

# Time series and cross section (TSCS) analysis

ELECDEM

Istanbul, session #4

**Christopher Wlezien**

with special thanks to **Mark Franklin**

## The plan for the three sessions

Today: An overview of time series modeling. A sampling platter of sorts—wide and not too deep, i.e., technical. It is important to do and follow as we will build on it tomorrow. Mostly will be done at the break, but will use the period afterward to finish up, address any questions and then put our hands on the computer. I'll also ask you to do a (very) little homework assignment for tomorrow.

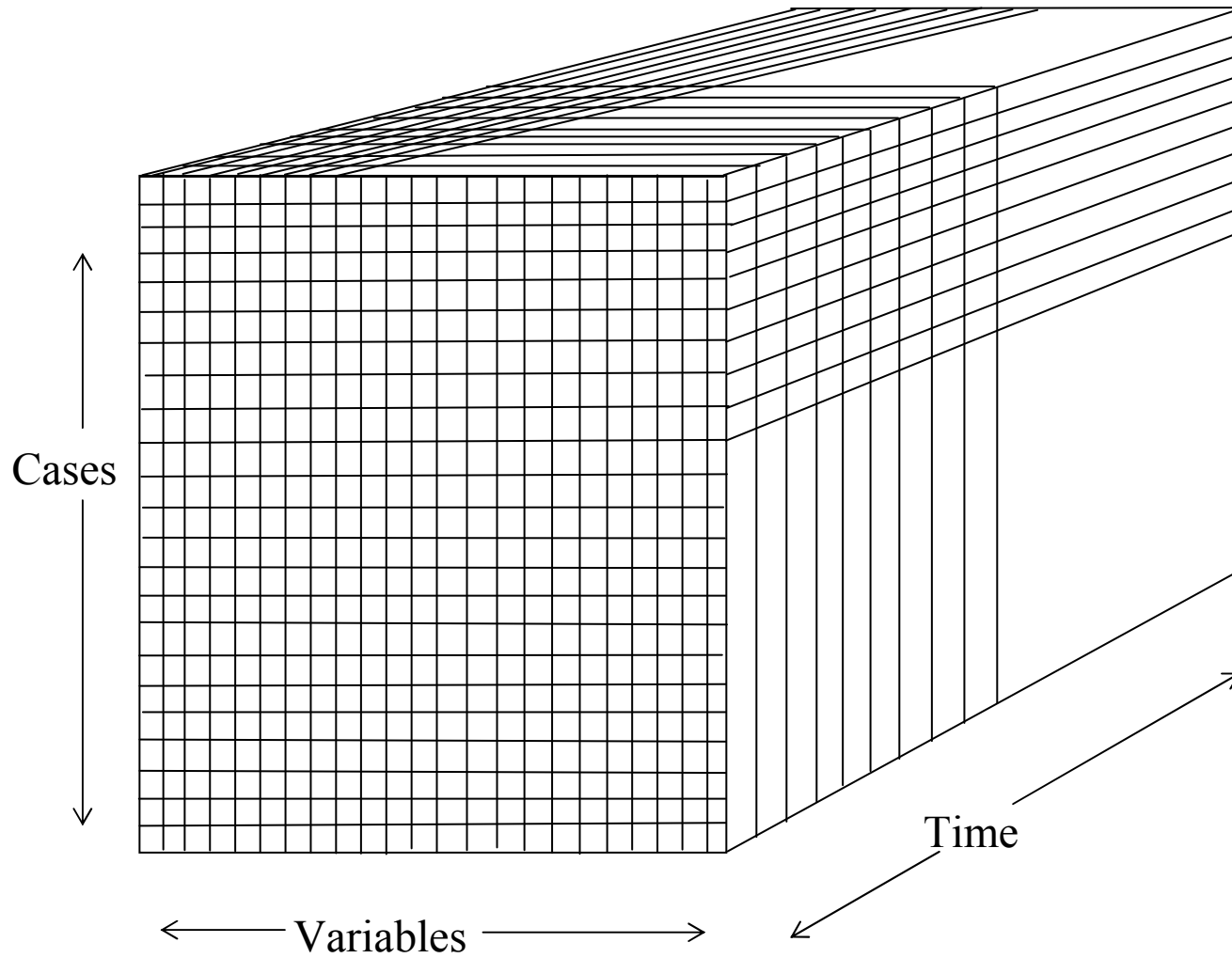
Tomorrow morning:

We'll do two general things: (1) explore autoregressive distributed lag (ADL) and error correction model (ECM) approaches to analyzing time serial data. Then we'll introduce TSCS models. Our most demanding session.

Tomorrow afternoon:

We'll build on and extend what we did in the second half of the morning session to focus on advanced topics in TSCS modeling. Then we'll have some Stata time address questions.

# The Data Box



The full data box—all three dimensions

For TSCS or cross-sectional time-series (CSTS) analysis.  
(multiple cases and multiple variables over multiple time points)

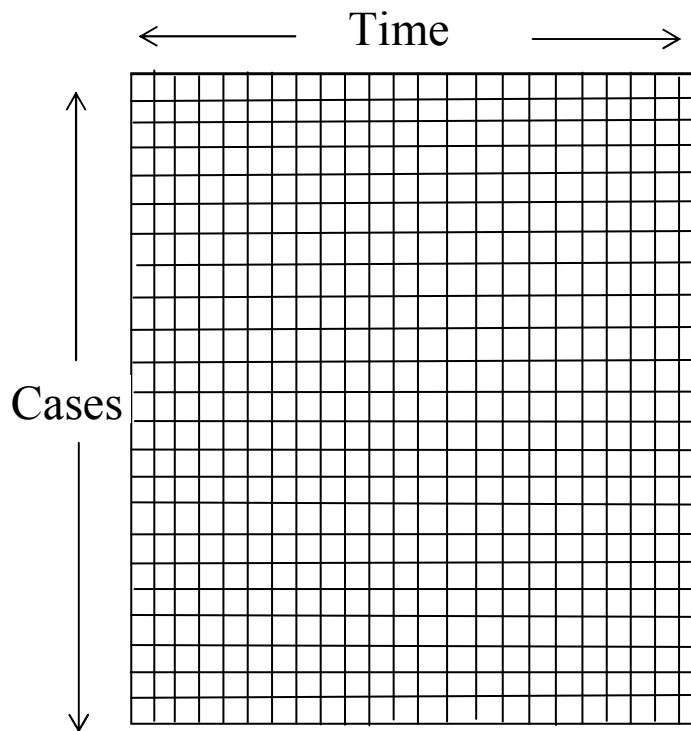
More on this tomorrow.

