Unit Roots, Cointegration, VARs and VECMs

L. Magee   Winter, 2013

_____

1 Vector Autoregressions (VAR)

1.1 Notation and assumptions

An m-variable VAR with p lags is written unsing vector-matrix notation as:

$$y_t \;=\; \mu + \sum_{i=1}^{p} \Gamma_i y_{t-i} + \epsilon_t = \mu + \Gamma_1 y_{t-1} + \ldots + \Gamma_p y_{t-p} + \epsilon_t \tag{1}$$

where y_t is an $m \times 1$ vector and the Γ_i's are $m \times m$ coefficient matrices.

The $m \times 1$ vector ϵ_t is assumed to have mean zero ($E\epsilon_t = 0$), with no autocorrelation ($E\epsilon_t \epsilon'_{t-s} = 0$ for all $s \neq 0$) but can be correlated across equations. That is, $E\epsilon_t \epsilon'_t = \Omega$, and Ω may have non-zero off-diagonal elements.

Let $m = 3$, where the three variables' values at time t are $y_{1t} = \log$ GDP, $y_{2t} = \log$ CPI, and $y_{3t} = $ interest rate. Then

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix}$$

is a 3×1 vector, its first variable, y_{1t}, being the tth observation of log GDP, y_{2t} is the tth observation of log CPI, and y_{3t} is the tth observation of the interest rate.

Suppose $p = 4$. Then the VAR (1) represents a three equation dynamic model, where each equation

contains an intercept and the same $p \times m = 3 \times 4 = 12$ regressors:

$$
\begin{aligned}
y_{1t} &= \mu_1 + \gamma_{1(11)} y_{1,t-1} + \ldots + \gamma_{4(11)} y_{1,t-4} + \gamma_{1(12)} y_{2,t-1} + \ldots + \gamma_{4(12)} y_{2,t-4} \\
&\quad + \gamma_{1(13)} y_{3,t-1} + \ldots + \gamma_{4(13)} y_{3,t-4} + \epsilon_{1t} \\
y_{2t} &= \mu_2 + \gamma_{1(21)} y_{1,t-1} + \ldots + \gamma_{4(21)} y_{1,t-4} + \gamma_{1(22)} y_{2,t-1} + \ldots + \gamma_{4(22)} y_{2,t-4} \\
&\quad + \gamma_{1(23)} y_{3,t-1} + \ldots + \gamma_{4(23)} y_{3,t-4} + \epsilon_{2t} \\
y_{3t} &= \mu_3 + \gamma_{1(31)} y_{1,t-1} + \ldots + \gamma_{4(31)} y_{1,t-4} + \gamma_{1(32)} y_{2,t-1} + \ldots + \gamma_{4(32)} y_{2,t-4} \\
&\quad + \gamma_{1(33)} y_{3,t-1} + \ldots + \gamma_{4(33)} y_{3,t-4} + \epsilon_{3t}
\end{aligned}
$$

Coefficient $\gamma_{i(ab)}$ is the $(a,b)^{th}$ element of Γ_i. The slope coefficients in the first equation, the "GDP equation," are in the first rows of the Γ matrices,

$$
\begin{aligned}
y_{1t} &= \mu_1 + \begin{bmatrix} \gamma_{1(11)} , & \gamma_{1(12)} , & \gamma_{1(13)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{bmatrix} + \begin{bmatrix} \gamma_{2(11)} , & \gamma_{2(12)} , & \gamma_{2(13)} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \\ y_{3,t-2} \end{bmatrix} + \\
&\quad \begin{bmatrix} \gamma_{3(11)} , & \gamma_{3(12)} , & \gamma_{3(13)} \end{bmatrix} \begin{bmatrix} y_{1,t-3} \\ y_{2,t-3} \\ y_{3,t-3} \end{bmatrix} + \begin{bmatrix} \gamma_{4(11)} , & \gamma_{4(12)} , & \gamma_{4(13)} \end{bmatrix} \begin{bmatrix} y_{1,t-4} \\ y_{2,t-4} \\ y_{3,t-4} \end{bmatrix} + \epsilon_{1t} \\
&= \mu_1 + \Gamma_{1(1\cdot)} y_{t-1} + \Gamma_{2(1\cdot)} y_{t-2} + \Gamma_{3(1\cdot)} y_{t-3} + \Gamma_{4(1\cdot)} y_{t-4} + \epsilon_{1t}
\end{aligned}
$$

where $\Gamma_{i(a\cdot)}$, a $1 \times m$ row vector, is the ath of Γ_i. The second and third (log CPI and interest rate) equations are, respectively,

$$
\begin{aligned}
y_{2t} &= \mu_2 + \Gamma_{1(2\cdot)} y_{t-1} + \Gamma_{2(2\cdot)} y_{t-2} + \Gamma_{3(2\cdot)} y_{t-3} + \Gamma_{4(2\cdot)} y_{t-4} + \epsilon_{2t}, \quad \text{and} \\
y_{3t} &= \mu_3 + \Gamma_{1(3\cdot)} y_{t-1} + \Gamma_{2(3\cdot)} y_{t-2} + \Gamma_{3(3\cdot)} y_{t-3} + \Gamma_{4(3\cdot)} y_{t-4} + \epsilon_{3t}
\end{aligned}
$$

They contain slope coefficients from the second and third rows of the Γ matrices, respectively.

1.2 Advantages and disadvantages of reduced form VARs

A good feature of VARs is that they can be written and analyzed as one algebraic equation, even though they represent a system of m equations. This simplification arises from its linearity, lack of structural assumptions (coefficient restrictions), and the common assumption that each variable in each equation appears with the same number of lags. The researcher need only choose which

variables to enter (i.e. the elements of y_t), and the number of lags, p.

This simplicity comes at a cost. Unless both m and p are very small, each equation has a large number of regressors, and they usually are highly correlated. Consequently, the estimates have high standard errors, as do the impulse response functions (IRFs), which are discussed in section 1.4.

