

Time series and pooled analysis

ELECDEM

Istanbul, session #5

**Christopher Wlezien**

with special thanks to **Mark Franklin**

Homework...

To begin with, recall where we were last time...

## **On the identification of time series processes**

*Autoregression:*

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

*Moving Average process:*

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

AR and MA process are *stationary*. They have a constant mean and variance. (They also are 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  $t-s-1$ .)

*Integration.* Where the process integrates or sums shocks to  $Y$ . If first order,  $\phi=1$ . 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. 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 whether or not something new happens, that is, because old shocks decay (if the variable is not in equilibrium.)

Now, after identification, estimation and diagnosis, we can try a more general “meta” diagnosis. Here we're interested in rejecting other alternative ARIMA processes, e.g., AR(2), MA(3). This is an important step as our methods for identification and diagnosis are limited, but now let's consider “combined” processes.

*Combined processes:*

For instance, I plus AR (or MA). Seemingly common. That is, we expect a portion of shocks to persist indefinitely and the rest to dissipate. We should observe a stationary series around a nonstationary equilibrium. Statistical theory tells us that such “combined” processes, for lack of a better word, are integrated in the long run. In theory, the Integrated process dominates, but this depends on the rate of decay in the AR process and the ratio between the I and AR components.

Consider the combined series (**comb1**), which is the aggregation of I and AR processes. Specifically, **comb1** is  $.5 * cum + .5 * ara$ , where **ara** is an AR(1) process with parameter = .8. Now, we want to assess its time-serial characteristics.

Let's look at the autocorrelations.

**.corrgram comb1, lags(20)**

The ACF implies an integrated process. The PACF? More complicated. Additional spikes. What are these things? AR? Or MA? What is the identification?

Telling, but there is a more definite test. A random walk can be represented using a simple AR(1) model:

$$Y_t = a + b Y_{t-1} + e_t .$$

A unit root process is one where the parameter  $b$  in this equation equals 1. The value of  $Y$  at any point in time is exactly equal to the value of  $Y$  in the last time period plus the new shock,  $e_t$ . If it is less than 1, the series is autoregressive. The value of  $Y$  is some fraction of the previous value together with the new shock. (If the coefficient is greater than 1, the series is explosive, whereby past shocks have a bigger impact than current ones, which seemingly is not common. Regardless, it is a subject for another day.) Given the importance of distinguishing stationary and integrated

processes, a huge literature has mounted focusing on unit root testing.

Seems easy to tell whether  $b = 1$ ; just regress  $Y_t$  on  $Y_{t-1}$  and test against null hypothesis  $H_0: b=1$ . Given that stat package defaults are to test null of 0, one may want to estimate the following equation instead:

$$\begin{aligned}\Delta Y_t &= a + (b-1) Y_{t-1} + e_t . \\ &= a + c Y_{t-1} + e_t .\end{aligned}$$

The test then is against the null of  $c = 0$ . If not rejected, we infer that  $b = 1$ , by definition. This is the standard way to conduct what is called the Dickey-Fuller test. The test is not entirely straightforward, however. The problem is that critical values for significance are non-standard, which leads us to over-reject the null hypothesis. Thus, the D-F test involves estimating the above regression and then calculating a standard t but assessing significance using not so standard (MacKinnon) critical values.

**.dfuller comb1, regress**  
**.dfuller comb1, regress lags(20) ...**  
**.dfuller comb1, regress lags(10) trend**

From this analysis we can tell that the series is integrated. It is not a pure series, however, as we know by construction that fully half of the shocks to the series decay over time. This is important, and it puts a very fine point on what unit root tests reveal: All we really can conclude is that some sufficient portion of shocks to the series cumulates over time.

*In the short run:*

Over long stretches of time, series are easier to diagnose. Over finite time, things get complicated, though particularly when dealing with combined series, e.g., when combining an I process with an AR one.

Let's return to our combined I and AR series. Recall that we had 10,000 "time points." Now let us focus on the first 100 observations.

**.dfuller comb1 if x<101, regress**

Here we see that the coefficient is sizable (-.17) and highly reliable ( $p = .025$ ). Do we conclude that the series is stationary? This would be an incorrect conclusion, of course, as the series is nonstationary by construction. It's just that this isn't absolutely clear from the DF tests.

This is true using the Augmented Dickey-Fuller (ADF) tests, which involves including lagged differences, and even when including trend.

**.dfuller comb1 if x<101, regress lags(3) [trend]**

What to do? Well, one thing is to consult  $cf$ 's.

**.ac comb1 if x<101, lags(10)**

**.pac comb1 if x<101, lags(10)**

These are what we would expect of an integrated series over finite time. With such evidence we have good reason to think that the series contains an integrated component, and that a meaningful portion of the effects on our series stand the test of time. This is of obvious importance.

