Longitudinal (Panel and Time Series Cross-Section) Data

Nathaniel Beck
Department of Political Science
University of California, San Diego
La Jolla, CA 92093
beck@ucsd.edu
http://weber.ucsd.edu/∼nbeck

Summer, 2001

Day 1a: What is longitudinal data?

Observed over time as well as over space.

Pure cross-section data has many limitations (Kramer, 1983). Problem is that only have one historical context.

(Single) time series allows for multiple historical context, but for only one spatial location.

Longitudinal data - repeated observations on units observed over time

Subset of *hierarchical data* — observations that are correlated because there is some tie to same unit.

E.g. in educational studies, where we observe student i in school u. Presumably there is some tie between the observations in the same school.

In such data, observe $y_{j,u}$ where u indicates a unit and j indicates the j'th observation drawn from that unit. Thus no relationship between $y_{j,u}$ and $y_{j,u'}$ even though they have the same first subscript. In true longitudinal data, t represents comparable time.

Types of longitudinal data

- "Panel study" (NES, PSID, Congressional election outcomes by CD and year)
- Often use panel data as a single "enriched" cross-section, with info on prior behavior
- "Time-Series–Cross-Section" (political economy data on 15 OECD nations observed annually)
- Event history data
- Dyad year design in IR

- Data combining different surveys taken at different times (eg Markus article on Kramer)
- Rolling Cross-Section (Canadian Election Study)
- "Pseudo Panel" (group respondents by cohort) based on "Repeated Cross Section Data" (eg Family Expenditure Surveys)
- Binary dependent variable - estimation of transition matrix - Markov process

Texts

- Hsaio, *Analysis of Panel Data*, 1986
- Baltagi, *The Econometric Analysis of Panel Data*, 1995
- Matyas and Sevestera, *The Econometrics of Panel Data*, 1996 (handbook)

Panels vs TSCS data

Logically TSCS data looks like panel data, but panels have large number of cross-sections (big N) with each unit observed only a few times (small T); TSCS data has reasonable sized T and not very large N. For panel data, asymptotics are in N, T is fixed. For TSCS data, asymptotics in T, N is fixed.

This distinction is critical. Many of the panel methods are designed to deal with what is known as the "incidental parameters" problem, that is, as the number of parameters goes to ∞, one loses consistency. As we shall see this is a problem only for panel, not TSCS data.

Furthermore, with small T there is no hope of saying anything about the time series structure of the data; with "biggish" T there is.

We also care about the units in TSCS data; they are states or countries. We do not usually care about the units in panel models; they are just a sample, and we care about the population parameters, not the sample.

ALWAYS KEEP THIS DISTINCTION IN MIND

Statistical Issues - Estimation Technique

- OLS
- GLS (and FGLS)
- Full ML (and condition, or REML)
- How examine the likelihood
- Current - find mode and make asymptotic curvature assumptions
- Bayesian MCMC - explore entire likelihood - same church, different sect)

Issues that always arise in longitudinal data

- How model non-independent observations?
  - Repeated observations on same unit are seldom independent
  - Assumption of independence should be tested, not assumed

- So standard likelihood trick of breaking up likelihood of the sample into product of individual likelihoods does not work EASILY
- How model homogeneity?
  - Complete heterogeneity (area studies)
  - Complete homogeneity (econometrics)
  - Both positions silly, what is good compromise? Fixed effects?

Day 1b: Time-Series–Cross-Section Data

The generic cross-national panel model we consider has the form:

$$y_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t}; \qquad i = 1, \ldots, N \tag{1}$$
$$t = 1, \ldots, T$$

where $\mathbf{x}_{i,t}$ is a K vector of exogenous variables and observations are indexed by both unit (i) and time (t). Let $\mathbf{\Omega}$ to be the $NT \times NT$ covariance matrix of the errors with typical element $E(\epsilon_{i,t}\epsilon_{j,s})$.

(Note: we are assuming a "rectangular" structure of the data; this is not critical, but makes notation simpler.)

ASSUME THAT y IS CONTINUOUS, NOT DISCRETE.

Since this is TSCS data, the units are fixed, not sampled. While there is no real bounds on N, in typical applications it will be between 10 and 100.

Assume T is large enough so that time averages make sense (say at least 10). In applications, T's of 20-50 are common.