Another difficulty lies in the interpretation of the IRFs themselves. The IRF describing the effect of a shock in y_2 on y_1, for example, derives from changing ϵ_{2t} while holding constant the other elements of the disturbance vector ϵ_t. The contemporaneous "what is causing what?" issues, which are the major concerns of structural models, are buried here in the covariance matrix $E(\epsilon_t \epsilon_t')$. The IRF "identifies" the effects simply by imposing the above thought experiment on the reduced form estimates, and this thought experiment itself may be misleading. (E.g. a change in ϵ_{2t} may cause ϵ_{1t} to change as well, but the thought experiment holds ϵ_{1t} constant.)

Various other methods of imposing structure on the VAR have been proposed. One is to put a causal ordering to the variables, and impose the shocks in the same order, which involves a Cholesky decomposition of $E(\epsilon_t \epsilon_t')$. Other methods involve modeling the contemporaneous relations among the y_t's, in effect by replacing y_t on the left side of (1) by $\Gamma_0 y_t$, say, where Γ_0 is another coefficient matrix, and imposing enough restrictions on Γ_0 to make estimation possible. This approach is called a *structural VAR*.

1.3   Estimation

A reduced form VAR is estimated by *equation-by-equation OLS*, and so are the t statistics for the μ and Γ_i estimates are the usual ones. This can be motivated by seeing that each individual equation in the VAR is a dynamic regression model of the sort described in the previous notes, where it was mentioned that they typically are estimated by OLS.

There is a bit more to it, though. Earlier it was noted that $E\epsilon_t \epsilon_t'$ is not assumed to be diagonal, that is, a nonzero correlation between ϵ_{it} and ϵ_{jt}, $i \neq j$, called *contemporaneous correlation*, is allowed. A set of time series regression equations with contemporaneously correlated errors is one of the early models of econometrics, known as the SUR (seemingly unrelated regression) model of Zellner (1962). He proposed an estimator with an asymptotic variance that is less than or equal to that of OLS. However, when the explanatory variables are the same in each equation of a SUR model, the SUR and OLS estimators happen to be equal, and so are their estimated variances. So, OLS is used even though a reduced-form VAR is a SUR model. When testing restrictions involving coefficients from more than one equation, for example when using `varwle` described below in section 1.5.2, the cross-equation correlation in the errors must be accounted for in the construction of the

test statistic. The SUR variance and covariance estimators are used when these cross-equation covariances are required.

1.4 MA representation and impulse response functions, (IRFs)

The Γ coefficient matrices in (1) are difficult to interpret, even more so than for one-equation models, because in a VAR the lags of each variable appear in all of the equations. Over time, everything affects everything. The most common way to interpret these coefficients is through *impulse response functions (IRFs)*. These functions show the effect of a one-unit, or sometimes a one standard deviation, shock in variable j (i.e. a one-unit increase in ϵ_{jt}) on variable k (that is, the change in $y_{k,t+s}$) s periods ahead, as a function of s. They are obtained by forming the MA representation of the estimated model (1). To motivate an MA representation, rewrite (1) using a matrix lag polynomial,

$$(I - \sum_{i=1}^{p} \Gamma_i L^i) y_t \quad = \quad \mu + \epsilon_t$$

Assuming that the lag polynomial is invertible, express the MA representation using $m \times m$ MA coefficient matrices Θ_i,

$$y_t \quad = \quad (I - \sum_{i=1}^{p} \Gamma_i L^i)^{-1} \mu + (I - \sum_{i=1}^{p} \Gamma_i L^i)^{-1} \epsilon_t$$

$$y_t \quad = \quad \mu^* + (I + \sum_{i=1}^{\infty} \Theta_i L^i) \epsilon_t$$

where the Θ_i's are defined so that $(I + \sum_{i=1}^{\infty} \Theta_i L^i) = (I - \sum_{i=1}^{p} \Gamma_i L^i)^{-1}$. The Θ_i matrices can be derived as functions of the Γ's using the identity

$$(I - \sum_{i=1}^{p} \Gamma_i L^i)(I + \sum_{i=1}^{\infty} \Theta_i L^i) \quad = \quad I$$

$$= \quad I + 0 \times L + 0 \times L^2 + \dots$$

$$= \quad I + (-\Gamma_1 + \Theta_1)L + (\Gamma_2 - \Gamma_1 \Theta_1 + \Theta_2))L^2 + \dots$$

The Γ_i's can be solved iteratively, by multiplying the left-hand side lag polynomials and successively setting the coefficient matrices of the products to zero for the L term, then the L^2 term, and so

on. The first few matrices are shown below.

$$
\begin{aligned}
-\Gamma_1 + \Theta_1 &= 0 \\
\Rightarrow \Theta_1 &= \Gamma_1 \\
-\Gamma_2 - \Gamma_1\Theta_1 + \Theta_2 &= 0 \\
\Rightarrow \Theta_2 &= \Gamma_2 + \Gamma_1\Theta_1 \\
&= \Gamma_2 + \Gamma_1(\Gamma_1) \\
&= \Gamma_2 + \Gamma_1^2 \\
-\Gamma_3 - \Gamma_2\Theta_1 - \Gamma_1\Theta_2 + \Theta_3 &= 0 \\
\Rightarrow \Theta_3 &= \Gamma_3 + \Gamma_2\Theta_1 + \Gamma_1\Theta_2 \\
&= \Gamma_3 + \Gamma_2(\Gamma_1) + \Gamma_1(\Gamma_2 + \Gamma_1^2) \\
&= \Gamma_3 + \Gamma_2\Gamma_1 + \Gamma_1\Gamma_2 + \Gamma_1^3
\end{aligned}
$$

1.4.1 Numerical example of an IRF

Successive terms become more cumbersome, but the numerical implementation is not as bad. For example, suppose $m = 2$, $p = 2$, and the estimated equations are

$$
\begin{aligned}
y_{1t} &= 3 + .4y_{1,t-1} + .2y_{1,t-2} + -3y_{2,t-1} + 0.5y_{2,t-2} + \epsilon_{1t} \\
y_{2t} &= 7 + .07y_{1,t-1} + .03y_{1,t-2} + .8y_{2,t-1} - 0.3y_{2,t-2} + \epsilon_{2t}
\end{aligned}
$$

