}
   \]

   where the quantity demanded depends negatively on the current price and the quantity supplied depends positively on the previous price (think of supply of agricultural product). The third equation is market-clearing condition.
• This model implies a first order difference equation for $P_t$

$$P_t = \left( -\frac{d}{b} \right) P_{t-1} + \frac{a-c}{b}$$

• For this equation we know

$$y^e = \frac{a-c}{b+d}$$

$$y^c_t = k \left( -\frac{d}{b} \right)^t$$

$$y_t = \left( y_0 - \frac{a-c}{b+d} \right) \left( -\frac{d}{b} \right)^t + \frac{a-c}{b+d}$$

• We exam $\phi = -\frac{d}{b}$ closely. First, $\phi \neq 1$ so the equilibrium is a fixed point. Second, whether the price converges or not depends on if $| -\frac{d}{b} | < 1$. We can see if $d < b$ the price converges. Moreover, because $-\frac{d}{b} < 0$ the price will converges in oscillation. That means, in short run the price will overshoot its equilibrium level. But eventually the price will reach its equilibrium.

11. (Homework) Consider a market-clearing model with adaptive expectation. This model consists of three equations:

$$Q^D_t = a - bP_t, \ (a > 0, b > 0)$$

$$Q^S_t = c + dE_p_t, \ (c > 0, d > 0)$$

$$E_p_t = E_{p_{t-1}} + \lambda(P_{t-1} - E_{p_{t-1}}), \ (0 < \lambda < 1)$$

where $E_{p_t}$ is the expected price level at period $i$. The third equation shows how individuals form the expectation about the price. In this case, people add adjustment (the adjustment speed is measured by $\lambda$) to their previous expectation. Please show the condition under which the price converges.

12. Now consider a general nonlinear first order deterministic difference equation

$$y_t = f(y_{t-1}) + \phi_0,$$

where $f()$ is a general nonlinear function.
13. \( y^e \) can be obtained by solving the equation

\[
y^e = f(y^e) + \phi_0.
\]

This equation may have multiple roots, indicating multiple equilibria.

14. For the \( i \)-th equilibrium \( y^e_i \), \( y_t \) converges to it only when

\[
|f'(y^e_i)| < 1,
\]

where \( f' \) denotes the first-order derivative.

(a) If \( 0 < f'(y^e_i) < 1 \), \( y_t \) converges monotonically.

(b) If \( -1 < f'(y^e_i) < 0 \), \( y_t \) converges in oscillation.

15. Alternatively, we can use the phase diagram to analyze the difference equation.

(a) We put \( y_{t-1} \) on the horizontal axis, and \( y_t \) on the vertical axis.

(b) Drawing a 45 degree line, representing \( y_{t-1} = y_t \).

(c) Drawing a curve representing \( y_t = f(y_{t-1}) \), called the phase curve.

(d) The intersection of phase curve and the 45 degree line gives the equilibrium.

(e) The absolute value of the slope of the phase curve at the equilibrium determines whether convergence occurs or not.

(f) The sign of the slope of the phase curve at the equilibrium determines whether oscillation occurs or not.

16. (Homework) Please use the phase diagram to analyze the first order linear difference equation.

**Second Order Deterministic Difference Equation**

1. Consider a linear second order deterministic difference equation

\[
y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_0.
\]

2. First we obtain the equilibrium, \( y^e \), and there are three possibilities.
(a) If \( \phi_1 + \phi_2 \neq 1 \), \( y^e = \frac{\phi_0}{1 - \phi_1 - \phi_2} \). This is a point equilibrium.

(b) If \( \phi_1 + \phi_2 = 1 \), \( \phi_2 \neq -1 \), \( y^e = \left( \frac{\phi_0}{\phi_1 + 2\phi_2} \right) t \). This is a linear trend equilibrium.

(c) If \( \phi_1 + \phi_2 = 1 \), \( \phi_2 = -1 \), \( y^e = \phi_2^2 t^2 \). This is a quadratic trend equilibrium.

3. The complementary (homogeneous) form is

\[ y'_t = \phi_1 y'_{t-1} + \phi_2 y'_{t-2}, \]

for which the characteristic equation is

\[ x^2 - \phi_1 x - \phi_2 = 0. \]

This quadratic equation has two roots \( x_1 \) and \( x_2 \).

(a) If \( x_1 \neq x_2 \), \( y'_t = k_1 x_1^t + k_2 x_2^t \)

(b) If \( x_1 = x_2 \), \( y'_t = k_1 x_1^t + k_2 t x_1^t \)

where \( k_1 \) and \( k_2 \) are determined by \( y_0 \) and \( y_1 \).

4. \( y_t \) converges to its equilibrium only when both characteristic roots are less than one in absolute values (or lies inside the unit circle), i.e.,

\[ |x_1| < 1, |x_2| < 1, \quad \text{(stability condition)} \quad (3) \]

5. Now we look more closely at the complementary form \( y'_t = \phi_1 y'_{t-1} + \phi_2 y'_{t-2} \)

(a) The characteristic equation is \( x^2 - \phi_1 x - \phi_2 = 0 \)

(b) If \( \phi_1^2 + 4\phi_2 > 0 \), then the characteristic roots are real, distinct, and given by

\[ x_{1,2} = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2} \]

(c) If \( \phi_1^2 + 4\phi_2 = 0 \), there are duplicate roots, and \( x_1 = x_2 = \frac{\phi_1}{2} \)

(d) (Optional) If \( \phi_1^2 + 4\phi_2 < 0 \), then the characteristic roots are complex, conjugate, and given by

\[ x_{1,2} = a \pm bi, \quad a = \frac{\phi_1}{2}, \quad b = \frac{\sqrt{-\phi_1^2 - 4\phi_2}}{2} \]
Alternatively, we can write the complex roots using polar coordinates as

\[ x_{1,2} = R(\cos \theta \pm i \sin \theta), \quad R = \sqrt{a^2 + b^2}, \quad \tan \theta = \frac{b}{a} \]

Using the Ruler Relations

\[ e^{\pm i\theta} = \cos \theta \pm i \sin \theta \]

we can also write the complex roots as

\[ x_{1,2} = Re^{\pm i\theta} \]

Therefore

\[ x_{1,2}^t = R^t e^{\pm it\theta} = R^t(\cos(t\theta) \pm i \sin(t\theta)) \quad (4) \]

6. In general, a characteristic root that lies inside the unit circle (less than one in absolute value) can take three forms:

(a) \( 0 < x < 1 \), then \( x^t \) converges to zero exponentially with unchanged sign

(b) \( -1 < x < 0 \), then \( x^t \) converges to zero exponentially with alternating sign (or in oscillation)

(c) \( x = a + bi, \; |x| = R = \sqrt{a^2 + b^2} < 1 \), so \( x^t = R^t[\cos(t\theta) + i \sin(t\theta)] \). So \( x^t \) is a damped mixture of cosine and sine wave

The figure below plots \( |x|^t \) for those cases

7. To sum up, the solution for the linear second order deterministic difference equation takes the form of

\[ y_t = k_1 x_1^t + k_2 x_2^t + y^e. \]

Only under the stability condition the solution converges, i.e.,

\[ \lim_{t \to \infty} y_t = y^e \quad \text{only if} \quad |x_1| < 1, |x_2| < 1, \]

where \( x_i \) is the (characteristic) root for the characteristic equation:

\[ x^2 - \phi_1 x - \phi_2 = 0. \]
Figure 1: Time Path of Characteristic Roots
The equilibrium, $y^e$, can be a point, a line or a curve. If convergence occurs, it may show oscillation (overshooting).

8. (Homework) Consider a market model with inventory, which consists of the two equations:

$$Q^D_t - Q^S_t = -a(P_t - P_{t-1}), \ (a > 0)$$
$$P_t - P_{t-1} = b(Q^D_{t-1} - Q^S_{t-1}), \ (b > 0)$$

where the first equation says that excess demand will respond negatively to change in price. The second equation says that price change depends positively on previous excess demand. Please show that this model implies a second order difference equation for $P_t$. Please show the condition under which the $P_t$ converges to its equilibrium level.

**First Order Stochastic Difference Equation: AR and MA Form**

1. The solution of the deterministic difference equation is a deterministic function of time. In order to introduce randomness, we add a white noise shock (or disturbance, error term) $e_t$ into the deterministic difference equation. The result is a stochastic difference equation.

2. The white noise has the properties of

   (a) $Ee_t = 0, \forall t$
   
   (b) $Ee_t^2 = \sigma^2, \forall t$
   
   (c) $Ee_te_{t-i} = 0, \forall t,i$

So the white noise has zero mean, variance of $\sigma^2$, and lacks serial correlation.

3. Consider a first-order stochastic difference equation

$$y_t = \phi y_{t-1} + \phi_0 + e_t.$$ 

By recursive substitution we find the solution as

$$y_t = \phi^t y_0 + \phi_0 (1 + \ldots + \phi^{t-1}) + [\phi^{t-1} e_1 + \ldots + e_t]. \quad (5)$$
We see $\phi^t y_0 + \phi_0 (1 + \ldots + \phi^{t-1})$ approaches the solution for the first order deterministic difference equation. So the new part is $\phi^{t-1} e_t + \ldots + e_t$, which captures the effects of all random shocks.

4. We know that a first order deterministic difference equation is stable (or its solution converges) if $|\phi| < 1$. For stochastic difference equation, this stability condition is relabeled as stationarity condition, under which the effect of past shock on present value dies out eventually, i.e.,

$$\lim_{j \to \infty} \frac{\partial y_t}{\partial e_{t-j}} = \lim_{j \to \infty} \phi^j = 0.$$ 

The ratio of $\frac{\partial y_t}{\partial e_{t-j}}$ is called the impulse response. It is a function of lag, $j$. So it is also called impulse response function (IRF).

5. In terms of time series analysis, the stochastic difference equation is also called the autoregressive (AR) form for the time series $y_t$, or, just AR process for short. Accordingly, formula (5) is called the moving-average (MA) form with truncation (at $y_0$). If the series does not stop at $y_0$, then we can keep moving backward, and obtain the infinite-order MA form (MA($\infty$))

$$y_t = y^c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots$$ 

$$= y^c + e_t + \sum_{i=1}^{\infty} \theta_i e_{t-i}. \quad (6)$$

where

$$y^c = \phi_0 (1 + \phi + \phi^2 + \ldots).$$

The limit of $y^c$ may or may not exist.

6. We can show $y^c < \infty$ only when $|\phi| = |x| < 1$. In that case

$$y^c = \phi_0 (1 + \phi + \phi^2 + \ldots) = \frac{\phi_0}{1 - \phi}.$$ 

Then the mean of $y_t$ exists, and equals

$$\mu_y = E y_t = E \left( y^c + e_t + \sum_{i=1}^{\infty} \theta_i e_{t-i} \right) = y^c = \frac{\phi_0}{1 - \phi}.$$
Put differently the mean of \( y_t \) does not exists when the stationarity (stability) condition, \( |\phi| < 1 \), fails.

7. We can also show under the stationarity condition the variance of \( y_t \) exists.

(a) Note \( \theta_j = \phi^j \) for AR(1) process.

(b) The properties of white noise indicate that

\[
\text{var}(y_t) = E(y_t - \mu_y)^2 = \sigma^2(1 + \theta_1^2 + \theta_2^2 + \ldots) = \sigma^2(1 + \phi^2 + \phi^4 + \ldots) = \sigma^2 \frac{1}{1 - \phi^2} < \infty
\]

The last equality follows because \( |\phi| < 1 \) implies that \( \phi^2 = |\phi||\phi| = 1 \), so the series of \( (1 + \phi^2 + \phi^4 + \ldots) \) converges.

8. Finally, we can show that under the stationary condition,

\[
\rho_j = \text{cov}(y_t, y_{t-j}) = E(y_t - \mu_y)(y_{t-j} - \mu_y) = \sigma^2(\theta_{t-j}\theta_0 + \theta_{t-j-1}\theta_1 + \ldots) \quad (\theta_0 \equiv 1) = \sigma^2(\phi^j + \phi^{j+2} + \ldots) = \sigma^2 \frac{\phi^j}{1 - \phi^2}
\]

So the (auto)covariance is finite, and is function of \( j \), the lag. The covariance is not function of \( t \).

9. The fact that \( \rho_j \neq 0 \) implies that \( y_{t-j} \) and \( y_t \) are (linearly) related, so we can use the former to predict the latter.

10. However, the linear association between the current and past dies off eventually. Thus long term prediction is ought to be less precise than short term prediction. Mathemat-
ically, the following is true

\[
\lim_{j \to \infty} |\rho_j| = 0
\]

\[
\sum_j |\rho_j| = \sum_j \sigma^2 \left| \frac{\phi^j}{1 - \phi^2} \right| = \frac{\sigma^2}{1 - \phi^2} \sum_j |\phi^j| < \infty
\]

Later we can use this summability of autocovariance to derive the asymptotic theory.

11. A time series is (weakly) stationary if it has finite variance, and if its covariance is independent of \(t\). So we just show \(y_t = \phi y_{t-1} + \phi_0 + e_t\) is stationary if \(|\phi| < 1\).

12. (Optional) A time series is ergodic if its covariance goes to zero sufficiently fast as lag rises. Again, \(y_t = \phi y_{t-1} + \phi_0 + e_t\) is ergodic because \(\sum_j |\rho_j| < \infty\) (for which a necessary but not sufficient condition is \(\lim_{j \to \infty} |\rho_j| = 0\)).

13. (Homework) Consider the Cobweb Model with random supply shock:

\[
\begin{align*}
Q^D_t &= a - bP_t, \ (a > 0, b > 0) \\
Q^S_t &= c + dP_{t-1} + e_t, \ (c > 0, d > 0) \\
Q^D_t &= Q^S_t
\end{align*}
\]

where \(e_t\) is a white noise supply shock (think of weather shock for agricultural product). Please show that this model implies a first order stochastic difference equation for the price (or the price is an AR(1) process). What is the condition under which the random price has finite mean and variance (weakly stationary)? Does this model imply predictability of the price?

\textbf{\(p\)-th Order Stochastic Difference Equation: AR\((p)\) and MA\((\infty)\) Form}

1. Consider a \(p\)-th order stochastic difference equation, which is also called AR\((p)\) process, given by

\[
\begin{align*}
y_t &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + e_t \\
&= \phi_0 + \sum_{i=1}^p \phi_i y_{t-i} + e_t.
\end{align*}
\]
2. We first need to address the issue that whether $y_t$ is statistically well-behaved (stationary). Toward that end, we focus on the characteristic equation

$$x^p - \phi_1 x^{p-1} - \phi_2 x^{p-2} - \ldots - \phi_p = 0 \quad (10)$$

Let $x_i, (i = 1, \ldots, p)$, be the characteristic roots. Then the solution for the (deterministic) difference equation takes the form of

$$y_t^c = k_1 x_1^t + \ldots + k_p x_p^t.$$

(a) Stationary condition is that all characteristic roots are less than one in absolute values, i.e.,

$$|x_i| < 1, \ (i = 1, \ldots, p).$$

(b) On the other hand, if one characteristic root equals one, called unit root, then it must be case that

$$\phi_1 + \phi_2 + \ldots + \phi_p = 1.$$

3. Now it is time to show formally why the stationarity condition matters.

(a) If $1 - \phi_1 - \phi_2 - \ldots - \phi_p \neq 0$, let $y^c = \frac{\phi_0}{1 - \phi_1 - \phi_2 - \ldots - \phi_p}$. We can always rewrite the original equation (8) as

$$(y_t - y^c) = \phi_1(y_{t-1} - y^c) + \phi_2(y_{t-2} - y^c) + \ldots + \phi_p(y_{t-p} - y^c) + e_t.$$

(b) By definition $y_t - y^c = y_t^c$. So the above equation is just the complementary form (with white noise shock):

$$y_t^c = \phi_1 y_{t-1}^c + \phi_2 y_{t-2}^c + \ldots + \phi_p y_{t-p}^c + e_t \quad (11)$$

(c) By recursive substitution we know the MA form for the complementary equation is

$$y_t^c = e_t + \sum_{i=1}^{\infty} \theta_i e_{t-i}.$$

Then it follows that

$$y_t = y^c + e_t + \sum_{i=1}^{\infty} \theta_i e_{t-i} \quad (12)$$
This is the MA(∞) form for the AR(p) process.

(d) The MA form (12) is convenient to compute the impulse response function (of $j$). We can show

$$\text{irf}_j = \frac{\partial y_t}{\partial e_{t-j}} = \theta_j, \quad (j = 1, 2, \ldots). \quad (13)$$

For a linear process, the impulse response depends on the lag $j$ only, and is independent of the time $t$. If you think that is not the case, you may try nonlinear model.

(e) Equation (13) indicates that the impulse response at lag $j$ is the same as the $j$-th MA coefficient $\theta_j$. This fact suggests that we may obtain the MA coefficient by simulating the impulse response. Explicitly, the so-called method of simulating impulse response works as follows. Let $e_t = 1, e_{t-j} = e_{t+j} = y_{t-j} = 0, (j = 1, 2, \ldots)$, and then simulate $y_t, y_{t+1}, y_{t+2}, \ldots$ using equation (11). The resulting sequence is the impulse response function. It is also the set of MA coefficients $\theta_0, \theta_1, \theta_2, \ldots$. This method is recommended for computing the MA coefficients in practice.

(f) The first $p$ MA coefficients have no particular pattern. However, equation (11) indicates that, starting at $p + 1$, the MA coefficient follows the deterministic difference equation of

$$\theta_j = \phi_1 \theta_{j-1} + \phi_2 \theta_{j-2} + \ldots + \phi_p \theta_{j-p}, \quad (j = p + 1, p + 2, \ldots). \quad (14)$$

Therefore, the MA coefficients eventually take the form of

$$\theta_j = k_1 x_j^1 + \ldots + k_p x_j^p \Rightarrow |\theta_j| \leq K|x_{max}|^j$$

where $x_{max}$ is the largest (in absolute value) characteristic root.

(g) The last inequality shows that $\theta_j$ is bounded in absolute value by a geometric series.

