Introduction to Panel Data

Krishna Pendakur

January 8, 2016

1 Panels

Panel data are no different from regular data except that they have an extra subscript (or, maybe many extras). That is, each observation has an i subscript ($i = 1, ..., N$) which indexes the unit number and a t subscript ($t = 1, ..., T$) which indexes the time period. (Or, different subscripts.)

1. In fact, we can use panel-type methods for any data with more than one subscript. For example, biologists often do multiple treatments on multiple plants, so you'd have plant numbers i and treatment numbers t. There could be in addition a subscript for the lab in multi-lab experiments.

2. The data for panels look just like regular data, except that each i, t in the sample has a variable called UNIT (or something like that) which takes on values from $1, ..., N$ and a variable called TIME (or year or something like that) which takes on values from $1, ..., T$. (Of course, all that matters here is that we know what values i and t can take on, eg, t could go from $1950, ..., 2000$ as in the Penn World Tables.)

3. The data generating process for a panel is the same old thing

$$Y_{it} = f(X_{it}, \beta, \varepsilon_{it})$$

   If we make the ε_{it} mean zero and additive, then we get

$$Y_{it} = f(X_{it}, \beta) + \varepsilon_{it}$$

   and if we make f linear in the parameters, we get the structure that is typically used

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

   Note that if X contains a vector of variables interacted with the vector of time dummies, this is equivalent to

$$Y_{it} = X_{it}\beta_t + \varepsilon_{it}$$

4. where the coefficients are different in each period. Or, if X contains a vector of variables interacted with unit dummies, you get

$$Y_{it} = X_{it}\beta_i + \varepsilon_{it}$$

(a) In Stata, with I being the unit number:

(b) `bysort I: reg Y X`

5. We are often concerned about two types of effects on Y: those that vary with t but not i and those that vary with i but not t. (Those that vary with both are embodied in ε_{it}.)

   (a) Unit Effects. What if X contains unit effects, which we will call θ_i, $i = 1, ..., N$? Here,
   $$Y_{it} = X_{it}\beta + \theta_i + \varepsilon_{it}$$
   is a data generating process where there are unit (time-invariant) effects that differ across every unit i, and these are embodied in θ_i.

   (b) Time effects. Let δ_t, $t = 1, ..., T$ be a set of time effects, one for each time period. Then, the DGP might be
   $$Y_{it} = X_{it}\beta + \delta_t + \varepsilon_{it}$$

   (c) Naturally, we often think that both might be present:
   $$Y_{it} = X_{it}\beta + \theta_i + \delta_t + \varepsilon_{it}$$

6. Fixed Effects vs Random Effects. Models where we do not assume anything about the distribution of θ are called "fixed effects models"; models where we make assumptions about this distribution are often called "random effects models".

   (a) A different way to contrast fixed from random effects is to ask whether or not they are not drawn from a stochastic process. In random effects models, they are assumed drawn from some distribution. In fixed effects models, they may not be. Typical candidates for fixed effects are states or provinces. We often use fixed effects for things like person-identifier in a random sample of people. But, since these are a draw for each person, we ultimately are not going to be interested in any particular estimated θ_i, because we are interested in the features of the population, not of the particular sample drawn. We would instead be interested in perhaps the distribution of θ_i.

7. [*Fixed Effects Estimators*] Consider first models that do not impose restrictions on the distribution of θ_i. Without loss of (much) generality, just consider the model with unit effects and with exogenous errors and without time effects:
   $$Y_{it} = X_{it}\beta + \theta_i + \varepsilon_{it}$$
   and
   $$E[\varepsilon_{it}]|_{(X_{it}\,\theta_i)} = 0.$$

(a) In fact, this is too much exogeneity. Writing the model in matrix form, we could use

$$Y = X\beta + \theta + \varepsilon$$

where Y is the $NT-$vector of Y_{it}, X is the $NTxK$ matrix comprised of X_{it} , θ is the $NTx(N-1)$ matrix comprised of T replications of an $Nx(N-1)$ matrix equal to the stacked row vectors of θ_i and ε is the $NT-$vector of ε_{it}, and the regressors $(X\ \theta)$ are exogenous satisfying

$$E[\varepsilon'\ (X\ \theta)] = 0_{K+N-1}.$$

The restriction $E[\varepsilon_{it}]|_{(X_{it}\ \theta_i)} = 0$ is sufficient for this, but massively more restrictive.

(b) To get estimates of both β and θ_i, one may use Least Squares with Dummy Variables (LSDV), where you regress Yon X and dummies for all but one of the units.

i. Letting D be a vector of $N-1$ dummies for unit identifiers, in Stata,

A. `reg Y X D`

ii. Since $E[\varepsilon_{it}]|_{(X_{it}\ \theta_i)} = 0$ implies $E[\varepsilon'\ (X\ \theta)] = 0_{K+N-1}$, the estimates of β are unbiased. That is, the true model is linear in X and D, and the regressors (including D) are exogenous, so the OLS estimates are unbiased.

iii. The estimates of θ_i are unbiased, but are not N-consistent. That is, as you increase N, the estimate of any particular θ_i does not get closer to its true value. However, the estimates of θ_i are T-consistent, because as you increase T, these estimates do get closer to their true values.

iv. The LSDV approach has a severe drawback. It can be hard to estimate. If you have a panel where N is big, but T is small, then estimating N fixed effects plus K coefficients for β requires inverting an $(N+K)x(N+K)$ matrix. If $N+K$ gets really big (say, in the thousands), then this matrix may get very ill-conditioned and numerically unstable. That is, you might run the regression on different computers and get different answers, or on different software and get different answers. You get this problem whenever the RHS variable matrix is of high rank (has a lot of columns) and sparse (has a lot of zeroes).