In VAR matrix notation, this is

$$
\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix} + \begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} .2 & .5 \\ .03 & -.03 \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}
$$

so that $\Gamma_1 = \begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix}$ and $\Gamma_2 = \begin{bmatrix} .2 & .5 \\ .03 & -.03 \end{bmatrix}$. Using the approach described above, the MA coefficient matrices are

$$
\Theta_1 = \Gamma_1 = \begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix}
$$

$$\Theta_2 = \Gamma_2 + \Gamma_1\Theta_1 = \begin{bmatrix} .2 & .5 \\ .03 & -.03 \end{bmatrix} + \begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix}\begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix} = \begin{bmatrix} .15 & -3.1 \\ .114 & .4 \end{bmatrix}$$

$$\Theta_3 = \Gamma_3 + \Gamma_2\Theta_1 + \Gamma_1\Theta_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} .2 & .5 \\ .03 & -.03 \end{bmatrix}\begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix} + \begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix}\begin{bmatrix} .15 & -3.1 \\ .114 & .4 \end{bmatrix}$$
$$= \begin{bmatrix} -.167 & -2.64 \\ .1116 & -.011 \end{bmatrix}$$

$$\Theta_4 = \Gamma_4 + \Gamma_3\Theta_1 + \Gamma_2\Theta_2 + \Gamma_1\Theta_3 = \Gamma_2\Theta_2 + \Gamma_1\Theta_3$$
$$= \begin{bmatrix} .2 & .5 \\ .03 & -.03 \end{bmatrix}\begin{bmatrix} .15 & -3.1 \\ .114 & .4 \end{bmatrix} + \begin{bmatrix} .4 & -3 \\ .07 & .8 \end{bmatrix}\begin{bmatrix} -.167 & -2.64 \\ .1116 & -.011 \end{bmatrix}$$
$$= \begin{bmatrix} -0.3146 & -1.443 \\ .07867 & -.2986 \end{bmatrix}$$

and so on. Impulse response functions are obtained by reading off the appropriate element of the Θ matrices. For example, the first few values of the IRF for the effect of a one-unit shock to y_1 at time t on the values of y_2 at times $t + s, s = 0, 1, 2, \ldots$, are given by the $(2, 1)$ elements of the Θ matrices: $\Delta y_{t+1} = .07$; $\Delta y_{t+2} = .114$; $\Delta y_{t+3} = .1116$; and $\Delta y_{t+4} = .07867$.

The above effects are for a one-time-only change, and would die out to zero as $s \to \infty$. The effects of a permanent change are given by the cumulative sums of the above IRFs. For example, the effects on future y_2 values of a permanent one-unit upward shift in y_1 are: $\Delta y_{t+1} = .07$; $\Delta y_{t+2} = .184$; $\Delta y_{t+3} = .2956$; and $\Delta y_{t+4} = .37427$.

1.5 Specification tests

The following specification tests relate to specific Stata commands. They are the most common tests conducted after estimating a VAR.

1.5.1 Autocorrelation

The Stata command `varlmar` tests the null that there is no autocorrelation in any of the m equations at a given lag, resulting in a table of test statistics and P values with m degrees of freedom. Rejecting the null (i.e. getting a P value less than a chosen significance level α) is evidence of autocorrelation,

and may indicate that a greater number of lagged variables (a larger p) is needed.

1.5.2 Lag length

The Stata command `varwle` tests the null hypothesis that a set of coefficients at a specific lag all equal zero. The first tables give results for the m equations separately. Each H_0 corresponds to a row of a Γ_i equalling zero, hence has m restrictions and degrees of freedom. The last table produced by this command lists the results of tests of the null hypotheses that all of the coefficients at a given lag equal zero. It will then be testing m^2 restrictions and have m^2 degrees of freedom. These are testing H_0 that an entire Γ_i matrix equals zero.

In a testing-down framework that imposes the same number of lags in each equation, the most relevant H_0 produced by `varwle` is the one pertaining to the highest-order lag across all equations, $\Gamma_p = 0$.

Model selection criteria also can be applied to a VAR in order to guide the choice of lag length. They typically are referred to by terms such as AIC, FPE, HQIC, and SBIC. Sometimes a simpler approach is used where the lag length is imposed in an *ad hoc* fashion, e.g. 4 or 8 lags with quarterly data.

1.5.3 Granger causality

The statement "y_2 does not Granger-cause y_1," means that all of the coefficients on the lagged y_2 variables in the y_1 equation are equal to zero. This is the null hypothesis of the Granger causality test. In the above notation, it is $H_0 : \gamma_{1(1,2)} = \gamma_{2(1,2)} = \ldots = \gamma_{m(1,2)} = 0$, or that the $(1, 2)$ elements in all of the Γ_i's equal zero. If the test rejects H_0, you have evidence that "y_2 Granger-causes y_1."

In a single-equation context, these usually are F-tests. In a VAR, the Stata command `vargranger` produces a set of Wald tests that have a χ^2 distribution with p degrees of freedom asymptotically when H_0 is true. In the Stata output table, the H_0 being tested is that the variable listed in the `Excluded` column does not Granger-cause the variable listed in the `Equation` column.

2 Unit Roots

2.1 Random walks and trend stationarity

A basic example of a nonstationary time series process is the *random walk (without drift)*

$$y_t \;=\; y_{t-1} + u_t \tag{2}$$

where u_t is white noise. This process is nonstationary because $\text{Var}(y_t) = \text{Var}(y_{t-1}) + \text{Var}(u_t) \neq \text{Var}(y_t)$, so that the unconditional variance of y_t is not constant across values of t.

With an intercept, (2) becomes a *random walk with drift*

$$y_t \;=\; \mu + y_{t-1} + u_t \tag{3}$$

in which both the mean and unconditional variance change with t.

These processes each have a *unit root*. When their AR component is expressed as a lag polynomial, $A(L) = 1 - L$, the "root", or value of z that solves the equation $A(z) = 0$, has a magnitude equal to one. In (3), that is just a way of saying that $\rho_1 = 1$ in the AR(1) process $y_t = \mu + \rho_1 y_{t-1} + u_t$, but the unit root definition is more general as in section 2.3.

