Graduate Macro Theory II:
Notes on Time Series

Eric Sims
University of Notre Dame

Spring 2013

1 What is a Time Series?

A time series is a realization of a sequence of a variable indexed by time. The notation we will use to denote this is x_t, $t = 1, 2, \ldots, T$. A variable is said to be "random" if its realizations are stochastic. Unlike cross-sectional data, time series data can typically not be modeled as independent across observations (i.e. independent across time). Rather, time series are persistent in the sense that observations of random variables are typically positively correlated across time. Most of what we do in macro involves variables with such dependence across time.

Before discussing the unique issues presented by time series data, we quickly review expectations and first and second moments.

2 Expectations

The expected value of x_t is denoted by $E(x_t)$ and is the weighted average of possible realizations of x_t. Denote the set of possible realizations by Ω_x, and the probability density function of x as $p(x)$. Essentially the expected value is just possible realizations times the probability of each realization.

$$E(x_t) = \sum_{x \in \Omega_x} xp(x) \tag{1}$$

This is a linear operator. As such, it has the following properties, where a is a constant:

$$E(a) = a \tag{2}$$

$$E(ax_t) = aE(x_t) \tag{3}$$

$$E(x_t + y_t) = E(x_t) + E(y_t) \tag{4}$$

Non-linear operators cannot "go through" an expectation operator:

$$E(x_t y_t) \neq E(x_t)E(y_t) \tag{5}$$

$$E(g(x_t)) \neq g(E(x_t)) \tag{6}$$

We are often interested in *conditional* expectations, which are expectations taken conditional on some information. Letting I_t denote the information set available at time t, the conditional expectation of a random variable can be written: $E(x_t \mid I_t)$. It is common, however, to use the shorthand notation $E_t(x_t)$ to refer to the expectation of x_t conditional on information available at time t.

For two arbitrary random variables y and z, the Law of Iterated Expectations says that $E(y) = E(E(y \mid z))$. In words, the unconditional expectation of the conditional expectation of y conditional on z is equal to the unconditional expectation of y. This has the following implication for a time series:

$$E_t(E_{t+1}(x_{t+2})) = E_t x_{t+2} \tag{7}$$

In other words, your current best guess of your best guess next period of the realization of x two periods from now is equal to your current best guess of x two periods from now.

3 Higher Moments

The expectation is sometimes called the "first moment". We are often interested in "higher" moments, and in particular second moments. The variance is defined as:

$$\text{var}(x_t) = E(x_t - E(x_t))^2 \tag{8}$$

In words, the variance is equal to the expected (or average) squared deviation of x_t about its mean. The standard deviation is the square root of the variance. The variance can also be written:

$$\text{var}(x_t) = E(x_t^2) - (E(x_t))^2 \tag{9}$$

For mean zero random variables (such as white noise processes; see below) the variance will just be equal to $E(x_t^2)$. The following are properties of variance:

$$\text{var}(a) = 0 \tag{10}$$

$$\text{var}(ax_t) = a^2\text{var}(x_t) \tag{11}$$

$$\text{var}(x_t + y_t) = \text{var}(x_t) + \text{var}(y_t) + 2\text{cov}(x_t, y_t) \tag{12}$$

The covariance is a measure of the linear relationship between two random variables:

$$\text{cov}(x_t, y_t) = E((x_t - E(x_t))(y_t - E(y_t))) = E(x_t y_t) - E(x_t)E(y_t) \tag{13}$$

We say that two random variables are independent if knowing the realization of one of the variables does not alter one's expectation for the other. Mathematically, this means that $E(x_t \mid y_t) = E(x_t)$. If two variables are independent, then $\text{cov}(x_t, y_t) = 0$. The converse is not true – the covariance being zero does not imply two series are independent, since the dependence could be non-linear.

The following are some properties of covariance:

$$\begin{aligned}
\text{cov}(x_t, c) &= 0 \\
\text{cov}(x_t, x_t) &= \text{var}(x_t) \\
\text{cov}(x_t, y_t) &= \text{cov}(y_t, x_t) \\
\text{cov}(ax_t, by_t) &= ab\text{cov}(x_t, y_t)
\end{aligned}$$

The units of covariance depend on the units of of the underlying series. So, for example, $\text{cov}(x_t, y_t) > \text{cov}(z_t, v_t)$ does not imply that x and y are "more" strongly related than z and v. The correlation coefficient provides a means by which to make such statements. It is equal to the covariance divided by the product of the standard deviations, and is bound between -1 and 1:

$$\text{corr}(x_t, y_t) = \frac{\text{cov}(x_t, y_t)}{\sqrt{\text{var}(x_t)}\sqrt{\text{var}(y_t)}} \tag{14}$$

4 Markov Processes

There are two common ways to model time series: as Markov processes or as autoregressive moving average processes built on white noise. The so-called "Markov Property" says that the current state of the system (i.e. x_t) is a sufficient statistic to form forecasts about the future of the system. In other words, knowing x_{t-j} for $j > 0$ provides no additional information on future values x_{t+j} for $j > 0$ that x_t does not.

Let $\bar{x}$ be a $n \times 1$ vector of possible realizations of x_t. Let P be $n \times n$ matrix known as a probability or transition matrix. Its elements are the probabilities of transitioning from state i to state j between periods t and $t + 1$. Hence:

$$P_{i,j} = \text{prob}(x_{t+1} = \bar{x}_j \mid x_t = \bar{x}_i) \tag{15}$$

In words, the row tells you the current state, and the column tells you probabilities of transitioning to each possible state in the next period. As such, the rows must sum to one (i.e. the system has to transition to some value next period).

5 ARMA Processes

In contrast to Markov processes, which are discrete, autoregressive moving average (ARMA) processes are typically continuous. The building block of an ARMA process is a white noise process. A white noise process has zero mean, constant variance, and is uncorrelated across time. Let ε_t be a white noise process. This means it has the following properties:

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ \text{var}(\varepsilon_t) &= \sigma^2 \\ \text{cov}(\varepsilon_t, \varepsilon_{t+j}) &= 0 \ \forall j \end{aligned}$$

Given a white noise process, an ARMA(p,q) process is:

$$x_t = a + \rho_1 x_{t-1} + \rho_2 x_{t-2} + \ldots \rho_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots \theta_q \varepsilon_{t-q} \tag{16}$$

$$x_t = a + \sum_{j=1}^{p} \rho_j x_{t-j} + \varepsilon_t + \sum_{j=1}^{q} \theta_q \varepsilon_{t-j} \tag{17}$$

Use of lag operators makes working with ARMA processes significantly easier. Formally, the lag operator, L, is defined such that $L^j x_t = x_{t-j}$. Hence, $Lx_t = x_{t-1}$, $L^2 x_t = x_{t-2}$, and so on. We can write a lag polynomial of order p as:

$$a(L) = a_0 L^0 + a_1 L^1 + \ldots a_p L^p \tag{18}$$

As such, we can write an ARMA(p,q) process as:

$$a(L) x_t = b(L) \varepsilon_t \tag{19}$$

Using the notation above, $a(L) = 1 - \rho_1 L - \rho_2 L^2 \cdots - \rho_p L^p$ and $b(L) = 1 + \theta_1 L + \theta_2 L^2 \cdots + \theta_q L^q$.

