

1 Panels

Panel data are no different from regular data except that they have an extra subscript. That is each panel datum has an i subscript ($i = 1, \dots, N$) which indexes the unit number and a t subscript ($t = 1, \dots, T$) which indexes the time period.

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 $1, \dots, N$ and a variable called TIME (or year or something like that) which takes on values from $1, \dots, T$. (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

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

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

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

4. 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 not identifiable: they are embodied in ε_{it} .)

- (a) Unit Effects. What if X contains unit effects, which we will call θ_i , $i = 1, \dots, N$? 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) Let us assume that

$$\begin{aligned} E[\theta_i \varepsilon_{it}] &= 0 \quad \forall i, t \\ E[\varepsilon_{it}] &= 0 \quad \forall i, t \\ E[(\varepsilon_{it})^2] &= \sigma^2 \quad \forall i, t \\ E[\varepsilon_{it} \varepsilon_{js}] &= 0 \quad \forall i \neq j, t \neq s \end{aligned}$$

- (c) These assumptions imply that ε_{it} is mean zero no matter what value of X_{it} , which is sufficient for orthogonality of X and ε : $E[X_{it} \varepsilon_{it}] = 0_K$.
- (d) We also include the assumption that ε_{it} is orthogonal to θ_i .
- (e) In addition, we assume that ε_{it} is spherical with $E[\varepsilon \varepsilon'] = \sigma^2 I_{NT}$.
- (f) There are at least five ways to estimate β , and one way to estimate θ_i , $i = 1, \dots, N$.
5. To get both β and θ_i , one may use Least Squares with Dummy Variables (LSDV), where you regress Y on X and dummies for all but one of the units.
- (a) The estimates of β are unbiased, consistent and efficient.
- (b) 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.
- (c) 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 $(N + K) \times (N + K)$ 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, 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 has a lot of zeroes.
6. To get just β , there are two other options that don't involve assuming things about the parameters θ_i . These approaches, including the LSDV are called Fixed Effects Models. What makes the effects fixed is that we assume that they are not drawn from a stochastic process. (In random effects models, they are assumed drawn from some distribution.) Typical candidates for fixed effects are states or provinces. Lousy choices for fixed effects would be things like person-identifier in a random sample of people.

- (a) First, 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.

- (b) Could we regress ΔY_{it} on ΔX_{it} to obtain a good estimator for β (notice that β did not change its meaning when we differenced the model)?

i.

$$\begin{aligned} E[\Delta X_{it}\Delta \varepsilon_{it}] &= E[\Delta X_{it}\varepsilon_{it}] - E[\Delta X_{it}\varepsilon_{i,t-1}] \\ &= E[X_{it}\varepsilon_{it}] - E[X_{i,t-1}\varepsilon_{it}] - E[X_{it}\varepsilon_{i,t-1}] + E[X_{i,t-1}\varepsilon_{i,t-1}] \\ &= 0 - 0 - 0 + 0 = 0 \end{aligned}$$

because $E[\varepsilon_{it}] = 0 \forall i, t$. The value of X_{it} doesn't affect its mean.

- ii. ΔX_{it} is exogenous to the disturbance term $\Delta \varepsilon_{it}$, so OLS regression of ΔY_{it} on ΔX_{it} is *unbiased*.

- iii. Is it *efficient*? What is the square of this model disturbance?

$$\begin{aligned} E[\Delta \varepsilon_{it}\Delta \varepsilon_{js}] &= E[\varepsilon_{it}\Delta \varepsilon_{js} - \varepsilon_{i,t-1}\Delta \varepsilon_{js}] \\ &= E[\varepsilon_{it}\varepsilon_{js} - \varepsilon_{it}\varepsilon_{j,s-1} - \varepsilon_{i,t-1}\varepsilon_{js} + \varepsilon_{i,t-1}\varepsilon_{j,s-1}] \\ &= \sigma^2 \text{ if } t = s, i = j \\ &= 0 \text{ if } i \neq j \\ &= -2\sigma^2 \text{ if } i = j, t = s + 1 \end{aligned}$$

This has some off-diagonal terms that are not zero. So, OLS is not efficient.

- (c) Second, 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:

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)\beta + \theta_i - \theta_i + \varepsilon_{it} - \bar{\varepsilon}_i$$

and the unit effect disappears:

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)\beta + \tilde{\varepsilon}_{it}.$$

- (d) If the composite error term

$$\tilde{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$$

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 .

7. The LSDV model is often referred to as the *Fixed Effects Model*, to contrast with a model (given below) called the random effects model. This is not great terminology. I prefer to ask "is θ_i a random variable, drawn from a distribution whose parameters I'd like to learn about?, or is it a fixed vector that I want to estimate?". However, you will most often find that "fixed effects" refers to inclusion of a set of dummy variables corresponding to the units.
8. The *Random Effects Model* assumes that the θ_i are random variables, and not a fixed vector of parameters. Assume that

$$\begin{aligned} E[\theta_i] &= 0, \forall i \\ E[(\theta_i)^2] &= \sigma_\theta^2, \forall i \end{aligned}$$

The fact 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.

- (a) Given the assumption on θ_i , we get

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

where

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

is a composite error term.

- (b) It satisfies exogeneity:

$$E[X_{it}(\theta_i + \varepsilon_{it})] = E[X_{it}\theta_i] + E[X_{it}\varepsilon_{it}] = 0 + 0 = 0,$$

So OLS regression of Y on X provides an unbiased estimator of β .

- (c) How about efficiency?

$$\begin{aligned} E[(\theta_i + \varepsilon_{it})(\theta_j + \varepsilon_{js})] &= E[\theta_i\theta_j] + E[\theta_i\varepsilon_{js}] + E[\theta_j\varepsilon_{it}] + E[\varepsilon_{it}\varepsilon_{js}] \\ &= 0 \text{ if } i \neq j \text{ or } t \neq s \\ &= \sigma_\theta^2 \text{ if } i = j \text{ and } t \neq s \\ &= \sigma_\theta^2 + \sigma^2 \text{ if } i = j \text{ and } t = s. \end{aligned}$$

There are off-diagonal terms that are nonzero, so OLS is not efficient.

- (d) One could use OLS of 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.

- (e) 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_θ^2 , and knew the variance of the true disturbances, σ_ε^2 , we could take advantage of this fact.
- (f) 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 $I[\cdot]$ 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

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

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.

- (g) 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 :

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

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.

- (h) 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.

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

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

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 .

10. 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.

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

where ϕ_t are year effects, $t = 1, \dots, T$.

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

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

where

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

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.