For now, consider two-dimensional representations of  
the data box.

# Two or three approaches to two dimensions

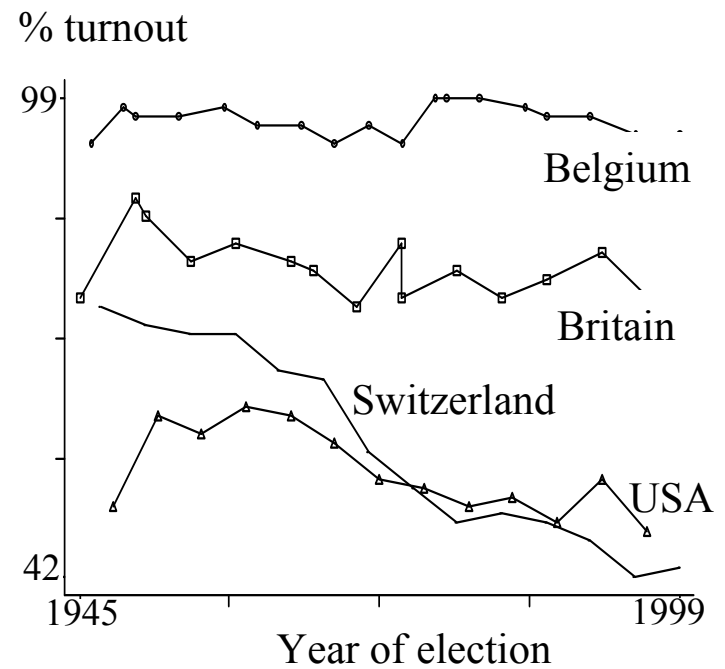
(a) Trend analysis (a single variable – or average)

Data matrix (unusual)



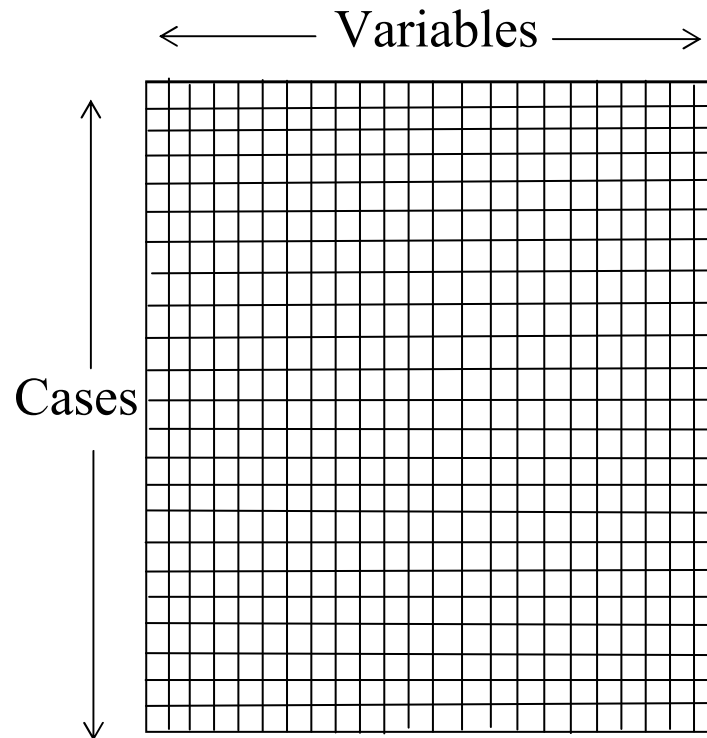
Exemplary chart (often seen)

Turnout over time for four countries

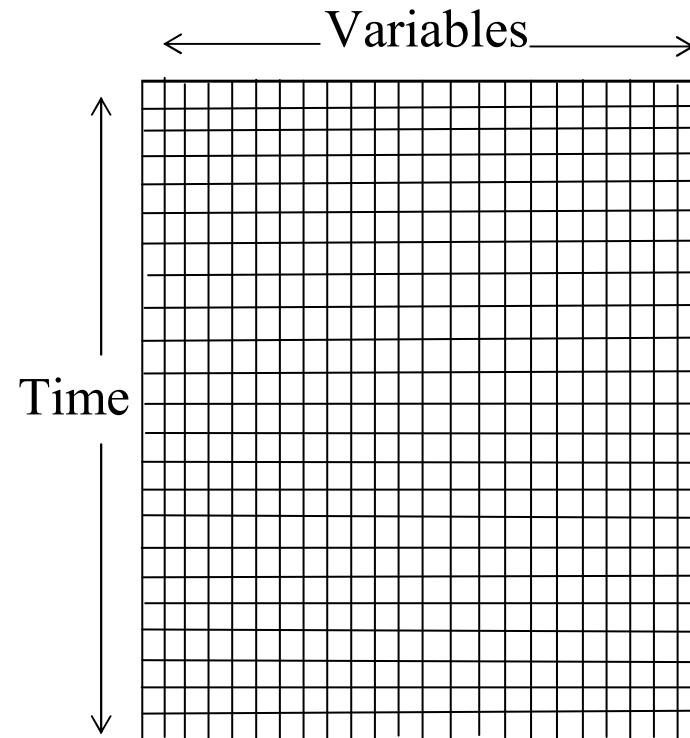


## (b) More customary representations of the data box

For cross-sectional analysis  
(a single time-point – or average over time)



For classic time-series  
(a single case – or average case)



Of course, both representations can be extended in hierarchical fashion to represent units embedded within higher-level units (countries, schools, or whatever).

Time series analysis To effectively conduct TSCS analysis, one needs an appreciation for time series. Some basics...

A time series:  $N$  time-ordered observations of a process.

Ideally, interval level measurement  
Time separating successive observations should  
be constant.

Minor violations are acceptable.