The nice thing about ARMA processes is that they are not unique. In particular, we can (usually, at least) "invert" them to express the AR component as MA terms and vice versa. We'll show this via an example. Consider an AR(1) process:

$$x_t = \rho x_{t-1} + \varepsilon_t \tag{20}$$

This can be written in lag operator notation as:

$$(1 - \rho L) x_t = \varepsilon_t \tag{21}$$

In terms of the notation above, $a(L) = (1 - \rho L)$, a lag polynomial of order 1. We can invert this as follows:

$$x_t = (1 - \rho L)^{-1}\varepsilon_t \tag{22}$$

Now why is this helpful? Recall a certain trick for infinite sums. Suppose that $\mid \alpha \mid < 1$. Let:

$$S = 1 + \alpha + \alpha^1 + \alpha^2 + \dots \tag{23}$$

Manipulate and solve:

$$\begin{aligned} \alpha S &= \alpha + \alpha^1 + \alpha^2 + \dots \\ S - \alpha S &= 1 \\ S &= \frac{1}{1-\alpha} \end{aligned}$$

I can make this substitution because I assumed $\mid \alpha \mid < 1$. Letting ρL play the role of α, I see that $(1-\rho L)^{-1} = 1 + \rho + \rho^2 + \dots$. Therefore, I can write the AR(1) as an MA(∞), with $b(L) = a(L)^{-1}$:

$$x_t = b(L)\varepsilon_t = (1 + \rho L + \rho^2 L^2 + \dots)\varepsilon_t \tag{24}$$

$$x_t = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j} \tag{25}$$

We can also go from MA(1) to AR(∞) as well. Suppose we have:

$$x_t = \varepsilon_t + \theta\varepsilon_{t-1} \tag{26}$$

$$x_t = (1 + \theta L)\varepsilon_t \tag{27}$$

$$(1 + \theta L)^{-1} x_t = \varepsilon_t \tag{28}$$

By the same logic used above, $(1 + \theta L)^{-1} = 1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + \dots$. Hence, we can write the MA(1) process as:

$$\sum_{j=0}^{\infty} (-\theta)^j x_{t-j} = \varepsilon_t \tag{29}$$

This is an important result. It means that we can estimate MA processes by approximating them as AR processes. As long as the number of lags, p, is sufficiently high and the "roots" of $b(L)$ are not too close to one (basically that means that the coefficients decay from one pretty quickly), estimating an AR process will provide a reliable approximation to the MA process. We care about this because dynamic economic models often have MA representations, and MA terms are "hard" to estimate, whereas AR processes can be estimated consistently via OLS.

An "impulse response function" is the change in the current and expected future values of a

random variable conditional on the realization of uncertainty at some point in time. Formally:

$$\text{IRF} = E_t x_{t+j} - E_{t-1} x_{t+j} \ \forall j \geq 0 \tag{30}$$

Suppose that we have an ARMA process, which we can write as $x_t = b(L)\varepsilon_t$. Applying the expectations operator to the process, we have $E_t x_{t+j} = b(L)E_t\varepsilon_{t+j} = b(L)\varepsilon_t$. This follows from the fact that $E_t\varepsilon_{t+j} = 0 \ \forall j > 0$. By the same logic, then $E_{t-1}\varepsilon_{t+j} = 0 \ \forall j \geq 0$. Hence, the impulse response function, using the definition above, is just $b(L)\varepsilon_t$. In other words, the MA representation and the impulse response function are the same thing.

This presentation has all been in terms of scalars. ARMA processes apply equally well to vectors and matrixes. The typical notation for vectors and matrixes is capital letters; e.g. X_t is a vector, and $A(L)$ is matrix lag polynomial. But the basic stuff is all the same.

6 Econometrics

Suppose we estimate a classical linear regression, where Y_t is a $T \times 1$ vector and X_t is $T \times k$, which includes a constant:

$$Y_t = X_t\beta + \varepsilon_t \tag{31}$$

The OLS estimator is:

$$\begin{aligned} \hat{\beta} &= (X_t'X_t)^{-1}X_t'Y_t \\ \hat{\varepsilon}_t &= Y_t - X_t\hat{\beta} \\ \sigma^2 &= \frac{1}{T-k}\hat{\varepsilon}_t'\hat{\varepsilon}_t \\ \text{var}(\hat{\beta}) &= \sigma^2(X_t'X_t)^{-1} \end{aligned}$$

The fact that the observations are indexed by t as opposed to i presents no special problems per se, as long as all of the variables are stationary, in a sense to be defined below. But what are the properties of the OLS estimator here?

We say that an estimator is unbiased if $E(\hat{\beta}) = \beta$. This requires that ε_t and X_t are independent. In a time series context, this assumption is likely to be violated. To be independent, it must be the case that $E(X_{t+j} \mid \varepsilon_t) = E(X_{t+j})$. This is unlikely to happen, particularly if a lagged dependent variable appears in X_t (which would happen if you were estimating an AR term). This is because a positive realization of ε_t likely means, other things being equal, that you expect more positive realizations of Y_t in the future. If lags of Y_t are in X_t, then the independence assumption will fail.

What about consistency? An estimator is consistent if, $\lim_{T\to\infty} E(\hat{\beta}) = \beta$ and $\lim_{T\to\infty} \text{var}(\hat{\beta}) = 0$. This only requires that X_t is uncorrelated with the *current* realization of ε_t, which is likely to hold. Hence, OLS slope estimates of time series regressions with stationary variables will typically

produce consistent estimates, though biased. This of course assumes that there are not other sources of endogeneity. Construction of "correct" standard errors is a little more complicated; but we won't worry about that for now.

7 Stationarity, Trends, and Unit Roots

Many macroeconomic series have trends – for example, output. A trend means that a series grows or declines over time. This presents some unique challenges.

A related concept is that of stationarity. We say that a time series is stationary if it has a time-invariant mean, variance, and autocovariance. Formally:

$$E(x_t) = E(x_{t-j}) = \mu \ \forall \ j \tag{32}$$

$$\text{var}(x_t) = \text{var}(x_{t-j}) = \sigma^2 \ \forall \ j \tag{33}$$

$$\text{cov}(x_t, x_{t-j}) = \gamma_j \ \forall \ t \tag{34}$$

A series that has a trend clearly is not going to be stationary according to our definition. Why? If a series is growing or declining at a non-zero rate, then its mean is going to be different depending on when you look at it – i.e. the unconditional expectation of GDP in the 1950s is much different than the unconditional mean of GDP in the 1990s. That being said, a series that can also be non-stationary without having a trend (a random walk). We'll return to this in a minute.

This all matters for a couple of reasons. Macroeconomists are often interested in "business cycle" statistics – things like standard deviations (volatility) and correlations (co-movement). These statistics aren't well defined for non-stationary series. Secondly, standard econometric inference (i.e. construction of standard errors and hypothesis tests) breaks down when series become non-stationary. As such, we are often interested in transformations that render data stationary. Unfortunately, what is the appropriate transformation depends on what drives the non-stationarity.

Conventional wisdom in macro (at least until the 1980s) was that most macro series were trend stationary. A trend stationary series features some kind of deterministic time trend plus a stochastic component which is stationary. An example specification for log output, say, y_t, would be:

$$y_t = at + u_t \tag{35}$$

$$u_t = \rho u_{t-1} + \varepsilon_t \quad 0 < \rho < 1 \tag{36}$$

Here at is the trend specification, with a equal to the trend growth rate. u_t represents deviations about the trend (i.e. the "cycle"), and ε_t is a random business cycle disturbance. These shocks have effects on y_t which eventually die out. If $a = 0$, then y_t would be stationary.

Another specification that is very different but turns out to the be nearly observationally equivalent is the random walk with drift. Take the same process but let $\rho = 1$. Make this substitution

into (35) and simplify:

$$y_t = at + u_{t-1} + \varepsilon_t$$
$$y_t = at + (y_{t-1} - a(t-1)) + \varepsilon_t$$
$$\Rightarrow$$
$$y_t = a + y_{t-1} + \varepsilon_t \tag{37}$$

The second step follows by lagging (35) one period and solving for u_{t-1}. Here, a is still the trend growth rate (remember we are working in logs, so first differences are growth rates). (37) is called a random walk with drift. It is a special case of (35)-(36) when $\rho = 1$.

The non-stochastic version of the deterministic trend and random walk with drift processes are identical (up to a constant): $y_t^{ns} = at$. But the series have very different economic implications. For the deterministic trend case, stochastic shocks have transitory effects. For the random walk with drift case, shocks have effects that don't die out. Even if $a = 0$ here, the random walk series is still non-stationary. Sometimes this kind of process is called a "stochastic trend" model, as opposed to "deterministic trend" above.

Formally, a process is said to have a "unit root" if one of the roots of its characteristic equation is one. For the specification given above it means that the AR(1) coefficient is 1. A good rule of thumb for more complicated processes is this: if stochastic disturbances don't have effects that eventually die out, then you have a unit root.