(h) Now we can prove that under the stationary condition ($|x_i| < 1, (i = 1, \ldots, p)$) the bounding geometric series is declining. So

$$\sum_j |\theta_j| \leq \sum_j K|x_{max}|^j = \frac{K}{1 - |x_{max}|} < \infty$$

14
\[
\sum_j \theta_j^2 \leq \left( \sum_j |\theta_j| \right)^2 < \infty
\]
\[
\sum_j \theta_j \theta_i \leq \infty
\]

4. Finally we can use the above three inequalities to show that under the stationary condition, the AR(p) process \( y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + e_t \) is stationary (with finite variance, and time-independent covariance)

5. We can also show that under the stationary condition, the AR(p) process is ergodic.

**Lag Operator**

1. Alternatively, we can obtain the MA coefficient using the method of undetermined coefficients.

2. Define the lag operator \( L \) as

\[
Ly_t \equiv y_{t-1}, \quad L^d y_t \equiv y_{t-d}, \quad Lc = c. \quad (15)
\]

The method of undetermined coefficients works as follows

(a) Define the polynomial in lag operator (treat \( L \) as a number) for the AR form as

\[
\phi(L) \equiv 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p.
\]

Then the AR form can be rewritten as

\[
\phi(L)y_t = e_t.
\]

(b) Dividing \( \phi(L) \) on both sides and we get the MA form

\[
y_t = \phi(L)^{-1}e_t.
\]

We also let the MA form be

\[
y_t = \theta(L)e_t,
\]
where $\theta(L) \equiv 1 + \theta_1 L + \theta_2 L + \ldots$. Therefore

$$\theta(L) \equiv \phi(L)^{-1}.$$ 

(c) The method of undetermined coefficients is based on the identity

$$\theta(L)\phi(L) \equiv 1$$

$$\Rightarrow (1 + \theta_1 L + \theta_2 L + \ldots)(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) = 1, \quad (16)$$

and equates the coefficients of $L^j$ to zero for $j = 1, 2, \ldots$.

3. (Homework) Obtain the MA coefficient for the AR(2) process $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$. Compare the answer you get by using the method of undetermined coefficient and the method of simulating impulse response.

4. Identity (16) shows clearly that there is no special pattern for the first $p$ MA coefficients. But after $\theta_p$, the MA coefficient follows the same deterministic difference equation as $y_t$, i.e.,

$$\theta_j = \phi_1 \theta_{j-1} + \phi_2 \theta_{j-2} + \ldots + \phi_p \theta_{j-p}, \quad (j = p + 1, p + 2, \ldots). \quad (17)$$

So $\theta_j$ can be obtained recursively.

5. (Optional) We can also obtain the MA coefficient (and IRF) using partial fraction:

$$\theta(L) \equiv \frac{1}{\phi(L)}$$

$$= \frac{c_1}{1 - x_1 L} + \frac{c_2}{1 - x_2 L} + \ldots + \frac{c_p}{1 - x_p L}$$

$$= c_1(1 + x_1 L + x_1^2 L^2 + \ldots) + \ldots + c_p(1 + x_p L + x_p^2 L^2 + \ldots).$$

where $x_i$ is the characteristic root for (11), and $c_i$ can be solved for. Eventually we can show

$$\text{irf}_j = c_1 x_1^j + c_2 x_2^j + \ldots + c_p x_p^j. \quad (18)$$

In words, the irf is dominated by the exponential of the largest characteristic root (if the root is complex, then damped sine wave).
Stationary AR(p) Process

Let’s summarize. For the AR(p) process given by (8)

1. The process is stationary if all the characteristic roots obtained from equation (10) are less than 1 in absolute value (stationary condition)

2. Under the stationary condition
   (a) The MA coefficients in the MA(∞) form are bounded in absolute value by geometrically declining series.
   (b) The impulse response is bounded in absolute value by geometrically declining series.
   (c) The process has finite mean and variance

Non-Stationary AR(p) Process

1. \( y_t \) given by (8) is non-stationary if at least one characteristic root equals one. A nonstationary series has the following properties
   (a) \( \phi_1 + \phi_2 + \ldots + \phi_p = 1 \)
   (b) The mean and variance are ill-defined. So computing sample average makes no sense.
   (c) The series is very smooth (or with long swing). If \( \phi_0 \neq 0 \), the series is trending.
   (d) The theoretical IRF never decays, so past shock has permanent effect.

2. \( y_t \) has one unit root (or is integrated of order one) if only one characteristic root (say \( x_1 \)) equals one, while all other roots are less than one in absolute value. We can show

\[
\phi(L)y_t = e_t \quad (19)
\]

\[
\Rightarrow (1 - x_1 L)(1 - x_2 L)\ldots(1 - x_p L)y_t = e_t \quad (20)
\]

\[
\Rightarrow (1 - L)(1 - x_2 L)\ldots(1 - x_p L)y_t = e_t \quad (21)
\]

\[
\Rightarrow (1 - x_2 L)\ldots(1 - x_p L)\Delta y_t = e_t, \quad (22)
\]

where the difference operator \( \Delta \) is defined as

\[
\Delta \equiv 1 - L, \quad \Delta^d \equiv (1 - L)^d. \quad (23)
\]
Because $|x_i| < 1$ for $i = 2, 3, \ldots, p$, the last equation shows that we can take difference once and end up with a stationary series if the original series has just one unit root.

3. For example, consider the AR(2) process given by $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$.
   (a) The characteristic equation is
   $$x^2 - \phi_1 x - \phi_2 = 0.$$  
   (b) Let $L \equiv x^{-1}$, the above equation can be equivalently written as
   $$1 - \phi_1 L - \phi_2 L^2 = 0.$$  
   (c) Suppose one characteristic root is one, while the other root is less than one in absolute value: $x_1 = 1, |x_2| < 1$.
   (d) Thus we can factor the polynomial $x^2 - \phi_1 x - \phi_2$ as
   $$x^2 - \phi_1 x - \phi_2 \equiv (x - x_1)(x - x_2) = (x - 1)(x - x_2).$$
   Because $L \equiv x^{-1}$, we can factor the polynomial $\phi(L) = 1 - \phi_1 L - \phi_2 L^2$ as
   $$\phi(L) = (1 - L)(1 - x_2 L).$$
   (e) So
   $$\phi(L)y_t = \epsilon_t \Rightarrow (1 - x_2 L)(1 - L)y_t = \epsilon_t \Rightarrow (1 - x_2 L)\Delta y_t = \epsilon_t.$$  
   This means that after taking difference once, the original non-stationary AR(2) process, which has one unit root, becomes a stationary AR(1) process given by
   $$\Delta y_t = x_2 \Delta y_{t-1} + \epsilon_t, \quad (|x_2| < 1)$$
   in which the AR coefficient is the second characteristic root.

4. In general, a series is integrated of order $d$, denoted by $I(d)$, if it becomes stationary after being taken difference $d$ times.

5. All integrated series are nonstationary, for which the theoretical irf never decays, whereas the estimated irf decays very slowly. The irf for stationary series decays very
quickly.

6. (Homework) Consider the AR(2) process of \( y_t = 0.8y_{t-1} - 0.12y_{t-2} + \epsilon_t \).

   (a) Find the AR lag polynomial \( \phi(L) \)
   (b) Find the roots for the characteristic equation
   (c) Is \( y_t \) stationary?
   (d) Using the method of undetermined coefficients, find \( \theta_2 \)
   (e) Using the method of simulating impulse response function, find \( \theta_3 \)

7. (Homework) Consider another AR(2) process of \( y_t = 1.2y_{t-1} - 0.2y_{t-2} + \epsilon_t \).

   (a) Find the AR lag polynomial \( \phi(L) \)
   (b) Find the roots for the characteristic equation
   (c) Is \( y_t \) stationary?
   (d) Using the method of undetermined coefficients, find \( \theta_2 \)
   (e) Using the method of simulating impulse response function, find \( \theta_3 \)

Simulation

The diagram below shows four simulated series.

1. The first diagram shows 100 observations of white noise: \( \epsilon_t, t = (1, \ldots, 100) \). The series fluctuates around 0, the mean. This phenomenon is called mean-reverting, a signal for stationarity. The series is choppy (another signal for stationarity), so knowing \( \epsilon_{t-i} \) does not help forecast \( \epsilon_t \). This is because \( E\epsilon_t\epsilon_{t-i} = 0 \).

2. The second diagram shows 100 observations of a stationary AR(1) process given by \( y_t = \phi y_{t-1} + \epsilon_t, \phi = 0.6 \). This AR process is stationary because \( |x| = |\phi| = 0.6 < 1 \). It is shown that the stationary AR(1) process is mean-reverting, and smoother (or less choppy) than the white noise. Observations with the same sign tend to cluster. This is because \( E\epsilon_t y_{t-i} > 0 \). Notice that \( \lim_{i \to \infty} E\epsilon_t y_{t-i} = 0 \) for this series. So serial correlation dies out eventually.

3. The third diagram shows 100 observations of a non-stationary AR(1) process (or unit-root process, or random walk) given by \( y_t = \phi y_{t-1} + \epsilon_t, \phi = 1 \). The series is shown not
Figure 2: Simulated Series
to be mean-reverting. Instead, the series just wanders away from the initial value (or has long swing). This AR process is non-stationary because $|x| = |\phi| = 1$. It is shown that the unit root process is even smoother than the stationary AR(1) process. This is because $\lim_{i \to \infty} E[y_t y_{t-i}] \neq 0$. For the unit root process, past shock has permanent effect on the present and future.

4. The fourth diagram shows 100 observations of a non-stationary AR(1) process (or unit-root process, or random walk with drift) given by $y_t = \phi_0 + \phi y_{t-1} + e_t$, $\phi = 1$, $\phi_0 = 0.2$. This AR process is non-stationary because $|x| = |\phi| = 1$. It is shown that the random walk with drift process is both trending and smooth (both are signals for non-stationarity). The trend is generated by the drift term $\phi_0$, and the smoothness is caused by the unit root.

In summary, choppiness and mean-reverting signify stationarity, whereas smoothness (long swing) and trend signify non-stationarity. The R codes to generate this figure is

n=100  # sample size  
wn=round(rnorm(n), 2)  # white noise; building block

sar1 = rep(0, n)  # initialize stationary AR(1)  
for (i in 2:n) {  
sar1[i] = 0.6* sar1[i-1] + wn[i]  
}

nsar1 = rep(0, n)  # nonstationary (unit-root) AR(1)  
for (i in 2:n) {  
nsar1[i] = 1* nsar1[i-1] + wn[i]  
}

nsar2 = rep(0, n)  # nonstationary (unit-root) AR(1) with intercept  
for (i in 2:n) {  
nsar2[i] = 0.2 + 1* nsar2[i-1] + wn[i]  
}
par(mfrow=c(2,2)) # plot series side-by-side
plot(wn,type="o", main = "white noise")
plot(sar1,type="o", main = "stationary AR(1)")
plot(nsar1,type="o", main = "nonstationary AR(1)")
plot(nsar2,type="o", main = "nonstationary AR(1) with intercept")
(Advanced, Optional) Linear Process

1. Define the finite-order filter \( \theta_n(L) \equiv 1 + \theta_1 L + \ldots + \theta_n L^n \).

2. A finite-order linear process, MA(n), is defined as

\[ y_{t,n} = \theta_n(L)e_t, \]

where \( e_t \) is white noise.

3. An MA(n) process is always stationary because its mean and variance are always finite, and its autocorrelation only depends on lag.

4. An indefinite-order linear process, MA(\( \infty \)), is defined as the mean-square limit of \( y_{t,n} \)

\[ \lim_{n \to \infty} y_{t,n} \equiv \theta_\infty(L)e_t = y_t. \]

5. Under the condition that

\[ \sum_j \theta_j^2 < \infty, \]

\( y_t \) is well defined.

6. A sufficient condition for \( \sum_j \theta_j^2 < \infty \) is

\[ \sum_j |\theta_j| < \infty, \text{ (absolute summability)}. \]

Using the fact that a sequence converges to a finite limit if and only if it is a Cauchy sequence:

\[ a_n \to a \iff |a_m - a_n| \to 0, \text{ as } m, n \to \infty. \]

Let \( a_n = \sum_{j=1}^n |\theta_j| \), then

\[ \sum_j |\theta_j| < \infty \Rightarrow \lim_{m,n \to \infty} \sum_{j=m}^n |\theta_j| = 0 \Rightarrow \lim_{m,n \to \infty} \sum_{j=m}^n \theta_j^2 = 0 \Rightarrow \sum_j \theta_j^2 < \infty. \]

7. Thus

\[ \lim_{m,n \to \infty} E(y_{t,n} - y_{t,m})^2 = 0, \]
which implies \( y_t \) converges in mean square to a random variable. That is,

\[
y_t = \lim_{n \to \infty} y_{t,n}.
\]

8. Therefore

\[
Ey_n = \lim_{n \to \infty} Ey_{t,n},
\]

and

\[
Ey_n^2 = \lim_{n \to \infty} Ey_{t,n}^2.
\]

9. Finally, we can show both \( Ey_n \) and \( Ey_n^2 \) are finite under the condition \( \sum_j |\theta_j| < \infty \).

10. For the MA(\( n \)) process we can show

\[
r_k = \text{cov}(y_t y_{t-k}) = \mathbb{E}\left( \sum_{i=0}^{\infty} \theta_i e_{t-i} \sum_{j=0}^{\infty} \theta_j e_{t-k-j} \right) = \sigma^2 \sum_{i=0}^{\infty} \theta_i \theta_{k+i},
\]

which is a function of \( k \), not a function of \( t \).

11. Using the condition that \( \sum_j |\theta_j| < \infty \) we can show

\[
\sum_k |r_k| = \sigma^2 \sum_{k=\infty}^{\infty} \left| \sum_{i=0}^{\infty} \theta_i \theta_{k+i} \right| \leq \sigma^2 \sum_{k=\infty}^{\infty} \sum_{i=0}^{\infty} |\theta_i| |\theta_{k+i}| \leq \sigma^2 \left( \sum_{i=0}^{\infty} |\theta_i| \right) \left( \sum_{j=\infty}^{\infty} |\theta_j| \right) \leq \infty
\]

This result can be used to prove the time average of a linear process with absolutely summable coefficients is ergodic for the mean.

12. Actually the above result is a special case of the general fact that the convolution of two absolutely summable sequence \( \{a_j\} \) and \( \{b_j\} \) defined by

\[
c_j = \sum_{k=\infty}^{\infty} a_k b_{k-j}
\]

is absolutely summable, see page 30 of Fuller (1996).
13. $\sum |r_k| < \infty$ implies that $r_k \to 0$ as $k \to \infty$.

14. A stationary AR(p) process is a special case of linear process. The condition $\sum |\theta_j| < \infty$ holds under the stationarity condition. Because when all characteristic roots lie inside the unit circle, $|\theta_j|$ is bounded by a geometrically declining series. So under the stationarity condition an AR(p) process is stationary.

(Advanced, Optional) Weak Law of Large Number (WLLN)

1. Consider a sample $(y_1, \ldots, y_n)$. If $y_t$ is governed by a stationary AR(p) process, then
   (a) $\mu = E y_t < \infty$
   (b) $\rho_0 = \text{var}(y_t) < \infty$
   (c) $|\rho_j| = |\text{cov}(y_t, y_{t-j})| \leq c \lambda^j, (0 < \lambda < 1)$
   (d) $\sum |\rho_j| < \infty$

2. The sample mean is defined as
   $$\bar{y}_n = \frac{\sum_{t=1}^{n} y_t}{n} = \sum_{t=1}^{n} \frac{1}{n} y_t.$$

3. Essentially, the sample mean is a filter for $y_t$:
   $$\bar{y}_n = \psi_n(L)y_n,$$
   where the filter coefficients are absolute summable:
   $$\psi_n(L) = \psi_0 + \psi_1 L + \ldots + \psi_{n-1} L^{n-1},$$
   $$\psi_i = \frac{1}{n}, (i = 0, 1, \ldots, n-1), \Rightarrow \sum |\psi_j| = 1 < \infty.$$

4. We can show
   $$E \bar{y}_\infty = \lim_{n \to \infty} E \bar{y}_n = \mu.$$

25
\[ \text{var}(\bar{y}_\infty) = \lim_{n \to \infty} \text{var}(\bar{y}_n) \]
\[ = \lim_{n \to \infty} \left( \frac{1}{n} \right) \left[ \rho_0 + 2 \left( 1 - \frac{1}{n} \right) \rho_1 + \ldots + 2 \left( 1 - \frac{n - 1}{n} \right) \rho_{n-1} \right] \]
\[ \leq \lim_{n \to \infty} \left( \frac{2}{n} \right) \sum_{j=0}^{n-1} |\rho_j| \]
\[ = \lim_{n \to \infty} \left( \frac{2}{n} \right) \lim_{n \to \infty} \sum_{j=0}^{n-1} |\rho_j| = 0 \]

The last equality follows because \( \lim_{n \to \infty} \sum_{j=0}^{n-1} |\rho_j| = \sum_{j=0}^{\infty} |\rho_j| < \infty. \)

5. It follows that
\[ \lim_{n \to \infty} \mathbb{E} (\bar{y}_n - \mu)^2 = \lim_{n \to \infty} \mathbb{E} (\bar{y}_n - \mathbb{E} \bar{y}_n + \mathbb{E} \bar{y}_n - \mu)^2 \]
\[ = \lim_{n \to \infty} \text{var}(\bar{y}_n) = 0 \]

So \( \bar{y}_n \) converges in mean square to \( \mu. \)

6. It follows that
\[ \lim_{n \to \infty} \mathbb{P} (|\bar{y}_n - \mu| > \epsilon) \leq \frac{\lim_{n \to \infty} \mathbb{E} (\bar{y}_n - \mu)^2}{\epsilon^2} = 0 \]
so \( \bar{y}_n \) converges in probability to \( \mu, \) i.e.,
\[ \bar{y}_n = \mu + o_P(1), \quad \text{(WLLN)} \]

7. We can generalize this WLLN to any stationary linear process
\[ y_t = \phi(L)e_t, \]
where a sufficient condition for stationarity is
\[ \sum_j |\phi_j| < \infty, \]
which implies that
\[ \sum_j |\rho_j| < \infty \]
which is the key condition that the variance of sample mean goes to zero in limit.
8. Actually it is possible that
\[ \frac{1}{n} \sum_{j=0}^{n-1} |\rho_j| \to 0 \]
under condition weaker than \( \sum_j |\rho_j| < \infty \). The weaker condition is that
\[ \rho_j \to 0 \Rightarrow \frac{1}{n} \sum_{j=0}^{n-1} |\rho_j| \to 0. \]

Proof: if \( \rho_j \to 0 \) then for any \( \frac{\epsilon}{2} > 0 \), there exist \( N(\epsilon) \) so that all \( |\rho_j| \leq \frac{\epsilon}{2} \), \( \forall j \geq N \). So
\[
\frac{1}{n} \sum_{j=0}^{n-1} |\rho_j| = \frac{1}{n} \left( \sum_{j=0}^{N-1} |\rho_j| + \sum_{j=N}^{n-1} |\rho_j| \right)
\leq \left( \frac{NM}{n} \right) + \left( \frac{n-N}{n} \right) \left( \frac{\epsilon}{2} \right), \quad (M = \max(|\rho_1|, \ldots, |\rho_{N-1}|) < \infty)
\leq \left( \frac{NM}{n} \right) + \left( \frac{\epsilon}{2} \right)
\]
By taking \( n \) large enough, \( \frac{NM}{n} \) can be made less than \( \frac{\epsilon}{2} \).

9. So for WLLN to work, we only need to check if \( \rho_j \to 0 \). WLLN does not require \( \sum_j |\rho_j| < \infty \).