A process that is stationary around a trend, or *trend stationary*, is also nonstationary, because its mean changes with time. The most common version uses a linear trend,

$$y_t = \alpha + \beta t + \epsilon_t \tag{4}$$

ϵ_t is a stationary disturbance term. The slope coefficient β in (4) is seen to be like the drift term μ in (3) after first-differencing the y_t in expression in (4):

$$
\begin{aligned}
Dy_t &= y_t - y_{t-1} \\
&= (\alpha + \beta t + \epsilon_t) - (\alpha + \beta(t-1) + \epsilon_{t-1}) \\
&= \beta + (\epsilon_t - \epsilon_{t-1}), \text{ so that} \\
y_t &= \beta + y_{t-1} + (\epsilon_t - \epsilon_{t-1})
\end{aligned}
$$

2.2 Order of integration, I(d)

When y_t must be differenced d times before it is stationary, then it is said to be *integrated of order* d, or I(d). A stationary variable is I(0). The three processes (2), (3), and (4) all are I(1), because: (i) they are not stationary; and (ii) their first differences, listed below, are stationary.

$$
\begin{aligned}
Dy_t &= u_t & (y_t \text{ is a random walk}) \\
Dy_t &= \mu + u_t & (y_t \text{ is a random walk with drift}) \\
Dy_t &= \beta + (\epsilon_t - \epsilon_{t-1}) & (y_t \text{ is trend stationary})
\end{aligned}
$$

Most econometric models treat variables as I(0) or I(1), but in principle, variables also can be I(2), I(3), etc.. Here are three examples where y_t is I(2). Their first differences then are I(1), and their second differences are I(0).

(i) The second difference of y_t is white noise:

$$
\begin{aligned}
y_t &= 2y_{t-1} - y_{t-2} + u_t \\
Dy_t &= y_{t-1} - y_{t-2} + u_t = Dy_{t-1} + u_t \\
D^2 y_t &= Dy_t - Dy_{t-1} = u_t
\end{aligned}
$$

(ii) The first difference of y_t is trend stationary:

$$
\begin{aligned}
y_t &= \alpha + y_{t-1} + \beta t + u_t \\
Dy_t &= \alpha + \beta t + u_t \\
D^2 y_t &= Dy_t - Dy_{t-1} = \alpha + \beta t + u_t - (\alpha + \beta(t-1) + u_{t-1}) = \beta + (u_t - u_{t-1})
\end{aligned}
$$

(iii) y_t is trend stationary around a quadratic trend:

$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 t + \beta_2 t^2 + u_t \\
Dy_t &= (\beta_0 + \beta_1 t + \beta_2 t^2 + u_t) - (\beta_0 + \beta_1(t-1) + \beta_2(t-1)^2 + u_{t-1}) \\
&= \beta_1 + \beta_2(2t - 1) + (u_t - u_{t-1}) = (\beta_1 - \beta_2) + 2\beta_2 t + (u_t - u_{t-1}) \\
D^2 y_t &= 2\beta_2 + (u_t - u_{t-1})
\end{aligned}
$$

2.3   Unit roots and characteristic roots

In a univariate ARMAX process

$$A(L)y_t = \alpha + x_t'\beta + B(L)u_t$$

consider the *characteristic equation*

$$A(z) = 0$$

This is a pth-order polynomial in z, and has p solutions or *roots*. Some of the roots may be equal to each other, and some may be complex numbers, $z = a + b\iota$, where $\iota^2 = -1$.

Suppose the roots are real numbers, $z_1, z_2, \ldots, z_p$. Then one way to determine if a process is stationary is to look at the roots of the *characteristic equation*. This equation is obtained by expressing the AR(p) part of the process in lag polynomial notation:

$$A(L) = 1 - \rho_1 L - \rho_2 L^2 - \ldots - \rho_p L^p$$

To get the characteristic equation, replace the lag operator L by a variable (call it z), and set the resulting polynomial equal to zero:

$$1 - \rho_1 z - \rho_2 z^2 - \ldots - \rho_p z^p = 0$$

The *characteristic roots* are the values of z that solve this equation. There are p of them, although some of them may be equal. y_t is *stationary* if all of the roots "lie outside the unit circle". This phrase reflects the fact that some of these roots may be complex numbers. If the roots all are real numbers (that is, none of the roots are complex numbers), then y_t is stationary if the absolute values of all of these real roots are greater than one. If a root equals one or minus one, it is called a *unit root*. If there is at least one unit root, or if any root lies between plus and minus one, then the series is not stationary.

For example, the AR(1) process: $y_t = \rho_1 y_{t-1} + \epsilon_t$ has a characteristic equation: $1 - \rho_1 z = 0$ and its one characteristic root is $z^* = 1/\rho_1$. The series is stationary as long as $|\rho_1| < 1$ which is the same condition as $|z^*| > 1$ (i.e. "z^* lies outside the unit circle").

For $ARMA(p, q)$ processes, the $MA(q)$ part is irrelevant for determining stationarity, as long as q is finite.

2.4 Spurious regression problem

Granger and Newbold (1974) report a result that makes it hard to ignore the unit root issue in econometric time series models. Suppose y_t and x_t are in fact statistically independent random walk processes and a researcher, not knowing this, uses OLS to estimate

$$y_t = \alpha + \beta x_t + \text{error}$$

and then tests $H_0 : \beta = 0$ by the usual t test at, say, the 5% level. Since H_0 is true (because y_t and x_t are independent), we would like the t test to reject H_0 about 5% of the time, and expect this probability to approach 5% as $n \to \infty$ when the test is well-behaved. When y_t and x_t are random walks, however, the t test rejects H_0 much more than 5% of the time. Alarmingly, as $n \to \infty$ this probability approaches 100%! This result undermines the credibility of standard inference methods whenever variables in a time series regression are suspected to have been generated by I(1) processes.

2.5 How to proceed?

Researchers commonly investigate the order of integration of time series variables before modeling their relationships. Many tests have been suggested. One of the more common methods is the Dickey-Fuller test (see section 2.6).