NOTE: It's even hard to find integration when the rho is smaller and decay is greater. Consider when using comb2, which combines a I(1) and AR(1) process with rho=.5.

**.dfuller comb2 if x<101, regress lags(x) [trend]**

For more on issues relating to analysis of 'combined' time series in political science, see "An essay on 'combined' time series processes," Electoral Studies, vol. 19, issue 1, pps. 77-94, 2000.

*On the importance of identification...*

The importance of nonstationarity if it's there—is type I error here (assuming null is stationarity) more tolerable than type II error? What if both X and Y variables are nonstationary? We can't just regress one on the other. We seemingly can't just difference one and regress on the other. The traditional ARIMA approach is to difference both and then diagnose residuals. If white noise, then regress differenced Y on differenced X. But there are other approaches—error correction modelling (ECM). More on this in a bit. First let's consider stationary data.

## **Modeling stationary data**

*The basic ADL model (a really simple VAR, for Vector AutoRegression)*

In social science we are most often interested in how two concepts may relate to each other. Does a good economy increase public support for the prime minister? Do single-parent children tend to get divorced more often? Time series analysis affords several ways to investigate relationships between variables.

One method is to construct an autoregressive distributed lag model (ADL). The basic ADL model—ADL(1,1)—would be written:

$$Y_t = \mu + \phi Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$$

(Note: Assumes exogeneity of X.) The number of lags indicates the order of the model, in this case, first order. Higher order ADL models are rare in sociology and political science, except when doing full-fledged VAR. These types of models can be estimated with OLS, but diagnostic checks are needed to ensure that residual autocorrelation is not a problem.

The ADL(1,1) is a good *general* time series modeling approach. It does not make assumptions about the nature of the relationship and dynamics. It lets the data do the talking, and keeps the researcher from assuming. The other good general modeling approach is the error correction model, which we'll turn to shortly.

Specific forms:

Static model:

$$Y_t = \mu + \beta_0 X_t + \varepsilon_t$$

When  $\phi = 0$  and  $\beta_1 = 0$

Differenced model:

$$\Delta Y_t = \mu + \beta_0 \Delta X_t + \varepsilon_t,$$

where  $\Delta$  means “differenced,” e.g.,  $\Delta Y_t = Y_t - Y_{t-1}$ , when  $\phi = 1$  and  $\beta_0 = -\beta_1$

Thus, don't be surprised if you find coefficients alternating in sign—it tells you something about the nature of the effect(s) of  $X$  on  $Y$ .

The “partial adjustment” model:

This is a lagged dependent variable (LDV) model:

$$Y_t = \mu + \phi Y_{t-1} + \beta_0 X_t + \varepsilon_t$$

Changes in  $Y$  decline as  $Y$  tends toward its asymptote.

NOTE: There are differences in the long-run asymptotes for LDV and specific ADL with  $X_{t-1}$ .

For LDV:

$$\text{Y tends to: } \frac{\beta_0}{1 - \phi}$$

For ADL:

$$\text{Y tends to: } \frac{\beta_0 + \beta_1}{1 - \phi}$$

Of course it may be that  $\beta_0 = 0$ , what is sometimes referred to as a “dead start” model. Here

$$Y_t = \mu + \phi Y_{t-1} + \beta_1 X_{t-1} + \varepsilon_t$$

and so Y tends toward:  $\beta_1$

$$\frac{\quad}{1 - \phi}$$

There are other models, e.g., the various distributed lag (DL) models, where  $Y$  is modelled as a function of current and lagged values of  $X$  (but not  $Y$ )—Koyck’s exponential distributed lag (EDL) and the finite distributed lag (FDL).

One observation: There are many possible models relating  $X$  and  $Y$ .

One lesson: Theory is important but it is a starting point, not a final resting place. Scholars typically just pick one model and estimate it. This may not be right, however. It is important to see what the data “say.” After all, the actual structure is best settled empirically, not by assumption. Of course, settling things empirically is not always easy.

## **Modeling nonstationary data**

Thus far we have focused on stationary data. But what if our data are nonstationary?