(Advanced, Optional) Central Limit Theorem (CLT)

1. We can show that
\[ \text{var} \left( \sqrt{n} \bar{y}_n \right) = \rho_0 + 2 \sum_{j=1}^{n-1} \left( 1 - \frac{j}{n} \right) \rho_j \]

2. A sufficient condition for \( \lim_{n \to \infty} \text{var} \left( \sqrt{n} \bar{y}_n \right) < \infty \) is that
\[ \sum_j \rho_j < \infty. \]
Notice that this condition is weaker than \( \sum_j |\rho_j| < \infty \), but stronger than \( \rho_j \to 0 \).

3. We only need to prove that
\[ \sum_j \rho_j < \infty \Rightarrow \frac{1}{n} \sum_j j \rho_j \to 0. \]
Proof:

\[ \sum_j j \rho_j = \rho_1 + 2 \rho_2 + \ldots + n \rho_n \]

\[ = \sum_{j=1}^n \rho_j + \sum_{j=2}^n \rho_j + \ldots + \sum_{j=n}^n \rho_j \]

\[ = \sum_{j=1}^n \sum_{i=j}^n \rho_i \leq \sum_{j=1}^n \left| \sum_{i=j}^n \rho_i \right| \]

\[ = \sum_{j=1}^N \left| \sum_{i=j}^n \rho_i \right| + \sum_{j=N+1}^n \left| \sum_{i=j}^n \rho_i \right| \]

\[ \leq NM + (n - N) \varepsilon \]

The last inequality follows because \( \sum_j \rho_j < \infty \) implies that for any \( \varepsilon > 0 \), there exist \( N(\varepsilon) \) so that all \( \sum_{i=j}^n \rho_i \leq \frac{\varepsilon}{2}, \forall j \geq N \). Therefore

\[ \frac{1}{n} \sum_j j \rho_j \leq \frac{NM}{n} + \frac{n - N}{n} \frac{\varepsilon}{2} \leq \varepsilon, \]

by taking \( n \) large enough.

**Theorem 1 (CLT for MA(\( \infty \)))** Let \( y_t = \mu + \sum_{j=0}^{\infty} \theta_j e_{t-j} \) where \( e_t \) is white noise and \( \sum_j |\theta_j| < \infty \), then

\[ \sqrt{n}(\bar{y}_n - \mu) \Rightarrow N \left( 0, \sum_{j=-\infty}^{\infty} \rho_j \right), \]

where \( \sum_{j=-\infty}^{\infty} \rho_j \) denotes the long run variance.
Lecture 2: ARIMA Model

Identification

1. Time series is a set of repeated observations of the same variable.

2. We use a parametric model (with finite number of unknown parameters), i.e., linear ARMA model, to describe the stationary time series. For nonstationary series (with trend or smoothness) we use ARIMA model, where letter “I” in the middle stands for “integrated”. ARIMA = Taking difference first + ARMA for the differenced series.

3. The process of selecting a right ARMA model for a time series is called identification. Identification is guided by the autocorrelation function (ACF) and partial autocorrelation function (PACF). Later we will show PACF is function of ACF.

4. The rational for identification is that if two processes have the same ACF, they are the same process. Identification involves (i) first computing ACF and PACF for the real series, and then (ii) comparing them to those of candidate ARMA models.

5. The autocovariance function (of lag $j$), $r_j$, is defined as

$$ r_j = \text{cov}(y_t, y_{t-j}) = E[(y_t - E y_t)(y_{t-j} - E y_{t-j})], \quad j = 0, 1, 2, \ldots $$

We see $r_0 = \text{var}(y_t)$.

We get the autocorrelation function (ACF), $\rho_j$, by dividing $r_j$ by $r_0$:

$$ \text{ACF}_j = \rho_j = \frac{r_j}{r_0}, \quad j = 0, 1, 2, \ldots $$

We see $\rho_0 = 1$ by construction.

6. The partial autocorrelation at lag $j$, denoted by $\phi_{jj}$, is defined as the coefficient of $y_{t-j}$ of a $j$-th order autoregression.

(a) Find $\phi_{11}$

(b) Find $\phi_{22}$
(c) Find $\phi_{pp}$

7. Exercise: Consider the white noise $e_t \sim \text{iidn}(0, 1)$

(a) Find $r_0$

(b) Find $r_p$ for $p = 1, 2, 3$

(c) Find $\rho_p$ for $p = 1, 2, 3$

(d) Plot ACF

8. Exercise: Consider MA(2) process $y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}$

(a) Find $r_0$

(b) Find $r_p$ for $p = 1, 2, 3$

(c) Find $\rho_p$ for $p = 1, 2, 3$

(d) Plot ACF

9. Exercise: Consider AR(1) process $y_t = 0.6 y_{t-1} + e_t$

(a) Is $y_t$ stationary?

(b) Find $r_0$ by imposing stationarity

(c) Find $r_p$ for $p = 1, 2, 3$ by imposing stationarity

(d) Find $\rho_p$ for $p = 1, 2, 3$

(e) Plot ACF

10. In general for a stationary AR(p) process we can obtain ACF using Yule-Walker (system) equations. For example, consider an AR(3) of $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + e_t$. Multiplying both sides by $y_{t-1}, y_{t-2}, \ldots$, taking expectation, and dividing by $r_0$ we obtain

$$
\begin{align*}
\rho_1 &= \phi_1 + \phi_2 \rho_1 + \phi_3 \rho_2 \\
\rho_2 &= \phi_1 \rho_1 + \phi_2 + \phi_3 \rho_1 \\
\rho_3 &= \phi_1 \rho_2 + \phi_2 \rho_1 + \phi_3 \\
\rho_k &= \phi_1 \rho_{k-1} + \phi_2 \rho_{k-1} + \phi_3 \rho_{k-2} & k = 4, 5, \ldots
\end{align*}
$$

The first, second, and third equations can be solved for $\rho_1, \rho_2, \rho_3$. Next we can get $\rho_4, \rho_5, \ldots$ recursively using the fourth equation. Notice that the $\rho$s follows the same
(deterministic) difference equation as the original series. That means the characteristic equations are the same for \( \rho \) and \( y \). So if \( y_t \) is stationary, then the \( \rho s \) must eventually converge.

11. Exercise: Consider AR(2) process \( y_t = 0.8y_{t-1} - 0.12y_{t-2} + e_t \)

   (a) Is \( y_t \) stationary?
   (b) Find \( \rho_1, \rho_2 \) using Yule-Walker equation
   (c) Find \( \rho_3, \rho_4 \).
   (d) Plot ACF

12. Now we want to show PACF is function of ACF. This fact allows some computer packages to compute PACF using ACF. For example, consider an stationary AR(1) process \( y_t = \phi y_{t-1} + e_t, \ (|\phi| < 1) \)

   (a) Find \( \rho_0, \rho_1, \rho_2 \)
   (b) Find \( \text{PACF}_1 = \phi_{11} \) based on the regression \( y_t = \phi_{11}y_{t-1} + e_t \).
   (c) Find \( \text{PACF}_2 = \phi_{22} \) based on the regression \( y_t = \phi_{21}y_{t-1} + \phi_{22}y_{t-2} + e_t \). Hint: we apply the same procedure for Yule-Walker equation and get

\[
\begin{pmatrix}
1 & \rho_1 \\
\rho_1 & 1
\end{pmatrix}
\begin{pmatrix}
\phi_{21} \\
\phi_{22}
\end{pmatrix}
= \begin{pmatrix}
\rho_1 \\
\rho_2
\end{pmatrix}
\]

Next we apply the Cramer’s rule and get

\[
\phi_{22} = \frac{| 1 \ \rho_1 \ \\
\rho_1 \ \rho_2 \ |}{| 1 \ \rho_1 \ \\
\rho_1 \ \rho_2 \ |} = 0,
\]

because \( \rho_1 = \phi, \rho_2 = \phi^2 \) so the determinant in the numerator equals zero.

(d) In a similar fashion, we can show \( \phi_{33} = 0, \phi_{44} = 0, \ldots \) Thus, for an AR(1) process, PACF cuts off (equals zero) after lag 1.

13. Summary for a stationary AR or MA process

   (a) ACF cuts off after lag \( q \) for MA(\( q \)) process
(b) PACF cuts off after lag \( p \) for AR(\( p \)) process

(c) Past \( p \) lags, the ACF for AR(\( p \)) process is mixture of damped sine waves or exponentials

(d) Past \( q \) lags, the PACF for MA(\( q \)) process is mixture of damped sine waves or exponentials

14. In practice we need to estimate ACF and PACF using real observations. The formulae are

\[
\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T} \quad (26)
\]

\[
\hat{r}_j = \frac{\sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y})}{T}, \quad j = 0, 1, 2 \quad (27)
\]

\[
\hat{\rho}_j = \frac{\hat{r}_j}{\hat{r}_0}, \quad j = 0, 1, 2 \quad (28)
\]

The estimated PACF is the OLS coefficient of \( \hat{\phi}_{pp} \).

The diagram below shows ACF and PACF for three series.

1. The first row contains ACF and PACF for the (always stationary) MA(1) process \( y_t = e_t + 0.6 e_{t-1} \). Notice that \( \rho_0 = 1 \) by construction. The ACF cuts off at lag 1 (goes beyond the two dashed horizontal lines, which represent the standard error), a signal for MA(1). The estimated \( \hat{\rho}_1 \) is about 0.5, close to the true value of \( 0.6/(1 + 0.6^2) = 0.44 \).

2. The second row is for the stationary AR(1) process of \( y_t = 0.6 y_{t-1} + e_t \). Now the PACF cuts off at lag 1, a signal for AR(1). The ACF decays relatively fast, a signal for stationarity.

3. The third row is for the non-stationary AR(1) process of \( y_t = y_{t-1} + e_t \). Notice that \( \phi_{11} \approx 1 \), a signal for unit root (non-stationarity). The ACF decays very slowly, another signal for non-stationarity.

The R codes for this diagram is

```r
sma1 = rep(0, n) # initialize stationary MA(1)
for (i in 2:n) {
  sma1[i] = wn[i]+0.6* wn[i-1]
}
```
Figure 3: ACF and PACF
par(mfrow=c(3,2)) # plot the ACF and PACF side-by-side
acf(sma1, 24, main = "ACF of stationary MA(1)")
pacf(sma1, 24, main = "PACF of stationary MA(1)")
acf(sar1, 24, main = "ACF of stationary AR(1)")
pacf(sar1, 24, main = "PACF of stationary AR(1)")
acf(nsar1, 24, main = "ACF of nonstationary AR(1)")
pacf(nsar1, 24, main = "PACF of nonstationary AR(1)")

In practice, we should keep take difference of nonstationary series until ACF becomes fast decaying. The following diagram shows the ACF and PACF for the nonstationary AR(3) process: $y_t = 2.2y_{t-1} - 1.4y_{t-2} + 0.2y_{t-3} + e_t$, and its first and second differenced series. The R codes for this diagram is

```r
n = 200
wn=round(rnorm(n), 2)
nsar3 = rep(0, n) # nonstationary (unit-root) AR(3)

for (i in 4:n) {
  nsar3[i] = 2.2*nsar3[i-1] - 1.4*nsar3[i-2]+0.2*nsar3[i-3]+ wn[i]
}
d1nsar3 = diff(nsar3, 1)
d2nsar3 = diff(d1nsar3, 1)
```

par(mfrow=c(3,3)) # plot the ACF and PACF side-by-side
plot(nsar3,type="o", main = "nonstationary AR(3)")
acf(nsar3, 24, main = "ACF of nonstationary AR(3)")
pacf(nsar3, 24, main = "PACF of nonstationary AR(3)")
plot(d1nsar3,type="o", main = "first difference (FD) of nonstationary AR(3)")
acf(d1nsar3, 24, main = "ACF of FD nonstationary AR(3)")
pacf(d1nsar3, 24, main = "PACF of FD nonstationary AR(3)")
plot(d2nsar3,type="o", main = "second difference (SD) of nonstationary AR(3)")
acf(d2nsar3, 24, main = "ACF of SD nonstationary AR(3)")
pacf(d2nsar3, 24, main = "PACF of SD nonstationary AR(3)")
Figure 4: ACF and PACF of Nonstationary AR(3)
Estimation

1. Consider a general ARMA\((p,q)\) process

\[ \alpha(L)y_t = \beta(L)e_t, \]  

(29)

where \(\alpha(L) = 1 - \alpha_1L - \ldots - \alpha_pL^p\) and \(\beta(L) = 1 + \beta_1L + \ldots + \beta_qL^q\). The AR form for \(29\) is given by

\[ \phi(L)y_t = e_t, \quad \phi(L) = \alpha(L)\beta(L)^{-1}, \]

and the MA form is

\[ y_t = \theta(L)e_t, \quad \theta(L) = \alpha(L)^{-1}\beta(L). \]

2. After identification is done we know the values of \(p, q\). Now it is time to estimate the coefficient \(\alpha, \beta\).

3. Model \((29)\) is estimated by Maximum Likelihood (MLE). If MA term is missing, MLE is same as OLS. I would recommend using AR model as often as possible (and avoid MA term).

4. ARMA model is fitted by R command \texttt{arima}. For a stationary series, set the parameter for the command as “order = c(p, 0, q)”. The value zero in the middle of parentheses indicates the order of integration = 0, so the series is stationary.

5. (Optional). Consider an ARMA\((1,1)\) process: 

\[ y_t = \alpha y_{t-1} + e_t + \beta e_{t-1}. \]

We assume \(e_0 = 0, y_0 = 0\), and all the error terms follow normal distributions of \(N(0, \sigma^2)\).

(a) For the first observation we have \(y_1 = e_1\). So the density function for \(e_1\) is

\[ f(e_1) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{e_1^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{y_1^2}{2\sigma^2}} \]

(b) For the second observation \(y_2 = \alpha y_1 + e_2 + \beta e_1\). So the density function for \(e_2\) is

\[ f(e_2) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{e_2^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(y_2-\alpha y_1-\beta e_1)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(y_2-\alpha y_1-\beta y_1)^2}{2\sigma^2}} \]

(c) In a similar fashion we can derive \(f(e_i), (i = 3, 4, \ldots, n)\).
(d) The log likelihood function is defined as

\[ l = \log \left( \prod_{i=1}^{n} f(e_i) \right). \]

(e) The maximum likelihood estimators for \((\alpha, \beta, \sigma)\) maximize the log likelihood function. In practice we obtain the maximum likelihood estimates using numerical methods (such as Newton-Raphson Method).

6. There is a reminder about using R command \texttt{arima} to model a trending series. For example, an ARIMA\((1,1,1)\) with drift can be estimated by using commands of (i) \texttt{arima(diff(y), order = c(1, 0, 1))} or (ii) \texttt{arima(y, order = c(1, 1, 1), xreg = (1:length(y)))}. By default, ARIMA\((1,1,1)\) estimates a model without a drift.

7. We can use option \texttt{FIXED} to impose a restricted value for certain coefficient. For example, \texttt{arima(y, order = c(3, 1, 0), fixed = c(0, NA, NA))} will estimate an ARIMA\((3, 1, 0)\) model in which the first autoregressive coefficient (the coefficient of \(\Delta y_{t-1}\)) equals zero. The final model is

\[ \Delta y_t = \beta_2 \Delta y_{t-2} + \beta_3 \Delta y_{t-3} + e_t. \]

8. For series showing seasonality, we can use ARMA model for the seasonally-differenced series. For example, \texttt{arima(diff(y, lag=4), order = c(1, 0, 0))} estimates the model

\[ \Delta^4 y_t = \beta_0 + \beta_1 \Delta^4 y_{t-1} + e_t, \quad (\Delta^4 y_t = y_t - y_{t-4}). \]

**Forecast**

1. Denote the expectation of \(y_{t+j}\) conditional on the present information as

\[ E_t y_{t+j} = E(y_{t+j} | \Omega_t), \]

where \(\Omega_t\) denotes the information set available at \(t\). Typically, \(\Omega_t = \{y_t, y_{t-1}, \ldots, e_t, e_{t-1}, \ldots\}\).

2. Similarly we define the conditional variance as

\[ V_t y_{t+j} = E[(y_{t+j} - E_t y_{t+j})^2 | \Omega_t]. \]
3. **Theorem:** the optimal (minimizing mean square error) forecast for \( y_{t+j} \) at time \( t \) is 
\[ E_t y_{t+j} \]

Proof: without loss of generality, let’s focus on forecasting \( y_{t+1} \)

(a) The conditional mean has the property that
\[
E [(y_{t+1} - E_t y_{t+1}) f(\Omega_t)] = E \{E[(y_{t+1} - E_t y_{t+1}) f(\Omega_t)] | \Omega_t] = 0
\]

Essentially this result means that when using the conditional mean as the forecast, the forecasting error, \( y_{t+1} - E_t y_{t+1} \), is uncorrelated with any function of given information. Put differently, the conditional mean has made use of all relevant information contained in the information set.

(b) Using the above result we can show that
\[
E [y_{t+1} - f(\Omega_t)]^2 = E [y_{t+1} - E_t y_{t+1}]^2 + E(f(\Omega_t) - E_t y_{t+1})^2 \geq E [y_{t+1} - E_t y_{t+1}]^2.
\]

We reach the equality only when \( f(\Omega_t) = E_t y_{t+1} \). So the variance of forecasting error is minimized when the conditional mean is used as the forecast.

The theorem provides a general guidance for finding the optimal forecast: we just need to find the mean of future value conditional on the given (current) information.

4. Consider the easiest case AR(1): \( y_t = \phi y_{t-1} + e_t \). Using the recursive substitution and definition of conditional expectation, we can show that

(a) \( y_{t+j} = \phi^j y_t + \phi^{j-1} e_{t+1} + \ldots + e_{t+j} \)

(b) \( E_t e_{t+i} = 0, \quad i = 1, \ldots, j \)

(c) \( E_t \phi^j y_t = \phi^j y_t \)

(d) \( E_t y_{t+j} = \phi^j y_t \)

(e) \( V_t y_{t+j} = (1 + \phi^2 + \ldots + \phi^{2(j-1)}) \sigma^2 \)

In words, the forecast decays exponentially if \(|\phi| < 1\). Consider the limit when \( j \to \infty \)

(a) \( \lim_{j \to \infty} E_t y_{t+j} = \lim_{j \to \infty} \phi^j y_t = 0 = E_t y_t \)

(b) \( \lim_{j \to \infty} V_t y_{t+j} = \lim_{j \to \infty} (1 + \phi^2 + \ldots + \phi^{2(j-1)}) \sigma^2 = \frac{\sigma^2}{1-\phi^2} = \text{var}(y_t) \)
In words, the long run forecast converges to unconditional mean for a stationary series, and the long run forecast error variance converges to unconditional variance.

5. For predicting AR(p) process with $p > 1$ we can use the state space representation (or vector AR(1) representation).

6. For example, consider an AR(2) $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$. Define

$$Y_t = \begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix}, \quad A = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} \sigma \\ 0 \end{pmatrix}.$$  

Then the state space or vector AR(1) representation for the univariate AR(2) model is

$$Y_t = AY_{t-1} + Cv_t,$$

where $v_t \sim \text{iidn}(0, 1)$. Now we can show

(a) $E_t Y_{t+j} = A^j Y_t$

(b) $E_t y_{t+j} = RA^j Y_t$ where $R = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is the selecting vector picking out $y_{t+j}$.