The process for producing a stationary series from either of these series with trends is different. For the deterministic trend case, you would (i) estimate an OLS regression of a time dummy and (ii) take the residuals. You would be left with u_t, which is stationary. For the random walk with drift case, removing a deterministic trend would *not* render the series stationary (verify this on your own). Here, you would want to first difference the series – i.e. construct $\Delta y_t = y_t - y_{t-1}$. First differencing the trend stationary series will yield a stationary series, but it won't correspond exactly to u_t. Hence, if you think the series is a random walk with drift, difference it. If you think the series is trend stationary, remove a deterministic time trend (e.g. linear, quadratic, etc..).

8 Unit Root Econometrics

Unit roots are problematic for econometrics because the usual assumptions about error terms break down. If you have a deterministic trend series, you would remove the deterministic trend and could do econometrics (OLS, etc..) on the cycle component just as in Section 6 above. But if you think your series may have a stochastic trend, then removing the deterministic trend doesn't remove the unit root.

Not removing trends can be problematic. Suppose that we have two independent random walks (say x_t and z_t). Regressing x_t on z_t will tend to produce spurious results, in the sense that you

will find regression coefficients that are "significantly" different from zero when doing conventional OLS-based hypothesis tests. Put differently, you're likely to find relationships in the data where there truly are none. What should one do to deal with this issue? You should either detrend or difference each serious (whichever is appropriate), provided the series are not cointegrated, which we come back to in a minute. For example, if both x_t and z_t have stochastic trends, the appropriate thing to do is to difference each series, and then regress Δx_t on Δz_t. This is known as the "spurious regression" problem.

Because of all this, it is important to pre-test variables to see if there are unit roots, so that we know how to render them stationary. In practice, this is difficult to do. The intuition is straightforward. $x_t = x_{t-1} + \varepsilon_t$ is non-stationary, whereas $x_t = 0.9999x_{t-1} + \varepsilon_t$ is stationary. In finite samples, it's essentially impossible to differentiate 0.9999 from 1.

But we proceed anyway. Much of what follows is based on Dickey and Fuller (1979). Take the process given by (35)-(36), but allow for the possibility that $\rho = 1$. Substitute (36) into (35):

$$y_t = at + \rho u_{t-1} + \varepsilon_t$$

Now eliminate u_{t-1} by lagging this equation one period:

$$y_t = at + \rho(y_{t-1} - a(t-1)) + \varepsilon_t$$

Now first difference the equation (i.e. subtract y_{t-1} from both sides) and simplify:

$$\Delta y_t = \rho a + (1-\rho)at + (\rho - 1)y_{t-1} + \varepsilon_t \tag{38}$$

Consider running the following regression based on (38): $\Delta y_t = \alpha_0 + \beta t + \gamma y_{t-1} + \varepsilon_t$. The null hypothesis that the series is stationary about a deterministic trend corresponds to $\gamma < 0$ and $\beta \neq 0$. The null hypothesis that the series follows a random walk with drift corresponds to $\gamma = 0$ and $\beta = 0$. The null hypothesis that the series follows a random walk with no drift corresponds to $\gamma = \beta = \alpha_0 = 0$.

You can estimate this regression by OLS under the null of a unit root (i.e. under the null that $\gamma = 0$), but inference is not standard. Dickey and Fuller (1979) did some Monte Carlo experiments to numerically construct critical values for a t test (testing the hypothesis that $\gamma = 0$) and F tests (testing joint hypotheses that $\gamma = \beta = 0$ or $\gamma = \beta = \alpha_0 = 0$). The critical values are much more "stringent" than under normal circumstances; for example, to reject $\gamma = 0$ at 95 percent significance you would need a t statistic of -3.45.

If one is interested in estimating regressions on non-stationary variables, one should proceed as follows:

- Pre-test the variables using something like a Dickey-Fuller test to see if the series are (i) stationary, (ii) are stationary about a deterministic trend, or (iii) are unit roots (i.e. have

stochastic trends)

- If (i), estimate regressions and do inference as normal
- If (ii), then fit deterministic trends to the series and then estimate regressions on the detrended series and do inference as normal
- If (iii), first difference the series and then estimate regressions and do inference as normal

9 Cointegration

Many macro variables have trends. In the previous section I just said that we would want to get rid of the trend (either by removing a deterministic trend or by first differencing) before doing econometrics. It turns out that there is a very important caveat to that prescription. This occurs if two variables are cointegrated.

Two variables are said to be cointegrated if they are each unit root processes, but if a linear combination of them is stationary. In such a case first differencing is not appropriate.

Take two unit root processes (abstract from constants for now):

$$x_t = x_{t-1} + \varepsilon_t \tag{39}$$

$$z_t = z_{t-1} + \upsilon_t \tag{40}$$

Suppose that $x_t - \theta z_t$ is, however, stationary. We say then that x_t and z_t are cointegrated, with cointegrating vector $[1 \ \theta]$. Suppose that we estimate the following regression (again ignoring constants):

$$x_t = \beta z_t + e_t \tag{41}$$

OLS, in a large enough sample, will pick $\hat{\beta} = \theta$. That is, OLS will produce a consistent estimate the cointegrating vector, even though both variables are non-stationary. Furthermore, and perhaps oddly, OLS turns out to the "super" consistent in this case, which means that the OLS estimate of $\hat{\beta}$ converges to θ *faster* than it would if the series were stationary. In fact, OLS is going to produce good estimates even if z_t is correlated with e_t, so that under normal circumstances we would have an endogeneity problem. What's the intuition for this? Recall that OLS tries to minimize the sum of squared residuals. Hence OLS is trying to minimize $x_t - \hat{\beta} z_t$. If it picks something other than $\hat{\beta} = \theta$, then the residuals are non-stationary, which means that they will get arbitrarily big or small. Hence, if the sample size is big enough, OLS will hone in on $\hat{\beta} = \theta$. Hence, the spurious regression problem does *not apply* if variables are cointegrated with one another.

When variables are cointegrated, first differencing them is *not* appropriate. To see this, suppose that you estimate:

$$\Delta x_t = \gamma \Delta z_t + u_t \tag{42}$$

What does γ measure? To see this, start with the true process, (41). Subtract x_{t-1} from both sides:

$$\Delta x_t = -x_{t-1} + \theta z_t + e_t$$

Now add and subtract θz_{t-1} from the right hand side:

$$\Delta x_t = -x_{t-1} + \theta z_{t-1} + \theta z_t - \theta z_{t-1} + e_t$$

Simplify:

$$\Delta x_t = -(x_{t-1} - \theta z_{t-1}) + \theta \Delta z_t + e_t$$

If you estimate (42), you will not get a consistent estimate of θ. This is because there is an omitted term in the error, and that term is $-(x_{t-1} - \theta z_{t-1})$. In other words, in (42) $u_t = e_t - (x_{t-1} - \theta z_{t-1})$. Δz_t is correlated with this, and so you have a bias.

The representation above is called an "error correction" representation. The term $(x_{t-1} - \theta z_{t-1})$ is the "error", or deviation from the long run equilibrium. Intuitively, if $x_t - \theta z_t$ is stationary, then $x_t > \theta z_t$ means that x_t must be expected to fall so as to restore equilibrium over time. Regressing the first difference of x_t on the first difference of z_t ignores this "long run" relationship, and thus introduces a bias.

This insight about cointegration has a couple of other implications for doing econometrics with variables that have a unit root. Suppose again that I have two independent random walks. Suppose I want to estimate equation (39) (i.e. I don't want to impose a unit root). Conceptually, you can think of x_t and x_{t-1} as different variables, each a random walk. Would the spurious regression problem apply? NO. Essentially, x_t cointegrates with itself lagged one period (i.e. the linear combination $x_t - x_{t-1}$ is stationary). So if you estimate (39) in levels, not only will OLS provide consistent estimates of the autoregressive coefficient (true value of one), it will be super consistent.

Finally, suppose that I have a third variable that follows an independent random walk, $y_t = y_{t-1} + u_t$, and suppose that it is not cointegrated with x_t. Suppose I want to estimate the following regression:

$$x_t = \beta_1 x_{t-1} + \beta_2 y_t + e_t$$

Will OLS "work" here? If we didn't include x_{t-1} on the right hand side, we would be in the spurious regression case. But we are including it. Once again, x_t essentially cointegrates with itself.

OLS will produce super consistent results for β_1 and β_2, with again faster than normal convergence.