By this definition, a time series is discrete. Of course, it could be a measure of some underlying continuous process.

Measurement is a big deal in time series because it can change over time. The companies on stock markets have changed. The things we consume have changed. The meaning of unemployment has changed. Does crime mean the same thing today that it did fifty years ago? Is the crime rate higher or lower?

Differences between time-series and conventional cross-sectional analysis. In time series:

- 1) (Typically) fewer cases
- 2) Lack of independence between cases, i.e., “dependent data”
- 3) Greater sensitivity to model specification
- 4) Greater sensitivity to “influential cases”
- 5) Greater premium on theory
  - Theory is what tells us how to correctly specify a model, at least to begin with.
  - Theory also may tell us whether an influential case is an aberration or a critical exemplar.
- 6) We also need methodological theory and practice.

7) Change and persistence. In contrast with cross-sectional analysis, we can actually assess (a) whether and the extent to which a variable changes and (b) whether and the extent to which it persists. These are things we theorize about and really want to know. Consider the following (autoregressive) equation for variable  $Y$ :

$$Y_t = \rho Y_{t-1} + \varepsilon_t.$$

In this equation, the new shocks to  $Y$  are represented by the disturbance term  $\varepsilon_t$ . The persistence is represented by the parameter  $\rho$ —if less than 1, shocks decay, if equal to 1, they persist. If greater than 1, the process explodes. (The parameter  $\rho$  can be less than 0, with similar, though oscillating, consequences.) Note that the actual change in  $Y$  between two points in time will be a function of the variance of the disturbance term ( $\varepsilon_t$ ) and the parameter  $\rho$ , as well as the intercept ( $\alpha$ ) if there is one. This equation is the basis for much of what we will be doing in this workshop, as we will see.

## From cross-sectional to time-serial analysis.

The cross-sectional model—different units measured at the same point in time; multiple units, a single observation of each. This is useful for answering many research questions. For example, why is one person more likely to vote than another person? Why is the welfare regime in one country more generous than that in another country? Why are some countries more peaceful than others?

$$Y_i = a_0 + B_1 X_{1i} + B_2 X_{2i} + \dots + e_i$$

The time series model—the same unit measured at different points in time; a single unit, multiple observations.

$$Y_t = a_0 + B_1 X_{1t} + B_2 X_{2t} + \dots + e_t$$

Why bother?

(1) Our research interests often are explicitly time serial. For example, (why) does a person's propensity to vote change over

time? Why do welfare regimes change? Why do some countries become more peaceful?

(2) Time allows us more causal leverage.

(a) We can see whether change in an independent variable produces a change in the dependent variable. E.g., Is a voter more likely to vote when her level of education increases?

(b) We can partly mitigate issues of endogeneity via lags. (This depends a lot on measurement and the timing of causality. Sometimes it works out well. Consider opinion-policy dynamics.)

(3) As we already have discussed, we can assess (change and) persistence.

Of course, we can do both cross-sectional and time-serial analysis, i.e., “pooled” or “panel”—both  $i$ 's and  $t$ 's—and we will, just not today. And we can add additional levels  $j$ .

For today, a *single* time series:

A unit, e.g., an individual, aggregation of individuals, a country  
N time-ordered observations  
Time gaps separating successive observations are constant  
Traditionally, interval-level measurement

Time series are discrete but they may reflect some underlying continuous process.

So, how to analyze a time series?

**Econometricians** began by treating time series like cross-sections and applied OLS. Deal with violations of assumptions, e.g., serially correlated errors, heteroskedasticity. (NOTE: These are the Gauss-Markov assumptions.)

E.g., for autocorrelation, the Durbin-Watson (d) statistic. (If the model includes a lagged DV, use Durbin's h or m test.)  
Breusch-Godfrey for a generalized test.

Estimate model with OLS, conduct DW test or Breusch-Godfrey, if significant estimate rho and transform variables accordingly, and then re-estimate the original model with OLS. Called “generalized least squares” or GLS. Still pretty common in many circumstances, particularly where we have small N’s and dynamics are not of primary interest or concern.

Note that autocorrelation can have different sources, for instance, inertia, delayed influence of shocks, data smoothing (e.g., moving averages), specification error.

In the face of autocorrelation (or heteroskedasticity), OLS estimates are not unbiased (mean estimate is true value) or inconsistent (sample size does not matter) but they are inefficient (not best, optimal).

Positive. Very common and consequential. Understates standard errors.

Negative. Less common and less consequential.

NOTE: With time series analysis we often will find some serial correlation—it thus is important to keep in mind that *a small amount of serial correlation causes few problems* but that a large amount can cause big problems.

In the face of autocorrelation or heteroskedasticity, OLS estimates are not unbiased (mean estimate is true value) or inconsistent (sample size does not matter) but inefficient (not best, optimal)—the coefficients track the (error-laden) data too much. NOTE: With time series analysis we often will find some serial correlation—it thus is important to keep in mind that a small amount of serial correlation causes few problems but that a large amount can cause big problems.

Detection:

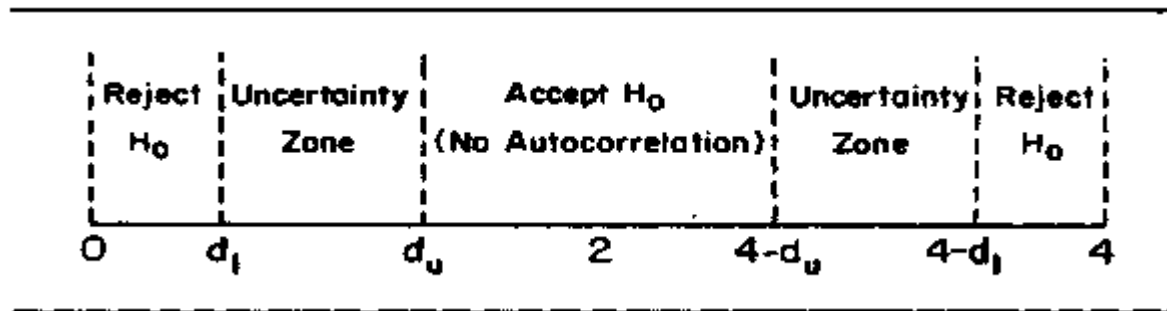
*Visualization*—a big deal in time series.