(c) $V_t Y_{t+j} = CC' + ACC' A' + \ldots + A^{j-1} CC' A'^{j-1}$.

(d) $V_t y_{t+j} = R(CC' + ACC' A' + \ldots + A^{j-1} CC' A'^{j-1}) R'$.

7. The last two equations indicate that the conditional mean and variance can be computed recursively as

$$E_t Y_{t+j} = A E_t Y_{t+j-1} \quad (30)$$

$$V_t Y_{t+j} = CC' + AV_t Y_{t+j-1} A' \quad (31)$$

These two results are handy for programming.

8. Consider an MA(1) model $y_t = e_t + \theta e_{t-1}$, for which the vector AR(1) representation $Y_t = AY_{t-1} + Cv_t$ is obtained if we define

$$Y_t = \begin{pmatrix} y_t \\ e_t \end{pmatrix}, \quad A = \begin{pmatrix} 0 & \theta \\ 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} \sigma \\ \sigma \end{pmatrix}.$$
There is one catch: $e_t$ is unobservable, but can be obtained recursively as

$$

e_1 = y_1 \\
\varepsilon_2 = y_2 - \theta \varepsilon_1 \\
\varepsilon_3 = y_3 - \theta \varepsilon_2 \\
\vdots = \vdots \\
\varepsilon_t = y_t - \theta \varepsilon_{t-1}
$$

It turns out any ARMA model has a vector AR(1) representation. So the above method works for all ARMA models. See page 35 of Cochrane for another example.

**(Advanced, Optional) Linear Projection and Wold Theorem**

1. Let $\Omega$ be a symmetric, positive definite matrix. Then there exist unique matrix of $A$ and $D$ so that

$$
ADA' = \Omega,
$$

where $A$ is a lower triangular matrix with 1 on its diagonal line, and $D$ is a diagonal matrix. $ADA'$ is called triangular factorization of $\Omega$. The Cholesky factorization is given by

$$
CC' = \Omega,
$$

where

$$
C \equiv AD^{1/2}.
$$

2. For example, for the equation

$$
\begin{pmatrix}
1 & 0 \\
a_{21} & 1
\end{pmatrix}
\begin{pmatrix}
d_{11} & 0 & 0 & d_{22}
\end{pmatrix}
\begin{pmatrix}
1 & a'_{21} & 0 & 1
\end{pmatrix}
= 
\begin{pmatrix}
\Omega_{11} & \Omega_{12} \\
\Omega_{21} & \Omega_{22}
\end{pmatrix}
$$

we get the unique solution of

$$
A = \begin{pmatrix}
1 & 0 \\
\Omega_{21}\Omega_{11}^{-1} & 1
\end{pmatrix}, \quad D = \begin{pmatrix}
\Omega_{11} & 0 \\
0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}
\end{pmatrix}
$$
3. Let \( Y = (Y_1, \ldots, Y_n)' \), \( \Omega = E(YY') \). Then define

\[
e \equiv A^{-1}Y.
\]

We can show that

\[
Eee' = A^{-1}\Omega A'^{-1} = A^{-1}ADA'A'^{-1} = D.
\]

So \( e \) is serially correlated. This result is important. It shows how to transform a vector of correlated variables into a vector of uncorrelated components.

4. Conversely, we can show how to build a vector of correlated variables using a vector of uncorrelated components.

\[
Ae = Y \Rightarrow \begin{pmatrix} 1 & 0 & \ldots & 0 \\ a_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & 1 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}
\]

We can see that

(a) \( e_1 = Y_1 \)

(b) \( e_2 = Y_2 - a_{21}e_1 = Y_2 - a_{21}Y_1 = Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1 = Y_2 - \hat{P}(Y_2|Y_1) \). Here we use \( \hat{P} \) to denote linear projection. The linear projection of \( y \) on \( x \) is by definition a linear combination of \( x \) so that

\[
E[(y - ax)x] = 0.
\]

We can see \( \Omega_{21}\Omega_{11}^{-1}Y_1 \) is a linear projection of \( Y_2 \) on \( Y_1 \) because

\[
E[(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1)Y_1] = 0.
\]

(c) In a similar fashion we can show

\[
e_3 = Y_3 - a_{31}e_1 - a_{32}e_2 = Y_3 - \hat{P}(Y_3|Y_1) - a_{32}(Y_2 - \hat{P}(Y_2|Y_1))
\]

\[
a_{31} = \Omega_{31}\Omega_{11}^{-1}
\]

\[
a_{32} = d_{22}^{-1}(\Omega_{23} - a_{21}d_{11}a_{31})
\]
Then we can show that 

\[ \hat{P}(Y_3|Y_1) + a_{32}(Y_2 - \hat{P}(Y_2|Y_1)) \]

is indeed the linear projection of \( Y_3 \) on \( Y_2 \) and \( Y_1 \) because it satisfies \( E e_3 Y_2 = 0 \) and \( E e_3 Y_1 = 0 \). Now we can write

\[ \hat{P}(Y_3|(Y_2, Y_1)) = \hat{P}(Y_3|Y_1) + a_{32}(Y_2 - \hat{P}(Y_2|Y_1)). \]

This shows how to use the newly arrived information of \( Y_2 \) to update the forecast for \( Y_3 \). More explicitly, the new forecast, \( \hat{P}(Y_3|(Y_2, Y_1)) \), equals the old forecast, \( \hat{P}(Y_3|Y_1) \), plus an adjustment term. The adjustment term, \( a_{32}(Y_2 - \hat{P}(Y_2|Y_1)) \), captures the new information about \( Y_2 \) which is unexplained by \( Y_1 \).

(d) The above equality also indicates the law of iterated projection. That is,

\[ \hat{P} \left[ \hat{P}(Y_3|(Y_2, Y_1)) | Y_1 \right] = \hat{P}(Y_3|Y_1). \]

This result holds because the linear projection of the adjustment term, \( a_{32}(Y_2 - \hat{P}(Y_2|Y_1)) \), on \( Y_1 \) is zero. For an general AR(p) process, the law of iterated projection implies that

\[ \hat{y}_{t+p|t} = \phi_1 \hat{y}_{t+p-1|t} + \phi_2 \hat{y}_{t+p-2|t} + \ldots + \phi_p y_t, \]

where \( \hat{y}_{t+i|t} = \hat{P}(y_{t+i}|y_t) \). Thus in practice we can generate \( \hat{y}_{t+1|t}, \hat{y}_{t+2|t} \) and so on in a recursive manner. In other words, when forecasting \( y_{t+p} \) we just replace the unknown \( y_{t+p-i} \) with its linear projection \( \hat{y}_{t+p-i|t} \). We start with \( i = p - 1 \), or the 1-step ahead forecast.

(e) In general,

\[ \hat{P}(Y_i|(Y_{i-1}, Y_{i-2}, \ldots, Y_1)) = a_{i1} e_1 + a_{i2} e_2 + \ldots + a_{i-1} e_{i-1} \]

\[ e_i = Y_i - \hat{P}(Y_i|(Y_{i-1}, Y_{i-2}, \ldots, Y_1)). \]

(f) \( E e_i e_j = 0, \forall i \neq j \)

(g) \( Y_n = a_{n1} e_1 + a_{n2} e_2 + \ldots + e_n \). This is called Wold Theorem. Basically Wold Theorem shows that how to express a series as a sum of serially uncorrelated shocks, which is called MA form before.
5. The requirement for linear projection is

\[ E \left[ (Y - \hat{P}(Y|X))X \right] = 0, \]

which is much weaker than the condition for conditional mean:

\[ E \left[ (Y - E(Y|X))f(X) \right] = 0 \]

where \( f(\cdot) \) can be any function. So in general linear projection is not optimal.

6. However, under the condition that data are multivariate normal or data are generated by a linear process, the linear projection is the same as conditional expectation, so is optimal.

7. The Wold Theorem implies that any covariance-stationary process can be written as a linear process (in which the error term is the projection error). So linear projection is optimal for covariance-stationary process.

**Box-Jenkins Methodology**

You may follow the following steps to build a ARIMA model.

1. Read and plot data. Data are probably nonstationary if

   (a) The data is trending
   (b) The data is highly smooth (or has ‘long swing’)
   (c) The ACF dies down very slowly, and sometimes dies down nearly linearly with lag.

   The above is just signal for non-stationarity. Only unit root test can formally confirm that. We will learn unit root test later.

2. If the data is nonstationary, Box-Jenkins method always differences data first.

3. Now suppose the data is already stationary (i.e., we use the differenced data if original data is nonstationary). Denote the stationary data by \( y_t \).

4. By Wold Decomposition Theorem, any stationary process has an ARMA representation. This justifies the widely-used ARMA model. The model selection consists of the following steps
(a) Compute ACF
   i. ACF of an AR process eventually dies out
   ii. There is cut off after lag $q$ for ACF of of MA($q$) process

(b) Compute PACF
   i. PACF of a MA process eventually dies out
   ii. There is cut off after lag $p$ for PACF of of AR($p$) process

(c) The ACF and PACF within the two-standard-error bound are *not* significantly different from zero. So only ACF and PACF *outside* the two-standard-error bound matter.

(d) Most likely, ACF and PACF can only help narrow down a *pool* of candidate models. For example, your candidate models may include AR(1), AR(2) and ARMA(1,1). Now use these tips:
   i. pick the parsimonious (simple) model, which typically outperforms complicated models in terms of *out-of-sample forecasting*
   ii. pick the model with *smallest* AIC or BIC
   iii. Everything else equal, AR model is always better than MA and ARMA models because the latter are difficult to estimate and forecast.

5. Next you should run diagnostic checking. A good (or adequate) model is the one with serially *uncorrelated* (or white noise) residual. Your model is adequate if you get a Ljung-Box test with $p$-value *greater* than 0.10 for the residuals.

6. Forecasting is next, provided that the model passes the diagnostic checking.

(a) Forecasting an AR model is always easy. For instance, consider the forecast for an AR(2) model

\[
\hat{y}_{t+1} = \hat{\phi}_1 y_t + \hat{\phi}_2 y_{t-1} \\
\hat{y}_{t+2} = \hat{\phi}_1 \hat{y}_{t+1} + \hat{\phi}_2 y_t \\
\vdots = \vdots \\
\hat{y}_{t+s} = \hat{\phi}_1 \hat{y}_{t+s-1} + \hat{\phi}_2 y_{t+s-2}
\]

Here we apply the law of iterated projection. The predicted value *converges to the unconditional mean* of $y_t$ as $s \to \infty$. 

44
(b) Forecasting MA model is trickier. You need to first express (approximate) the unobserved error term $e_t$ in terms of observed $y_t$ and its lags in a recursive manner. For instance, you can obtain $e_i, (i = 1, 2, \ldots, t)$ for a MA(2) model $y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}$ as

\[
\begin{align*}
e_1 &= y_1 \\
e_2 &= y_2 - \theta_1 e_1 \\
e_3 &= y_3 - \theta_1 e_2 - \theta_2 e_1 \\
& \vdots \\
e_t &= y_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}
\end{align*}
\]

Next the forecast is

\[
\begin{align*}
\hat{y}_{t+1} &= \theta_1 e_t + \theta_2 e_{t-1} \\
\hat{y}_{t+2} &= \theta_2 e_t \\
\hat{y}_{t+3} &= 0 \\
& \vdots \\
\hat{y}_{t+s} &= 0 \quad \forall s > 2
\end{align*}
\]

Again we see that $\hat{y}_{t+s} \to Ey_t$, but the convergence is much faster than an AR process

(c) Forecasting an ARMA model is based on the most general Wiener-Kolmogorov formula.

(d) If the true model is unknown or is not of main interest, you may always use an AR(m) model for forecasting. There are two steps to get $\hat{y}_{t+s}$.

i. Run regression

\[
y_t = \hat{\phi}_1 y_{t-s} + \hat{\phi}_2 y_{t-s-1} + \ldots + \hat{\phi}_m y_{t-s-m+1} + \hat{\epsilon}_t.
\]
We know the coefficient can be obtained via
\[
\begin{pmatrix}
\hat{\phi}_1 \\
\hat{\phi}_2 \\
\vdots \\
\hat{\phi}_m
\end{pmatrix} = \begin{bmatrix}
\gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\
\gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0
\end{bmatrix}^{-1}
\begin{bmatrix}
\gamma_s \\
\gamma_{s+1} \\
\vdots \\
\gamma_{s+m-1}
\end{bmatrix},
\]
where \( \gamma_j = \text{COV}(Y_t Y_{t-j}) \).

ii. The forecast is
\[
\hat{y}_{t+s} = \hat{\phi}_1 y_t + \hat{\phi}_2 y_{t-1} + \ldots + \hat{\phi}_m y_{t-m+1}.
\]

This method is called *forecasting based on finite number of observations* for \( y_t \).

7. Your job is *not* done if your data is transformed. You need to recover the forecast of original data from that of transformed data. For instance, suppose \( y_t \) is the nonstationary original data. We use an AR(1) model for the differenced data \( z_t = y_t - y_{t-1} \). Then
\[
\hat{z}_{t+1} = \phi z_t \\
\Rightarrow (\hat{y}_{t+1} - y_t) = \phi (y_t - y_{t-1}) \\
\Rightarrow \hat{y}_{t+1} = (\phi + 1) y_t - \phi y_{t-1} \\
\Rightarrow \hat{y}_{t+2} = (\phi + 1) \hat{y}_{t+1} - \phi y_t \\
\vdots \\
\Rightarrow \hat{y}_{t+s} = (\phi + 1) \hat{y}_{t+s-1} - \phi \hat{y}_{t+s-2} \quad \forall s > 3.
\]

Again, we use the law of iterated projection repeatedly.
1. There are two evidences that the unemployment rate series $y_t$ is nonstationary: (1) $y_t$ is smooth with long swing; (ii) the ACF of $y_t$ decays very slowly (and almost linearly with lags). Hence we need to build ARMA model for differenced series $\Delta y_t$. Put differently, we need to consider $\text{ARIMA}(p, 1, q)$ for the original series, where $p$ and $q$ are selected according to the ACF and PACF of $\Delta y_t$.

2. The ACF of $\Delta y_t$ decays quickly, meaning $\Delta y_t$ is stationary already (or no need to take difference second time). The PACF of $\Delta y_t$ seems to cut off after lag 5. So we may consider $\text{ARIMA}(5, 1, 0)$. Meanwhile we want to consider the parsimonious $\text{ARIMA}(1, 1, 1)$.

3. The AIC for ARIMA(5, 1, 0) is less than that for ARIMA(1, 1, 1). So the former outperforms the latter. We notice that the coefficient of $\Delta y_{t-1}$ is insignificant. Dropping $\Delta y_{t-1}$ improves the AIC from -273.63 to -275.55.
4. The Ljung-Box test is insignificant at lag 1, 4, 8, failing to reject the hypothesis that the residual is white noise (no serial correlation). So the final and adequate model is

$$\Delta y_t = 0.2354\Delta y_{t-2} + 0.1555\Delta y_{t-3} + 0.0826\Delta y_{t-4} + 0.0966\Delta y_{t-5} + e_t.$$ 

The Ljung-Box test is a formal test for model adequacy. We can use it to show a model, say, ARIMA(3, 1, 0) is not adequate because it rejects the null of white noise residual at lag 8.

5. We can see the overall fit by plotting the true and predicted (in sample or out of sample) values together.

6. The forecast is made based on the final and adequate model. The R command is `predict`
Sample Test 1

Each question is worth 7 points

Consider an AR(2) process of \( y_t = 0.4y_{t-1} - 0.03y_{t-2} + e_t \), where \( e_t \) is white noise. Answer Questions 1, 2, 3, 4

1. Find the characteristic roots for \( y_t \). Is \( y_t \) stationary? Why?

2. Find \( \gamma_0 \) and \( \rho_2 \) for \( y_t \).

3. Find \( \theta_3 \) which is the coefficient of \( e_{t-3} \) in the MA representation (MA form) for \( y_t \).

4. Find the 2-step-ahead forecast \( E_t y_{t+2} \) and its variance \( V_t y_{t+2} \).

5. Please specify how to obtain the second order partial autocorrelation function (PACF) for \( y_t \).
Lecture 3: GARCH Model

1. Typically we apply GARCH model to high-frequency data, such as daily stock market closing price. Denote the series by $y_t$.

2. We use ARIMA model to describe the persistence in the level of $y_t$ (or $\Delta y_t$ in case of non-stationarity). So the ARIMA model is concerned with the first moment.

3. In order to describe the persistence in the second moment (or conditional variance), we use GARCH model.

4. Jointly, we consider

$$\phi(L)y_t = \mu + \epsilon_t$$  \hspace{1cm} (33)
$$\epsilon_t = \sqrt{h_t}\epsilon_t, \text{ } \epsilon_t \sim \text{white noise}$$ \hspace{1cm} (34)
$$h_t = \omega + b_1h_{t-1} + \ldots + b_ph_{t-p} + a_1\epsilon^2_{t-1} + \ldots + a_q\epsilon^2_{t-q}.$$ \hspace{1cm} (35)

5. Equation (33) is an AR model, which describes $y_t$. The residual is $\epsilon_t$. We include enough lags in (33) so that there is no serial correlation in $\epsilon_t$. However, it is possible that $\epsilon^2_t$ are serially correlated.

6. Graphically, the plot of $\epsilon_t$ shows volatility clustering if $\epsilon^2_t$ are serially correlated. Volatility clustering means volatile (peaceful) observation tends to be followed by another volatile (peaceful) observation. In statistics we use variance to measure volatility. So volatility clustering signifies correlation (persistence) in variance.

7. According to Equation (34) we can show

(a) $E(\epsilon_t | \epsilon_{t-1}, \ldots) =$
(b) $\text{var}(\epsilon_t | \epsilon_{t-1}, \ldots) =$

8. So $h_t$ is the variance of $\epsilon_{t-1}$ conditional on its past values.

9. Equation (35) is called GARCH($p$,q) model. There are $p$ GARCH terms ($h_{t-1}, \ldots, h_{t-p}$), and $q$ ARCH terms ($\epsilon^2_{t-1}, \ldots, \epsilon^2_{t-q}$). Again we include sufficient lags so that there is no correlation left in the squared residual of GARCH model.

10. Most often we let $p = 1, q = 1$, and we may find $a_1 + b_1 \approx 1$. Roughly speaking, this indicates a unit root (or high persistence) in the conditional variance.
11. We estimate GARCH models using the method of MLE. Because numerical method is involved, it is sometimes difficult for convergence to occur.

12. To estimate GARCH model using R, you should load the tseries package first. The key command is garch. Type ?garch in the R console for instruction.
Lecture 4: Unit-Root Test

Why do we care?

The test for unit root (or non-stationarity) is important because

1. It helps determine $d$, the number of times a series needs to be differenced, before a stationary ARMA model can be applied.
2. It helps specify VAR.
3. It helps avoid spurious regression.
4. It helps prove long-run relationship (cointegration)
5. It helps policy debates, e.g., about global warming.

