Time series and pooled analysis

ELECDEM

Istanbul, session #6

Christopher Wlezien

with special thanks to **Mark Franklin**

More on TSCS modelling.

On modelling TSCS data

So far, we have done the basics of estimation and also focused on spatial issues.

There are time series issues, especially, dynamics and nonstationarity and consequences.

To begin with, we discussed yesterday that one's modelling approach is a function of interest, but interest isn't everything. The approach we adopt also depends on the number of units (N) we observe and the number of observations (T) of each unit. If T is small, we can't very effectively assess time-serial processes even if we want to; if T is large, we can. It's not that time series issues don't apply to the former, just that T does not allow a serious examination. Kennedy highlights this in Chapter 18,

recalling that with small T, TSCS regressions mostly capture the effects of cross-sectional variation simply because there will tend to be less time-serial variation to begin with.

More generally, with panel data, one just can't get fancy.

Consider first fixed effects estimation. Using it, one sacrifices degrees of freedom and increases inefficiency, so that particular unit effects are over-estimated. One also introduces collinearity with independent variables that move little over a handful of observations, i.e., less than they would over longer stretches of time.

Consider next LDV estimation. Using it, one loses N cases, one precious observation for each (short) time series. There also is little basis for dynamics over short stretches. (Plus, estimates using LDV will pick

up the effects of both cross-sectional and time-serial variation in Y, i.e., not just dynamics.)

What to do with panel data? Not the subject of the course, but note that most panel researchers view the data as providing a basis for controlling for unobserved heterogeneity, and so will advocate fixed effects. It will help us deal with questions of endogeneity. But not perfect. Xs and e's might still be correlated. And there might be time-varying endogeneity.

What about where we have larger T, TSCS?

Analyzing sources of variance: Total, cross-unit, temporal, error.

-Plotting, especially the "box plot" of variables by units.

-Statistical summaries of variance, e.g., ANOVA.

(NOTE: Both of these are useful with panel data too.)

Analyzing time-serial dependency.

-Correlograms (if a little trickier, recalling from last time, i.e., no **xt** "ac" or "pac" Stata commands).

-Autoregression.

Modeling TSCS

Let's start with a very simple ("constrained") pooled model relating x to y across units i and years t:

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

Assume a rectangular data structure, where we have observations for all i and t. What if panels are unbalanced, starting and ending at different points? If differences are not major, it's not a major problem e.g., Stata can handle this. What if observations are missing inside the box? If not huge in number, then also not a problem, e.g., imputation.

Then what?

Specification: The right X variable(s) and the proper functional form.

Assuming this, then Gauss-Markov assumptions (i.e., uncorrelated and homoscedastic) met?

If yes, OLS.

If no, manipulations are required.

Time-serial issues

Let's begin with stationary data.

The main issue is autocorrelated errors. The identification of the problem is the same as with a single time series. The solutions are too.

The classic (GLS) approach, e.g., a LM test.

Run OLS.
Generate residuals.
Regress residuals on lagged residuals and all X variables.
Test significance of coefficient on lagged residuals (using usual t-test).

The alternative modelling approach, e.g., a LDV.

$Y_{it} = \gamma Y_{it-1} + \beta X_{it} + e_{it}$

(This is ok only if the resulting residuals are not serially correlated.)

Recall that the two approaches are similar but differ in regards to the effects of lagged values of X. In the AR(1) approach, Y reacts to X instantaneously but the effects of errors decay; in the LDV approach, Y adjusts geometrically. Why would we expect there to be a difference in the temporal effects of X and the errors?

The ADL allows for both possibilities, as both X_t and X_{t-1} are in the model.

$Y_{it} = \square Y_{it-1} + \beta X_{it} + \gamma X_{it-1} + e_{it}$

Estimating it can settle whether either approach is correct. It is the flexible solution.

If $\gamma = 0$ then LDV; if $\gamma = -\beta$ □ then AR.

To see the latter, consider that the coefficient indicates that the estimated effects of lagged X and Y go to 0 but the effect of lagged errors remain, with coefficients □, $□^2$, $□^3$ and so on. (Substitute for Y_{it-1} and do the algebra.)

FYI: As regards GLS (AR1) and LDV, the difference matters little when □ is small, as there is little difference between the immediate and long-run impact of X. As □ increases, the differences widen. When □ approaches 1.0, note that neither approach is correct. Here the data are nonstationary.

Now, with nonstationary TSCS data, the same problems apply as with a single nonstationary time series, though solutions are less developed—almost all of the work in the area focuses on the latter.