### *Error correction models (ECMs)*

An advantage of ARIMA (and transfer function models) 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. (VAR less so than ARIMA, which just differences away equilibria.) The ECM explicitly models the effects of equilibria and disequilibria as follows:

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

where  $B_3$  is the coefficient relating  $Y_t$  and  $X_t$  and  $B_2 < 0$ . (If you are not sure that  $X$  at time  $t$  is exogenous to  $Y$  at time  $t$ , use  $\Delta X_{t-1}$  instead of  $\Delta X_t$ .) As with ARIMA, this is a differenced model. In contrast with ARIMA, long-run information provided by the level data is explicitly modelled.  $Y$  changes when  $X$  changes AND when the previous values of  $X$  and  $Y$  are out of equilibrium, i.e., if  $Y$  is higher or lower than it should be given  $X$ . When it is too high, we expect  $Y$  to decline at time  $t$ ; when it is too low, we expect  $Y$  to increase. This is why  $B_2 < 0$ . The ECM was designed by Granger/Hendry to deal with cointegrated variables—nonstationary variables that maintain an equilibrium relationship with each other. (Either  $X$  or  $Y$  responds to disequilibria or else both respond.) Cointegrated variables will be integrated but a linear combination of them will be stationary. Granger showed that cointegrated variables must have an ECM representation.

*The ECM and stationary data*

The ECM also is useful for modelling stationary data, and it is powerful as well. It does not assume that only short term changes have effects, that only long-term changes have effects, or that they both have the same effects. It doesn't assume persistence or decay. While designed for (co-) integrated variables it can be used with stationary data. Indeed, the ADL and ECM are mathematically equivalent.

$$Y_t = \mu + \beta_1 Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

Subtracting  $Y_{t-1}$  from both sides:

$$Y_t - Y_{t-1} = \mu - Y_{t-1} + \beta_1 Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

Simplifying the left-hand side and combining into a single  $Y_{t-1}$  on the right-hand side:

$$\Delta Y_t = \mu + (\beta_1 - 1) Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

Rewriting  $X_t$  as the sum of the *lagged* level and *current* difference:

$$\Delta Y_t = \mu + (\beta_1 - 1) Y_{t-1} + \beta_2 \Delta X_t + \beta_2 X_{t-1} + \beta_3 X_{t-1} + \varepsilon_t$$

Combining into a single  $X_{t-1}$  on the right-hand side:

$$\Delta Y_t = \mu + (\beta_1 - 1) Y_{t-1} + \beta_2 \Delta X_t + (\beta_2 + \beta_3) X_{t-1} + \varepsilon_t$$

Reassembling as an error correction model:

$$\Delta Y_t = \mu + \beta_2 \Delta X_t + (\beta_1 - 1) [Y_{t-1} + (\beta_2 + \beta_3) / (\beta_1 - 1) X_{t-1}] + \varepsilon_t$$

Now let's see how it reduces to some of the specific forms traced above. To simplify things moving forward, let's return to our original ECM from above:

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

Then, let's multiply terms out as follows:

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

The easiest to see is the

Differenced model:

$$\Delta Y_t = \mu + \beta_1 \Delta X_t + \varepsilon_t$$

When  $\beta_2 = 0$

Also can demonstrate the other specific forms discussed above re: ADL model, though with a bit more work.

Which model to use—the ADL or ECM?

If one is pretty sure the data are stationary, ADL might be most familiar and straightforward; if not sure about stationarity, ECM is the safer bet.

For more on this topic, see Suzanna De Boef (Linn) and Luke Keele, “Taking Time Seriously,” *American Journal of Political Science*, 52, 184-200.

*Beyond a Single Time Series, but not Panel just yet*

What to do when we have data over time but not real time series—a time series of cross sections. Two general cases.

First, where we don't have real time series. Consider polls of electoral preferences during election campaigns. We have a lot of polls regarding an upcoming general election in many countries. Consider aggregate voter preferences for a particular election campaign within a particular country. Let's call the variable  $Poll_t$ . In practice, the variable is not a real time series. There are missing data. And the data we do have overlap. We just don't have regular independent readings of preferences for reasonable periods of time. The situation is worse in previous years.

An alternative approach—in effect, a time series of cross sections, that is, across election years within a particular country

Instead of modelling  $Poll_t$  as a function of  $Poll_{t-1}$  in each year, one can model the Election Day Vote in different election years as a function of the Poll result at different points in time during the election years.

$$\text{VOTE}_j = a_T + b_T \text{Poll}_{jT} + e_{jT}$$

To get estimates of  $\text{Poll}_j$  at each point in time  $T$ , one simply interpolates—what is a problem for a strict time series analysis within election years (because it reduces temporal independence) is actually a benefit when estimating cross-sectional regressions across election years (where temporal dependence is of little concern). We can observe patterns of  $b_T$  and  $R$ -squareds as we vary  $T$ . In our example, this will tell us whether campaign-induced changes in preferences last. (If  $b = 1$ , then the poll result is the best prediction of the Election Day vote; if the  $R$ -squared also = 1, then polls perfectly predict the vote.) Now, if there is persistence, then both statistics will increase as the campaign unfolds; how much is a function of the variance of shocks to preferences and the degree of persistence. For an example, see my article with Robert S. Erikson on “The Timeline of Presidential Election Campaigns,” *Journal of Politics*, 2002;

We also can assess relations between polls over periods of time where we have independent readings, say, fortnights or months.

$$\text{Poll}_{jT} = a_T + b_T \text{Poll}_{jT-1} + e_{jT}$$

across different values of  $T$ . NOTE: here, we clearly cannot interpolate  $\text{Poll}_j$ , as we need independent readings at adjacent time points.