Fixed Effects

Equation 1 assumes that all countries are fit by same model. And easy (though not trivial and also probably not enough) of a fix is to simply adjoin to the equation country specific intercepts α_i. These are simply dummy variables added to the OLS and so cause no estimation problems. Note that each effect is estimated with T observations, so no problem as long as $T \rightarrow \infty$, even with large (very large) N. We return to this issue tomorrow.

Can test for whether need fixed effects by standard F-test, just compare the SSEs in usual way between Equation 1 and the specification with all the country dummies.

Some leave in only significant country dummies, but that is probably less than best practice (there are large se's on the dummies, so how much do you care if $t = 1.6$ or 1.7?

Spherical errors

Assuming that the errors in Equation 1 are spherical (that is, satisfy the Gauss-Markov assumptions), OLS is optimal if the model is appropriately specified.

The errors are spherical if all errors are independent and identically distributed so that

$$E(\epsilon_{i,t}\epsilon_{j,s}) = \begin{cases} \sigma^2 & \text{if } i = j \text{ and } s = t \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

But even if we like spherical errors, Equation 1 ASSUMES:

- all differences between units are accounted for by differences in the independent variables, that is, no "unmodeled heterogeneity"
- no effects of other units on each other - no spatial effects
- homogeneity (all units obey same equation) or "pooling"
- no dynamics (temporal dependence)

That is, G-M theorem assures us that under stated conditions that OLS of Equation 1 is optimal, but it doesn't tell us the Equation 1 has anything to do with how we think the world works.

Having said that, let us assume that Equation 1 represents our social science!

Non-spherical errors

It is unlikely that cross-national panel errors will meet the assumption of sphericality.

The usual OLS formula for standard errors will (always? often? sometimes?) provide misleading indications of the sampling variability of the coefficient estimates UNLESS THE ERRORS ARE SPHERICAL. The correct formula is given by the square roots of the diagonal terms of

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\{\mathbf{X}'\mathbf{\Omega}\mathbf{X}\}(\mathbf{X}'\mathbf{X})^{-1}. \tag{3}$$

OLS estimates this by

$$\widehat{\text{Cov}(\hat{\beta})} = (\mathbf{X}'\mathbf{X})^{-1}\left\{\left(\frac{\sum_i\sum_t e_{i,t}^2}{NT-k}\right)\mathbf{X}'\mathbf{X}\right\}(\mathbf{X}'\mathbf{X})^{-1} \tag{4}$$

which then simplifies to the usual OLS estimate of the variance-covariance matrix of the estimates (the e's are OLS residuals). OLS standard errors are incorrect insofar as the middle terms in the two equations (in braces) differ.

ASIDE: What does it mean for standard errors to be incorrect, why not "inconsistent," and how would we know?

The OLS errors will be wrong if the errors show any of

- panel heteroskedasticity (Equation 13)
- contemporaneous correlation of the errors (Equation 14)
- serially correlated errors (to be dealt with Friday)

Assuming observations are stacked by unit (not time), that is the first observation is unit 1, time 1, then unit 2, time 1, then unit 3, time 1, etc., we can write the VCV matrix for panel heteroskedastic and contemporaneously correlated (but temporally independent errors as)

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma} & \dots & \mathbf{0} \\ & & & \cdot & \\ & & & \cdot & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Sigma} \end{pmatrix} = \mathbf{\Sigma} \otimes \mathbf{I_N} \tag{5}$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,N} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,N} \\ & & & \cdot & \\ & & & \cdot & \\ \sigma_{1,N} & \sigma_{2,N} & \sigma_{3,N} & \cdots & \sigma_N^2 \end{pmatrix} \tag{6}$$

Models with temporally independent errors

NOTE: We deal with dynamics shortly.