Signals for Unit Root (Non-stationarity)

1. The series shows smoothness or long swing
2. The series is trending (Be Careful!)
3. The ACF decays very slowly
4. Run regression
   \[ y_t = \phi y_{t-1} + \epsilon_t, \]
   and find that $\hat{\phi}$ is close to one.

Augmented Dickey Fuller (ADF) Unit Root Test for Non-Trending Series

1. Consider an AR(2) process
   \[ y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t. \]

2. The corresponding characteristic equation is
   \[ x^2 - \phi_1 x - \phi_2 = 0. \]
3. If one unit root exists, i.e., \( x_1 = 1 \), then
\[
1 - \phi_1 - \phi_2 = 0, \quad \text{or} \quad \phi_1 + \phi_2 = 1.
\]

4. So the test for one unit root is the same as testing for the hypothesis that \( H_0 : \phi_1 + \phi_2 = 1 \).

5. To test this hypothesis, we rewrite the original regression as
\[
\Delta y_t = \beta y_{t-1} + \alpha \Delta y_{t-1} + \epsilon_t.
\]

Using the method of undetermined coefficient, we can show
\[
\beta = \phi_1 + \phi_2 - 1 \\
\alpha = -\phi_2.
\]

6. Hence the ADF test is the t-statistic (or t-ratio) for \( \hat{\beta} \).

7. In general, for an AR(p) process
\[
y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t,
\]
the null hypothesis that one unit root is present can be formulated as
\[
H_0 : \phi_1 + \phi_2 + \ldots + \phi_p = 1.
\]

This null hypothesis is equivalent to
\[
H_0 : \beta = 0,
\]
where \( \beta \) is the coefficient in the transformed regression of
\[
\Delta y_t = \beta y_{t-1} + \alpha_1 \Delta y_{t-1} + \ldots + \alpha_{p-1} \Delta y_{t-p+1} + \epsilon_t.
\]

8. The ADF test is based on (39) (augmented with an intercept) and is the t-ratio for \( \hat{\beta} \).
ADF Test for Trending Series

1. For an AR(p) process with one unit root, we can add a constant to make the series to appear trending. This is how.

2. Consider a unit-root AR(2) of

   \[ y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t, \quad \phi_1 + \phi_2 = 1. \]

3. Under \( \phi_1 + \phi_2 = 1 \), we can rewrite the above equation as

   \[ \Delta y_t = c - \phi_2 \Delta y_{t-1} + e_t. \]

   Using lag operator we can show

   \[(1 + \phi_2 L)\Delta y_t = c + e_t, \]

   or

   \[ \Delta y_t = \frac{c}{1 + \phi_2} + v_t, \quad v_t \equiv (1 + \phi_2 L)^{-1} e_t. \]  \hspace{1cm} (40)

   Recall that \( \Delta \equiv 1 - L \), and \( (1 - L)^{-1} = 1 + L + L^2 + \ldots \). It follows that

   \[ y_t = \left( \frac{c}{1 + \phi_2} \right) t + \sum_{i=1}^{t} v_i + y_0. \]  \hspace{1cm} (41)

4. Define \( z_t = \sum_{i=1}^{t} v_i \). We can show

   \[ z_t = z_{t-1} + v_t. \]

   So \( z_t \) is a random walk. A random walk is a nonstationary series that has one unit root.

5. Similarly \( y_t = \left( \frac{c}{1 + \phi_2} \right) + y_{t-1} + v_t \), so \( y_t \) is a random walk with drift. The drift term is \( \left( \frac{c}{1 + \phi_2} \right) \). The drift term introduces a linear trend of \( \left( \frac{c}{1 + \phi_2} \right) t \) into the series, and makes the series appear trending.

6. In general, a unit-root AR(p) process with a constant is called random walk with drift, and can be expressed as the sum of a linear trend and a random walk without drift.
Trend-Stationary VS Difference-Stationary

1. There are two types of trending series: (i) trend-stationary series, and (ii) difference-stationary series.

2. Graphically both series are trending, but there are important differences
   
   (a) The trend-stationary series is trend-reverting, but difference-stationary series is not
   
   (b) The effect of past shock is transitory for trend-stationary series, whereas permanent for difference-stationary series.

Both differences have important policy implication.

3. A trend-stationary series becomes stationary if we regress it onto a linear trend. The residual is called the detrended series, and is stationary.

4. A difference-stationary series becomes stationary if we take difference. Equation (41) shows that a random walk with drift cannot become stationary if we detrend it. However, equation (40) shows a random walk with drift becomes stationary if we take difference. So random walk with drift is difference-stationary, not trend-stationary. This example also shows that a difference-stationary series cannot become stationary if we detrend it.

5. We can show the trend-stationary series becomes stationary if we difference it. That means the ARIMA model can be applied to the trend-stationary series.

6. We can use the ADF test to tell a difference-stationary series from a trend-stationary series. The regression is

\[ \Delta y_t = c_0 + c_1 t + \beta y_{t-1} + \alpha_1 \Delta y_{t-1} + \ldots + \alpha_{p-1} \Delta y_{t-p+1} + \epsilon_t. \]  

(42)

7. The ADF test for a trending series is based on (42) and is the t-ratio for \( \hat{\beta} \).

A Technical Note

The ADF test DOES NOT follow the student-T distribution or normal distribution, even in large sample. R takes this fact into account already.
Practical Suggestions

1. Plot the series first.

   (a) If it is trending, use (42). Rejecting the null means the series is trend-stationary. Failing to reject means the series is difference-stationary.

   (b) If it is not trending, use (42) but drop the trend term (so only intercept is left). Rejecting the null means the series is stationary. Failing to reject means the series is nonstationary.

2. ADF test is sensitive to the number of lagged terms included in the regression. Use AIC to help select the optimal model.

3. You must apply the Ljung-Box test to the residual of (42). The model is adequate if the null of white noise residual cannot be rejected.

More about R Command ARIMA

1. For a non-trending series

   (a) Use ARIMA(p, 0, q) if the ADF test rejects the null

   (b) Use ARIMA(p, d, q), d = 1, 2, . . . , if the ADF test cannot reject the null

2. For a trending series, you can always apply the ARIMA model no matter it is trend-stationary or difference-stationary. There is a catch though. The R command ARIMA(p, 1, q) returns no intercept (drift) term. In other words, the R command ARIMA(p, 1, q) CANNOT be applied directly to a trending series. Instead, the right way is to apply ARIMA(p, 0, q) to the differenced series, or force ARIMA(p, 1, q) to be augmented with a trend.

3. For example, an ARIMA(1,1,1) with drift can be estimated by using commands of

   (a) (i) arima(diff(y), order = c(1, 0, 1))

   (b) (ii) arima(y, order = c(1, 1, 1), xreg = (1:length(y)))
Example: Testing Unit Root for U.S. GNP

1. First, the plot of log GNP (lgnp) is trending. So we should include trend in the testing regression, i.e., use (42) to run ADF test.

2. The ACFs of lgnp and detrended lgnp decay slowly, but decay quickly for differenced lgnp (dlgnp). This is signal for difference-stationarity.

3. Neither of the ADF tests using 1, 2, 3, 4 lags can reject the null hypothesis of unit root (the \( p \)-values are all greater than 0.05). So lgnp is indeed difference-stationary (or is random walk with a drift, or has at least one unit root).

4. Difference-stationarity means (i) GNP is not trend-reverting, and (ii) past shock has permanent effect on present value. Macroeconomists use this fact to statistically justify the use of monetary or fiscal policy during recession since the economy cannot go back to its normal trend without help (no trend-reverting).
5. Based on the PACF of dlgnp, the candidate model is ARIMA(1,1,0). We see the R command `arima(lgnp, order = c(1, 1, 0))` returns no intercept (drift), so is WRONG. The right way is to use R command `arima(diff(lgnp), order = c(1, 0, 0))`. The estimated intercept is 0.0083, and is significant.

6. The Ljung-Box test cannot reject the white noise residual. So ARIMA(1,1,0) is adequate. The final model is

\[ \Delta \log(gnp)_t = 0.0083 + 0.3467 \Delta \log(gnp)_{t-1}. \]
Appendix (optional): Unit Root Test

The null data generating process

1. Consider an AR(2) process without intercept

\[ y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t. \]

We can always rewrite the equation as

\[ \Delta y_t = \beta_1 y_{t-1} + \beta_2 \Delta y_{t-1} + e_t, \]

\[ \beta_1 = \phi_1 + \phi_2 - 1 \]

\[ \beta_2 = -\phi_2 \]

2. Under the null hypothesis that there is one unit root, i.e.,

\[ H_0 : \phi_1 + \phi_2 - 1 = 0, \text{ or } \beta_1 = 0 \]

the data generating process is given by

\[ \Delta y_t = \beta_2 \Delta y_{t-1} + e_t, \]

or

\[ y_t = y_{t-1} + (1 - \beta_2 L)^{-1} e_t \equiv y_{t-1} + u_t. \]

3. So under the null hypothesis \( y_t \) becomes a random walk with serially correlated error

\[ u_t \equiv \phi(L) e_t = (1 - \beta_2 L)^{-1} e_t = (1 + \phi_2 L)^{-1} e_t. \]

4. In short, there are two equivalent ways to model a general process with unit root

(a) random walk with serially correlated error, This approach is used by the Phillips and Perron Test.

(b) an adequate AR(p) model with serially uncorrelated error. This approach is used by the ADF test.
The Beveridge-Nelson Decomposition and Invariance Principle

1. We can show that for \( u_t = \phi(L)e_t, e_t \sim \text{iid}(0, \sigma^2) \)

\[
 u_1 + u_2 + \ldots + u_t = \phi(1)(e_1 + e_2 + \ldots + e_t) + \eta_t - \eta_0, \tag{43}
\]

where

\[
 \phi(1) = \frac{1}{1 + \phi_2} = \frac{1}{1 - \beta_2},
\]

for the AR(2) model and \( \eta_t = O_p(1), \eta_0 = O_p(1) \).

2. Consider one example of the Beveridge-Nelson Decomposition. Let \( u_t = e_t + \phi e_{t-1} \), then

\[
 u_1 + u_2 + \ldots + u_t = e_1 + (e_2 + \phi e_1) + \ldots + (e_t + \phi e_{t-1})
\]

\[
 = (1 + \phi)(e_1 + e_2 + \ldots + e_t) - \phi e_t
\]

\[
 = \phi(1)(e_1 + e_2 + \ldots + e_t) + \eta_t - \eta_0, \quad (\eta_t - \eta_0 = -\phi e_t)
\]

3. It follows that

\[
 y_t = y_{t-1} + u_t
\]

\[
 = \sum_{i=1}^{t} u_i + y_0
\]

\[
 = \phi(1) \sum_{i=1}^{t} e_i + \eta_i - \eta_0 + y_0 = O_p(T^{1/2})
\]

4. From calculus we know the integral is the limit of an infinite series given by

\[
 \int_{0}^{1} f(x)dx = \lim_{T \to \infty} \sum_{t=1}^{T} \left( \frac{1}{T} \right) f \left( \frac{t}{T} \right).
\]

(a) Example 1: \( \sum_{t}^{T} \sqrt{t} = T^{3/2} \sum_{t}^{T} \left[ \frac{1}{T} \sqrt{\frac{t}{T}} \right] = T^{3/2} \int x^{1/2}dx = \frac{2T^{3/2}}{3} \).

(b) Example 2: \( \sum_{t}^{T} t^2 = T^3 \sum_{t}^{T} \left[ \frac{1}{T} \left( \frac{t}{T} \right)^2 \right] = T^3 \int x^2dx = T^3/3 \)
5. By the invariance principle, we have

\[ T^{-1/2} \sum_{i=1}^{[Tr]} u_i \Rightarrow \sigma \phi(1)W(r) = \lambda W(r), \quad \lambda = \sigma \phi(1) \quad (44) \]

\[ T^{-3/2} \sum_{t=1}^{T} y_t = \sum \left( \frac{1}{T} \right) \left( \frac{y_t}{T^{1/2}} \right) \Rightarrow \lambda \int_0^1 W(r)dr \quad (45) \]

\[ T^{-2} \sum_{t=1}^{T} y_t^2 = \sum \left( \frac{1}{T} \right) \left( \frac{y_t}{T^{1/2}} \right)^2 \Rightarrow \lambda^2 \int_0^1 W^2(r)dr \quad (46) \]

\[ T^{-1} \sum_{t=1}^{T} v_{t-1} e_t = (1/2) \left( \frac{V_T^2}{T} - \frac{\sum e_t^2}{T} \right) \Rightarrow \frac{\sigma^2}{2} [W^2(1) - 1], \quad v_t = e_1 + \ldots + e_t. \quad (47) \]

\[ T^{-1} \sum_{t=1}^{T} y_{t-1} e_t = \phi(1)T^{-1} \sum v_{t-1} e_t + o_p(1) \Rightarrow \frac{\sigma^2 \phi(1)}{2} [W^2(1) - 1] \quad (48) \]

\[ T^{-1} \sum_{t=1}^{T} y_{t-1} u_t = (1/2) \left( \frac{y_T^2}{T} - \frac{\sum u_t^2}{T} \right) \Rightarrow \frac{1}{2} [\lambda^2 W^2(1) - Eu_t^2] \quad (49) \]
Asymptotic Theory

1. Now go back to the regression
\[
\Delta y_t = \beta_1 y_{t-1} + \beta_2 \Delta y_{t-1} + e_t.
\]

We want to derive the distribution of \( \hat{\beta} \) under the null hypothesis \( H_0 : \beta_1 = 0 \). Toward that end, define the scaling matrix as
\[
\Gamma = \begin{pmatrix} T^{-1} & 0 \\ 0 & T^{-1/2} \end{pmatrix}.
\]

2. It follows that
\[
\Gamma X' e = \begin{pmatrix} T^{-1} \sum y_{t-1} e_t \\ T^{-1/2} \sum u_{t-1} e_t \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{\sigma^2 \phi(1)}{2} [W^2(1) - 1] \\ \text{normal} \end{pmatrix}
\]
\[
\Gamma X' \Gamma = \begin{pmatrix} T^{-2} \sum y_{t-1}^2 & T^{-3/2} \sum y_{t-1} u_{t-1} \\ T^{-3/2} \sum y_{t-1} u_{t-1} & T^{-1} \sum u_{t-1}^2 \end{pmatrix} \Rightarrow \begin{pmatrix} \lambda^2 \int_0^1 W^2(r) dr & 0 \\ 0 & Eu_{t-1}^2 \end{pmatrix}
\]

It’s instructive to note that \( \Gamma X' \Gamma \) is asymptotically diagonal. The inverse of a diagonal matrix is also diagonal.

3. Finally,
\[
\Gamma^{-1}(\hat{\beta} - \beta) = (\Gamma X' \Gamma)^{-1}(\Gamma X' e).
\]

More explicitly, due to the fact that \( \Gamma X' \Gamma \) is diagonal
\[
T \hat{\beta}_1 \Rightarrow \left( \lambda^2 \int_0^1 W^2(r) dr \right)^{-1} \frac{\sigma^2 \phi(1)}{2} [W^2(1) - 1].
\]
\[
T^{1/2}(\hat{\beta}_2 - \beta_2) \Rightarrow \text{normal}
\]

4. Under the null hypothesis \( \beta_1 = 0 \), we have
\[
\hat{\phi}_2 = -\hat{\beta}_2 \Rightarrow \text{normal}
\]
\[
T^{1/2}(\hat{\phi}_1 - \phi_1) = T^{1/2}(\hat{\beta}_1 - \beta_1) + T^{1/2}(\hat{\beta}_2 - \beta_2) \Rightarrow \text{normal}
\]

This means the standard inference can be applied to \( \hat{\phi} \) in the un-transformed regression.
5. All linear combinations of $\hat{\phi}$ are normal except $\hat{\phi}_1 + \hat{\phi}_2 - 1$ which equals the non-standard $\hat{\beta}_1$.

6. Because $y_t = y_{t-1} + u_t$ under the null, it follows that the following regression will not suffer the bias due to serially correlated error:

$$y_t = \phi y_{t-1} + u_t, \quad (u_t = \phi(L)e_t).$$

It is easy to show that $\hat{\phi}$ is super-consistent (converging to a non-degenerate distribution at a rate faster than $T^{1/2}$):

$$T(\hat{\phi} - 1) = \frac{T^{-1} \sum y_{t-1} u_t}{T^{-2} \sum y_{t-1}^2} \Rightarrow \frac{0.5[\lambda^2 W(1)^2 - Eu_t^2]}{\lambda^2 \int W^2(r) dr}.$$  

The standard inference is invalid to $\hat{\phi}$, though.

7. So PACF $\approx 1$ is always a signal for unit root.

**ADF Test, Form 1**

1. First recall that

$$\lambda = \sigma \phi(1) = \frac{\sigma}{1 - \beta_2},$$

which implies that

$$\frac{\lambda}{\sigma} = \frac{1}{1 - \beta_2}.$$  

2. Second,

$$T(\lambda/\sigma) \hat{\beta}_1 \Rightarrow \left( \int_0^1 W^2(r) dr \right)^{-1} \frac{1}{2} [W^2(1) - 1].$$

3. Thus the first form of ADF test is

$$\frac{T\hat{\beta}_1}{1 - \beta_2} \Rightarrow \left( \int_0^1 W^2(r) dr \right)^{-1} \frac{1}{2} [W^2(1) - 1]. \quad (50)$$

**ADF Test, Form 2**

1. The t-ratio is defined as

$$t = \frac{\hat{\beta} - \beta}{s.e(\beta)} = \frac{\Gamma^{-1}(\hat{\beta} - \beta)}{\sqrt{s^2(\Gamma X'X\Gamma)^{-1}}},$$

63
2. We know already

\[ T\hat{\beta}_1 \Rightarrow \frac{\sigma}{2\lambda} \left( \int_0^1 W^2(r)dr \right)^{-1} [W^2(1) - 1] \]

\[ s^2 = \sigma^2 + o_p(1) \]

\[ T^{-2} \sum y_{t-1}^2 \Rightarrow \chi^2 \int W^2(r)dr \]

\[ T^{-3/2} \sum y_{t-1} \Delta y_{t-1} = o_p(1) \]

3. So the second form of ADF test is

\[ t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} \Rightarrow \frac{1}{2} \left( \int_0^1 W^2(r)dr \right)^{-1/2} [W^2(1) - 1] \quad (51) \]
ADF Test with Intercept

1. Now consider the new testing regression with an intercept term

\[ \Delta y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 \Delta y_{t-1} + \epsilon_t. \]

2. Under the joint null hypothesis

\[ H_0 : \beta_0 = 0, \beta_1 = 0, \]

\( y_t \) is a random walk without drift, but with serially correlated errors.