Second, where we have time series but not real panels. Again, consider polls voter preferences over a campaign in a particular country. In this case, consider the preferences of individuals in a particular year. Again we don't have time series, in this case because the respondents change. That is, we have repeated cross sections of the population, not panels of individuals.

What to do? One solution: a form of stacking. This is a fairly common strategy. (Mark Franklin is one practitioner.) An

analysis of individual-level vote choice taking into account the full range of independent variables AND time. In effect, one can assess the time-dependent effects of individual-level variables, e.g.,

$$V_{it} = a_i + \text{Demographics}_{it} + \text{Time}_t + \text{Demographics}_{it} \text{Time}_t + e_{it}$$

A useful strategy if not perfect. Implicit matching. Cannot directly address dynamics. ADL and ECM models just don't apply.

### *Time series cross section (TSCS or CSTS) data*

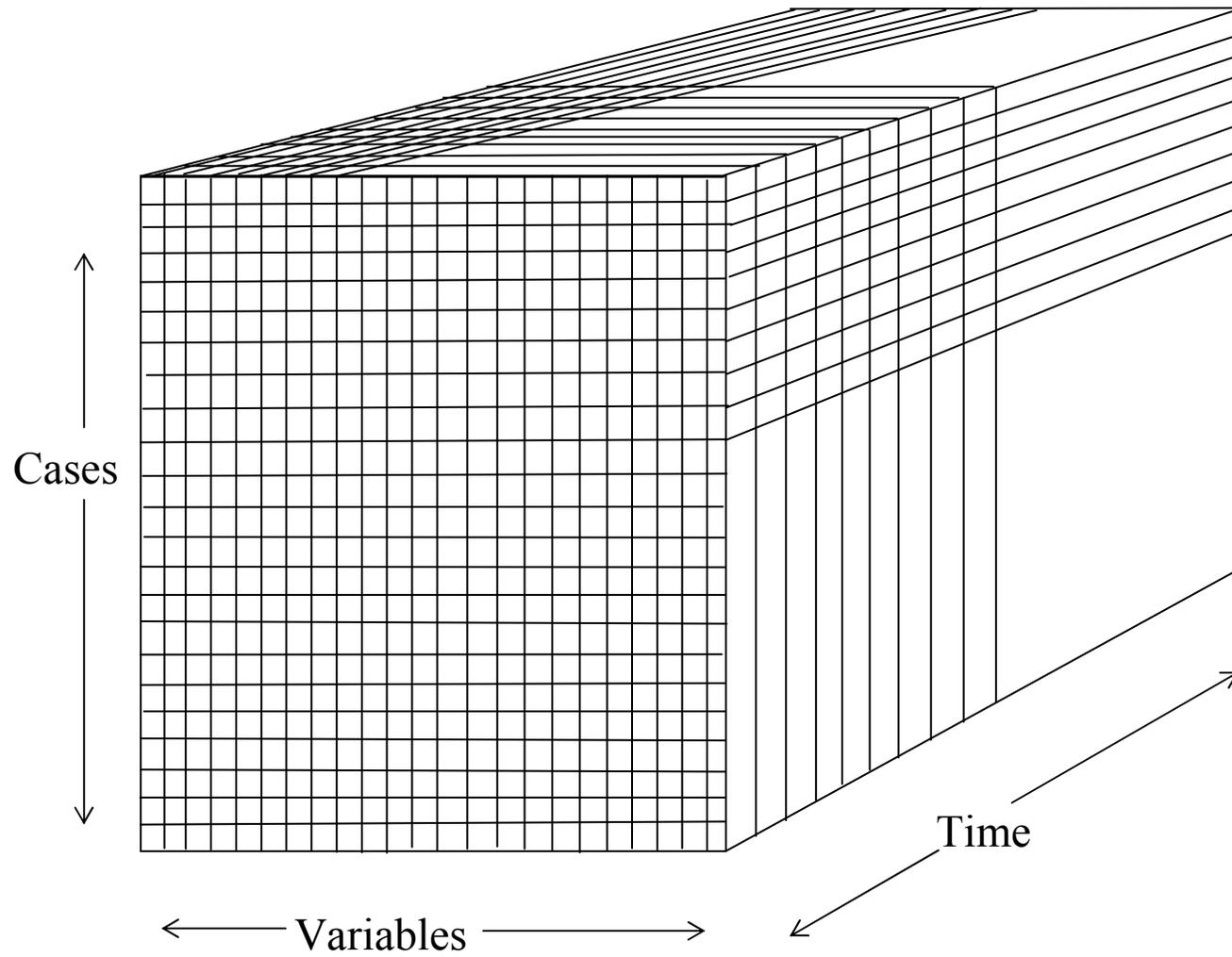
With time series cross section data (also known as “pooled time” series or “cross section time series” panel data) we disaggregate our time series to a lower level of analysis – often the country or, within countries, the state or province. Or we create duplicate time series involving the same variables, often for a