Statistical tests.

The Durbin Watson  $d$  statistic:

$$d = \frac{\sum_t (x_t - x_{t-1})^2}{\sum_t x_t^2}$$

A brilliant idea but the statistic runs from 0 to 4. No autocorrelation is 2.0—in smaller samples just below 2.0—positive if closer to 0, negative if closer to 4. (Why not create a stat centered at 0?) In any event, you need tables of critical values easily found at the end of stat books or online.



In Stata,

**.estat dwatson**

With a lagged dependent variable, the DW h statistic.

In Stata, durbina.

Across multiple lags. Breusch-Godfrey test, in Stata:

**.estat bgodfrey.**

A Lagrange Multiplier, LM, test, where regress residual on X variables and lagged residual(s) and test significance of latter.

If you have significant autocorrelation, what to do? Not OLS but GLS:

Cochrane-Orcutt (estimated or feasible GLS).

Ingenious.

1. Regress Y on X.
2. Generate residuals, i.e.,  $Y - a - bX$ .
3. Regress current residuals on lagged residuals to produce estimate of rho ( $\rho$ ).
4. Transform Y and X using  $\rho$ — $Y_t^* = Y_t - \rho Y_{t-1}$  and  $X_t^* = X_t - \rho X_{t-1}$ .
5. Regress  $Y^*$  on  $X^*$ .

Return to 2 and continue to cycle until convergence.

Prais-Winsten, reintroduces first case with imputation.

Looking ahead: C-O transformation is much like the autoregressive distributed lag (ADL) model:

$$Y_t = a_0 + B_1 Y_{t-1} + B_2 X_{t-1} + e_t$$

It is *autoregressive* because of lagged Y, *distributed lag* because of lagged X.

Heteroskedasticity detection:

Visualization.

Testing.

ARCH (AutoRegressive Conditional Heteroskedasticity)

Robert Engle. (Big TS name along with Cliver Granger and David Hendry.)

Regresses squared residuals on lag(s) of squared residuals.

A big deal in financial econometrics, less so in pol sci (and other parts of econ too).

## A (very brief) Introduction to ARIMA.

Along came **Box and Jenkins** and ARIMA modeling. ARIMA stands for AutoRegressive, Integrated, Moving Average.

Represented as ARIMA (p,d,q). The goal: to reduce a variable to “white noise,” i.e., a stochastic series, and then—if you want to—focus on how it relates to other variables. (Stochastic is a pure random variable, i.e., not deterministic.) ARIMA is designed to identify and eliminate systematic sources of error, such as an AR and MA process, and “difference” a variable to remove nonstationarity.

### *Stationary vs Nonstationary variables:*

A variable is stationary if the mean and variance are constant, nonstationary if the mean and variance change over time. **See figures.**

### *On Trend and Drift:*

Trend changes the level of a series. The series thus is nonstationary in levels. A drifting series is not. Time series typically drift up and down—this is stochastic. Many also trend—this is deterministic.

$$Y_t = a_0 + B_1 \text{Time}_t + B_2 Y_{t-1} \dots + e_t$$

Trend is evident from coefficient  $B_1$ . De-trending can make a variable stationary. In other cases, a variable may need to be differenced, which means subtracting  $Y_{t-1}$  from  $Y_t$  to make it stationary. It depends on the coefficient  $B_2$  and the nature of the process. De-trending a variable that is nonstationary but does not have a deterministic trend, will not make it stationary; it needs to be differenced— one needs to subtract  $Y_{t-1}$  from  $Y_t$ . More on this in a bit.

### *Autoregression:*

$$Y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + e_t$$

AR(1) only  $\phi_1$  (phi) is significant; AR(2)  $\phi_1$  and  $\phi_2$  significant. An AR process describes geometric decay in the effects of shocks. As such some portion lasts forever. AR processes are fairly common in practice. **See Figure.**

The variable equilibrates toward the equilibrium, which is represented by the intercept ( $a$ ) and  $\phi_1$ . Specifically, the equilibrium equals:  $a/(1 - \phi_1)$ . Consider an example where  $a = 10$  and  $\phi_1 = .8$ . The equilibrium is  $10/(1 - .8) = 10/.2 = 50$ .

Seemingly less common is the *moving average*:

$$Y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots$$

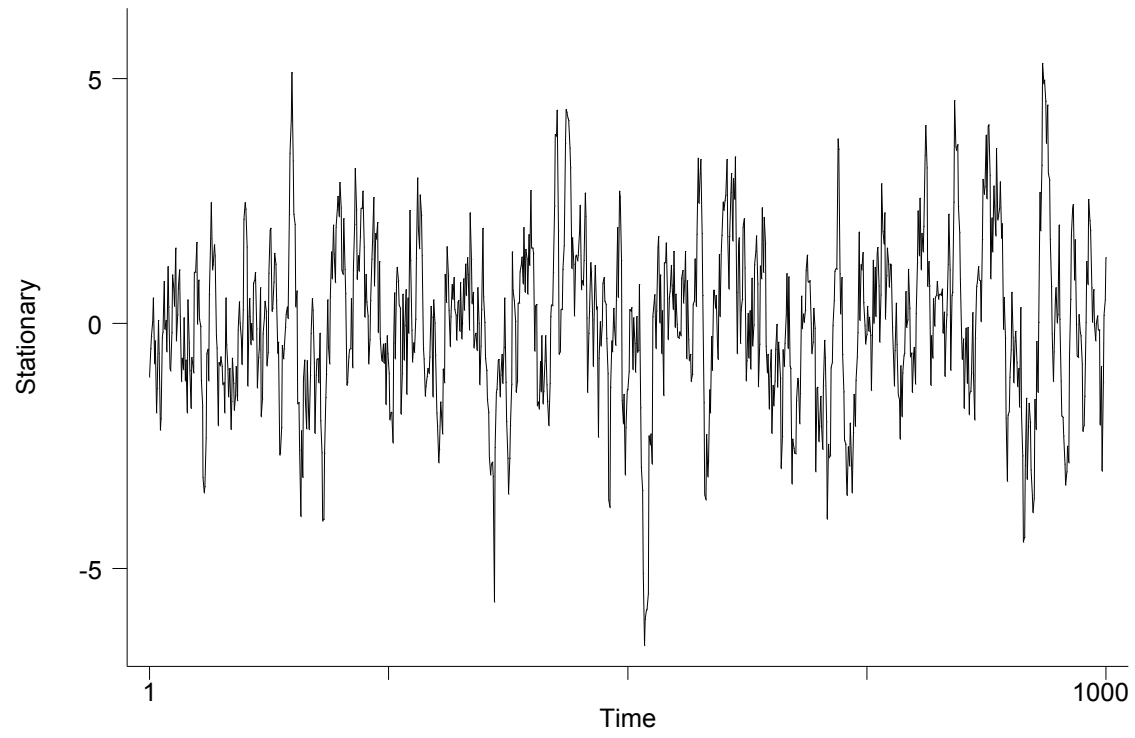


Figure 1: An autoregressive time series.

(Often expressed using subtraction.) MA(1) only  $\theta_1$  (theta) is significant. In contrast with an AR process, in the MA process shocks have finite effects—effects come and then stay for a period or two and then disappear.

If AR and MA, ARMA. E.g., ARMA (1,1) contains an AR(1) process and a MA(1) process.

AR and MA processes are *stationary*. That is, they have a constant mean and variance. **See figure again.** Also covariance-stationary, which refers to the constancy of autocorrelations at different points in the series, i.e., that the correlation between  $Y$  at times  $t$  and  $t-1$  is the same as the correlation between  $Y$  at times  $t-s$  and  $t-s-1$ .

How to tell? One way is the autocorrelation function (ACF), or correlogram. ACFs correlate each observation with its lags. So if  $t$  is high, where is  $t-1$ , and then  $t-2$  and so on. **Example.**

The partial autocorrelation function (PACF). PACF show the correlation between observation  $t$  and  $t-2$ , controlling for the

relationship between  $t$  and  $t-2$ , and so on. **Example.** The PACF does help us identify the ARIMA form but this mostly is discernible from the ACF. The PACF is critical for identifying higher order processes, e.g., AR(2).

What to do if we identify an AR or MA process? Estimation. Remember the goal: white noise residuals. It's all about the residuals. Thus estimate and assess residuals.

Now, an *integrated* process. This is where  $\phi_1 = 1$ . (The process integrates or sums shocks to  $Y$ .) An integrated process is *nonstationary*—shocks do not decay but cumulate and last indefinitely. The mean and variance are not constant. Indeed, the variance increases over time. **See figure 2 again.** Put differently, an integrated variable doesn't change unless something happens to change it. This is not true of an AR process or an MA process, where the variable changes because whether something new happens, that is, because old shocks decay.

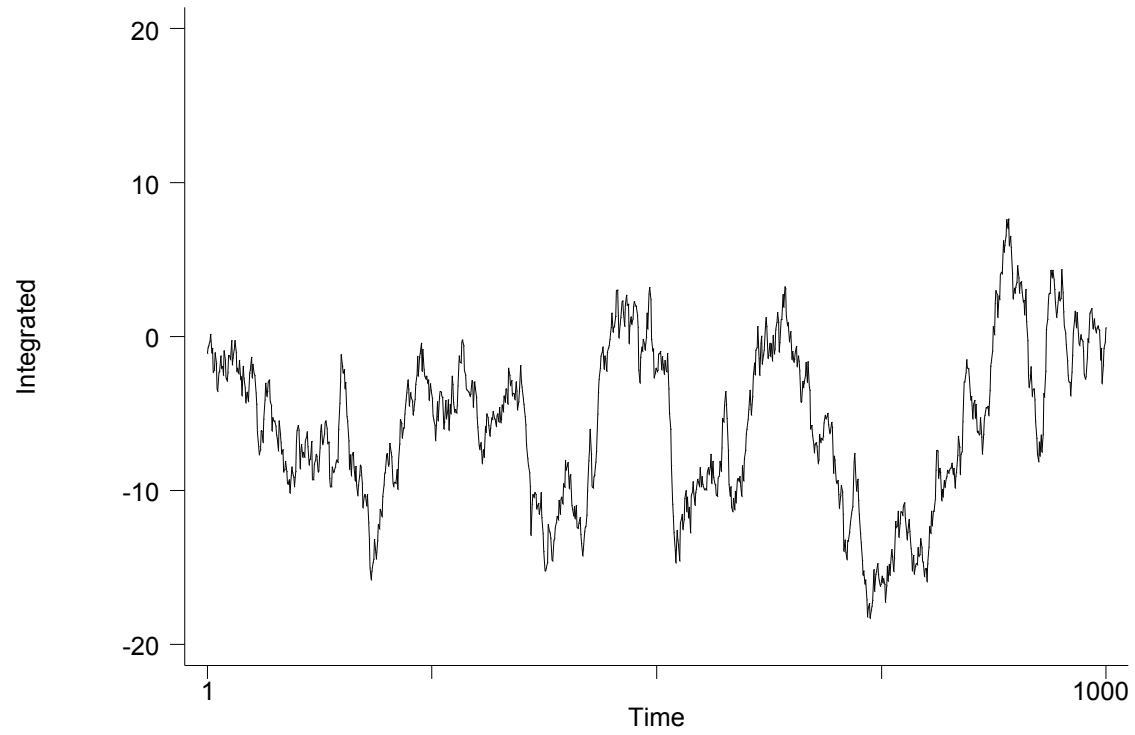


Figure 2: An integrated time series.

I(1) is first order integration. Differencing one time makes the series stationary. If a second differencing is required, I(2). Take the series of squared integers: 1, 4, 9, 16, 25, 36 ... First differencing gives: 3, 5, 7, 9, 11, .... Still nonstationary. Second differencing gives: 2, 2, 2, 2, ... I(3) is very rare. Even I(2) is not all that common in political science research.

For detection, the ACF (and PACF) once again. There also are other more formal techniques, but these were developed later on, and there will be more on this later this session and tomorrow.

Nonstationarity is a big problem for analysis but didn't trouble econometricians, who mostly assumed that variables were stationary or *trend stationary*. OLS or GLS regressions of nonstationary variables can produce spurious results. Consider regressing two trending variables.