(c) An alternative approach is to difference the data to make the unit effects disappear.

i. If your panel has an uninterrupted sequence of t's for every i and if there are at least two t's for every i, then you could first-difference:

$$Y_{it} - Y_{it-1} = (X_{it} - X_{it-1})\,\beta + \theta_i - \theta_i + \varepsilon_{it} - \varepsilon_{it-1},$$

and the unit effect disappears:

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

where Δ is the first-difference-over-time operator.

ii. If the composite error term $\Delta\varepsilon_{it}$ is exogenous, you just use OLS and get β. A sufficient (and overly strong condition) for exogeneity of this composite error term is

$$E[\varepsilon_{it}]|_{X_{it},\theta_i} = 0, E[\varepsilon_{is}\varepsilon_{it}]|_{X_{it},X_{is},\theta_i} = 0 \forall s \neq t$$

If we assume that the disturbances ε_{it} are mean-zero and uncorrelated with each other for every value of X and for every i,t, s, then OLS delivers unbiased estimates of β.

iii. in Stata, we use the difference operator `D.` as follows:

A. `reg D.Y D.X`

(d) Or, if your panel has at least two t's for every i, you could subtract the mean of Y from the LHS and the mean of X from the RHS:

$$Y_{it} - \overline{Y_i} = \left(X_{it} - \overline{X_i}\right)\beta + \theta_i - \theta_i + \varepsilon_{it} - \overline{\varepsilon_i}$$

and the unit effect disappears:

$$Y_{it} - \overline{Y_i} = \left(X_{it} - \overline{X_i}\right)\beta + \widetilde{\varepsilon}_{it}.$$

i. If the composite error term

$$\widetilde{\varepsilon}_{it} = \varepsilon_{it} - \overline{\varepsilon}_i$$

is spherical, you just use OLS to get β. This composite error is spherical if the mean error for any observation i is zero and is uncorrelated with X.

ii. in Stata

A. `areg Y X, absorb(D)`

8. [*Random Effects Estimators*] If you are willing to assume things about the parameters θ_i, then other doors open. For example, assume that the unit effects θ_i are drawn from some distribution satisfying

$$E\left[\theta_i\right]|_{X_{it}} = 0,$$
$$E\left[\left(\theta_i\right)^2\right]|_{X_{it}} = \sigma_\theta^2$$

(Actually, this is a bit stronger than what is needed: you just need θ_i orthogonal to X_{it}, but the differing subscripts makes that assumption notationally cumbersome.)

(a) The restriction that θ_i are mean zero no matter what value X takes is strong. For example, if X includes education and θ_i is meant to capture smartness, we would expect correlation between them. We also need the variance of θ_i to be independent of X. For example, if half of all people are lazy and lazy people never go to college, then the variance of θ_i would covary positively with X observed post-secondary schooling.

(b) Given the assumption on θ_i, we get

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

where

$$u_{it} = \theta_i + \varepsilon_{it}$$

is a composite error term which satisfies exogeneity, but does not satisfy the spherical error term requirement for efficiency of OLS.

(c) One could regress Y on X and get unbiased consistent estimates of β. The reason is that the nonspherical error term only hurts the efficiency of the OLS estimator; it is still unbiased.

9. Difference-in-Difference

(a) Let

$$Y_{ijt} = X_{it}\beta + \theta_i + \delta_t + \varepsilon_{ijt}$$

so that there are many observations (j's) in each unit i and in each year t. Let X_{it} be a binary treatment, and let $s = 1, 2$ and $t = 1, 2$ and let ε_{ijt} be noise. Assume $X_{22} = 1$ and $X_{11} = X_{12} = X_{21} = 0$ so that only observations in unit 2 in time period 2 are treated.

(b) Two observations in the same i, t differ only by ε_{ijt}, which is noise.

(c) Assume that $j = 1, .., J_{st}$ in each s, t. Let $J_{st} \to \infty$ for all s, t, so that our asymptotics can be within each s, t.

(d) Let $\bar{Y}_{it} = \sum_{j \varepsilon it} Y_{ijt}$. I can construct a consistent estimate of population-level analog of this object for each i, t because there are many j's in each i, t.

(e) The over-time difference for observations in the treated unit $\left(\bar{Y}_{22} - \bar{Y}_{21}\right)$ gives an estimate of the sum of the treatment effect plus the time effect $\beta + \delta_2 - \delta_1$.