3. The new scaling matrix is defined as

\[
\Gamma = \begin{pmatrix}
T^{-1/2} & 0 & 0 \\
0 & T^{-1} & 0 \\
0 & 0 & T^{-1/2}
\end{pmatrix}.
\]

4. It follows that

\[
\Gamma X'e = \begin{pmatrix}
T^{-1/2} \sum e_t \\
T^{-1} \sum y_{t-1}e_t \\
T^{-1/2} \sum u_{t-1}e_t
\end{pmatrix} \Rightarrow \begin{pmatrix}
\sigma W(1) \\
\frac{\sigma^2}{2} (W^2[1] - 1) \\
\text{normal}
\end{pmatrix}
\]

\[
\Gamma X'e \Gamma = \begin{pmatrix}
1 & T^{-3/2} \sum y_{t-1} \\
T^2 \sum y_{t-1}^2 & T^{-1} \sum u_{t-1}
\end{pmatrix} \Rightarrow \begin{pmatrix}
1 & \lambda \int W(r) & 0 \\
\lambda \int W(r) & \lambda^2 \int W^2(r) & 0 \\
0 & 0 & Eu_{t-1}^2
\end{pmatrix}
\]

Again, it’s important that \( \Gamma X'e \Gamma \) is asymptotically diagonal.

5. Finally

\[
\begin{pmatrix}
T^{3/2} \hat{\beta}_0 \\
T \hat{\beta}_1
\end{pmatrix} \Rightarrow \begin{pmatrix}
\sigma & 0 \\
0 & \sigma/\lambda
\end{pmatrix} \begin{pmatrix}
1 & \int W(r) \\
\int W(r) & \int W^2(r)
\end{pmatrix}^{-1} \begin{pmatrix}
W(1) \\
W^2(1) - 1
\end{pmatrix} \quad (52)
\]

6. In particular,

\[
T (\lambda/\sigma) \hat{\beta}_1 = \frac{T \hat{\beta}_1}{1 - \hat{\beta}_2} \Rightarrow \frac{0.5(W^2(1) - 1) - W(1) \int W(r)}{\int W^2(r) - [\int W(r)]^2}.
\]
Therefore, including the intercept term changes the distribution.

Demeaned Brownian Motion

ADF test with a trend

Null Data Generating Process

1. Consider an AR(2) regression with an intercept

\[ y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t, \]

which can be rewritten as

\[ \Delta y_t = c + \beta_1 y_{t-1} + \beta_2 \Delta y_{t-1} + e_t, \]

\[ (\beta_1 = \phi_1 + \phi_2 - 1, \beta_2 = -\phi_2). \]

2. The null is that there is one unit root, so

\[ H_0 : \phi_1 + \phi_2 - 1 = 0, \text{ or } \beta_1 = 0. \]

Under the null, it follows that

\[ \Delta y_t = c + \beta_2 \Delta y_{t-1} + e_t. \]

3. So the null generating process is

\[ \Delta y_t = \mu + u_t = \frac{c}{1 - \beta_2} + u_t, \]

where \( \mu = \frac{c}{1 - \beta_2} \) is the drift, and

\[ u_t \equiv \phi(L)e_t = (1 - \beta_2 L)^{-1}e_t \]

is serially correlated. In words, \( y_t \) is random walk with drift and serially correlated errors.

4. There is another way to figure out the drift term \( \mu \). Applying the method of undetermined coefficient to

\[ (\Delta y_t - \mu) = \beta_2 (\Delta y_t - \mu) + e_t, \]
we get 

\[ \mu = \frac{c}{1 - \beta_2}. \]

Notice that 

\[ 0 = EU_t = E(\Delta y_t - \mu). \]

5. So the null process is 

\[ \Delta y_t - \mu = u_t, \quad (u_t = \phi(L)e_t = (1 - \beta_2 L)^{-1}e_t). \]

Next it’s easy to see that \( y_t \) has a trend

\[ y_t = \mu t + \sum_{i=1}^{t} u_i + y_0. \]

6. We can define the detrended series as 

\[ \eta_t \equiv y_t - \mu t = \sum_{i=1}^{t} u_i + O_p(1) = \phi(1) \sum_{i=1}^{t} e_i + O_p(1). \]

It follows that 

\[ T^{-3/2} \sum_{t=1}^{T-1} \eta_{t-1} \Rightarrow \lambda \int W(r)dr \]

\[ T^{-2} \sum_{t=1}^{T-1} \eta_{t-1}^2 \Rightarrow \lambda^2 \int W^2(r)dr \]

\[ T^{-1} \sum_{t=1}^{T-1} \eta_{t-1} e_t = T^{-1} \phi(1) \sum_{t=1}^{T-1} e_t \Rightarrow \frac{\sigma^2 \phi(1)}{2} (W^2(1) - 1) \]

**Canonical Form of Testing Regression**

1. The testing regression is given by

\[ \Delta y_t = \alpha + \delta t + \rho y_{t-1} + \beta_2 \Delta y_{t-1} + \epsilon_t. \]  \hspace{1cm} (53)

2. The joint null hypothesis is 

\[ H_0: \rho = 0, \delta = 0. \]
Under $H_0$ the series is

$$y_t = \mu t + \sum u_i + y_0,$$

where $\mu = \alpha/(1 - \beta_2)$, and $u_t = \phi(L)e_t = (1 - \beta_2)^{-1}e_t$.

3. Consider transforming variables as

$$u_t = \Delta y_t - \mu$$
$$\mu = \frac{\alpha}{1 - \beta_2}$$
$$\eta_{t-1} = y_{t-1} - \mu(t - 1) = \sum u_i + O_p(1)$$

Note that $Eu_t = 0$.

4. Using the method of undetermined coefficient, the canonical form (or transformed) of the testing regression is

$$\Delta y_t = \mu^* + \delta^* t + \rho \eta_{t-1} + \beta_2 u_{t-1} + e_t. \quad (54)$$
$$\mu^* = (1 - \rho)\mu \quad (55)$$
$$\delta^* = \delta + \rho \mu \quad (56)$$

5. We can show that

$$\begin{pmatrix} T^{1/2}(\hat{\mu} - \mu^*) \\ T^{3/2}(\hat{\delta} - \delta^*) \\ T\hat{\rho} \\ T^{1/2}(\hat{\beta}_2 - \beta_2) \end{pmatrix} \Rightarrow \begin{pmatrix} V1 & 0 \\ 0 & V2 \end{pmatrix}^{-1} \begin{pmatrix} E1 \\ E2 \end{pmatrix},$$

where

$$V1 = \begin{pmatrix} T^{-1} \sum 1 & T^{-2} \sum t & T^{-3/2} \sum \eta_{t-1} \\ .. & .. & .. \\ T^{-3} \sum t^2 & T^{-5/2} \sum t\eta_{t-1} & T^{-2} \sum \eta^2_{t-1} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 1/2 & \lambda \int W \\ .. & 1/3 & \lambda \int rW \\ .. & .. & \lambda^2 \int W^2 \end{pmatrix} \quad (57)$$

$$E1 = \begin{pmatrix} T^{-1/2} \sum e_t \\ T^{-3/2} \sum te_t \\ T^{-1} \sum \eta_{t-1}e_t \end{pmatrix} \Rightarrow \begin{pmatrix} \sigma W(1) \\ \sigma(W(1) - \int W) \\ \frac{\sigma \lambda W^2(1-1)}{2} \end{pmatrix} \quad (58)$$
6. The fact that $\Gamma X'X\Gamma$ is diagonal is due to

$$T^{-1} \sum u_i = Eu_t + o_p(1)$$

$$T^{-2} \sum t e_i = T^{-1/2} \sum \left( \frac{t}{T} \right) d \left( \frac{v_i - v_{i-1}}{T^{1/2}} \right) = T^{-1/2} \int rdW = o_p(1)$$

$$T^{-2} \sum t u_i = o_p(1)$$

$$T^{-3/2} \sum \eta_{t-1} u_t = T^{-1/2} \sum \left( \frac{\eta_{t-1}}{T^{1/2}} \right) d \left( \frac{\eta_t - \eta_{t-1}}{T^{1/2}} \right) = T^{-1/2} O_p(1) = o_p(1)$$

7. Finally, we can show the distribution of the ADF test of

$$ADF = \frac{\hat{\rho}}{s.e(\hat{\rho})}.$$
Spurious Regression

1. Consider the simple regression of

\[ y_t = \beta x_t + u_t, \]

where both \( y_t \) and \( x_t \) are nonstationary. If there does not exist population value for \( \beta \) for which the residual \( u_t = y_t - \beta x_t \) is stationary, then OLS is likely to produce spurious results.

2. For example, let

\[
\begin{align*}
y_t &= y_{t-1} + e_{1t} \\
x_t &= x_{t-1} + e_{2t} \\
\hat{\beta} &= \frac{T^{-2} \sum y_t x_t}{T^{-2} \sum x_t^2} \Rightarrow \int W_1 W_2 \equiv B
\end{align*}
\]

where \( W_1 \) and \( W_2 \) are two independent Brownian motions.

3. In words, when the dependent and independent variables are independent I(1) series, \( \hat{\beta} \) converges to a nonstandard random variable.

4. In contrast, if data are stationary, then \( \hat{\beta} \) converges in probability to the true population value, which is zero if two series are independent. \( T^{1/2} \hat{\beta} \) converges to a normal distribution.

5. It is easy to show that the residual in the spurious regression is nonstationary since

\[
T^{-1/2} \hat{u}_t = T^{-1/2} y_t - \hat{\beta}(T^{-1/2} x_t) \Rightarrow W_1(r) - BW_2(r) = O_p(1)
\]

\[
\hat{u}_t = O_p(T^{1/2})
\]

6. If there exists \( \beta \) so that \( u_t \) is I(0), then the data are cointegrated. In that case we can show that

\[
T(\hat{\beta} - \beta) = \frac{T^{-1} \sum x_t e_t}{T^{-2} \sum x_t^2} \Rightarrow \int W_2 dW_1.
\]

7. So in case of cointegration, the estimated cointegrating vector converges to the true value at a rate faster than \( T^{1/2} \) (super-consistency), and the limit distribution is non-
standard. The latter fact implies that standard inference is invalid for the estimated cointegrating vector.

8. We need to apply unit root test to the residual to guard against the spurious regression.

9. One cure to spurious regression is running the following regression

\[ y_t = \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + e_t. \]

So we include the lagged values of both dependent and independent variables. We can show

(a) the above regression is equivalent to

\[ y_t = \beta_1 y_{t-1} + \beta_2 \Delta x_t + \beta_3^* x_{t-1} + e_t, \]

where \( \beta_3^* = \beta_3 + \beta_2. \)

(b) \[
\begin{pmatrix}
T(\hat{\beta}_1 - 1) \\
T^{1/2}\hat{\beta}_2 \\
T\hat{\beta}_3^*
\end{pmatrix}
\begin{pmatrix}
\text{nonstandard} \\
\text{standard} \\
\text{nonstandard}
\end{pmatrix}
\]

(c) The individual t-test for \( H_0 : \beta_2 = 0 \) is asymptotically normal.

(d) The individual t-test for \( H_0 : \beta_3 = 0 \) is also asymptotically normal because \( \hat{\beta}_3 = \hat{\beta}_3^* - \hat{\beta}_2 \) is asymptotically dominated by the normally distributed \( \hat{\beta}_2. \)

(e) The F test for the joint hypothesis \( H_0 : \beta_3 = 0, \beta_2 = 0 \) is nonstandard because this is equivalent to testing \( \beta_3^* = 0, \) which follows nonstandard distribution.
Lecture 5: VAR

Motivation

So far we only cover univariate time series model, which cannot account for the interaction among different series (such as long term and short term interest rates, the gold price and stock market index). VAR is a multivariate model, which can take that interaction into account.

Stationary Vector Processes

Consider a first order vector autoregression (in reduced form), denoted by VAR(1), given by

\[ y_t = \phi y_{t-1} + e_t, \quad (59) \]

where \( y_t = (y_{1t}, y_{2t})' \), \( e_t = (e_{1t}, e_{2t})' \), \( Ee_t = 0 \), and

\[ \phi = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \quad \omega = Ee_t e_t' = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}. \]

In general \( \omega_{12} \neq 0 \).

- Notice that \( y \) is a column vector, and \( \phi \) is a matrix
- \( y_{1t} = \)
- \( y_{2t} = \)
- So each variable depends on the lagged values of itself and the lagged values of other variable.
- Using recursive substitution, the solution (or MA form with truncation) for equation (59) is

\[ y_t = \phi^t y_0 + \phi^{t-1} e_1 + \ldots + e_t. \]

- The series is stationary if the past shock has transient effect. So the stationarity condition states that

\[ \lim_{t \to \infty} \phi^t = 0, \quad \text{or} \quad |\lambda_i| < 1, i = 1, 2, \quad (60) \]

where \( \lambda_i \) is the eigenvalue of \( \phi \).
By definition, the eigenvalue $\lambda$ satisfies
\[
\phi v = \lambda v \quad (61)
\]
\[
\Rightarrow |\lambda I_2 - \phi| = 0 \quad (62)
\]
\[
\Rightarrow \begin{vmatrix} \lambda - \phi_{11} & -\phi_{12} \\ -\phi_{21} & \lambda - \phi_{22} \end{vmatrix} = 0 \quad (63)
\]
\[
\Rightarrow (\lambda - \phi_{11})(\lambda - \phi_{22}) - \phi_{12}\phi_{21} = 0, \quad (64)
\]
where $v$ is the eigenvector, and $I_2$ denotes a $2 \times 2$ identity matrix.

- The above equation is the characteristic equation for VAR(1).

- In general, the characteristic equation for a VAR(P) process is
\[
|\lambda^p I_n - \lambda^{p-1}\phi_1 - \ldots - \phi_p| = 0, \quad (65)
\]
where $n$ is the number of elements in $y_t$, and $|.|$ denotes the determinant of the matrix. Equation (65) is a polynomial equation with order $np$.

- Exercise: find the the characteristic equation for VAR(2) with $y_t = (y_{1t}, y_{2t})'$.

- The stationarity condition requires that all roots of (65) are less than one in absolute value (or lie inside unit circle).

- Define the MA form for VAR(p) as
\[
y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots. \quad (66)
\]

- In practice, the MA coefficients are obtained by simulating the impulse response function.

Appendix: Eigenvalue and Eigenvector

AR

1. Consider a univariate AR(2) process
\[
y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t.
\]
2. The characteristic equation is given by
\[ x^2 - \phi_1 x - \phi_2 = 0. \]

3. Now we show we can obtain the same characteristic equation by deriving the equation to solve for the eigenvalues of the vector AR(1) coefficient matrix.

(a) Rewrite the AR(2) as a vector AR(1)
\[ Y_t = \Phi Y_{t-1} + E_t, \]
where
\[ Y_t = \begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix}, \Phi = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}, E_t = \begin{pmatrix} e_t \\ 0 \end{pmatrix}. \]

(b) Let \( \lambda_1, \lambda_2 \) be the eigenvalues of \( \Phi \), and \( v_1, v_2 \) the corresponding eigenvectors. By definition of eigenvectors, we have
\[ \begin{vmatrix} \lambda - \phi_1 & -\phi_2 \\ -1 & \lambda \end{vmatrix} = 0 \Rightarrow \lambda^2 - \phi_1 \lambda - \phi_2 = 0 \]

(c) The roots for this equation are same as those for the characteristic equation. So the eigenvalue is also called characteristic value (root).

4. We can use the eigenvalues to diagonalize a matrix as follows

(a) By definition
\[ \Phi(v_1, v_2) = (v_1, v_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}. \]

(b) Let \( V = (v_1, v_2), \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \). Then the matrix \( \Phi \) is diagonalized as
\[ \Phi = V \Lambda V^{-1}. \]  \hspace{1cm} (67)

5. To see why the stationarity condition requires that \( |\lambda_i| < 1 \), rewrite the vector AR(1)
as
\[ V^{-1}Y_t = \Lambda V^{-1}Y_t + V^{-1}E_t, \]
or
\[ \zeta_t = \Lambda \zeta_{t-1} + \xi_t, \]
where \( \zeta_t = V^{-1}Y_t \) and \( \xi_t = V^{-1}E_t \). There are three remarks

(a) \( \Lambda \) is diagonal.
(b) \( y_t \) is a linear combination of elements of \( \zeta_t \):

\[ y_t = V^{(1)} \zeta_t = \zeta_{1,t} v_{11} + \zeta_{2,t} v_{12}, \]

where \( V^{(1)} \) is the first row of \( V \), and \( v_{11}, v_{12} \) the \((1,1)\) and \((1,2)\) element of \( V \).
(c) whether \( \zeta_t \) (and \( y_t \)) is stationary depends on \( \Lambda \), and so the eigenvalues of \( \Phi \). There are three possibilities:

i. \( |\lambda_1| < 1, |\lambda_2| < 1 \), then \( y_t \) is stationary.
ii. \( \lambda_1 = 1, |\lambda_2| < 1 \), then \( y_t \) is I(1). In this case \( |\phi_2| < 1 \).
iii. \( \lambda_1 = 1, \lambda_2 = 1 \), then \( y_t \) is I(2). In this case \( \phi_1 = 2, \phi_2 = -1 \).

6. So \( y_t \) is nonstationary if \( \lambda = 1 \) (and \( 1 = \phi_1 + \phi_2 \)). To see the implication on the transformation to achieve stationarity, rewrite the AR(2) model as

\[ \Delta y_t = \rho_1 y_{t-1} + \rho_2 \Delta y_{t-1} + e_t, \]

where \( \rho_2 = -\phi_2, \rho_1 = (\phi_1 + \phi_2 - 1) \).

(a) if \( y_t \) is stationary, \( \rho_1 \neq 0 \).
(b) if \( y_t \) is I(1), \( \rho_1 = 0 \), but \( \rho_2 < 1 \). So \( \Delta y_t \) is I(0).
(c) if \( y_t \) is I(2), \( \rho_1 = 0 \), and \( \rho_2 = 1 \). So \( \Delta y_t \) is I(1).

**VAR**

1. VAR can also be analyzed using eigenvalue and eigenvector.
2. Consider a VAR(1) given by

\[ y_t = \Phi y_{t-1} + e_t, \]
where \( y_t = (y_{1,t}, y_{2,t})' \) and \( \Phi = \{ \phi_{ij} \} \) is a 2 \( \times \) 2 matrix.

3. We can diagonalize the coefficient matrix as
\[
\Phi = V \Lambda V^{-1}.
\]

4. So we can rewrite the VAR(1) as
\[
\zeta_t = \Lambda \zeta_{t-1} + \xi_t,
\]
where \( \zeta_t = V^{-1}y_t \). Let \( V^{(-1)} \) and \( V^{(-2)} \) be the first and second row of \( V^{-1} \). Then
\[
V^{(-1)}y_t = \zeta_{1t},
\]
\[
V^{(-2)}y_t = \zeta_{2t}.
\]

5. By construction \( y_t = V\zeta_t \). So \( y_t \) is linear combination of \( \zeta_t \). Let \( V^{(1)} \) and \( V^{(2)} \) be the first and second row of \( V \). Then
\[
y_{1t} = V^{(1)}\zeta_t
\]
\[
y_{2t} = V^{(2)}\zeta_t
\]

6. There are three possibilities

(a) \(|\lambda_1| < 1, |\lambda_2| < 1\), then \( \zeta_{1t} = I(0), \zeta_{2t} = I(0), y_{1t} = I(0), y_{2t} = I(0) \).