Then it became clear that ARIMA models were outperforming econometric forecasting models, which led to change.

Later it became clear that many variables being modelled actually were not stationary and this could lead to spurious results! *Thus, determining whether or not a variable is stationary is critical.*

On “transfer functions.” Some of you are or will be interested in the effects of particular events. In the Box-Jenkins ARIMA context this is known as *transfer function* modeling. The words from engineering. In our world, more commonly known as *impact assessment* or in the modern day *intervention analysis*. Useful for some of you studying the effects of exogenous shocks on political variables like government approval, vote intention, turnout, etc.

It may be the case, that as you look at your time series, or more importantly, think about the process you are modelling, that you view some event as changing the level of the series. This change could be temporary, permanent or a combination of the two.

Events and their effects:

- Abrupt, permanent (or “jump,” or “bump”?):
- Gradual, permanent:
- Abrupt, temporary (or “Pulse,” or “bounce”?)

There are other types, e.g., gradual, temporary. These are trickier to model but other approaches, like vector autoregression (VAR), can help here (and more generally).

An event. Something happens. Off and on: 00000111111.  
Possibly off again: 000000111110000. The a priori expectation is critical, even after the fact. That is, the technology is good for “confirmatory” analysis, not “exploratory” analysis. Why? Easy to “find” event effects that aren’t real, e.g., because of sampling error.

The basics of modeling event effects...

$$Y_t = f(I_t) + e_t$$

0 order models.

$$f(I_t) = \omega_0 I_t.$$

$$Y_{t-1} = \omega_0 I_t + e_t.$$

NOTE: No reference to  $Y_{t-1}$ .

That's abrupt, permanent.

What about gradual, permanent?

One way—a reformulated I variable, e.g., 0 0 0 .25 .50 .75 1.0  
1.0 1.0 1.0

May not be quite right. What if nonlinear? A more general modeling strategy.

1<sup>st</sup> order models.

$$Y_t = \phi Y_{t-1} + \omega_0 I_t + e_t.$$

Think through the effect over time.

What if  $\phi=1$ ?

If  $1 > \phi > 0$ ?

$\phi=0$ ?

Choosing among the models? Statistical tests may help but theory is important too.

Thus far, all permanent effects. What about temporary effects?

Differencing I: 0000001111111 becomes 0000001000000.

NOTE: Differenced  $Y_t = Y_t - Y_{t-1}$

$$Y_t = \phi Y_{t-1} + \omega_0 \Delta I_t + e_t.$$

Think through effects...

What if  $\varphi=1$ ?

What if  $1>\varphi>0$ ?

$\varphi=0$ ?

The great thing about the 1<sup>st</sup> order transfer function is that it can model the abrupt, permanent step, the gradual permanent shift, and abrupt temporary change, including the pulse.

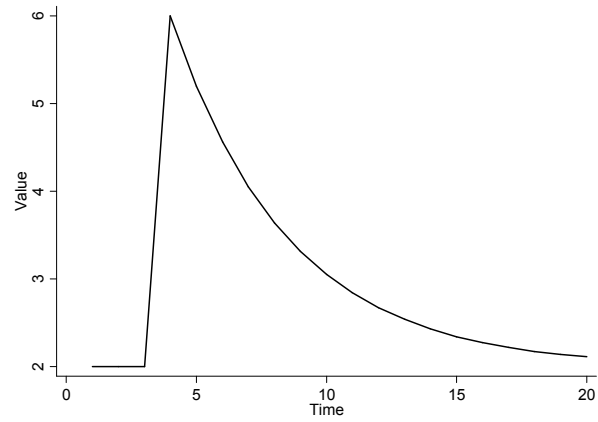
Higher order models.

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \omega_0 I_t + e_t.$$

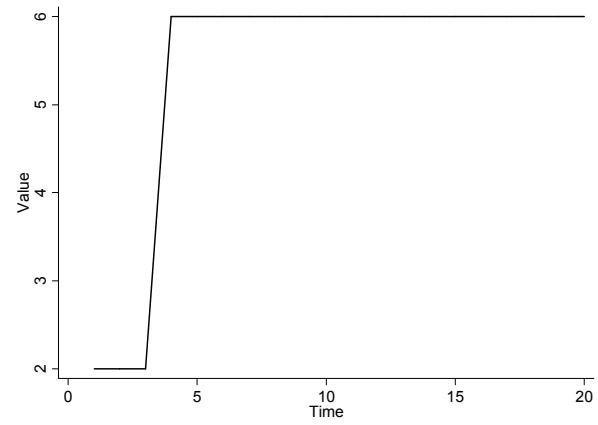
More important—compound effects.

Perhaps two different types of effects. A “combined” process!  
Perhaps an event has a lasting effect but also a temporary one.

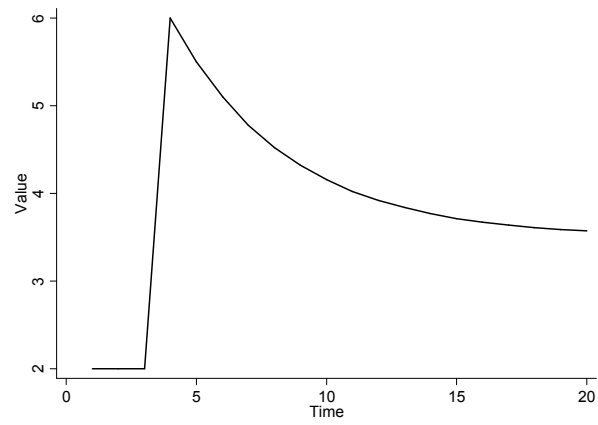
From work with Bob Erikson on the “timeline” of election campaigns.



**Figure 3a: A Bounce**



**Figure 3b: A Bump**



**Figure 3c: A Hybrid Effect**

## **Vector autoregression (VAR)**

VAR began as structural econometric times series approach (SEMTSA). Very complicated.

In, VAR, one models the DV as a function of its own lagged values and lagged values of all of the other variables in the system, where the number of lags is to be determined. It's really ARIMA without the MA.