If it is decided that some of the variables are I(1), the next step may be to check if they are *cointegrated* – that is, if their I(1) components cancel out (sections 2.7 and 2.9). The spurious regression problem mentioned earlier is a case where two I(1) variables are modeled that are not cointegrated. If instead they are cointegrated, then the situation is better econometrically, and an error correction model can be estimated (section 2.8).

Another approach, which is especially useful in a multiple equation context, is to specify a *vector error correction model* (VECM) and use the estimation results to check for cointegration properties (section 3).

2.6 Dickey-Fuller (DF) test

The basic idea of a DF test is to test if $\gamma = 1$ in the AR equation

$$y_t = \mu + \gamma y_{t-1} + u_t$$

The regression usually is specified a bit differently though, so that the t ratio on y_{t-1}'s coefficient can be used as the DF statistic. This is done by subtracting y_{t-1} from both sides:

$$
\begin{aligned}
y_t - y_{t-1} &= \mu + \gamma y_{t-1} - y_{t-1} + u_t \\
Dy_t &= \mu + \gamma^* y_{t-1} + u_t
\end{aligned}
\tag{5}
$$

where $\gamma^* = \gamma - 1$. (An aside: equations like (5) are sometimes called *artificial regressions*, meaning that the regression equation itself is not of interest, it is only a device used to compute a test statistic.) The hypotheses are

$$
\begin{aligned}
H_0 : \quad & \gamma^* = 0 \quad (y_t \text{ has a unit root}) \\
H_a : \quad & \gamma^* < 0 \quad (y_t \text{ does not have a unit root})
\end{aligned}
$$

The alternative is left-sided. The case $\gamma^* > 0$ implies $\gamma > 1$, which is an explosive process that is ruled out *a priori*. Implicitly, the case $\gamma^* < -2$ also is ruled out. If $\gamma^* < -2$, then $\gamma < -1$, which is an even more bizarre sign-switching explosive process. So when the DF test rejects H_0, that is interpreted as evidence in favour of $|\gamma| < 1$ (i.e. $0 < \gamma^* < 2$), which would imply that y_t does not have a unit root.

Although it is a t ratio, the DF statistic does not have a t distribution under the null hypothesis. Furthermore, the null distribution of DF changes when trend regressors are included in (5). DF test statistics can also be based on these regressions,

$$
\begin{aligned}
Dy_t &= \mu + \gamma^* y_{t-1} + \beta t + u_t \tag{6} \\
Dy_t &= \mu + \gamma^* y_{t-1} + \beta_1 t + \beta_2 t^2 + u_t \tag{7}
\end{aligned}
$$

Using (6), a DF test rejection favours y_t being stationary around a linear trend, and using (7), a DF rejection favours y_t being stationary around a quadratic trend. Each of these tests requires different critical values and P values, which are supplied by most econometric software.

A common variant of the DF test is the Augmented Dickey Fuller (ADF) test. It adds lagged dependent variable regressors to the DF regression in order to get a better match between the assumed null distribution (which is derived assuming no autocorrelation in the errors of the DF regression) and the actual null distribution. This will make the P values more accurate and the null rejection rate closer to the stated one. The ADF regressions corresponding to the three DF

regressions given above are

$$Dy_t = \mu + \gamma^* y_{t-1} + \sum_{i=1}^{p} \phi_i Dy_{t-i} + u_t$$

$$Dy_t = \mu + \gamma^* y_{t-1} + \beta t + \sum_{i=1}^{p} \phi_i Dy_{t-i} + u_t$$

$$Dy_t = \mu + \gamma^* y_{t-1} + \beta_1 t + \beta_2 t^2 + \sum_{i=1}^{p} \phi_i Dy_{t-i} + u_t$$

2.7   I(0) and I(1) variables, and cointegration

The results in this section use the following properties, where a_t and b_t represent time series processes:

(i) If a_t is I(d_a) and b_t is I(d_b), then $a_t + b_t$ is I($\max(d_a, d_b)$)

(ii) If a_t is I(d_a), then for any nonzero finite constant κ, κa_t also is $I(d_a)$

We will use the following simple regression model to examine the implications of combining variables having various orders of integration in a regression equation.

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{8}$$

where we assume that x_t and ϵ_t are not correlated.

Consider the following cases.

(a) x_t is I(0) and y_t is I(0)

Then ϵ_t is I(0), since a weighted sum of a fixed number of I(0) variables (here it is $y_t - \beta x_t$) also is I(0). This is the usual stationary-variable situation.

(b) x_t is I(1) and y_t is I(0)

The left side of (8) is I(0). We must then have $\beta = 0$ and ϵ_t being I(0). ϵ_t is not correlated with x_t by assumption, so ϵ_t cannot cancel out the I(1) component of βx_t. β must then be equal to 0 for y_t to be I(0).

(c) x_t is I(0) and y_t is I(1)

ϵ_t is I(1), since something must be I(1) on the right-hand side to match y_t being I(1).

(d) x_t is I(1) and y_t is I(1)

    If y_t and x_t are *cointegrated* then there is a unique value of β at which the I(1) parts of y_t and βx_t are equal, leaving ϵ_t as I(0). If y_t and x_t are not cointegrated then ϵ_t is I(1).

2.7.1 Cointegration

Definition: Let z_t be a vector of variables that are I(d), $d \geq 1$. If there is a vector of constants θ such that $\theta' z_t$ is I(d^*) where $d^* < d$, then the variables in z_t are *cointegrated*, and β is the *cointegrating vector*.

Most of the time in econometrics, $d = 1$ and $d^* = 0$. Then cointegration means that there is a value of the vector θ for which the vector of I(1) variables, z_t, are linearly transformed to an I(0) term, as $z_t' \theta = \epsilon_t$.

In the two-variable example from the previous section, where the equation is $y_t = \alpha + \beta x_t + \epsilon_t$, let $z_t = \begin{bmatrix} y_t \\ x_t \end{bmatrix}$ and $\theta = \begin{bmatrix} 1 \\ -\beta \end{bmatrix}$. Then $z_t' \theta = y_t - \beta x_t = \epsilon_t + \alpha$. The intercept term α is of no consequence when evaluating the order of integration.