(b) \( \lambda_1 = 1, |\lambda_2| < 1\), then \( \zeta_{1t} = I(1) \), and \( \zeta_{2t} = I(0) \). Now \( y_{1t} \) and \( y_{2t} \) are both \( I(1) \) and cointegrated. Because \( V^{(-2)}y_t = \zeta_{2t} = I(0) \), \( V^{(-2)} \) is the cointegrating vector.

(c) \( \lambda_1 = 1, \lambda_2 = 1 \), then \( y_{1t} \) and \( y_{2t} \) are both \( I(1) \) but not cointegrated.

7. Now consider the VAR in difference
\[
\Delta y_t = (\Phi - I)y_{t-1} + e_t.
\]
Note the eigenvalues for matrix \( (\Phi - I) \) are \( \lambda_1 - 1, \lambda_2 - 1 \).

(a) \(|\lambda_1| < 1, |\lambda_2| < 1\), then \( \text{rank}(\Phi - I) = 2 \). In this case it does not matter running VAR in level or in difference.
(b) \( \lambda_1 = 1, |\lambda_2| < 1 \), then \( \text{rank}(\Phi - I) = 1 \). As a result,

\[
(\Phi - I) = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & \lambda_2 - 1 \end{pmatrix} \begin{pmatrix} V^{(-1)} \\ V^{(-2)} \end{pmatrix} = (\lambda_2 - 1)v_2V^{(-2)},
\]

where \( v_2 \) is the (column) eigenvector for \( \lambda_2 \) and \( V^{(-2)} \) is the second row of \( V^{-1} \). \( V^{(-2)} \) is the cointegrating vector, and \( v_2 \) is the vector of loading coefficients. In words, \( (\Phi - I) \) can be written as outer product of a column vector and a row vector. In this case we should run VAR in difference and impose the restriction that the rank of \( (\Phi - I) \) is one.

(c) \( \lambda_1 = 1, \lambda_2 = 1 \), then \( \Phi - I = 0 \), \( \text{rank}(\Phi - I) = 0 \). In this case we should run VAR in difference.

**Estimation of Stationary VAR**

1. VAR is one example of Seemingly Unrelated Regression (SUR).

2. For an unrestricted VAR, because the independent variables on the right hand side (RHS) are the same, the efficient estimation of GLS is the same as equation-by-equation OLS.

3. So we fit each regression by OLS, and keep the residual \( \hat{e}_t = (\hat{e}_{1t}, \hat{e}_{2t})' \). The variance matrix is estimated as

\[
\hat{\omega} = T^{-1} \sum \hat{e}_t \hat{e}_t'.
\]

The off-diagonal term measures the contemporaneous correlation between the errors (shocks).

4. In practice the number of lags \( p \) is selected by AIC. The optimal model minimizes AIC.

5. An adequate model has vector of white noise residuals. The Ljung-Box test can check that.

6. The R package for running VAR model is `vars`, and the main command is `VAR`. Notice that R is case-sensitive.
Granger Causality

1. Variable \( y_i \) Granger causes variable \( y_j \) if the null hypothesis of
\[
\beta_1 = \beta_2 = \ldots = \beta_p = 0
\]
can be rejected for the regression
\[
y_{jt} = \sum_{k=1}^{p} \beta_k y_{i,t-k} + \text{other regressors} \quad (68)
\]

2. Granger Causality \( \neq \) Causality.

3. At best, \( y_i \) Granger causes \( y_j \) only implies that \( y_i \) helps predict \( y_j \) in sample.

4. The variables should have the same integrated orders. If one is \( I(0) \) but the other is \( I(1) \), then we need to difference the \( I(1) \) variable before testing Granger Causality.

5. The Granger Causality test is sensitive to \( p \). We usually conduct the Granger Causality test based on the optimal (AIC-minimizing) model.

Impulse Response

1. Consider a VAR(1)
\[
y_t = \phi y_{t-1} + e_t.
\]

2. We can simulate the impulse response to a shock in \( e_{1t} \) by letting \( y_{t-j} = 0, e_{t-j} = e_{t+j} = 0, e_t = (1, 0) \), and computing \( y_t, y_{t+1}, \ldots \) recursively.

3. We can simulate the impulse response to a shock in \( e_{2t} \) by letting \( y_{t-j} = 0, e_{t-j} = e_{t+j} = 0, e_t = (0, 1) \), and computing \( y_t, y_{t+1}, \ldots \) recursively.

4. In theory, the impulse response function is the same as the MA form. For VAR, the \((i,j)\) element of the MA coefficient matrix \( \theta_s \) gives the effect of one-unit change of \( e_{jt} \) on \( y_{i,t+s} \).

5. More generally, there are two ways to get the impulse response function (irf).
(a) (Method 1, by MA representation) Suppose the estimated VAR(p) model is given by

\[
y_t = \hat{\phi}_1 y_{t-1} + \ldots + \hat{\phi}_p y_{t-p} + \hat{\epsilon}_t. \tag{69}
\]

The MA(∞) representation can be written as

\[
y_t = \hat{\epsilon}_t + \theta_1 \hat{\epsilon}_{t-1} + \theta_2 \hat{\epsilon}_{t-2} + \ldots,
\]

where the coefficients \( \theta_i \) can be obtained by the method of undetermined coefficients:

\[
(1 - \hat{\phi}_1 L - \ldots - \hat{\phi}_p L^p)(1 + \theta_1 L + \theta_2 L^2 + \ldots) = 1.
\]

Then

\[
\frac{\partial y_{t+s}}{\partial \hat{\epsilon}_t} = \theta_s.
\]

The effect of one-unit change of \( \epsilon_{jt} \) on \( y_{i,t+s} \) is given by the \((i,j)\) element of \( \theta_s \).

(b) (Method 2, by simulation). Let \( \hat{\epsilon}_t = (1, 0, \ldots, 0)' \), \( y_{t-1} = \ldots = y_{t-p} = 0 \), and \( \hat{\epsilon}_{t+j} = 0, \forall j > 0 \). Generate \( y_{t+j}, (j \geq 0) \), recursively by using (69). We then get irf when one unit change happens to the error of \( y_{1t} \). By changing \( \hat{\epsilon}_t \) we can get other irfs in a similar manner.

**IRFs of Orthogonal Innovations**

If \( \hat{\Omega} \) is not diagonal (or equivalently, the errors across equations are contemporaneously correlated), then the assumptions that only \( \epsilon_{jt} \) changes while errors in other equations are held constant is implausible. Suppose we have the following Cholesky factorization:

\[
\hat{\Omega} = PP',
\]

where \( P \) is a unique lower triangular matrix. Then the transformed innovations given by

\[
u_t = P^{-1} \epsilon_t
\]

is orthogonal, since

\[
\text{Var}(u_t) = P^{-1} \text{Var}(\epsilon_t) P'^{-1} = P^{-1} \hat{\Omega} P'^{-1} = P^{-1} PP' P'^{-1} = I.
\]
Now it makes sense to derive the irfs with respect to one-unit change of the new innovations, $u_t$. The new irfs is given by

$$\frac{\partial y_{t+s}}{\partial u_t} = \theta_s P. \quad (70)$$

The effect of one-unit change of $u_{jt}$ on $y_{i,t+s}$ is given by the $(i,j)$ element of $\theta_s P$. Note that the effect of one-unit change of $u_t$ on $y_t$ is given by $\theta_0 P = IP = P$. Because $P$ is a lower triangular matrix, that means only $u_{1t}$ has effect on $y_{1t}$, while both $u_{1t}$ and $u_{2t}$ have effects on $y_{2t}$. This fact is called **Sims Orthogonalization**.

More generally, using the matrix polynomial of lag operators we express a VAR($p$) as

$$A(L)y_t = e_t,$$

for which the MA($\infty$) representation is given by

$$y_t = B(L)e_t,$$

where $B(L) = A(L)^{-1}$. We can interpret $B(L)$ as the impulse response function for original innovations. Note that $A(0) = B(0) = I$. Since $\Omega = Ee_t e_t$ is not necessarily diagonal, we transform the original innovations in order to get orthogonal innovations. **Sims Orthogonalization** is based on the lower triangular matrix $P$ satisfying $\hat{\Omega} = PP'$, and the new irfs based on the transformed innovations are $C(L) \equiv B(L)P$. It is shown that

$$C(0) = B(0)P$$

is lower triangular. Sims believe this is a nice property.

**Blanchard-Quah orthogonalization** emphasizes different property of $C(L)$, and requires $C(1)$ is lower triangular. The orthogonalization matrix $Q$ (not lower triangular) should satisfy

1. $\hat{\Omega} = QQ'$,
2. $C(1) = B(1)Q$ is lower triangular.

Example: find the Blanchard-Quah orthogonalization for (59).

- $B(L) =$
- $B(1) =$
Let \( Q = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} \), then we can get a unique solution for those \( q_s \).

**Forecasting VAR**

Consider a VAR(1)

\[
y_t = \phi y_{t-1} + e_t.
\]

We can show that

- \( y_{t+1} = \)
- \( \hat{y}_{t+1} = \)
- \( \text{var}(\hat{y}_{t+1}) = \)

In general, by recursive iteration we can show for the VAR(1)

\[
\hat{y}_{t+s} = \phi^s y_t \quad (71)
\]

\[
\text{var}(\hat{y}_{t+s}) = \Omega + \phi \Omega \phi' + \phi^2 \Omega \phi'^2 + \ldots + \phi^{s-1} \Omega \phi'^{s-1}. \quad (72)
\]

To get forecast for a general VAR(p) process, we need to transform VAR(p) into VAR(1) using the matrix \( F \). In practice, both \( \phi \) and \( \Omega \) are estimated. Yet we ignore the estimation uncertainty when compute the forecast variance (by treating \( \hat{\phi} \) and \( \hat{\Omega} \) as the true parameters).

In general, suppose the MA representation of VAR(p) is given by

\[
y_t = \sum_{i=0}^{\infty} \Phi_i e_{t-i},
\]

where \( \Phi_0 = I \), and \( \Omega = E e_t e_t' \). This MA representation can be obtained via the method of undetermined coefficients. Then the variance for \( s \)-step ahead forecast is given by

\[
\text{var}(\hat{y}_{t+s}) = \Phi_0 \Omega \Phi_0' + \Phi_1 \Omega \Phi_1' + \ldots + \Phi_{s-1} \Omega \Phi_{s-1}' = \sum_{i=0}^{s-1} \Phi_i \Omega \Phi_i'.
\]

**Variance Decomposition**

Consider the Cholesky decomposition of \( \Omega \):

\[
PP' = \Omega.
\]
The above forecast variance can be rewritten as

$$\text{var}(\hat{y}_{t+s}) = \sum_{i=0}^{s-1} \Phi_i \Omega \Phi_i' = \sum_{i=0}^{s-1} \Phi_i PP\Phi_i' = \sum_{i=0}^{s-1} \Theta_i \Theta_i',$$

where $\Theta_i \equiv \Phi_i P$ is the coefficient for orthogonal errors of $u_t = P^{-1}e_t$. The (1,1) element of $\text{var}(\hat{y}_{t+s})$ is the forecast variance for $y_1$, and (2,2) element for $y_2$, and etc. Let $\Theta_i(jk)$ denote the $(j, k)$ element of $\Theta_i$. Then we can show that forecast variance for $y_1$ is the sum of two series

$$\text{var}(\hat{y}_{1,t+s}) = (\Theta_0^2(11)+\Theta_1^2(11)+\ldots+\Theta_{s-1}^2(11))+(\Theta_0^2(12)+\Theta_1^2(12)+\ldots+\Theta_{s-1}^2(12)) = S_{11} + S_{12}.$$

Similarly we can show

$$\text{var}(\hat{y}_{2,t+s}) = (\Theta_0^2(21)+\Theta_1^2(21)+\ldots+\Theta_{s-1}^2(21))+(\Theta_0^2(22)+\Theta_1^2(22)+\ldots+\Theta_{s-1}^2(22)) = S_{21} + S_{22}.$$

Now

1. $S_{11}/\text{var}(\hat{y}_{1,t+s})$ gives the proportion of forecast variance of $y_1$ that comes from $u_1$.
2. $S_{12}/\text{var}(\hat{y}_{1,t+s})$ gives the proportion of forecast variance of $y_1$ that comes from $u_2$.
3. $S_{21}/\text{var}(\hat{y}_{2,t+s})$ gives the proportion of forecast variance of $y_2$ that comes from $u_1$.
4. $S_{22}/\text{var}(\hat{y}_{2,t+s})$ gives the proportion of forecast variance of $y_2$ that comes from $u_2$.

**Structural VAR**

So far we discuss the VAR in **standard form**. It is standard because $y_{1t}$ does not contemporaneously affect $y_{jt}$. The structural VAR(1), given by the following, does allow for that contemporaneous feedback.

$$y_{1t} = \pi_1 y_{2t} + \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + u_{1t}$$
$$y_{2t} = \pi_2 y_{1t} + \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + u_{2t},$$

where

$$E u_{1t} u_{2t} = 0$$
by assumption. Let \( y_t = (y_{1t}, y_{2t})' \), \( u_t = (u_{1t}, u_{2t})' \) and

\[
\Pi = \begin{bmatrix}
1 & -\pi_1 \\
-\pi_2 & 1
\end{bmatrix}, \quad \phi = \begin{bmatrix}
\phi_{11} & \phi_{12} \\
\phi_{21} & \phi_{22}
\end{bmatrix}
\]

we can write the structural VAR(1) as

\[
\Pi y_t = \phi y_{t-1} + u_t. \tag{73}
\]

The structural VAR (73) cannot be estimated by OLS. To see this, note that in Appendix we show

\[
E(y_{2t} u_{1t}) \neq 0,
\]

meaning that the regressor in the first regression is correlated with the error. As a result, OLS estimator is inconsistent.

However, we can always get a consistent OLS estimate for \( \text{VAR in the standard or reduced form} \) obtained by pre-multiplying (73) by \( \Pi^{-1} \):

\[
y_t = \Pi^{-1} \phi y_{t-1} + \Pi^{-1} u_t \equiv \varphi y_{t-1} + e_t. \tag{74}
\]

We can show that in general \( Ee_te'_t \) is not diagonal, meaning that the reduced-form innovations are contemporaneously correlated though the structural innovations are contemporaneously uncorrelated. The indirect least squares (ILS) then are computed to recover parameters in structural form from the reduce-form parameter.

**Structural Impulse-Response Function**

In practice we are more interested in the irf with respect to the structural innovations, \( u_t \). Let’s call this irf the *structural impulse response function* (sirf). We have two equivalent ways to find sirf.

The first way is to compute the Cholesky decomposition of the covariance-variance matrix of reduced-form residuals: \( \Omega = E e_t e'_t = PP' \). Then the sirf is given by (70) in which \( \theta_s \) can be obtained either by simulation (easy) or solving the MA(\(\infty\)) representation of the VAR in standard form (or reduced form of structural VAR).
The second way is easier. Let \( u_t = (1, 0, \ldots) \) and solve \( e_t = Pu_t \). Then we compute the irf (by simulation) with respect to this \( e_t \). Now we get the sirf with respect to a one-unit change in \( u_{1t} \). The sirf with respect to a one-unit change in other elements of \( u_t \) can be obtained in a similar fashion by resetting \( u_t \).

**Appendix: Optional**

**SUR**

Consider two *seemingly unrelated regressions* (SUR):

\[
\begin{align*}
y_{1t} &= \beta_1' x_{1t} + e_{1t} \\
y_{2t} &= \beta_2' x_{2t} + e_{2t},
\end{align*}
\]

which can be written in matrix form as

\[
\begin{align*}
Y_1 &= X_1 \beta_1 + e_1 \\
Y_2 &= X_2 \beta_2 + e_2.
\end{align*}
\]

The estimators for \( \beta_i \) by equation-by-equation LS are consistent, but not efficient. To construct the efficient estimator, first stack regressions as

\[
\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},
\]

or

\[ Y = X \beta + e. \]

Note that

\[
\Sigma = Eee' = \begin{pmatrix} Ee_1 e_1' \\ Ee_2 e_1' \\ Ee_2 e_2' \end{pmatrix} = \begin{pmatrix} \omega_{11} I & \omega_{12} I \\ \omega_{21} I & \omega_{22} I \end{pmatrix} = \Omega \otimes I.
\]

The efficient GLS estimator, minimizing \( e' \Sigma^{-1} e \) instead of \( e' e \), is given by

\[
\hat{\beta}_{GLS} = (X' \Sigma^{-1} X)^{-1} (X' \Sigma^{-1} Y),
\]

where \( \Sigma^{-1} = \Omega^{-1} \otimes I \). We can show that \( \hat{\beta}_{GLS} = \hat{\beta}_{OLS} \) under the following two conditions.
1. when $X_1 = X_2 = x$,

\[
\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}(X'\Sigma^{-1}Y) = (\Omega^{-1} \otimes x'x)^{-1}[(I \otimes x')(\Omega^{-1} \otimes I)]Y = (I \otimes (x'x)^{-1})Y = \hat{\beta}_{GLS}.
\]

2. when $\omega_{12} = \omega_{21} = 0$, the covariance matrix is diagonal. The result follows directly.

In general, $\hat{\beta}_{GLS}$ is obtained by solving

\[
\hat{\beta}_{GLS} = \left(\begin{array}{cc}
\omega_{11}X_1'X_1 & \omega_{12}X_1'X_2 \\
\omega_{21}X_2'X_1 & \omega_{22}X_2'X_2
\end{array} \right)^{-1} \left(\begin{array}{c}
\omega_{11}X_1'Y_1 + \omega_{12}X_1'Y_2 \\
\omega_{21}X_2'Y_1 + \omega_{22}X_2'Y_2
\end{array} \right).
\]

Note for most VAR, the first condition (identical regressors) is satisfied. Hence equation-by-equation OLS is GLS and so efficient.

**SEM**

A simultaneous equation model (SEM) or structural model, in which the RHS variables are possibly endogenous, is given by

\[
y_{1t} = c_{10}y_{2t} + c'_{11}x_{1t} + e_{1t} \\
y_{2t} = c_{20}y_{1t} + c'_{21}x_{2t} + e_{2t}.
\]

The reduced form for $y_{1t}$, in which all RHS variables are exogenous, is given by

\[
y_{1t} = \frac{c'_{11}x_{1t}}{1 - c_{10}c_{20}} + \frac{c_{10}c'_{21}x_{2t}}{1 - c_{10}c_{20}} + \frac{c_{10}e_{2t}}{1 - c_{10}c_{20}} + e_{1t}.
\]

In general

\[
Ey_{1t}e_{2t} = E\left(\frac{c_{10}e_{2t}e_{2t}}{1 - c_{10}c_{20}} + e_{1t}e_{2t}\right) \neq 0,
\]

and this implies bias of simultaneity. So the OLS estimator for each regression is inconsistent. The following two estimators, however, are consistent.