What is the lesson to take from this? OLS with non-stationary variables will be *fine* in terms of coefficient estimates if you are careful – don't regress two non-stationary series on each other unless you think there is likely to be cointegration (economic theory often implies cointegration) and it is always a good idea to include lagged dependent variables on the right hand side. If a series is a unit root, it cointegrates with itself, and including a lag (or several) of it on the right hand side protects you from spurious regression problems on other coefficients. Inference is another issue; the persistence in the variables makes standard inference complicated. But, properly done, OLS with non-stationary variables is not going to result in large biases (contrary to some popular opinion).

Cointegration has potentially important implications for the estimation of multivariate time series systems. We turn to that issue next.

10 VARs

Vector autoregressions, or VARs, are the multivariate generalization of univariate ARMA processes. They are widely used as tools both for prediction and for model building and evaluation.

A VAR(p) is a vector autoregression where there are p autoregressive lags of each variable in each equation. That is, if there are n total variables in the VAR, then there are n separate regressions of each of the n variables on p of its own lags and p lags of all the other variables. That expression is typically referred to as the *reduced form* of a model. The VAR(p) can best be motivated by first considering the *structural form*, which is a system of simultaneous equations.

To make matters as clear as possible, consider two variables, x_t and z_t. Assume for now that these are stationary. The structural form can be written:

$$x_t = \alpha_x + a_{x,0} z_t + \sum_{j=1}^{p} a_{x,j}^{x} x_{t-j} + \sum_{j=1}^{p} a_{x,j}^{z} z_{t-j} + \varepsilon_{x,t}$$
$$z_t = \alpha_z + a_{z,0} x_t + \sum_{j=1}^{p} a_{z,j}^{x} x_{t-j} + \sum_{j=1}^{p} a_{z,j}^{z} z_{t-j} + \varepsilon_{z,t}$$

Here $\varepsilon_{x,t}$ and $\varepsilon_{z,t}$ are *structural* shocks. Structural in the econometric sense just means that they are mean zero and are uncorrelated with one another: $\text{cov}(\varepsilon_{x,t}, \varepsilon_{z,t}) = 0$. Each is drawn from some distribution with known variance.

Econometrically, estimating the above system of equations is a snake pit. Why? Because in the first equation, z_t will in general be correlated with $\varepsilon_{x,t}$, and in the second equation x_t will in general be correlated with $\varepsilon_{z,t}$, meaning that the assumptions for OLS to be consistent are violated. There is a pervasive simultaneity bias. Structural VAR methodology proceeds by making assumptions that allow this system of equations to be estimated consistently. Frequently this amounts to zero restrictions. For example, suppose that we assume $a_{x,0} = 0$. This means that $\varepsilon_{z,t}$ has no effect

on x_t within period (it will have effects at leads, because of the autoregressive parameters). But if $\varepsilon_{z,t}$ has no contemporaneous effect on x_t, then we can estimate the second equation consistently by OLS. Formally, this kind of restriction is an *exclusion* restriction, hopefully motivated by some economic theory, at least in the loose sense of the word theory.

It is helpful to write out this system in matrix form (abstracting from the constant):

$$\begin{pmatrix} 1 & -a_{x,0} \\ -a_{z,0} & 1 \end{pmatrix} \begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{pmatrix} a_{x,1}^x & a_{x,1}^z \\ a_{z,1}^x & a_{x,1}^z \end{pmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \ldots \begin{pmatrix} a_{x,p}^x & a_{x,p}^z \\ a_{z,p}^x & a_{x,p}^z \end{pmatrix} \begin{bmatrix} x_{t-p} \\ z_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{z,t} \end{bmatrix}$$

Simplifying on notation we can write this as:

$$A_0 \begin{bmatrix} x_t \\ z_t \end{bmatrix} = A_1 \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \ldots A_2 \begin{bmatrix} x_{t-p} \\ z_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{z,t} \end{bmatrix}$$

Pre-multiply everything by A_0^{-1}:

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = A_0^{-1}A_1 \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \cdots + A_0^{-1}A_p \begin{bmatrix} x_{t-p} \\ z_{t-p} \end{bmatrix} + A_0^{-1} \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{z,t} \end{bmatrix}$$

Make the following definition:

$$u_t \equiv A_0^{-1} \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{z,t} \end{bmatrix}$$

We will refer to u_t as the vector of *innovations*, and ε_t as the vector of *structural shocks*. Innovations are just forecast errors – the above specification is just a regression of each variable on lags of itself and the other variable, with u_t the vector of innovations. In general, $u_{x,t}$ will be correlated with $u_{z,t}$. The only condition under which that would not be the case is if the off-diagonal elements of A_0 are equal to zero, which would mean that neither structural shock affects the other variable within period. Denote the variance covariance matrix of innovations by Σ_u. This maps into the parameters of the structural model as follows:

$$\Sigma_u = A_0^{-1}\Sigma_\varepsilon A_0^{-1\prime}, \quad \Sigma_\varepsilon = E(\varepsilon_t \varepsilon_t{}') = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{pmatrix}$$

This follows because the structural shocks are, by assumption, uncorrelated. In fact, it is common to assume that the structural shocks have unit variance, so that the variance-covariance matrix of structural shocks is just an identity matrix, $\Sigma_\varepsilon = I$. This is just a normalization, and it can be incorporated by re-scaling all the coefficients. Let's work with that from now on. Hence, the variance-covariance matrix of innovations can then be written:

$$\Sigma_u = A_0^{-1}A_0^{-1\prime}$$

Even though I presented the VAR in its structural, simultaneous equation form, and then went

to the reduced form, in practice we go the other way around when estimating. You begin by estimating a VAR(p) on the data (either in first difference or levels if you think the series are cointegrated). Hence, you estimate:

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \widetilde{A}_1 \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \cdots + \widetilde{A}_p \begin{bmatrix} x_{t-p} \\ z_{t-p} \end{bmatrix} + u_t$$

I put tildes on the coefficient matrixes because these are reduced form; they are not the structural coefficients, but rather $A_0^{-1}A_j,\ \ j = 1, \ldots, p$. You can estimate this system equation by equation via OLS. This turns out to be both consistent (all the regressors are dated $t-1$ or earlier, and hence uncorrelated with the residuals) and efficient, since the regressors are the same in every equation (otherwise seemingly unrelated regressions would be more efficient). Use OLS to get a time series of residuals in each equation, and form $\widehat{u}_t$. Then get an estimate of the variance-covariance matrix of innovations, $\Sigma_{\widehat{u}} = \frac{1}{T-p-1}\widehat{u}_t\widehat{u}_t'$.

You can used the reduced form estimation to do forecasting: conditional on a current state of the system, the reduced form is a perfectly good way to forecast the future state of the system. Indeed, forecasting is one of the chief uses of VARs. To make economic conclusions, though, one needs to recover the structural form we discussed above. But, as discussed above in the derivation of the reduced form, we know the mapping from structural form to reduced form. It is given by:

$$A_0^{-1}A_0^{-1\prime} = \Sigma_{\widehat{u}} \tag{43}$$

Given an estimate of $\Sigma_{\widehat{u}}$, which we get from OLS equation by equation estimation above, this is just a non-linear system of equations. Does it have a unique solution? *No.* There are n variables in the VAR; hence $\Sigma_{\widehat{u}}$ is $n \times n$. How many unique elements are in $\Sigma_{\widehat{u}}$ though? Since it is a variance-covariance matrix, it is symmetric, and has only $\frac{n^2+n}{2}$ unique elements, but A_0^{-1} has n^2 elements. The way to think about this is as follows: there are $\frac{n^2+n}{2}$ "knowns" (unique elements of the variance-covariance matrix) and there are n^2 unknowns (the elements of A_0^{-1}). This system is overdetermined – there are more unknowns than knowns. To get a unique solution, we need these to match up. Hence, we need to impose $n^2 - \frac{n^2+n}{2} = \frac{n(n-1)}{2}$ restrictions on A_0^{-1} to get a unique solution. Once we do that, (43) is just a non-linear system of equations which can be solved using numerical optimization routines.