When there are two variables, there can be at most only one linearly independent cointegrating vector. That is, given the normalizing "1" in the first position of θ as defined in the above example, there can then be only one value of β that produces an I(0) result for $z_t' \theta$. To see this, let that unique value be β, let the unique cointegrating vector value be $\theta = \begin{bmatrix} 1 \\ -\beta \end{bmatrix}$, and let a different value of the vector be $\tilde{\theta} = \begin{bmatrix} 1 \\ -\tilde{\beta} \end{bmatrix}$. Then

$$
\begin{aligned}
z_t' \tilde{\theta} &= y_t - \tilde{\beta} x_t \\
&= y_t - \beta x_t - (\tilde{\beta} - \beta) x_t \\
&= I(0) - (\tilde{\beta} - \beta) \times I(1) \\
&= I(1)
\end{aligned}
$$

as long as $\tilde{\beta} \neq \beta$, or equivalently $\tilde{\alpha} = \alpha$. So $z_t' \theta$ can only be I(1) at the single value of β.

2.7.2 Superconsistency

Superconsistency is a property about the convergence rate (to zero) of the variance of a consistent estimator as the sample size goes to infinity. The OLS estimator of β (call it b) in the model

$y_t = \alpha + \beta x_t + \epsilon_t$ is *superconsistent* when x_t and y_t are cointegrated. This means that the variance of b converges to zero at a rate $1/n^2$ instead of the usual $1/n$.

Intuitively, convergence is faster in the cointegration case because as n increases, the variation in x_t grows at a faster rate, since it is nonstationary. Suppose that ϵ_t is white noise with variance σ^2. Then

$$\text{Var}(b) = \frac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

When x_t is I(0), the denominator grows at rate n, since $\bar{x}$ approaches $E(x)$. Then as n increases by one, the sum increases on average by $E(x_t - E(x))^2$. When x_t is I(1), however, the denominator grows at a faster rate, because as n increases, not only are there more terms in the denominator sum, but the terms themselves get larger on average, because $\text{Var}(x_t) = E(x_t - E(x_t))^2$ itself increases as t increases.

2.8 Error correction model (ECM)

Suppose x_t and y_t are I(1) and cointegrated. Then ϵ_t is I(0) in the cointegrating equation

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

These equations often are interpreted as long-run or equilibrium relationships between x_t and y_t.

A researcher will also be interested in the short-run dynamics - the way that x_t and y_t fluctuate around this long-run relationship, as in a business cycle. This is done by estimating an error correction model, which contains first differences of x_t and y_t, their lags, and an *error correction term*. An ECM is

$$Dy_t = \mu + \gamma_1 Dy_{t-1} + \ldots \gamma_p Dy_{t-p} + \omega_0 Dx_{t-1} + \ldots \omega_r Dx_{t-r} + \lambda EC_{t-1} + u_t$$

where $EC_t = y_t - (a + bx_t)$, the error correction term, is the lagged OLS residual from the cointegrating equation. The lag orders p and r are chosen in the usual ways discussed earlier.

2.8.1 From a dynamic model to an ECM

Normally λ is expected to be negative. Then λEC_{t-1} represents a force pulling y_t toward its long-run relationship, being negative when $EC_{t-1} > 0$, and positive when $EC_{t-1} < 0$.

The cointegrating equation and ECM can be shown to follow from a single dynamic regression

model in levels when y_t and x_t are cointegrated. For example, start with the dynamic model

$$y_t = \kappa_0 + \kappa_1 x_t + \kappa_2 x_{t-1} + \kappa_3 y_{t-1} + u_t$$

where y_t and x_t are I(1), and u_t is I(0). Subtract y_{t-1} from both sides:

$$Dy_t = \kappa_0 + \kappa_1 x_t + \kappa_2 x_{t-1} + (\kappa_3 - 1)y_{t-1} + u_t$$

Subtract and add $\kappa_1 x_{t-1}$ on the right hand side:

$$Dy_t = \kappa_0 + \kappa_1 Dx_t + ((\kappa_1 + \kappa_2)x_{t-1} + (\kappa_3 - 1)y_{t-1}) + u_t$$

The model now contains I(0) variables except for the term in parentheses involving x_{t-1} and y_{t-1}. But since every other term in the model is I(0), so must be the term in parentheses. Rewriting that term makes the equation look like

$$Dy_t = \kappa_0 + \kappa_1 Dx_t + (\kappa_3 - 1)\left(y_{t-1} - \left(\frac{-(\kappa_1 + \kappa_2)}{(\kappa_3 - 1)}\right)x_{t-1}\right) + u_t$$

Refreshing the notation for the parameters, this equation is

$$
\begin{aligned}
Dy_t &= \mu + \omega_1 Dx_t + \lambda\left(y_{t-1} - \beta x_{t-1}\right) + u_t, \quad \text{or} \\
Dy_t &= \mu^* + +\omega_1 Dx_t + \lambda\left(y_{t-1} - (\alpha + \beta x_{t-1})\right) + u_t \\
&= \mu^* + +\omega_1 Dx_t + \lambda EC_{t-1} + u_t
\end{aligned}
$$

where $EC_{t-1} = y_{t-1} - (\alpha + \beta x_{t-1})$ is the "true" error in the cointegrating equation. The intercept term from the cointegrating equation, α, can be included in the error correction term by adjusting the intercept in the ECM equation itself from μ to μ^* to compensate.

2.8.2   Estimation

Since the OLS estimate of β is superconsistent, the sampling error from estimating it in the cointegrating equation is less important than the sampling error of the ECM coefficient estimates asymptotically. This justifies a two-step procedure where the cointegrating equation is estimated first, followed by an ECM with the lagged OLS residual $(\widehat{EC}_{t-1})$ from the estimated cointegrating

equation serving as the error correction term in the ECM.

step 1, OLS: $\quad y_t = a + bx_t + \widehat{EC}_t$ $\qquad\qquad$ (cointegrating equation)

step 2, OLS: $\quad Dy_t = \hat{\mu}^* + + \hat{\omega}_1 Dx_t + \hat{\lambda}\widehat{EC}_{t-1} + \hat{u}_t$ $\quad$ (error correction model)

This approach can be varied by including time trend variables in the cointegrating equation, adding other I(0) regressors in the ECM, specifically, including more lags of Dx_t and Dy_t to capture the dynamic effects, or seasonal dummies, or other controls.