VAR is atheoretical but considered useful for some purposes, especially forecasting, which need not be guided by theory (and where endogeneity matters less). Many view VAR models as complements to, not substitutes for structural modeling. Plus the VAR structure helps analyzing nonstationary time series data, as we will see.

The autoregressive distributed lag (ADL), a special case of VAR:

$$Y_t = a_0 + B_1 Y_{t-1} + B_2 X_{t-1} + e_t$$

The first order ADL, along with the error correction model (ECM), is a very effective general modelling strategy, as we'll see tomorrow morning.

*On "Granger causality"*

Does X cause Y? Or does Y cause X? Clive Granger posited that if past values of X predict current values of Y beyond what we predict based on past values of Y, then X "Granger causes" Y. The null hypothesis is no causality. Given a two-equation system, there are four possibilities:

1. X does not cause Y and Y does not cause X.
2. X causes Y but Y does not cause X.
3. Y causes X but X does not cause Y.
4. X causes Y and Y causes X.

Of pretty obvious importance for identification. Consider what would happen if you think X causes Y and model it that way when X causes Y and Y causes X!

Testing with an F test the restriction that the coefficients for lagged values of the “other” endogenous variable(s) all equal 0.

For example:

$$Y_t = a_0 + B_1 Y_{t-1} + B_2 Y_{t-2} + B_3 Y_{t-1} + B_4 X_{t-1} + B_5 X_{t-2} + B_5 X_{t-2} + e_t$$

Test whether the contribution of the X's is significant—in Stata:

**.test B<sub>4</sub> B<sub>5</sub> B<sub>6</sub>**

If significant, X Granger causes Y. *Important:* Granger tests require stationary data.

## Error correction models

An advantage of ARIMA and VAR approaches is that they are dynamic. The old econometric approach was largely static. While flexible in their specification of the dynamic structure of the time series, ARIMA and VAR ignored the role of long-run equilibria. Enter the error correction model, or ECM.

$$\Delta Y_t = a_0 + B_1 \Delta X_t + B_2 (Y_{t-1} - B_3 X_{t-1}) + e_t,$$

where  $B_3$  is the coefficient relating  $Y_t$  and  $X_t$  and  $B_2 < 0$ . Like ARIMA, a differenced model. In contrast with ARIMA, long run information provided by the level data is explicitly modelled.

Notice the mixing of levels of differences. What if level variables are non-stationary? Interestingly, the ECM was invented precisely to model the effects of nonstationary variables.

## Unit roots (and testing for them)

We've talked a lot about nonstationary series. A nonstationary series is "integrated." Shocks to these series cumulate. The "order" of integration is the number of times (d) a variable needs to be differenced to become stationary. I(0) is stationary. I(1) means that a series needs to be differenced once. These are the most common but I(2) is not unheard of. The differencing characterization comes from Box-Jenkins approach, which assumed that differencing will make a nonstationary variable stationary. (Assumption seems based on empirical regularity.)

The simplest example of an I(1) is otherwise known as a "random walk." Recall the equation:

$$Y_t = \gamma y_{t-1} + e_t$$

In a random walk,  $\gamma = 1$ . ( $\gamma < 1$  indicates an I(0) process;  $\gamma > 1$ , explosive one)

Thus, in a random walk,

$$\Delta Y_t = e_t.$$

Dickey-Fuller test

$$\Delta Y_t = (\gamma - 1) y_{t-1} + e_t$$

Test whether  $(\gamma - 1)$  is less than 0. If stationary, it is ( $\gamma < 1$  and so  $\gamma - 1 < 0$ ); if nonstationary, it is not ( $\gamma = 1$  and so  $\gamma - 1 = 0$ ).

Nonstandard (MacKinnon) t-values.

On “fractional integration.” Mixing of stationary processes can produce a process that looks integrated.

## **Unit Roots and Cointegration and ECMs Again**

Why is diagnosis of unit roots important?

Obviously important for ARIMA modeling.

Also important more generally, as we cannot regress nonstationary variables on each other and believe the results.

What about the ECM, which contains integrated variables? Notice that they enter together via one error correction component (ECC). This is the focus of Granger's Nobel-winning insight: That the linear combination of two integrated variables could be stationary. The variables in such a relationship are said to be cointegrated. Indeed, Granger showed that cointegrated variables must have an ECM representation.

To assess cointegration, diagnose integration of the two variables, regress one on the other (called the "cointegrating regression"), and then test for a unit root in the residuals. If stationary, then include the lagged residuals from the cointegrating regression as the ECC in the ECM.

$$\Delta Y_t = a_0 + B_1 \Delta X_t + B_2 (Y_{t-1} - B_3 X_{t-1}) + e_t,$$

In small samples, also estimate a full ECM, which means directly estimating the effects of the lagged values of the integrated variables.

$$\Delta Y_t = a_0 + B_1 \Delta X_t + B_2 Y_{t-1} + B_3 X_{t-1} + e_t,$$

Not perfect. There may be concerns about endogeneity. It also may be that there is more than one cointegrating relationship. The latter has led to increasingly more general formulations, called vector error correction models (VECM).

### **Practicum: Diagnosing time series characteristics**

The data sets we will be using are quite small, so we will set the amount of memory allocated to Stata at about 10 MBs. You do this with the **set memory** command.

**.set memory 10m**

Some of the computations we will be doing today, however, are a bit tricky and will need large matrices within Stata to complete. Thus we set the largest “mat size” or matrix size to 150 with the **set matsize** command.

```
.set matsize 150
```

Datasets can be opened using either the pull-down file menu or the use command.

For example, let’s open the simulation database we will be using to begin with:

```
.use Session1.dta
```

### Time Series Specific Commands

After loading a dataset, recall that the first thing we need to do is tell Stata that we have time series data. This is done with the **tsset** command and the variable that marks the “timing” of the observations.

**.tsset x**

Where x is the time variable for the simulation data.

Often time we are interested in lagging or differencing a variable. In Stata, these are done with:

Lagging a variable by one unit: **L.var** = var[t-1]

Lagging a variable by two units: **L2.var** = var[t-2]

Differencing a variable: **D.var** = var[t]-var[t-1]

Twice difference: **D2.var**= (var[t]-var[t-1]) - (var[t-1] - var[t-2])

Seasonal difference: **SX.var** = var[t]-var[t-X]; where X is the lag

It is also useful to plot a series over time, to look at what you are dealing with.