1. If the SEM is just identified (the number of structural-form parameters equals the number of reduced-form parameters), the indirect least square (ILS) estimators are
consistent (although not unbiased): we use LS to estimate the reduced-form parameters, and then convert into estimates of the structural-form parameters. The variance of ILS is computed by Delta method due to nonlinearity.

2. We can also use \textit{two stage least square} method (2SLS). In the first stage we use LS to estimate the reduced-form and obtain fitted values. The second stage uses LS on the structural equation after the fitted values replace the observed values on the right hand side. Even though the fitted values are used in calculating coefficients, the original values are used in calculating residuals. For example, the residual in the regression for $y_{1t}$ is computed as

$$\hat{e}_{1t} = y_{1t} - \hat{c}_{10}y_{2t} - \hat{c}'_{11}x_{1t},$$

where $\hat{c}$ is obtained by applying LS to the regression

$$y_{1t} = c_{10}\hat{y}_{2t} + c_{11}'x_{1t} + \text{error},$$

and $\hat{y}_{2t}$ is the fitted value from reduced-form regression. The squared standard error of regression (SER) then is computed as $\hat{\sigma}^2 = n^{-1}\sum\hat{e}_{1t}^2$, and the variance of 2SLS is given by $\hat{\sigma}^2(Z'Z)^{-1}$, where $z_t = (\hat{y}_{2t}, x_{1t})$. This adjustment is done by SAS automatically.
Sample Test 2

You cannot get points by showing just final results. Show me the solving process.

1 (7 points). Find the 1-step-ahead and 2-step-ahead forecast of conditional variance \( \hat{h}_{t+1}, \hat{h}_{t+2} \) for the GARCH(1, 2) process \( h_t = \omega + a_1 \epsilon^2_{t-1} + a_2 h_{t-1} + a_3 h_{t-2} \). where \( \epsilon_t = \sqrt{h_t} v_t, v_t \sim \text{iidn}(0, 1) \)

2 (7 points). Please specify the testing regression based on which you run the unit root test for a *trending* variable. What is the null hypothesis?

3 (7 points). Tom is interested in the permanent income hypothesis. So he regresses total consumption onto GDP, and gets a significant coefficient less than unity for GDP. Then Tom believes that this result is consistent with the theory. Is Tom right or wrong? Under what condition does Tom’s regression make sense? How to check that condition? (Suppose both total consumption and GDP are I(1) variables).

Consider a two-variable, first-order VAR

\[
\begin{align*}
y_{1t} & = 0.4 y_{1t-1} + 0.1 y_{2t-1} + e_{1t} \\
y_{2t} & = 0.2 y_{1t-1} + 0.5 y_{2t-1} + e_{2t} \\
Ee^2_{1t} & = 1, Ee^2_{2t} = 1, Ee_{1t}e_{2t} = 0.
\end{align*}
\]

4 (4 points). Check the stationarity condition.

5 (5 points). Find the 3-step-ahead impulse response of \( y_{1t} \) after the shock \( (e_{1t}, e_{2t}) = (1, 0) \).

6 (5 points). What proportion of the 2-step-ahead forecasting variance for \( y_{2t} \) comes from \( e_{2t} \)?
Lecture 6: Cointegration

VAR with Nonstationary Series

1. Now we know how to fit a VAR model for stationary series. If some of the series are nonstationary, I(1), there are two possibilities
   
   (a) If they are not cointegrated, fit a VAR for differenced series.
   
   (b) If they are cointegrated, fit an Error Correction Model (ECM).

2. (a) Consider a VAR(2) model for $y_t = (y_{1t}, y_{2t})'$:

   $$y_t = \phi_1 y_{t-1} + \phi_2 y_{t+2} + e_t$$

   (b) We can always rewrite the model as

   $$\Delta y_t = \gamma_0 y_{t-1} + \gamma_1 \Delta y_{t-1} + e_t$$

   $\gamma_0 = \phi_1 + \phi_2 - 1$

   $\gamma_1 = -\phi_2$

   (c) The three possibilities are

   i. If $\gamma_0 = 0$, or equivalently, $\text{rank}(\gamma_0) = 0$, the series are nonstationary, and not cointegrated. We need to fit a VAR for differenced series:

   $$\Delta y_t = \gamma_1 \Delta y_{t-1} + e_t$$

   ii. If $\text{rank}(\gamma_0) = 1 < 2$, the series are nonstationary, but cointegrated. Then we can write $\gamma_0$ as the product of a row vector, $\alpha$ and a column vector, $\beta'$. In other words, we need to run the error correction model of

   $$\Delta y_t = \alpha z_{t-1} + \Delta y_{t-1} + e_t$$

   $$z_{t-1} = \beta' y_{t-1}$$

   In the error correction model all variables including $z_{t-1}$ are stationary. That implies (i) there is one linear combination of nonstationary variables given by $\beta' y_{t-1}$ that is stationary, (ii) $y_t$ is cointegrated, (iii) $\beta$ is the cointegrating vector, (iv) $z_{t-1}$ is the error correction term, measuring disequilibrium.
iii. If \( \text{rank}(\gamma_0) = 2 \), the series are stationary. Then just fit the original VAR(2) model.

3. In general let \( y_t = (y_{1t}, \ldots, y_{mt})' \), and consider a VAR(p) model

\[
y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + e_t, \tag{75}
\]

We can always rewrite the model as

\[
\Delta y_t = \gamma_0 y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{p-1} \Delta y_{t-p+1} + e_t. \tag{76}
\]

There are three possibilities

(a) If \( \text{rank}(\gamma_0) = 0 \), series are nonstationary and not cointegrated. We need to run VAR(p-1) for the differenced series, i.e., model (76) without \( y_{t-1} \).

(b) If \( 0 < \text{rank}(\gamma_0) < m \), series are nonstationary and cointegrated. We need to run error correction model.

\[
\begin{align*}
\Delta y_t &= \alpha z_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{p-1} \Delta y_{t-p+1} + e_t \tag{77} \\
z_{t-1} &= \beta y_{t-1} \tag{78}
\end{align*}
\]

(c) If \( \text{rank}(\gamma_0) = m \), series are stationary. We need to run the original VAR(p) model (75).

**Cointegration Test**

If at most one cointegration exists, use Engle-Granger Test. If the number of cointegrations is possibly greater than one, use Johansen Test.

1. Follow these steps to conduct the Engle-Granger Test

(a) Engle-Granger test is applicable if only one cointegration relation possibly exists.

(b) If the cointegrating vector \( \beta \) is known, then just apply the Augmented Dickey-Fuller test to \( z_t = \beta'y_t \). The variables are cointegrated if the null of non-stationarity can be rejected for \( z_t \). You can use the tables of critical values for ADF test.

(c) If the cointegrating vector \( \beta \) is unknown, then estimate the cointegrating vector first and obtain the residual

\[
\hat{\beta}' y_t = \hat{e}_t.
\]

89
Next conduct ADF test to $\hat{e}_t$. The Engle-Granger test does \textit{not} follow the same distribution as the ADF test.

(d) The result of Engle-Granger Test is sensitive to the choice of dependent variable in the regression $\hat{\beta}'y_t = \hat{e}_t$.

2. Follow these steps to conduct the Johansen Test

(a) Run model (76)

(b) Compute the ordered squared canonical correlation between $\Delta y_t$ and $y_{t-1}$ after controlling for $\Delta y_{t-1}, \ldots, \Delta y_{t-p+1}$, calling them $\rho_1^2 > \rho_2^2 > \ldots > \rho_m^2$.

(c) There are two tests for the null hypothesis

\[ H_0 : \text{There are } k \text{ or less cointegrating vectors} \]

i. The trace test is \textbf{Trace Test} $= -T \sum_{i=k+1}^{m} \ln(1 - \rho_i^2)$

ii. The maximal eigenvalue test is \textbf{Maximal Test} $= -T \ln(1 - \rho_{k+1}^2)$

(d) Compare the test statistic to the appropriate critical values. Under the null hypothesis $\rho_i^2, i = k + 1, \ldots, m$ are all close to zero, and both tests should be very small since $\ln(1 - \rho_i^2) \approx 0, i = k + 1, \ldots, m$. Hence the null is rejected by a large test statistic

\textbf{Johansen Methodology for Modeling Cointegration}

1. First test $H_0 : r = 0$ against $H_1 : r > 0$. No cointegration exists if the null is not rejected. At least one cointegration exists if the null is rejected. In that case, proceed to test $H_0 : r = 1$ against $H_1 : r > 1$, and so on. This sequential procedure is continued until the null is \textit{not} rejected.

2. Given the number of cointegration estimated by the above sequential procedure, impose normalization and identifying restriction on the cointegrating vectors.

3. Given the normalized cointegrating vectors, estimate the resulting cointegrated error correction model by MLE.
Guidance for Empirical Research

- **Step 1**: check stationarity of data by applying unit root test. If data is stationary, then running OLS for both VAR in level and VAR in difference is fine. If data is nonstationary, go to step 2.

- **Step 2**: check cointegration by applying cointegration test. If data is not cointegrated, then run OLS for VAR in *difference*. If data is cointegrated, go to step 3.

- **Step 3**: run error correction model.
Appendix: Optional

Multivariate Normal Distribution

1. Consider \( n \) independent standard normal random variables \( z = (z_1, \ldots, z_n)' \). Their joint density is the product of marginal densities

\[
f(z_1, \ldots, z_n) = (2\pi)^{-n/2} \exp \left( -\frac{\sum_{i=1}^{n} z_i^2}{2} \right) = (2\pi)^{-n/2} \exp \left( -\frac{z'z}{2} \right).
\]

2. Consider a linear nonsingular transformation of

\[
x = Pz, E(xx') = PE(zz')P' = \Omega,
\]

where

\[
P = \Omega^{1/2}
\]

is the Cholesky factorization of \( \Omega \).

3. Now the joint density for the new random vector is

\[
f(x_1, \ldots, x_n) = |J| f(z_1, \ldots, z_n)
\]

\[
= \left| \frac{dz}{dx} \right| (2\pi)^{-n/2} \exp \left( -\frac{(P^{-1}x)'(P^{-1}x)}{2} \right)
\]

\[
= |P^{-1}|(2\pi)^{-n/2} \exp \left( -\frac{(P^{-1}x)'(P^{-1}x)}{2} \right)
\]

\[
= (2\pi)^{-n/2} |\Omega|^{-1/2} \exp \left( -\frac{x'\Omega^{-1}x}{2} \right)
\]

where \( |J| \) denotes the determinant of Jacobian, and the last equality is due to \( |P^{-1}| = |\Omega^{-1/2}| = |\Omega|^{-1/2} \).

4. So the joint density for multivariate normal vector \( x \) with \( Ex = \mu, \text{var}(x) = \Omega \) is

\[
f(x_1, \ldots, x_n) = (2\pi)^{-n/2} |\Omega|^{-1/2} \exp \left( -\frac{(x - \mu)'\Omega^{-1}(x - \mu)}{2} \right)
\]

5. If \( \Omega \) is diagonal (data are uncorrelated) the joint density can be written as product of marginal densities. So uncorrelated normal variables are independent.
6. Suppose we have i.i.d sample of multivariate vectors \( \mathbf{x} = \{x_t\}_1^T \). The joint density for the random sample is given by

\[
f(\mathbf{x}) = (2\pi)^{-nT/2} |\Omega|^{-T/2} \exp \left( -\sum_{t=1}^T (x_t - \mu)' \Omega^{-1} (x_t - \mu) \right) = (2\pi)^{-nT/2} |\Omega|^{-T/2} \exp \left( -(\mathbf{x} - \mu \otimes I_T)' \Omega^{-1} (\mathbf{x} - \mu \otimes I_T) \right)\]

7. The MLE estimator for \( \mu \) minimizes

\[
\min (\mathbf{x} - \mu \otimes I_T)' \Omega^{-1} (\mathbf{x} - \mu \otimes I_T). 
\]

8. When the likelihood is evaluated at the MLE estimates, the likelihood is proportional to

\[
L \propto |\Omega|^{-T/2}. 
\]
ML Estimation of Cointegrated VAR

1. Consider an VAR(p) process for $y_t = (y_{1t}, \ldots, y_{nt})'$:

$$y_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + e_t, \quad (t = 1, \ldots, T)$$

The model is adequate so that $e_t$ is i.i.d.

2. We can rewrite the model as

$$\Delta y_t = \beta_0 y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_{p-1} \Delta y_{t-p+1} + e_t$$

where

$$\beta_0 = a_1 + a_1 + \ldots + a_p - 1.$$

3. Suppose $g$ of the eigenvalues of $\beta_0$ are 0, and $r = n - g$ of eigenvalues are non-zeroes (negative), then we can factorize $\beta_0$ as

$$\beta_0(c_1, c_2) = (c_1, c_2) \Lambda,$$

which implies that

$$\beta_0 = (c_1, c_2) \begin{pmatrix} 0_{g \times g} & 0 \\ 0 & \lambda_{r \times r} \end{pmatrix} \begin{pmatrix} c^{-1} \\ c^{-2} \end{pmatrix} = \lambda c_2 c^{-2},$$

where $c_2$ is the (column) eigenvectors for the nonzero eigenvalues, and $c^{-2}$ is the (row) vector of the inverse of $(c_1, c_2)$.

4. In short, under the hypothesis that

(a) $y_t$ consists of $g$ common stochastic trends, or equivalently

(b) there are $r$ linear combinations of $y_t$ are stationary, or equivalently

(c) the rank of $\text{rank}(\beta_0) = r$

the matrix $\beta_0$ can be written as (up to) a product of the $n \times r$ matrix $c_2$ and the $r \times n$ matrix $c^{-2}$

$$\beta_0 = \alpha \beta'.$$
5. Because all variables in the regressions, including \( \beta' y_t \), are I(0), \( y_t \) is cointegrated, and \( \beta \) is cointegrating vector.

6. The cointegrating vector \( \beta \) cannot be uniquely determined. But we can estimate the space spanned by \( \beta \). In other words, some normalization is necessary.

7. Now under the assumption that \( \text{rank}(\beta_0) = r \), the model is written as

\[
\Delta y_t = \alpha \beta' y_{t-1} + \beta_1 \Delta y_{t-1} + \ldots + \beta_{p-1} \Delta y_{t-p+1} + e_t,
\]

for suitable \( n \times r \) matrices \( \alpha \) and \( \beta \). This is a nonlinear regression.

8. It is incorrect to estimate the above regression by OLS, because doing so ignores the restriction that \( \text{rank}(\beta_0) = r \).

9. The right approach is to estimating the regression by MLE and impose the restriction that \( \text{rank}(\beta_0) = r \).

10. Denote the residual of regressing \( \Delta y_t \) onto \( (\Delta y_{t-1}, \ldots, \Delta y_{t-p+1}) \) by \( r_{0t} \), and denote the residual of regressing \( y_{t-1} \) onto \( (\Delta y_{t-1}, \ldots, \Delta y_{t-p+1}) \) by \( r_{1t} \). Then the concentrated likelihood function is proportional to

\[
L \propto |\Omega|^{-T/2} \exp \left[ -0.5 \sum_{t=1}^{T} (r_{0t} + \alpha \beta' r_{1t})' \Omega^{-1} (r_{0t} + \alpha \beta' r_{1t}) \right].
\]

11. For given \( \beta \), we can regress \( r_{0t} \) onto \( -\beta' r_{1t} \) and obtain LS estimates

\[
\hat{\alpha}(\beta) = -S_{01} \beta (\beta' S_{11} \beta)^{-1} \\
\hat{\Omega}(\beta) = S_{00} - S_{01} \beta (\beta' S_{11} \beta)^{-1} \beta' S_{10} \\
S_{ij} = T^{-1} \sum r_{it} r_{jt}', \quad (i, j = 0, 1)
\]

12. In particular, the estimate for \( \beta_0 \) is

\[
\hat{\beta}_0(\beta) = \alpha \beta' = -S_{01} \beta (\beta' S_{11} \beta)^{-1} \beta',
\]

and the maximized likelihood is proportional to

\[
L \propto \left| \hat{\Omega}(\beta) \right|^{-T/2} = |S_{00} - S_{01} \beta (\beta' S_{11} \beta)^{-1} \beta' S_{10}|^{-T/2}.
\]
13. Now the question becomes

\[ \min |S_{00} - S_{01}\beta'(S_{11}\beta)^{-1}\beta'S_{10}| = \min \left\{ \frac{|\beta'S_{11}\beta - \beta'S_{10}S_{00}^{-1}S_{01}\beta|}{|\beta'S_{11}\beta|} \right\}, \]

where the minimisation is over all \( n \times r \) matrices \( \beta \).

14. Basically we are minimizing \( 1 - R^2 = (|\beta'S_{11}\beta - \beta'S_{10}S_{00}^{-1}S_{01}\beta|)/(|\beta'S_{11}\beta|) \), or equivalently, maximizing \( R^2 \). This problem is called canonical regression, and \( R^2 \) is squared canonical correlation.

15. We estimate \( \beta \) to be the first \( r \) eigenvectors of \( S_{10}S_{00}^{-1}S_{01} \) with respect to \( S_{11} \). In other words, Let \( \lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \) be the eigenvalues of \( S_{11}^{-1}S_{10}S_{00}^{-1}S_{01} \), then

\[ \hat{\beta} = (c_{\lambda_1}, c_{\lambda_2}, \ldots, c_{\lambda_r}), \]

where \( c_{\lambda_i} \) is the eigenvector for \( \lambda_i \).

16. So \( \hat{\beta} \) is the first \( r \) canonical variates and the eigenvalues are the squared canonical correlations.

17. We normalize \( \hat{\beta} \) so that

\[ \hat{\beta}'S_{11}\hat{\beta} = I. \]

Under this normalization we have

\[ \hat{\alpha} = -S_{01}\hat{\beta} \]
\[ \hat{\Omega} = S_{00} - S_{01}\hat{\beta}'S_{10} = S_{00} - \hat{\alpha}\hat{\alpha}' \]
\[ \hat{\beta}_0 = -S_{01}\hat{\beta}\hat{\beta}'. \]

18. More importantly, the maximized likelihood is

\[ L_{\text{max}}^{-2/T} = |S_{00}|\prod_{i=1}^{r}(1 - \lambda_i), \]

or

\[ -2 \log L_{\text{max}} \propto T \sum_{i=1}^{r} \log(1 - \lambda_i). \]
19. Therefore the likelihood ratio test (or Johansen Cointegration Test) for

\[ H_0 : r \leq r_0 \quad \text{versus} \quad H_0 : r = n \]

is given by

\[ -2 \log(Q) = -T \sum_{i=r+1}^{n} \log(1 - \lambda_i), \]

where \( \lambda_{r+1}, \ldots, \lambda_n \) are the \( n - r \) smallest eigenvalues of \( S_{11}^{-1}S_{10}S_{00}^{-1}S_{01} \). \( r \) is the number of cointegration relations. \( r \) is also the rank of \( \beta_0 \).