Common approaches: (1) differencing, ala ARIMA; and (2) ECM's. The latter is recommended.

$$\Delta Y_{it} = a_{i0} + B_1 \Delta X_{it} + B_2 (Y_{it-1} - B_3 X_{it-1}) + e_{it,}$$

But note that if X and Y are *not* cointegrated, one only can model short run changes in Y as a function of short run changes in X, i.e., there is no pure equilibrium error correction.

To summarize:

-Assess TSCS dynamics as for single time series.
-T is limiting; if low frequency, little one can tell.
-Diagnose with LM test.
-Assess specific form (LDV or AR1) with ADL.
-If nonstationary or possibly so, an ECM.

But, what about cross-sectional unit effects?

Estimating fixed effects with a LDV biases the LDV coefficient, and by a specific amount—1/T. It is known as Hurwicz bias. Thus, the bias is especially bad for very short panels, e.g., 1/3 for 3-wave panels. (This only underscores the point about not getting fancy with panel data sets.) The bias disappears as T increases. (Asymptotics generally better in T for fixed effects.) It is not much of a problem for analysis of TSCS.

With small T, little alternative to random effects. Ditto with independent variables that are constant (or near constant) over time within units. Many of you will have both small T and variables that don't change much over time.

If data are multilevel, or effectively so, you probably are ok with random effects. It is useful thinking of TSCS models in those terms, as yearly data are nested

within countries. With TSCS, the observations are (or may be) connected because of their (possible) over-time dependence, i.e., dynamics. And can assess this.

In direct contrast with fixed effects, asymptotics are best in N for random effects.

When estimating random effects models, one should estimate using GLS and ML, i.e., with "re" and "mle" in **xtreg**. If results using the two approaches differ meaningfully, you most likely are in trouble.

Buyer beware: Using random effects, there is no avoiding the possibility that not estimating fixed effects produces biased estimates.

On estimating the effects of TIVs and slowly-changing variables

This is a real problem area. Cannot simultaneously estimate fixed effects and the effects of TIVs. Random effects one possibility but has limits.

There are alternatives to random effects. There are two main "solutions," an instrumental variables (IV) approach (Hausman-Taylor) and the multi-stage approach, see, e.g., Pluemper and Troeger.

The IV approach: find instruments for the variables. IVs are a solution to many issues, e.g., endogeneity, but has problems: (1) finding valid instruments; and (2) assessing exogeneity of the instruments themselves.

P&T's fixed effects vector decomposition (FEVD). This involves regressing the DV on the time-varying variables and fixed unit dummies and then regressing the dummies on the TIVs to isolate the "residual" part of the unit effects that is unrelated to the TIVs. The

TIVs and the residual part are then included into the equation in place of the dummies.

The latter is clever and is becoming quite popular, though it has been challenged. Not clearly different from LSDV, relies on certain assumptions, and asymptotic properties are not well-known.

There just is no magic statistical bullet.

Heterogeneity is not just for the intercepts

Fully constrained TSCS model assumes no heterogeneity, e.g., in the following LDV:

$$Y_{it} = a + \gamma Y_{it-1} + \beta X_{it} + e_{it,}$$

where intercepts are the same for all i, the coefficients are the same for all i, the dynamics are the same and the lag structures are too. This may not be true.

We have paid a lot of attention to the heterogeneity of intercepts. This may be the least important element of heterogeneity.

Dealing with some of the other elements can be pretty complicated, most notably dynamics. Disentangling differences in lag structure can be too.

One problem with TSCS data is that it is by no means certain that each time series will show the same relationship as the rest, so it is important to conduct "jack-knife" tests in which each unit, e.g., country, in turn is dropped from the dataset and the model re-run to see if the results are the same.

What about coefficients?

On random coefficients

Of course, as we saw earlier in the week, it is possible to assess variation in unit coefficients as a function of unit-level variables:

$\beta_i = f\{z_i\}$.

One can always assess whether effects vary across countries. Substituting variables for country names is For more on this, see Beck. For an explicitly Bayesian approach to multilevel TSCS, see Shor, Boris, Joseph Bafumi, Luke Keele and David Park. 2007. "A Bayesian Multilevel Approach to Time-Series Cross-Sectional Data." *Political Analysis* 15:165-181. (Also Gelman-Hill book on MLM/HLM.) Not perfectly straightforward but of

potential use when N and T are pretty large and we have good priors.

Dichotomous dependent variables

So far we have focused (sometimes only implicitly) on continuous dependent variables. Often our variables are binary, however—often referred to as BTSCS. Not that common with electoral and public opinion data, as we usually (at most) have a panel. But we have some BTSCS data sets and should have more. What to do with these data? How does the approach differ? Here's a short introduction.

Buyer beware: it's early days, and statistical knowledge is not well-developed.

User beware: what we have learned applies to binary DVs and does not generalize to more complicated limited DVs.