number of countries or states/provinces or, more rarely, individuals. In either case data are collected repeatedly, rather like a panel survey. We thus can model  $Y$  across units  $i$  and time points  $t$ . For many purposes this strategy is very powerful. Under some conditions, it also allows us to assess the generality of results and whether and to what extent they are conditioned by other factors, e.g., context.

Historically, research began by extending the cross-sectional model by adding multiple observations from the same units, i.e., “panels.” Here,  $N$ , the number of units, exceeds  $T$ , the number of time points, and by a lot. Statisticians learned a great deal about working with these types of data sets. (Actually led to the use of certain estimation approaches—the Parks procedure—with problems that political scientists Beck and Katz addressed, ultimately advocating panel corrected standard errors, or PCSEs, unless  $T \gg N$ .)

Increasingly, scholars have longer time series, and our understanding has evolved. It now is more common to look at “pooled” analyses as a set of different time series. The same issues we have been exploring yesterday and today apply to these time series. The approaches differ a little. The main issues are reflected in our Data Box:

Data Box:



Obviously, TSCS data contains variance not only over time but also across units (in our case countries). Analysis of TSCS data thus may require that you decide whether you are primarily interested in change over time, differences between countries, or both (in some sense of “both”). Is the estimated effect of X on Y a reflection of time-serial relations, cross-sectional relations, or both?

### *TSCS and the limits of OLS*

Panel unit effects:

Similar to problems identified by Marco on Monday morning. Consider the following equation:

$$Y_{it} = \gamma + \beta X_{it} + e_{it}$$

Consider further that the error term in the question can be written as follows:

$$e_{it} = \alpha_i + \mu_{it}$$

Here our error term consists of a spatial intercept ( $\alpha_i$ ) and a random error component ( $\mu_{it}$ ), i.e., the levels of our time series differ across units. What is the problem with estimating OLS? If  $\alpha_i$  is correlated with our  $X_{it}$ , estimates of  $\beta$  will be biased. It may be that the differences across units  $i$  reflect differences in the levels of  $X$ , of course. There is good reason to think they reflect other, unmeasured things as well.

Temporal “unit” effects:

Likewise, it may be that our error term consists of a temporal intercept ( $\tau_t$ ):

$$e_{it} = \tau_t + \mu_{it}$$

Here, the levels of our cross sections differ across time. As for spatial unit effects, if  $\tau_i$  is correlated with our  $X_{it}$ , estimates of  $\beta$  will be biased. Now, it may be that the differences across time  $t$  reflect differences in the levels of  $X$ , and this is a common hypothesis in TSCS analysis, i.e., that a particular variable changes over time to influence  $Y$ . For this reason, it usually is considered less of a problem. Indeed, analysis often employs time series variables that are common across units. Consider that it focuses analysis on the effect of the residual component ( $\mu_{it}$ ) from the following equation:

$$e_{it} = \alpha_i + \tau_t + \mu_{it}$$

A strong test that poses the consequence of Type II error against a null of no effect.

Serial correlation:

Same problem as for single time series (or cross section)—unbiased but inefficient estimates.

Heteroskedasticity:

Also same problem as for single time series or cross section—standard errors will be correspondingly wrong.

*Fixed effects, between effects, and random effects*

No discussion of TSCS is possible without a consideration of fixed effects and random effects and also between effects.

If interested in the over-time effects of X on Y (or worried about the contaminating effects of cross-

sectional unit effects), fixed effects is a preferred approach, as it controls for differences between cross-sectional units and thus isolates temporal change.

NOTE: “Fixed effects” here has different meaning to what it has in multi-level, where it refers to unit effects, usually cross-sectional units. In TSCS language, multi-level fixed effects can be thought of as “random intercepts.”

If interested in the cross-sectional effects of X on Y (or worried about the contaminating effects of temporal “unit” effects), between effects is preferred, as it controls for (fixed) differences across time and so isolates variation across units.

If interested in both, random effects is preferred, as it controls for neither temporal nor cross-sectional variation and so models the effects of both.