For panel models with contemporaneously correlated and panel-heteroskedastic (but temporally independent) errors, $\mathbf{\Omega}$ is an $NT \times NT$ matrix block diagonal matrix with an $N \times N$ matrix of contemporaneous covariances, $\mathbf{\Sigma}$ (having typical element $E(\epsilon_{i,t}\epsilon_{j,t})$), along the diagonal. To estimate Equation 4 we need an estimate of $\mathbf{\Sigma}$. Since the OLS estimates of Equation 1 are consistent, we can use the OLS residuals from that estimation to estimate $\mathbf{\Sigma}$. Let $e_{i,t}$ be the OLS residual for unit i at time t. We can estimate a typical element of $\mathbf{\Sigma}$ by

$$\hat{\mathbf{\Sigma}}_{i,j} = \frac{\sum_{t=1}^{T} e_{i,t}e_{j,t}}{T}. \tag{7}$$

Letting $\mathbf{E}$ denote the $T \times N$ matrix of the OLS residuals, we can estimate $\mathbf{\Sigma}$ by

$$\hat{\mathbf{\Sigma}} = \frac{(\mathbf{E}'\mathbf{E})}{T} \tag{8}$$

and hence estimate $\mathbf{\Omega}$ by

$$\hat{\mathbf{\Omega}} = \frac{(\mathbf{E}'\mathbf{E})}{T} \otimes I_T \tag{9}$$

where $\otimes$ is the Kronecker product.

We can then compute "Panel Correct Standard Errors" (PCSEs) by taking the square root of the diagonal elements of

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\frac{\mathbf{E}'\mathbf{E}}{T} \otimes \mathbf{I}_T\right)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{10}$$

These still use the OLS estimates of $\hat{\beta}$ but provide correct reports of the variability of these estimates.

Generalized Least Squares

An alternative is GLS. If $\mathbf{\Omega}$ is known (up to a scale factor), GLS is fully efficient and yields consistent estimates of the standard errors. The GLS estimates of β are given by

$$(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y} \tag{11}$$

with estimated covariance matrix

$$(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}. \tag{12}$$

(Usually we simplify by finding some "trick" to just do a simple transform on the observations to make the resulting variance-covariance matrix of the errors satisfy the Gauss-Markov assumptions. Thus, the common Cochrane-Orcutt transformation to eliminate serial correlation of the errors is almost GLS, as is weighted regression to eliminate heteroskedasticity.)

The problem is that $\mathbf{\Omega}$ is never known in practice (even up to a scale factor). Thus an estimate of $\mathbf{\Omega}$, $\hat{\mathbf{\Omega}}$, is used in Equations 11 and 12. This procedure, FGLS, provides consistent estimates of β if $\hat{\mathbf{\Omega}}$ is estimated by residuals computed from consistent estimates of β; OLS provides such consistent estimates. We denote the FGLS estimates of β by $\tilde{\beta}$.

In finite samples FGLS underestimates sampling variability (for normal errors). The basic insight used by Freedman and Peters is that $\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}$ is a (weakly) concave function of $\mathbf{\Omega}$. FGLS uses an estimate of $\mathbf{\Omega}$, $\hat{\mathbf{\Omega}}$, in place of the true $\mathbf{\Omega}$. As a consequence, the expectation of the FGLS variance, over possible realizations of $\hat{\mathbf{\Omega}}$, will be less than the variance, computed with the $\mathbf{\Omega}$. This holds even if $\hat{\mathbf{\Omega}}$ is a consistent estimator of $\mathbf{\Omega}$. The greater the variance of $\hat{\mathbf{\Omega}}$, the greater the downward bias.

This problem is not severe if there are only a small number of parameters in the variance-covariance matrix to be estimated (as in Cochrane-Orcutt) but is severe if there are a lot of parameters relative to the amount of data.

ASIDE: Maximum likelihood would get this right, since we would estimate all parameters and take those into account. But with a large number of parameters in the error process, we would just see that ML is impossible. That would have been good.

Panel Heteroskedasticity

If the errors follow the form

$$E(\epsilon_{i,t}\epsilon_{j,s}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

we have panel heteroskedasticity. It differs from simple heteroskedasticity in that error variances are constant within a unit.

The GLS correction for panel heteroskedasticity is to estimate σ_i^2 from the residuals in the obvious way and then use those estimates in a weighted least squares procedure.

The problem with this procedure is that it is basically weighting units by how well they fit the underlying regression, and so is simply downweighting those that fit poorly, clearly "improving" observed measures of fit.

Note how different this is from theoretically motivated weighted least squares. Panel weighted least squares is problematic, at best. Note that PCSEs correct the standard errors for panel heteroskedasticity.