The norm: ignore all time series issues and just do logit or probit. Pretty interchangeable at least with decent N's.

The problem: Time-serial dependence remains a problem. Estimates are consistent but not efficient.

Huber ("grouped") robust standard errors. But doesn't explicitly address serial correlation.

And not easy to address—we don't have a simple residual, and instead have to rely on latent errors.

What about unit effects?

As for continuous DVs, fixed effects are fine if T is large, not if T is small.

And estimating random effects also is complicated the reliance on latent errors—where we are observing

“0”’s and “1”’s, and so probabilities are not observed but are based on

Still, progress has been and is being made.

.xtprobit and **xtlogit**

There are yet other commands—**help xt**!

OK, it's a good time for more hands on.

Recall from last time,

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

If data are multilevel, or effectively so, you probably are ok with random effects. In Stata, one can use **xtmixed**. Even if you opt to not estimate using multilevel estimation, it is useful thinking of it in those terms, and when estimating you should estimate using GLS and ML, i.e., with "re" and "mle" in **xtreg**. If results using the two approaches really differ, you most likely are in trouble.

Now, a little more on Stata...

Last time we tried:

xtpcse

xtregar

xtunitroot fisher turnout, dfuller lags(1) [trend]

also, for diagnosing autocorrelation,

xtserial

There are yet other commands—**help xt**!

Another way to approach things:

reg iuni imem ntdiregss

Assess residuals.

Recall that with reg we can **predict xxxx , residuals**

Then we can run a corrgram.

A problem.

A simple solution—LDV and dummy variables.

reg iuni imem ntdiregss L.iuni cid1-cid12

Does it work?

Recall that this is the same as:

xtreg iuni imem ntdiregss L.iuni, fe

Adding year dummies:

reg iuni imem ntdiregss L.iuni cid1-12 year1-year27

Explicitly incorporating what we've learned about ADL and ECM models:

ADL:

$$Y_{it} = a_{i0} + B_1 Y_{it-1} + B_2 X_{it} + B_3 X_{it-1} + e_{it,}$$

reg iuni L.iuni imem L.imem ntdiregss L.ntdiregss cid1-12 year1-year27

How do residuals look?

The ECM produces very similar results, just making clear the effects of current (time t) changes and lagged (time t-1) levels of our X variables.

$$\Delta Y_{it} = a_{i0} + B_1 \Delta X_{it} + B_2 (Y_{it-1} - B_3 X_{it-1}) + e_{it,}$$

We just change all of the current (time t) level variables to differenced variables:

reg Duni L.iuni D.imem L.imem D.ntdiregss L.ntdiregss cid1-12 year1-year27

OK, in this example, T is pretty large—not great, but good. What if T is smaller, tending toward "panel"? There, are options are more limited. Random effects estimation becomes more practical. In Stata, one would not use **reg** and dummy up cross-sectional units. Instead, we'd be back with **xtreg** and the **re** and **mle** options.

When working with TSCS data, especially where it is difficult to estimate using fixed effects, recall that it is important to do everything you can to help you understand the data. That's what statistics are all about.

Now, it's time to bring things to an end.

Final questions? Parting shots?

Thank you once again and all the best. I'm around if you'd like to talk. Just let me know. Of course, I'm also available by e-mail when you're back home.

A little extra….

Duration modeling.

Also often referred to as "event history." Here we are interested in assessing the survival of a state or the hazard rate. How long will a regime last? What are the odds it will collapse? These questions are different to the ones we have been addressing in this class, at least until today. But they are important questions. How do we go about studying them? Can we just directly apply standard time series technology?

The answer—no. While survival models may seem like standard time series models, the estimates are complicated the censoring, particularly right-censoring. If we don't observe the event, e.g., a regime ending, what to do? Dropping it introduces (obvious) selection bias, retaining only those regimes that were more likely to experience the event. Another solution is to create a dummy variable tapping whether the regime experienced the event. But, this belies the time dependency in the data (and process)—it cannot discriminate between variance across cases. This is obviously true for those experiencing the event but more importantly for those not experiencing the event.

The real solution: Modeling duration. The resource in political science: Box-Steffensmeier and Jones's 2004 book. The basic focus:

Survival function:

The probability that the duration has not ended, i.e., the event has not occurred, by time t.

Hazard rate:

The rate at which a duration ends in an interval (t, t + Δt) given that the duration has not ended prior to the beginning of the interval.

Explicitly takes into account dynamics. Assesses the effects of other variables. Accommodates time-varying parameters.

Estimation takes a lot of work—there are classes on the subject. Thankfully, there is a recent (2007) book on the subject for Stata users by Blossfeld, et al.