If worried about the contaminating effects of both cross-sectional and temporal units, one can simultaneously estimate fixed and between effects. Recall that this is a very strong solution.

Discussions of TSCS analysis typically focus on the differences between fixed effects and random effects. Fixed effects is the strong approach, almost universal (and required!) in economics work.

Two approaches to fixed effects: (1) including dummy variables for cross-sectional units or (2) demeaning all of the variables. The approaches are identical and clearly control for differences across units. (And *while it might seem that demeaning uses less degrees of freedom, this is not right*. The reason is that we don't have  $N$  times  $T$  independent observations, because for every unit, the value for any particular observation is just the negative of the mean for all other time points. Most stat packages, like

Stata, use demeaning because it takes less computing power and time.)

Thus, using either approach, if the effect of X on Y is *entirely* driven by temporal variation in X, fixed effects estimation will have absolutely no effect on the estimated results. If the effect is driven entirely by cross-sectional differences in X, fixed effects estimation will produce a coefficient of zero for X. Now, if the effect is driven by both cross-sectional and time-serial variation, estimating using fixed effects will attenuate the coefficient for X, depending on the ratio of time-serial to cross-sectional variation.

This attenuation of the estimated effect of X is one of the drawbacks of using fixed effects, and especially so as N increases and T decreases, e.g., classic panel data.

Another important drawback of real significance, especially for comparative political science research, is that it does not allow estimating the effects of time-invariant variables (TIVs). If  $X$  is a constant, e.g., a political institutional variable that is the same over time in each country, fixed effects wipes out the explanatory power of the variable.

Yet another (even more obvious) drawback is that estimating fixed effects costs  $N-1$  degrees of freedom. This is consequential, and of special importance as  $N$  increases and  $T$  decreases.

In dealing with cross-sectional unit effects, some scholars just estimate first difference models:

$$\Delta Y_{it} = \beta \Delta X_{it} + \Delta e_{it}$$

Benefits: the cross-sectional intercepts are gone.

But, pretty big costs: (1) degrees of freedom; (2) inefficiency (why not differences between  $t$  and  $t-2$  or other lags?).

Random effects estimation. Here dummy variables are not used. Instead, the approach is to estimate a “composite” error term that consists of a “random intercept” component for each unit that is the same for all observations and a random error term.

Generate the intercepts using GLS or MLE (or Bayesian approaches). Then transform the data using the random intercepts and run OLS. Kennedy provides a pretty nice discussion. The simplest way is the GLS error components approach, GLSE in Stimson. Here, there estimated unit effects are based on the autocorrelations (or inter-class correlations, recalling Marco’s talk on Monday morning) from a first stage equation including our  $X$  variables and dummy variables for the different units. (It reflects the assumption of no correlation between  $X$  variables

and the units.) Then the data are transformed using the estimate. We then use OLS to estimate our model on the transformed data.

On the surface, there is a lot to recommend random effects. If we have a large  $N$  (panel data), random effects will be more efficient than fixed effects. It has  $N-1$  more degrees of freedom, and it also uses information from the “between” estimator. While random effects makes sense with panel data, i.e., where  $N \gg T$ ; with longer  $T$ , however, fixed effects is more appropriate. Indeed, there the difference between fixed effects and random effects largely disappear.

Random effects also makes sense if your model includes explanatory variables that don't change over time for a unit, such as government or electoral institutions. Here just cannot use fixed effects. Random effects estimation will work.

But there is a big assumption when estimating random effects, namely, that the intercepts are uncorrelated with our X variable(s). This may be true but it is a strong assumption to make. Kennedy has a nice example. The point is that to the extent the intercepts and X variable(s) are correlated, the more the effect of X will be biased. While fixed effects will tend to attenuate the coefficient on X, random effects will tend to inflate it. (NOTE: There are some diagnostic (Hausman) statistics that focus on whether fixed and random effects produce different results.) *This is one of the reasons why we cannot simply estimate OLS.* There will be more on this tomorrow—today we are just getting an overview of TSCS.

OK, now for some hands on.

The commands you need for working with TSCS data in Stata are:

**xtset panelvar timevar** [ for us, panelvar is nation, timevar is year]

**xtreg depvar indep indep indep ...** [, fe] or [,be] or [,re] or [,mle]

for fixed effects, between effects, random effects and maximum likelihood. Fixed focuses analysis on temporal effects, between effects on cross-sectional effects, and random effects captures both. Maximum likelihood estimation produces random effects estimates using maximum likelihood (instead of GLS in **re**).

NOTE: For the many other xt commands, **help xt**.