Now that we have all this, go back to the estimation of the reduced form. Define $A(L) = I - \widetilde{A}_1 L - \widetilde{A}_2 L^2 - \cdots - \widetilde{A}_p L^p$. Write the reduced form as:

$$A(L)X_t = u_t,\ \ X_t = \begin{bmatrix} x_t \\ z_t \end{bmatrix}$$

Now re-write this using the mapping between structural shocks and innovations:

$$A(L)X_t = A_0^{-1}\varepsilon_t$$

Now invert the matrix polynomial in the lag operator to write this out as the structural moving

average representation:

$$X_t = C(L)\varepsilon_t,$$

Here $C(L) = A(L)^{-1}A_0^{-1}$. Use the following notation: $C(j)$ is the coefficient at lag j. Hence, $C(0)$ is the coefficient matrix on impact, $C(1)$ at a one period lag, $C(2)$ at a two period lag, and so on.

To find the impulse response function of, say, x_t to $\epsilon_{1,t}$, we would set $\epsilon_{1,t} = 1$, $\epsilon_{2,t} = 0$, and all subsequent shocks are zero in expectation. The impulse response on impact would be $C_{1,1}(0)$, the response after two periods would be $C_{1,1}(1)$, and so on. We could do the same for variable 2. Our generic definition would be that $C_{i,j}(h)$ is the impulse response of variable i to shock j at horizon h. The matrix A_0^{-1} governs the impulse responses of the variables to the shock on impact – for this reason it is sometimes called the impact matrix or the matrix of impact multipliers.

The forecast error of a variable at time t is the change in the variable that couldn't have been forecast between $t-1$ and t. This is due to the realization of the structural shocks in the system, ϵ_t. We can compute the forecast error over many different horizons, h. The forecast error variance at horizon $h = 0$ for each variable is:

$$\begin{aligned} x_t - E_{t-1}x_t &= C_{1,1}(0)\epsilon_{1,t} + C_{1,2}(0)\epsilon_{2,t} \\ z_t - E_{t-1}z_t &= C_{2,1}(0)\epsilon_{1,t} + C_{2,2}(0)\epsilon_{2,t} \end{aligned}$$

The forecast error variances are just the squares of the forecast errors (since the mean forecast error is zero). Let $\Omega_i(h)$ denote the forecast error variance of variable i at horizon h. Then at $h = 0$, this is simply:

$$\begin{aligned} \Omega_1(0) &= C_{1,1}(0)^2 + C_{1,2}(0)^2 \\ \Omega_2(0) &= C_{3,1}(0)^2 + C_{3,2}(0)^2 \end{aligned}$$

The above follows from the assumptions that the shocks have unit variance and are uncorrelated. The forecast error of the variables at horizons $h = 1$ is:

$$\begin{aligned} x_{t+1} - E_{t-1}x_{t+1} &= C_{1,1}(0)\epsilon_{1,t+1} + C_{1,2}(0)\epsilon_{2,t+1} + C_{1,1}(1)\epsilon_{1,t} + C_{1,2}(1)\epsilon_{2,t} \\ z_{t+1} - E_{t-1}z_{t+1} &= C_{2,1}(0)\epsilon_{1,t+1} + C_{2,2}(0)\epsilon_{2,t+1} + C_{2,1}(1)\epsilon_{1,t} + C_{2,2}(1)\epsilon_{2,t} \end{aligned}$$

The forecast error variances are then:

$$\Omega_1(1) = C_{1,1}(0)^2 + C_{1,2}(0)^2 + C_{1,1}(1)^2 + C_{1,2}(1)^2$$
$$\Omega_2(1) = C_{3,1}(0)^2 + C_{3,2}(0)^2 + C_{1,1}(1)^2 + C_{1,2}(1)^2$$

To go to more periods, we can then define the forecast error variances recursively as follows:

$$\begin{aligned}\Omega_i(0) &= C_{i,1}(0)^2 + C_{i,2}(0)^2 \\ \Omega_i(1) &= C_{i,1}(1)^2 + C_{i,2}(1)^2 + \Omega_i(0) \\ &\vdots \\ \Omega_i(h) &= C_{i,1}(h)^2 + C_{i,2}(h)^2 + \Omega_i(h-1)\end{aligned}$$

More generally, for a n variable system, the total forecast error variance of variable i at horizon h in a n variable system is:

$$\Omega_i(h) = \sum_{k=0}^{h} \sum_{j=1}^{n} C_{i,j}(k)^2 \tag{44}$$

A forecast error variance decomposition – or just variance decomposition for short – is a way to quantify how important each shock is in explaining the variation in each of the variables in the system. It is equal to the fraction of the forecast error variance of each variable due to each shock at each horizon. Let $\omega_{i,j}(h)$ be the forecast error variance of variable i due to shock j at horizon h. This is:

$$\omega_{i,j}(h) = \sum_{k=0}^{h} C_{i,j}(k)^2$$

The fraction of the forecast error variance of variable i due to shock j at horizon h, denoted $\phi_{i,j}(h)$, is then the above divided by the total forecast error variance:

$$\phi_{i,j}(h) = \frac{\omega_{i,j}(h)}{\Omega_i(h)} = \frac{\sum_{k=0}^{h} C_{i,j}(k)^2}{\sum_{k=0}^{h} \sum_{j=1}^{n} C_{i,j}(k)^2} \tag{45}$$

10.1 Mechanics of How to Do This

Lag operator notation is convenient for writing things down, but it can be difficult to think about it and is computationally not something one can just implement. It turns out that there is a simple way to estimate a structural VAR and compute impulse responses. You effectively turn the system into a VAR(1). Continue with our example, which we can write in reduced form as:

$$X_t = \widetilde{A}_1 X_{t-1} + \widetilde{A}_2 X_{t-2} + \ldots \widetilde{A}_p X_{t-p} + u_t$$

Define:

$$Z_t = \begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p-1} \end{bmatrix}$$

We can re-write the VAR(p) as a VAR(1) using this expanded system:

$$\begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p-1} \end{bmatrix} = \begin{pmatrix} \widetilde{A}_1 & \widetilde{A}_2 & \ldots & \widetilde{A}_p \\ \underset{n\times n}{I} & \underset{n\times n}{0} & \ldots & \underset{n\times n}{0} \\ \vdots & \ddots & \vdots & \vdots \\ \underset{n\times n}{0} & \underset{n\times n}{0} & \underset{n\times n}{I} & \underset{n\times n}{0} \end{pmatrix} \begin{bmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-p} \end{bmatrix} + \begin{bmatrix} u_t \\ \underset{n\times 1}{0} \\ \vdots \\ \underset{n\times 1}{0} \end{bmatrix}$$

More compactly:

$$X_t = \Gamma X_{t-1} + U_t, \quad U_t = \begin{bmatrix} u_t \\ \underset{n\times 1}{0} \\ \vdots \\ \underset{n\times 1}{0} \end{bmatrix}$$

This is called the "companion" form – expressing a VAR(p) as a VAR(1). In structural form:

$$X_t = \Gamma X_{t-1} + \Upsilon_t, \;\; \Upsilon_t = \begin{bmatrix} A_0^{-1}\varepsilon_t \\ \underset{n\times 1}{0} \\ \vdots \\ \underset{n\times 1}{0} \end{bmatrix}$$

To get the impulse responses to, say, shock 1, set $\varepsilon_t(1) = 1$ and everything else equal to zero. Then $C_{:,1}(1) = A_0^{-1}(:, 1)$ – in other words, the impulse responses of the n variables on impact are given by the first column of A_0^{-1}. Call the impulse response of X_t in the first period then:

$$\widetilde{C}_{:,1}(1) = \begin{pmatrix} A_0^{-1}(:,1) \\ \underset{n(n-1)\times 1}{0} \end{pmatrix}$$

Because we've now re-written this as an AR(1), responses at subsequent horizons are just Γ times the response in the previous horizon:

$$\widetilde{C}_{:,1}(j) = \Gamma \widetilde{C}_{:,1}(j-1) = \Gamma^j \begin{pmatrix} A_0^{-1}(:,1) \\ \underset{n(n-1)\times 1}{0} \end{pmatrix}, \ \forall j \geq 1$$

You can recover the actual impulse response by simply dropping all but the first n elements of $\widetilde{C}_{:,1}(j)$. To get responses to a different shock, you simply set a different shock to 1 and all others to zero. For the kth shock, that would be:

$$\widetilde{C}_{:,k}(j) = \Gamma^j \begin{pmatrix} A_0^{-1}(:,k) \\ \underset{n(n-1)\times 1}{0} \end{pmatrix}, \ \forall j \geq 1$$

In other words, to get the responses to the kth shock, just start with the kth column of A_0^{-1}.