(f) The over-time difference for observations in the never-treated unit $\left(\bar{Y}_{12} - \bar{Y}_{11}\right)$ gives an estimate of the time effect $\delta_2 - \delta_1$.

(g) The difference in difference estimator of β is $\left(\bar{Y}_{22} - \bar{Y}_{21}\right) - \left(\bar{Y}_{12} - \bar{Y}_{11}\right)$.

(h) The difference in these two differences gives just the treatment effect β.

(i) Identification here rests crucially on the time effect being the same for treated and untreated. If it weren't, you couldn't difference it out.

10. If you are uncomfortable with the distributional assumptions embodied in random effects models, then you may prefer more recently developed *Mixed Models*. Here,

$$Y_{it} = X_{it}\left(\beta + \beta_i\right) + \theta_i + \varepsilon_{it}$$

where

$$\beta_i, \theta_i$$

are distributed jointly normally, with possible conditioning of the distribution on X.

(a) These models add a distributional assumption (normality), but relax the exogeneity assumption and relax the fixed coefficients assumption. They are (sometimes) estimable via maximum likelihood, and get around the big assumptions of random effects models. Of course, mixed models contain as cases the random effects model, and also the *Random Coefficients Model* where we assume that the coefficients on X are randomly distributed across the units i.

11. Time Effects

(a) Consider a model where there are both unit effects and time effects. For example, the LHS variable could be income and year effects might capture the business cycle.

$$Y_{it} = X_{it}\beta + \delta_t + \theta_i + \varepsilon_{it}$$

where δ_t are year effects, $t = 1, ..., T$.

(b) You could use the LSDV approach and regress Y on X with dummies for all but one of the years and dummies for all but one of the units.

(c) Let T be a vector of $T - 1$ dummies for each time period (except one).

(d) in Stata

i. `reg Y X D T`

ii. `reg D.Y D.X T`

iii. `areg Y X T, absorb(D)`

A. In the differenced regression, one of the $T - 1$ dummies for T will be dropped because it now represents differences in time effects across sequential periods.

(e) Usually when we worry about time (eg, in time-series models), we worry about correlations between errors across time. For example,

$$Y_{it} = X_{it}\beta + \delta_t + \theta_i + \varepsilon_{it}$$

where

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + \eta_{it}, \eta_{it} \tilde{} iid(0)$$

is an AR(1) error process. If one is using the LSDV approach, then the standard GLS solution for AR(1) errors can be applied. We will do time-series stuff later.

12. GLS estimation of the Random Effects Model

(a) However, this approach leaves out important information that could improve the precision of our estimate. In particular, we have assumed that the composite errors have a chunk which is the same for every t for a given i. There is a GLS approach to take advantage of this assumption. If we knew the variance of the θ_i terms, σ^2_θ, and knew the variance of the true disturbances, σ^2_ε, we could take advantage of this fact.

i. Under the model, we can compute the covariance of errors of any two observations:

$$E[(\theta_i + \varepsilon_{is})(\theta_j + \varepsilon_{jt})] = I[i = j]\sigma_\theta^2 + I[s = t]\sigma_\varepsilon^2$$

where $I[.]$ is the indicator function. This covariance matrix is block diagonal, where each block consists of the sum of the two variances σ_θ^2 and σ_ε^2 on the diagonal, and just σ_θ^2 off the diagonal. These blocks lie on the diagonal of the big matrix, and the off-diagonal blocks are all zero. (see Green around p 295 for further exposition). So, the model may be written

$$\begin{aligned} Y &= X\beta + \widetilde{\varepsilon}, \\ \widetilde{\varepsilon} &= \{\widetilde{\varepsilon}_{it}\}_{i=1,t=1}^{N,T} \\ \widetilde{\varepsilon}_{it} &= \theta_i + \varepsilon_{it} \\ E\left[X'\widetilde{\varepsilon}\right] &= 0 \\ E\left[\widetilde{\varepsilon}\widetilde{\varepsilon}'\right] &= \Omega, \end{aligned}$$

and Ω has diagonal elements equal to $\sigma_\theta^2 + \sigma_\varepsilon^2$ and within-person off-diagonal elements equal to σ_θ^2 and across-person off-diagonal elements equal to 0.

ii. The GLS transformation matrix is computed as $T = \Omega^{-1/2}$, which is the matrix square-root of this composite error covariance matrix. Then, FGLS regresses transformed Y on transformed X:

$$\begin{aligned} TY &= TX\beta + T\widetilde{\varepsilon} \\ E\left[XT'T\widetilde{\varepsilon}\right] &= 0 \\ E\left[T\widetilde{\varepsilon}\widetilde{\varepsilon}'T\right] &= 1_N \end{aligned}$$

This GLS approach is only easy to implement with a balanced panel in which each and every observation is observed for the same number of periods (so that T is well-defined). But, even with an unbalanced panel, you can still create the block diagonal matrix and invert it.

iii. FGLS requires an estimate of the two variances. A fixed effects model can be run in advance to get estimates of these variances. Or, one could run OLS and construct an estimate of the error covariance matrix directly.