Many of the things talked about in this class that one might want to do with single time series data can equally be done with TCSC data. In particular, one might want to visualize the data. For this, STATA provides *xtline*. The command is analagous to *tsline*, but there is an *overlay* option that makes all the separate series appear on a single plot. Let's try it.

First our data set, from research with Christine Arnold and Mark Franklin.

*If not already open, you may want to open a log file.*

**log using xxxx.**

**xtset panelvar timevar**

**xtline iuni**

**xtline iuni, overlay**

What does this tell us?

Unfortunately, **corrgram** does not work with panel data. But you can make your own corrgram pretty by hand.

**xtreg iuni L1.iuni, fe**

taking the square root of the *R*-squared ,

and then

**xtreg iuni L2.iuni, fe**

etc.

It takes a bit of typing, and you don't get the pretty graphs, but you get the same information. OK, is iuni integrated? Or do we see a regression to the mean?

Let's start with a basic static model from research with Mark Franklin:

**xtreg iuni imem ntdiregss, fe**

There is one major problem that has to be addressed before we can go to press with this model. Just like with hierarchical models, we have a problem with standard errors when using TSCS data. Why? It has to do with one of the fundamental assumptions of regression analysis, that residuals are uncorrelated.

Serial correlation is a big issue.

Let's look at the residuals from this our model...

**predict pred**

**gen resid = pred – iuni**

[NOTE: We cannot **predict ...**, resid]

**cor resid l1.resid l2.resid l3.resid l4.resid l5.resid**

These do not look perfect. A LDV?

**xtreg iuni L.iuni imem ntdiregss, fe**

The LDV works but the resulting residuals still do not look perfect. It's not absolutely clear how strong the correlation between residuals needs to be in order to violate OLS assumptions, but some of the correlations are over .3.

It is also possible with TSCS data to “clean up” the error term by correcting for time-serial autocorrelation, just as one does with **prais** when analyzing single time series. The command is called **xtregar**.

**xtregar iuni L.iuni imem ntdiregss, fe**

Recall that there are others that reflect the TSCS nature of the data:

Errors may be correlated across units (contemporaneous correlation).

Errors for different units may have different variances (panel heteroskedasticity). In a much-cited article in the *APSR* Beck and Katz (1995) proposed a way to deal with the problem by calculating “*panel-corrected standard errors*.” This article has been criticized of late (there was a Special Issue of *Political Analysis* devoted to the topic a few years ago). However, the method does do what it is supposed to do (more or less), as we will see.

First need to create country dummies.

**tab nation, generate (cid)**

```
xtpcse iuni imem ntdiregss L.iuni cid1-12  
predict pred  
gen resid = pred - iuni  
cor resid l1.resid l2.resid l3.resid l4.resid l5.resid
```

What we have been doing looks very much like what we did with simple time series. TSCS has two big differences from simple time-series, however. The first has to do with what kind of variance you are trying to explain, as mentioned in my introduction yesterday. So far we have been looking at fixed effects TSCS, which lets us do the closest thing to what we do with single time series.

Results look a lot like what we get using random effects.

```
xtreg iuni imem ntdiregss L.iuni, re
```

However, with TSCS data, in addition to *fixed effects* and random effects), there is the possibility of *between effects*.

The second big difference comes in the calculation of diagnostic tests. Many of these are not available in STATA and have to be done “by hand” (actually the machine does most of the work, but your hand has to tell the machine what to do) rather like our manual substitute for **corrgram**, but somewhat more complicated.

There are exceptions, a bit like the **xtline** command we used earlier. For example, the **xtunitroot** command is a convenience command that packages a group of commands one can use in single time series. The example below uses it to perform a *Dickey-Fuller* test.

```
xtunitroot fisher iuni, dfuller lags(1) [trend]
```

**xtunitroot** repeatedly calls **dfuller**, once for each panel, but does not present the individual results. The test is for a unit root in *any* panel. So you can conduct the test yourself, panel by panel, using an **if** suffix to your variable list in order to focus on each panel in turn. If you do this you can option to see the regression table. Probably you would do this if you wanted to find out which panel(s) contained the unit root diagnosed by **xtunitroot**.