2.9 Cointegration test

Suppose x_t and y_t are I(1). If they are cointegrated, then the error term in the cointegrating equation is I(0) by definition, and if they are not cointegrated, the error term is I(1). Engle and Granger's modified Dickey-Fuller test takes the OLS residual from the cointegrating equation, and uses the DF approach to test the null hypothesis that there is a unit root (implying no cointegration) against the alternative that there is no unit root (implying cointegratiion). The test statistic is the t ratio on γ^* in

$$De_t = \gamma^* e_{t-1} + \text{error}$$

and the procedure tests $H_0 : \gamma^* = 0$ against $H_a : \gamma^* < 0$. The null distribution of this statistic is different again from the other DF and ADF null distributions.

This test has *low power*, meaning that it does not reject H_0 when H_0 is false as often as one would like it to. This cointegration test then often will accept the null hypothesis' claim that the variables are not cointegrated, even when they actually are. This puts a researcher in a difficult situation when the test accepts H_0. If the acceptance is taken seriously, the model must be re-specified somehow, or the researcher must figure out how to proceed with an analysis using variables that are not cointegrated. Alternatively, the researcher could proceed with a cointegration assumption on the grounds that the test may have failed to reject due to low power, but then is left to wonder why the test was performed in the first place if the result is just ignored.

Other tests have been proposed, including ones where the null hypothesis implies cointegration rather than no cointegration. Another way to skirt this issue is to use a *vector error correction model* (VECM), which is like a VAR but consists of ECMs instead of dynamic regression equations in levels. This VECM, described below, leads to a different approach to the cointegration question, explained in section 3.2.

3 Vector Error Correction Model (VECM)

Like an ECM, a VECM is constructed from first differences of cointegrated I(1) variables, their lags, and some error correction terms. In vector-matrix notation, a VECM is

$$Dy_t \;=\; \mu + \Pi y_{t-1} + \sum_{i=1}^{p} \Gamma_i^* Dy_{t-i} + \epsilon_t \tag{9}$$

where: y_t is an $m \times 1$ vector of variables like in a VAR; Dy_t is an $m \times 1$ vector if the first differences of the variables in y_t; μ is an $m \times 1$ vector of intercept coefficients; Π and the Γ_i^*s are $m \times m$ coefficient matrices; ϵ_t is an $m \times 1$ error vector with contemporaneous correlation but no autocorrelation, like the error vector in a VAR.

3.1 The Π matrix

The coefficient matrix Π in (9) plays a special role and warrants further examination.

If y_t is I(1), then Dy_t in (9) is I(0), and by assumption so is ϵ_t. The only term left in (9) is Πy_{t-1}, so if the VECM is internally consistent, Πy_{t-1} must also be I(0), despite the fact that y_{t-1} is I(1). This can only happen if pre-multiplying y_{t-1} by Π produces linear combinations of y_{t-1} that are I(0). These linear combinations are the cointegrating relationships among the elements of y_t.

At this point it is convenient to think of Π as the product of an $m \times r$ matrix "α" and the transpose of an $m \times r$ matrix "β", as in $\Pi = \alpha\beta'$. Then this term of (9) is

$$\Pi y_{t-1} = (\alpha\beta')y_{t-1} = \alpha(\beta'y_{t-1})$$

$\beta'y_{t-1}$ is an $r \times 1$ vector containing the error correction terms. The r columns of β (rows of β') are the cointegrating vectors. The coefficients of α determine the size of the effects of the r error correction terms in the m equations of the VECM (like λ does in section 2.8.

In the one equation two variables ECM of section 2.7, $r = 1$, $m = 2$, and we are looking at the "y" equation.

3.2 Testing for the number of cointegrating relationships

r from section 3.1 is the number of cointegrating relationships and error correction terms. It is a nonnegative integer less than or equal to m. Since $\Pi = \alpha\beta'$, then r also is the rank of Π. We may

18

not want to assume that r is known beforehand. *Trace tests* are often used to set the value of r, where the null hypothesis is $H_0 : r \leq r^*$ for some hypothesized value r^* of r.

Suppose that $H_0 : r \leq 0$ is accepted. $r = 0$ implies that $\Pi = 0$, suggesting that there are no cointegrating relationships among the y's. Setting $|Pi = 0$ essentially converts the VECM to a VAR in the first differences. If $H_0 : r \leq 0$ is rejected, then one next tests $H_0 : r \leq 1$, and so on, until one of the tests does not reject. The value of r^* in that final test is used as the value of r in the VECM. In the Stata output of the `vecrank` command that carries out this test, the "$*$" is placed in the row corresponding to this value of r.

Suppose that you keep rejecting H_0 and increasing r^* until you get to $H_0 : r \leq m - 1$, and suppose that even this H_0 is rejected. This suggests that r equals m, which is the maximum possible value of r. If $r = m$, then Π has full rank. This leads to contradictions among the assumptions of the VECM. In the $m = 2$ case, the "unique value of β" result in section 2.7.1 is essentially saying that there can be only one cointegrating relationship between $m = 2$ variables. Extending this point to the general case, there can only be $m - 1$ independent cointegrating relationships between m variables. If Π has full rank, then the variables in Πy_{t-1} span the same vector space as the variables in y_{t-1} itself. (See Appendix A.3 of Greene's 5th edition for background material about vector spaces.) Therefore at least one element of Πy_{t-1} is I(1). But this is not possible given the assumptions since none of the other terms in the VECM model are I(1).

Another way to see the above point is to multiply (9) by Π^{-1}, which exists only if $r = m$. Every term in the resulting equation is I(0) except for $\Pi^{-1}\Pi y_{t-1} = y_{t-1}$, which is I(1). This inplies the left side is I(0) and the right side is I(1), which is impossible.