Follow these steps for easy estimation and construction of VAR impulse responses:

1. Decide what variables to include in the VAR
2. Decide how those variables should enter the VAR (levels, first differences, deviations from trend, etc.) – more on this below
3. Decide how many lags, p to include in the VAR. It is typical in time series setting to use a year's worth of lags – for quarterly data $p = 4$, for monthly data $p = 12$, etc.. You can determine the statistically best-fitting lag length through various information criteria, such as the AIC or BIC
4. Estimate the reduced form VAR, to get the $\widetilde{A}$ coefficient matrix and residuals, u_t. You can do this either by (i) estimating each equation separately by OLS, and then getting the variance-covariance matrix of residuals; or (ii) estimating the companion form directly. Either way is fine; estimating the companion form requires transforming the series a little bit but is easily doable
5. Form the companion form. You either have that already or you can construct it manually given the equation-by-equation OLS estimate
6. Impose $\frac{n(n-1)}{2}$ restrictions on A_0^{-1}; then solve for A_0^{-1} by setting $A_0^{-1}A_0^{-1\prime} = \Sigma_{\widehat{u}}$
7. Given A_0^{-1} and the companion matrix estimate Γ from above, form impulse response to each of the n structural shocks as described above
8. The estimates of the impulse responses can then be used to get the variance-decomposition

10.2 Where Do the Restrictions Come From?

The whole enterprise above requires imposing structure on the "impact matrix" A_0^{-1}. Where do these restrictions come from? Ideally they come from economic theory.

The most common set of restrictions are recursive – these amount to zero restrictions on A_0^{-1}. Economically, the (i, j) element of A_0^{-1} tells us how variable i responds "on impact" (i.e. immediately) to structural shock j. Recursive restrictions require certain variables to not respond within period to certain shocks. For example, suppose, because of fiscal delays, government spending or taxes can't react to technology or monetary policy shocks immediately. This imposes zero restrictions on A_0^{-1}. Alternatively, one might thing that monetary policy can react to other shocks immediately (they meet all the time and don't need to go through the legislative process), but that changes in monetary policy cannot affect the economy immediately (Friedman's "long and variable lags"). This would also impose zero restrictions on A_0^{-1}.

Recursive restrictions can be implemented with what is called a "Choleski decomposition" of the the variance-covariance matrix of innovations, $\Sigma_{\widehat{u}}$. A Choleski decomposition generates a lower triangular decomposition of the variance-covariance matrix that satisfies $BB' = \Sigma_u$, which happens to be exactly the mapping between innovations and structural shocks we saw above. In Matlab, for example, typing "B = chol(S)" will produce an *upper* triangular matrix, so to get it in lower triangular form simply transpose, "B = B'".

Economically, what does impose a lower triangular structure on A_0^{-1} mean? It means that the first variable doesn't respond within period to the last $n-1$ structural shocks, the second variable doesn't respond within period to the last $n-2$ shocks, and so on. Sometimes this kind of structural is referred to as a "causal ordering" – the first shock affects all variables within period, the second shock affects all but the first variable within period, and so on. It's also sometimes simply referred to as an "ordering." Variables "ordered" first are not allowed to react to other shocks within period; variables "ordered" last are allowed to respond to all shocks within period. Therefore, we typically think that "endogenous" variables (like monetary policy or output, things which react quickly to disturbances) should be "ordered" late, whereas "exogenous" variables (like government spending, things which don't quickly react) are "ordered" first. Note that I use "exogenous" and "endogenous" loosely here, as all variables are endogenous in the dynamic sense in a VAR.

There are many other kinds of restrictions that can be imposed. For example, you might guess that some shock has the same effects on two different variables – this would impose that certain elements of A_0^{-1} are equal to each other. Another celebrated example concerns "long run" restrictions. Economic theory suggests that "demand" shocks should have temporary effects on economic activity, whereas "supply" shocks might have permanent effects. Essentially what one can do is to search over the space of permissable A_0^{-1} matrixes (i.e. the set of matrixes satisfying $A_0^{-1}A_0^{-1\prime} = \Sigma_{\widehat{u}}$) for ones that produce impulse responses of variables that are zero after a long forecast horizon (e.g. $C_{i,k}(h) = 0$ as $h \to \infty$). One can also do "sign" and "shape" restrictions, sometimes only yielding set identification (a set of permissable A_0^{-1} matrixes, as opposed to a single unique one). These extensions are beyond the scope of the course.

10.3 How Should Variables Enter VARs?

As noted above, macroeconomic data are typically persistent and often have trends, and this can create some econometric challenges. Suppose you have settled on a set of variables to include in a VAR. One option would be to individually test them for unit roots, and then difference them if they have unit roots or leave them in levels otherwise (or remove a linear time trend). Then estimate the system using the transformed data. This approach will work as long as the variables are not cointegrated.

If the variables are cointegrated, simply differencing them is inappropriate and will result in a mis-specification (see the two variable example shown above). In this case, estimating the VAR in levels – even if the series are unit roots with trend – is much better than differencing, as it results in no mis-specification and is consistent (indeed, super consistent). There are two other options one can take if one thinks variables are cointegrated – either impose the cointegration and estimate a VAR, or estimate what is called a vector error correction model (VECM).

To see this, suppose that we have two series, x_t and z_t, each individually non-stationary, but where the ratio, $x_t - z_t$ is stationary (think of these variables as already being logged, so a ratio is a log first difference). Estimating a VAR on Δx_t and Δz_t would result in a mis-specification. Estimating a VAR on x_t and z_t will be consistent, but the non-stationarity of the variables makes inference difficult. One could alternatively estimate a VAR on Δx_t and $x_t - z_t$. These variables are both stationary. One can recover impulse responses by cumulating responses to Δx_t and then backing out the level response of z_t from the response of $x_t - z_t$. Alternatively, one could estimate the VECM. The VECM is essentially a VAR in first differences with an extra term in each regression: $x_{t-1} - z_{t-1}$ appears on the right hand side in both equations.

My recommendation is to *almost always* just estimate VARs in levels. This is the "safe" procedure. As long as there are lags on the right hand side the spurious regression problem isn't there, and if there is cointegration levels estimation is robust to that. Differencing or imposing cointegration results in efficiency gains, but only if it is appropriate. In practice it can be difficult to tell whether series are non-stationary or cointegrated, so imposing structure through differencing or cointegration can result in mis-specification. The safe route is to estimate in levels.

10.4 An Example: GDP and Fed Funds Rate

Suppose I am interested in learning about the effects of monetary policy shocks on economic activity. To take a very simply example, suppose I take data on real GDP and the effective Federal Funds Rate (the funds rate is aggregated to a quarterly frequency). The data run from 1960q1 to 2006q4. I estimate a VAR with $p = 4$ lags (i.e. a year's worth) with both variables in levels, even though output is clearly trending. The variance-covariance matrix of residuals is given by:

$$\Sigma_{\widehat{u}} = \begin{pmatrix} 0.0000514 & 0.000916 \\ 0.000916 & 0.809285 \end{pmatrix}$$