**dfuller iuni if cid==1, lags(1) trend reg**

There are also work-arounds – **xtregar** and **xtpcse** estimate rho for you (the unit root test also tells you whether your series is integrated)

But I'll make available the actual computation for the Breuch-Godfrey test. The easiest test for ARCH is an LM test in which the square of the OLS residual is

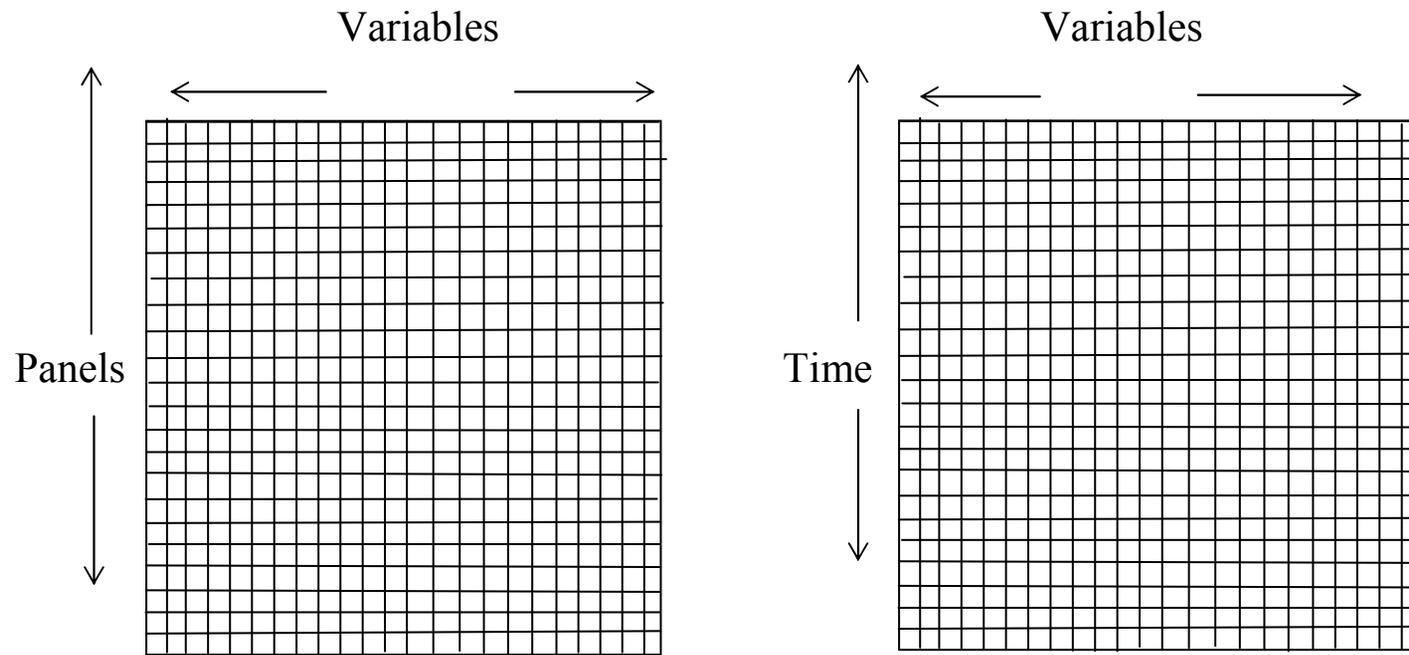
regressed on an intercept and lagged values of the same squared residual. The resulting  $R^2$  multiplied by the sample size is distributed as Chi square with degrees of freedom equal to the number of lags. The Breuch-Godfrey test is very similar.

As an example, to test for heteroskedasticity with five lags you

- (a) run the regression we have now run so many times, using **xtreg ..., fe**, then
- (b) save the residual as above (but let's call them "e" to save some typing)
- (c) square the residual by typing **"gen esq = e^2"**
- (d) type **"reg esq l1.esq l2.esq l3.esq l4.esq l5.esq"**  
[those l characters are the letter "l" for "lag"]
- (e) Multiply the resulting  $R^2$  by N and look up the result in a table of Chi square values.

## **Within, between and random effects**

**xtreg** (and **xtregar**) permit not just the **fe** option (for fixed effects) but also a **be** option. The vocabulary is a bit confusing, with **fe** instead of **we** (for “within effects”), but if you think again about the data box the difference between within and between effects corresponds to across vs into – or, using the other representation from the first day, vars by panels vs. vars by time.



Let's try **xtreg** ... , **be** and see what differences we see...

**xtreg iuni imem ntdiregss L.iuni, re**

Coming back to estimation, the fact that we have variation both cross-time and cross-panel means that we have the possibility of effects that are attributable to specific time-

points and effects that are attributable to specific panels (countries). Including relevant dummy variables in the models can be helpful in diagnosing such effect. For example...

```
reg depvar indeps id1-idj year1-yeark
```

Or, with the data we have been using and **xtpcse**...

```
xtpcse iuni L.iuni imem ntdiregss cid1-cid22 year1-year63,  
cor(ar)
```

The important thing is that all these methods give you a handle on what is going on with your data. Omitting the indeps, the two sets of dummies let you see if variance is within or between. The message is to use everything you can to help you understand the data.

That's enough for now, possibly more than enough. We'll dig more deeply into TSCS modelling after lunch.