Contemporaneously correlated errors

Errors are contemporaneously correlated if

$$E(\epsilon_{i,t}\epsilon_{j,s}) = \begin{cases} \sigma_{i,j}^2 & \text{if } s = t \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Note here the huge number of $\sigma_{i,j}^2$ to be estimated in GLS. The GLS procedure here, often known as Parks, thus provides horrible standard errors, off by several hundred percent unless $T >> N$. (Note: Parks doesn't work, that is matrices not invertible, if $T < N$.

Parks should be avoided unless either N is very small or T is very large. (Note that Parks is the same as Zellner's Seemingly Unrelated Regressions. Fortunately SUR is usually applied to a small number of equations with many time points per equation.)

Spatial ideas

There may be some relationship between units, with a bigger relationship between "nearby" units. Nearby could be geographical or, say, measured by trade.

Simplest thing would be, in a model of OECD political economy, to add overall OECD economic growth into a model of individual country GDP growth. Better would be to take the performance of related economies, using say the trade-weighted average of growth in all trading partners. This causes no econometric problems whatsoever (so long as the errors are spherical).

This is Garrett's OECD demand variable (though a more careful trade weighting would have been better).

Spatial Ideas to Improve Parks

The problem with Parks is that there are too many estimated covariances in the error matrix. We could parameterize those covariances, say by assuming that the contemporaneous correlation of the errors for units i and j is just $\lambda d(i, j)$ where $d(i, j)$ is the distance between the two units. This would enable us to use FGLS with only one additional parameter, and hence the estimates obtained should have good properties and the standard errors should be reasonably accurate (getting better as T increases).

Note that the distance measure need not be geographic, it could be the interrelationship of the two economies as measured by trade.

We could also use this idea to improve on PCSEs, simply estimating λ and then putting the estimated variance covariance matrix of the errors into Equation 3.

(Note: it is the repeated time observations and the assumption that all covariances are contemporary that makes this work.)

For lots of interesting econometrics of spatial concepts, see Anselim, "Spatial Econometrics." We will not purse these further.

TSCS Models with temporally dependent errors

Dynamics in TSCS data is very similar to dynamics in time series data.
The errors in cross-national panel models may show temporal dependence as well as spatial dependence.
GLS assumption - errors are serially correlated (usually AR1 with annual data, could easily generalize/test)

$$\epsilon_{i,t} = \rho\epsilon_{i,t-1} + \mu_{i,t} \tag{15}$$

where the μ are independently distributed across time.
(Aside: Old fashioned silliness - assumed that the temporal dependence of the errors could be modeled as a *unit-specific* AR1) process

$$\epsilon_{i,t} = \rho_i\epsilon_{i,t-1} + \mu_{i,t} \tag{16}$$

Simulations show that it is better to impose the additional assumption that the ρ are homogeneous across units, that is, $\rho_i = \rho$. (The time series are too short to get a good estimate of each ρ_i, and there are too many parameters to estimate. In addition, the assumption of a common ρ seems reasonable, especially when we assume common β.)
We can then correct for this serial correlation in the usual manner. First run OLS, compute the serial correlation of the residuals(that is, regress the residuals on the lagged residuals and take the coefficient on the lagged residual as $\hat{\rho}$.) Then transform by subtracting $\hat{\rho}$ of the prior observation from the current one, and run OLS on the transformed observations.

Testing

We can test for serially correlated errors (with or without a lagged dependent variable) via the TSCS analogue of the standard Lagrange multiplier test. Just run OLS, capture the residuals, and regress the residuals on all the independent variables (including the lagged dependent variable if present) and the lagged residual. If the coefficient on the lagged residual is significant (with the usual t-test), we can reject the null of independent errors.
We could test for, and model, second order and higher serial correlation, but most of our TSCS data is annual, and so short lags are often okay. But if you have monthly or quarterly data, worry about longer lags.

Lagged Dependent Variable

Just as with any time series, we could also model dynamics with a lagged dependent variable. The arguments for doing so with TSCS are identical to doing so for simple time series.

- they make the dynamics part of the model, not just a nuisance
- there is seldom any reason to prefer serially correlated errors to a lagged dependent variable
- lagged dependent variables usually simple to estimate and interpret (even if testing indicates SMALL remaining serial correlation)
- only problem is if resulting errors show serial correlation, almost never a problem in practice

You can then use the above LM test to test for remaining serial correlation, hoping you won't find any!
With an LDV in the model, all the error correlations across time period usually pretty much disappear, making life easy.