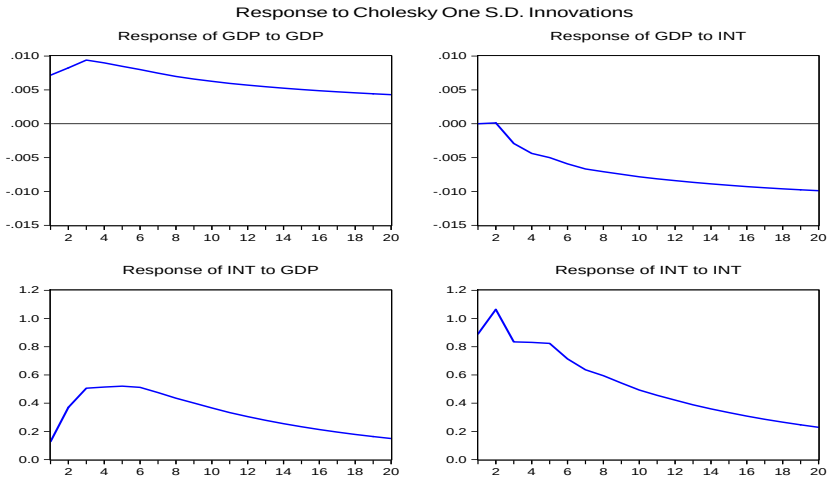
I'm going to "order" the Funds rate second – economically this means that the Fed can respond

immediately to economic conditions, but its actions have delayed effects on the economy. The estimate of A_0^{-1} is:

$$A_0^{-1} = \begin{pmatrix} 0.0072 & 0 \\ 0.1278 & 0.8905 \end{pmatrix}$$

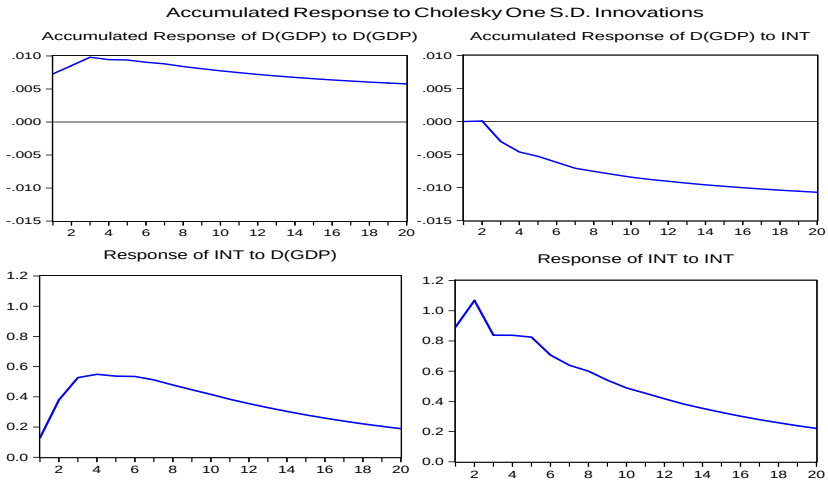
This means that an "output shock" (the first shock in the system) raises output by 0.72 percent on impact and the Fed reacts by raising the funds rate by 13 basis points – i.e. the Fed reacts to an expansion by tightening rates. In contrast, an exogenous increase in the funds rate (exogenous with respect to current output) causes the Funds rate to go up by 89 basis points and has no effect on output immediately.

Below is a plot of the dynamic responses:



This has the flavor of what our intuition suggests – the Fed reacts to improved economic activity (the first column) by raising interest rates. In contrast, an exogenous increase in interest rates (a monetary "shock") causes output to decline rather sharply after a period of delay.

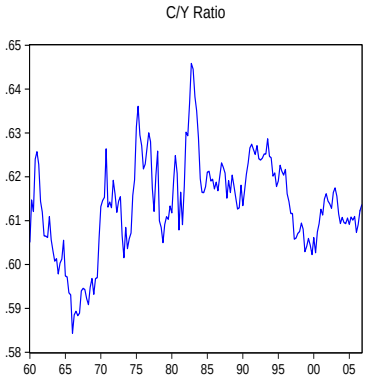
Now, so as to back up the claim I made before that estimating this in levels doesn't make much of a difference relative to estimating in first differences, I'm going to re-estimate the VAR with the interest rate in levels and GDP in the first differences (growth rates). Then I back out impulse responses and cumulate the GDP response – cumulating puts the GDP difference response back into levels. The responses are shown below:



These are almost identical to what you get when estimating in levels. The point is that you do not introduce a bias by estimating a system in levels when there are non-stationary variables. Inference (which we are ignoring) becomes more complicated, but it doesn't bias the response. Differencing when differencing isn't appropriate, however, does introduce a bias.

10.5 An Example: Cochrane (1994)

Cochrane (1994, *Quarterly Journal of Economics*), estimates a two variable VAR featuring non-durable and services consumption and real GDP. Both of these series are trending, but the ratio basically fluctuates around a constant (about 60 percent). This means that the series are likely cointegrated, which is also suggested by standard economic theory. Below is a graph of the consumption-income ratio.



There is no clear trend in this picture; there is some debate about whether it is stationary, but let's run with the stationarity assumption for now.

I go ahead and estimate the system in the log levels of both consumption and GDP. After

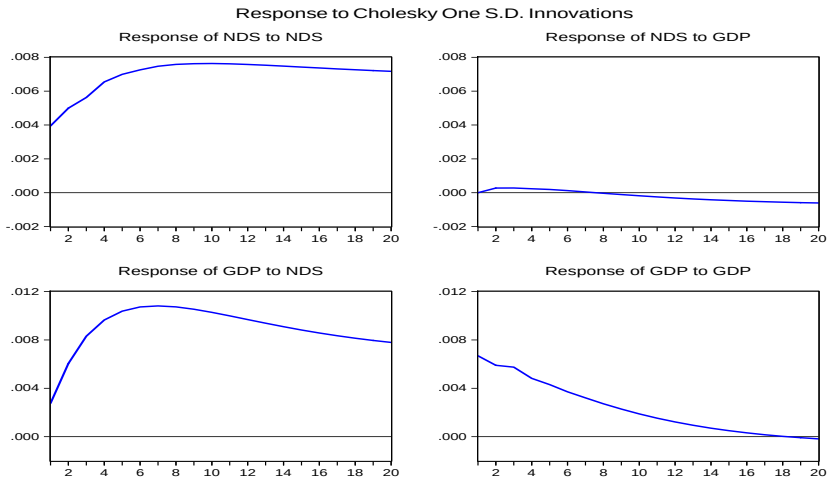
estimating a VAR(4) on the levels of the variables, I get the following estimate of the variance-covariance matrix of residuals:

$$\Sigma_{\widehat{u}} = \begin{pmatrix} 0.0000156 & 0.000011 \\ 0.000011 & 0.000052 \end{pmatrix}$$

Cochrane "orders" consumption first in a recursive identification. The restriction he is using comes from the very basic Friedman (1957) permanent income hypothesis: shocks to income which don't affect consumption (i.e. the second shock in the system with consumption ordered first) ought to have transitory effects on output; shocks to consumption ought to have permanent effects on income, since consumption is equal to permanent income. Imposing this restriction I get the following impact matrix:

$$A_0^{-1} = \begin{pmatrix} 0.0040 & 0 \\ 0.0028 & 0.0067 \end{pmatrix}$$

The dynamic impulse response are below:

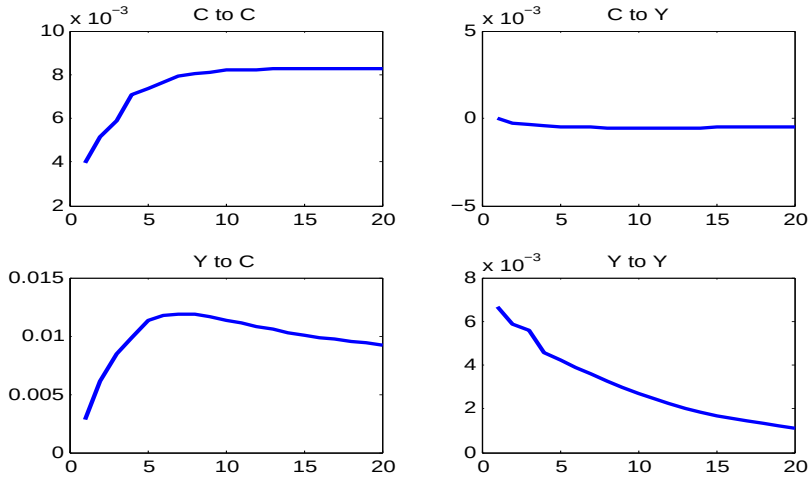


This is consistent with the permanent income story – increases in consumption today predict increases in income, and movements in income unrelated to consumption are transitory and never have any noticeable effect on consumption. Constructing a variance decomposition allows us to say something about how important permanent (consumption) and transitory (gdp) shocks are for the evolution of consumption and output.

Variable and Horizon	Due to Permanent Shock	Due to Transitory Shock
Consumption		
$h = 1$	1.00	0.00
$h = 4$	0.998	0.002
$h = 10$	0.999	0.001
$h = 20$	0.997	0.003
GDP		
$h = 1$	0.145	0.855
$h = 4$	0.603	0.397
$h = 10$	0.818	0.182
$h = 20$	0.892	0.108

The numbers in this table again conform to the basic theory. Consumption is basically completely explained (at all horizons) by the permanent shock. A significant fraction of the movements in GDP at short horizons are due to the transitory shock; at longer horizons it's mostly the permanent shock. This suggests that both "demand" (i.e. temporary) and "supply" (i.e. permanent) shocks are important in understanding the dynamics of GDP and consumption.