**.tsline random if x<=400**

→ Now also might be a good time to open a log file.

## **.log using ELECDEM**

If you have more than one series you can look at them together

**.tsline random cum if x<=400**

This is standard “exploratory” practice in the analysis of time series, in contrast with what we do with cross sections.

## **Identification**

*AR and MA processes*

One of the significant distinguishing characteristics of time series data is that it can be comprised of “dynamics”. By this, we mean that observations are related to each other across time. For example, spending today affects spending tomorrow, or a large unexpected increase in car sales in one month might lead to a

decrease in car sales the next month. We can model these dynamics using autoregressive and moving average terms:

$$Y_t = \mu + \phi Y_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t$$

Where  $\mu$  is a constant,  $\phi$  represents an autoregressive coefficient and  $\theta$  a moving average coefficient (this is an arma(1,1) model).

In order to determine whether a series has an autoregressive and/or moving average component(s), we have several tools. These include simple plots of the series, autocorrelation functions (acf), partial autocorrelation functions (pacf) and white noise tests (Q's). Recall that the ACF shows correlations between each observation with its lags. A PACF is the partial autocorrelation function, the relationship between observation  $t$  and earlier values, e.g., the relationship between observations at times  $t$  and  $t-2$  controlling for the relationship between  $t$  and  $t-1$ .

In Stata, you can create an acf with the command:

**.ac var**

and a pacf with the command:

**.pac var**

the command **corrgram**, gives us both the estimated acfs and the pacfs along with the statistical tests of whether those correlations are zero.

The **Session1.dta** file includes several time series that we can attempt to identify the characteristics of the time series. The variables are as follows:

*x*: time variable

*random*: each observation is a random shock

*cum*: the cumulative sum of all previous random shocks

*ara*:  $y_t = \text{random}(t) + .8y(t-1)$

*arb*:  $y_t = \text{random}(t) + .5y(t-1)$

*comb1*: combination of cum and ara

*comb2*: combination of cum and arb

*ma*:  $y_t = \text{random}(t) - .5\text{random}(t-1)$

*ma2*:  $y_t = \text{random}(t) + .5\text{random}(t-1)$   
*comb3*: combination of cum and ma

The first step in identification is to look at its acf and pacf. Let's look at random:

**.tsline random**  
**.ac random**  
**.pac random**  
**.corrgram random, lags(20)**

There do not appear to be any significant dynamics. Thus we call this series white noise. Notice that in the acf and pacf there are some spikes that fall outside of the confidence intervals. Yet, we know, since we created the data, that there is no systematic correlation at these lags. Why then do we get the spikes?  
Random chance.

We are testing whether a correlation (or partial) is statistically different from zero. Sometimes, we are going to get big spikes in the sample, even though the population parameter is zero. Out of

100 spikes, if we drawing 95% confidence intervals, 4 or 5 are going to appear significant.

Now look at the **ara** series. We know this was generated by a process with a dynamic, where  $y_t = .8y_{t-1}$ . Think about what the ACF and PACF should look like.

The ACF for an ar(1) dampens exponentially, and the PACF should have one big spike.

Take a look,

**.ac ara**

**.pac ara**

Now, we can see the first spikes dampen on the acf and there is one big spike (at .8) on the pacf.

Another AR process,

**.ac arb**

## **.pac arb**

You can do this using the other variables in the data set as well.

While we have focused on first order processes we can have second order ones too, but rarer. Recall that the pacf's are critical for distinguishing these.

### *Integrated Process*

AR (and MA) processes describe stationary time series. A stationary series is one in which equilibrium value or long-term mean is fixed. Shocks can set the series off equilibrium but then the series tends move back toward the equilibrium. That is, all effects on the variable decay. Sometimes we can see this pretty clearly from a simple graph of a series. Try the autoregressive process **arb** from above.

## **.tsline arb**

Note: For more info on using **graph** in STATA, type **help graph**.

Here we can see that the series tends to hover around a constant mean—recall from above that the equilibrium =  $a/(1 - \phi_1)$ . The hovering is characteristic of a *stationary* process.

Now let us consider an integrated time series. An integrated series is one in which the equilibrium value changes over time. Shocks to the series do not decay but persist forever. As effects cumulate, the series then tends to wander up and down over time. Consider the variable **cum**, which is integrated of order one or I(1). (Recall that the order of integration indicates the number of “differencing”s required to make a series stationary.) Such a series also is known as a “random walk” or “unit root” process. At any point in time, the variable is simply the sum of all past random shocks.

**.tsline cum**

Notice that the series does not hover around a constant mean but wanders over time. This is characteristic of a *nonstationary* series.

Now let us take a look at the acf.

### **.ac cum**

Here we can see that the autocorrelations decay only very slowly over time, that is, not in the geometric fashion associated with an autoregressive process. When the series takes on a large positive value at time  $t$  it is likely to take on a large positive value at times  $t+1$ ,  $t+2$ ,  $t+3$ ....

### **.pac cum**

As for the ar(1) process, we observe a single spike and then the autocorrelations are 0. This is basic identification. Let's (begin to) see whether we are right.

### **.arima cum, arima(0,1,0)**

Then predict residuals:

**.predict resid, residuals**

Then diagnose:

**.ac resid**

**.pac resid**

The arima(0,1,0) procedure simply differences the variable, in this case “cum.” Differencing an I(1) variable produces a stationary series, so we also can just do it by hand:

**.tsline D.cum**

**.ac D.cum**

**.pac D.cum**

OK, this has been a little taste of time series diagnosis. Tonight, I'd like you to do a little more with a data set that I will send right now. There are three variables: var1, var2, and var3.

Tomorrow:

The morning:

- Briefly revisit time series diagnostics, especially for combined time, e.g., I + AR.
- General modelling strategies: ADL and ECM.
- An introduction to CSTS.
- A CSTS practicum

The afternoon:

- More advanced CSTS modelling.
- Practicum

In the meantime, besides the homework, do feel free to begin working with your own time series data. I am happy to talk with you about your results (or other work) tonight or tomorrow.

Thank you.