Therefore if $r = m$ in a VECM, then the assumptions of the VECM must be modified. Every element of y_t is I(0), or else at least one element of ϵ_t is I(1), or the model must be re-specified somehow.

3.3   Connection between VARs and VECMs

A VAR containing I(1) cointegrated variables can be converted to a VECM in much the same way that a dynamic regression equation was converted to an ECM in section 2.8 as follows.

$$y_t = \mu + \Gamma_1 y_{t-1} + \ldots + \Gamma_p y_{t-p} + \epsilon_t$$

Subtract y_{t-1} from both sides,

$$y_t - y_{t-1} = Dy_t = \mu + (\Gamma_1 - I)y_{t-1} + \ldots + \Gamma_p y_{t-p} + \epsilon_t$$

Add and subtract $\Gamma_p y_{t-p+1}$ on the right side,

$$Dy_t = \mu + (\Gamma_1 - I)y_{t-1} + \ldots + \Gamma_{p-1}y_{t-p+1} + (\Gamma_p y_{t-p+1} - \Gamma_p y_{t-p+1}) + \Gamma_p y_{t-p} + \epsilon_t$$

Collect terms,

$$\begin{aligned}
Dy_t &= \mu + (\Gamma_1 - I)y_{t-1} + \ldots + (\Gamma_{p-1} + \Gamma_p)y_{t-p+1} + \Gamma_p(-y_{t-p+1} + y_{t-p}) + \epsilon_t \\
&= \mu + (\Gamma_1 - I)y_{t-1} + \ldots + (\Gamma_{p-1} + \Gamma_p)y_{t-p+1} - \Gamma_p Dy_{t-p+1} + \epsilon_t
\end{aligned}$$

Now repeat that last maneuver to convert the y_{t-p+2} term to a Dy_{t-p+2} term,

$$Dy_t = \mu + (\Gamma_1 - I)y_{t-1} + \ldots + \Gamma_{p-2}y_{t-p+2} + (\Gamma_{p-1} + \Gamma_p)y_{t-p+1} - \Gamma_p Dy_{t-p+1} + \epsilon_t$$

Add and subtract $(\Gamma_{p-1} + \Gamma_p)y_{t-p+2}$ on the right side,

$$\begin{aligned}
Dy_t &= \mu + (\Gamma_1 - I)y_{t-1} + \ldots + \Gamma_{p-2}y_{t-p+2} \\
&\quad + ((\Gamma_{p-1} + \Gamma_p)y_{t-p+2} - (\Gamma_{p-1} + \Gamma_p)y_{t-p+2}) + (\Gamma_{p-1} + \Gamma_p)y_{t-p+1} - \Gamma_p Dy_{t-p+1} + \epsilon_t
\end{aligned}$$

and collect terms,

$$\begin{aligned}
Dy_t &= \mu + (\Gamma_1 - I)y_{t-1} + \ldots + (\Gamma_{p-2} + \Gamma_{p-1} + \Gamma_p)y_{t-p+2} \\
&\quad + (-\Gamma_{p-1} - \Gamma_p)Dy_{t-p+2} + (-\Gamma_p)Dy_{t-p+1} + \epsilon_t
\end{aligned}$$

Repeating this approach right up to the y_{t-1} term causes the coefficient on a typical term, Dy_{t-i}, to be minus one times the sum of the coefficient matrices, Γ_{i+1} to Γ_p. It also ends up accumulating the sum of the coefficient matrices from Γ_2 to Γ_p and adds this sum to the existing $(\Gamma_1 - I)$ coefficient matrix on y_{t-1}. The VECM model is then

$$Dy_t = \mu + (\sum_{j=1}^{p}\Gamma_j - I)y_{t-1} + \sum_{i=1}^{p-1}(-\sum_{j=i+1}^{p}\Gamma_j)Dy_{t-i} + \epsilon_t$$

Assigning new notation for the coefficient matrices gives

$$Dy_t = \mu + \Pi y_{t-1} + \sum_{i=1}^{p-1}\Gamma_i^* Dy_{t-i} + \epsilon_t$$

where $\Pi = \sum_{j=1}^{p}\Gamma_j - I$ and $\Gamma_i^* = -\sum_{j=i+1}^{p}\Gamma_j$.

Note that the VECM has one less lag in the Dy_t's than there were in the y_t's from the original VAR. We don't need to know the Γ_j matrices from the VAR to interpret this model. The model is easier to interpret in VECM form. The Γ_i^*'s tell about the short-run dynamics, and Π tells about the cointegrating relationships. Granger causality (in the short run via the ECM) tests can be done, IRFs can be constructed to study short-run dynamics, and attempts can be made to interpret the estimated cointegration relationships.

3.4 Reference to an application

A paper by Gil *et al.* (2009) is an example of a VECM application. It has quite a few variables, and looks at the interaction between macro variables and the agricultural sector in Tunisia. It is not required reading for the course, but could be useful to look at if you plan to apply a VECM in a research project or future work.

Their section III, "Long Run Analysis," is where the authors look for meaningful cointegrating relationships hidden in the Π matrix, through imposing zero restrictions on some coefficients in the cointegrating equations. Their Table 4 has a nice way of representing the restrictions, with a "$*$" denoting unrestricted coefficients in the cointegrating matrix, which is the β' matrix from the notation in sections 3.1 and 3.2 of these notes.

Their Section IV looks at the short run dynamics through IRF functions.

Another innovative feature is the way they model the trend function on p.111, allowing for it to gradually shift using the regressor defined in their equation (3).

4   References

Gil, J.M., M. BenKaabiab and H.E. Chebbic (2009), Macroeconomics and agriculture in Tunisia, *Applied Economics*, 41, 105124.

Granger, C., and P. Newbold (1974), Spurious regressions in econometrics, *Journal of Econometrics*, 2, 111-120.

Greene, W. (2012), *Econometric Analysis*, Prentice Hall.

Zellner, A. (1962), An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias, *Journal of the American Statistical Association*, 57, 500-509.