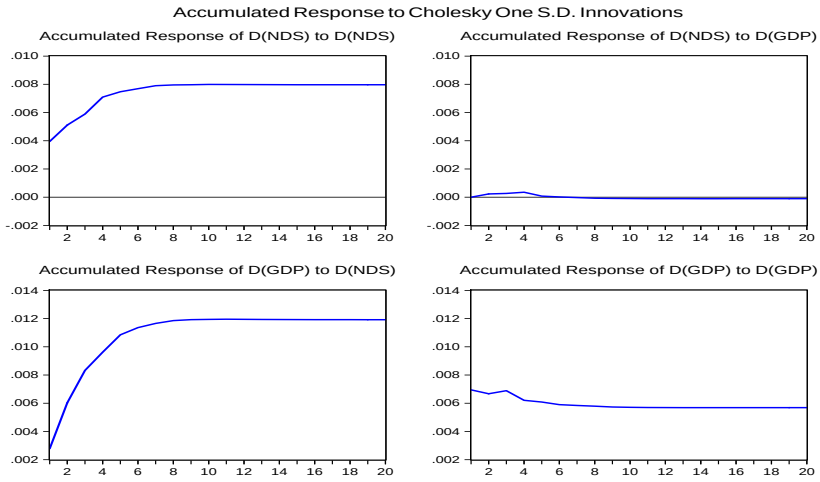
Now I'm going to re-estimate the VAR; this time not in levels, but rather imposing cointegration. That is, I estimate a VAR featuring Δc_t and $c_t - y_t$. I again "order" consumption first, and then cumulate the consumption response and back out the output response to get everything in levels. Here is what I get:



These response are almost exactly the same as what I get from estimating in levels, and the variance decomposition works out similarly as well. Point is – estimation in levels is robust to cointegration. It is more efficient to impose the cointegration – if it is there, but imposing that could create a mis-specification if it is not. Levels is safer.

So as to see that mis-specifications can create problems, I'm going to estimate a system in which

I ignore cointegration and estimate a VAR with both variables in first differences. The responses are below.



You can see that there is an important differences relative to what we found above – there is no transitory shock to GDP. We would *reject* (falsely, in this case) the basic implication of the PIH here, because consumption ought to respond to a permanent change in GDP. But this would be wrong – what has happened here is that we have introduced a mis-specification by inappropriate differencing.

Again, the bottom line is to *be careful* when working with non-stationary data. Differencing often feels like the right thing to do, but can result in serious mis-specifications if variables are cointegrated. Estimating in levels (provided there are lags of the dependent variable on the right hand side, which takes care of the spurious regression problem) is always safer.

11 Bootstrapping

As I've mentioned throughout the last several sections, standard econometrics with highly persistent and/or non-stationary data is usually fine from the perspective of getting asymptotically "right" coefficient estimates, but OLS inference can get screwed up. This means that hypothesis testing will generally be flawed without some correction to the standard errors. Doing this analytically is hard and requires assumptions. A simpler way is construct standard errors via re-sampling estimated residuals.

A convenient and popular way of constructing standard errors is via the bootstrap. The basic idea behind the bootstrap is as follows. Estimate a time series model and get the residuals. Then construct a large number of "fake" data sets by re-sampling (with replacement) from the observed residuals and using the estimated coefficients. Then re-apply your econometric procedure to the fake data and generate an entire distribution of estimates. Then construct standard errors using the bootstrap simulated distributions of coefficients.

I'll show this through an example. Suppose that we have a time series process:

$$x_t = 0.9x_{t-1} + \varepsilon_t \quad \varepsilon_t \sim N(0,1) \tag{46}$$

I generated 1000 different data sets of 200 observations each using this as the data generating process. Then on each data set I estimated an AR(1) process. My average OLS estimate of the AR coefficient was 0.88. This is ever so slightly downward biased, which is normal in time series data, but if I were to increase the sample size the bias would go away. My average OLS estimate of the standard error of this estimate is 0.034. But the standard deviation of the distribution of estimates is actually 0.037. In other words, the OLS standard errors are too small.

Now I do a bootstrap. For each simulation of the data, I estimate via OLS and get an estimate of $\hat{\rho}$ as well as a time series of residuals, $\hat{e}_t$. Then, I create N different re-sampled versions of $\hat{e}_t$ by drawing randomly (with replacement) from the observed empirical distribution of $\hat{e}_t$. You can use the "boostrp" command in Matlab to do this. Then, for each of these N different bootstrap samples, I construct another fake series of xs using my estimated ρ. Then I re-estimate an AR(1) on the bootstrap sample and save it. I repeat this N times – i.e. once for each bootstrap sample. Then I look at the standard deviation of the estimated $\hat{\rho}$ across each bootstrap sample. When I do this $N = 300$ times, my bootstrap standard deviation of $\hat{\rho}$ comes out to be 0.038. This is much closer to the true standard deviation of ρ than is the OLS estimate of the standard error.

You can use bootstrapping in a variety of different contexts. For example, you can use it to construct confidence intervals for impulse response functions in VARs. You would estimate the VAR and then construct N bootstrap samples of the VAR residuals. Then using the estimated AR coefficients, you would create N different samples of the variables in the VAR using the bootstrapped simulations of the errors. Then on each simulated data set, you would run the VAR and construct impulse responses. You would save the impulse responses. Then you would take the standard deviation (or percentiles) of the bootstrap distributions of the estimated impulse responses.

12 Filtering

A related concept to first differencing and detrending data is that of filtering. The basic premise is to break a series down into two components: "trend" and "cycle":

$$x_t = x_t^\tau + x_t^c \tag{47}$$

The basic approach is to come up with an estimate of the "trend" component, and then subtract that off from the actual series so as to get the "cycle" component. We've already seen one way to do this – estimate a linear time trend to get x_t^τ. But more generally, we may want to allow the trend to move around.

The Hodrick-Prescott filter (HP filter) is very common in empirical macro and does just this; it also has as a special case the linear time trend. Formally, let λ be an exogenous constant chosen by the researcher in advance. The HP filter chooses a sequence of trend, x_t^τ, to solve the following

minimization problem:

$$\min_{x_t^\tau} \quad \sum_{t=1}^{T} (x_t - x_t^\tau)^2 + \lambda \sum_{t=2}^{T-2} \left((x_{t+1}^\tau - x_t^\tau) - (x_t^\tau - x_{t-1}^\tau) \right)^2 \tag{48}$$

The first part is the cycle component (i.e. $x_t - x_t^\tau$). It represents a penalty for large cyclical components. The second part is that change in the change of the trend. Basically, this part is a penalty for the trend not being smooth. In words, then, the HP filter finds a trend that minimizes cyclical fluctuations subject to a penalty for the trend itself moving around.

Consider a couple of different possible values of λ. If $\lambda = 0$, then there is no penalty to the trend jumping around, and you would set $x_t^\tau = x_t$ – in other words, the trend would be the actual series and there would be no cyclical component. As $\lambda \to \infty$, you will want $(x_{t+1}^\tau - x_t^\tau) - (x_t^\tau - x_{t-1}^\tau) = 0$. This means that the change in the trend is constant, which means that you would pick out a linear time trend. For intermediate values of λ, the trend will move around some, but there will still be a cyclical component. For quarterly data, it is common to use $\lambda = 1600$. For annual data, people typically use $\lambda = 100$. For monthly data, it is common to use $\lambda = 14400$.

The bandpass filter is another popular statistical filter. It is aimed to isolate cycles with different "periodicities", where periodicity measures the amount of time it takes for a cycle to complete. To study the properties of this filter more fully, one needs to use frequency domain and spectral analysis, which is beyond the aim of the course. Business cycle frequencies are typically defined as having periodicities between 6-32 quarters (1.5 to 8 years). Hence a bandpass filter with periodicities between 6 and 32 quarters can be used to isolate the business cycle component of the data. In practice this turns out to be pretty similar to the HP filter.