Panel Data

Panel data is an important area in applied econometrics, simply because much of the available data has this structure.

Panel data combines cross sectional and time series data: we have a time series for each of the agents observed in a cross section.

- The addition of temporal information to a cross sectional model can in principle allow us to investigate issues such as persistence, habit formation, and dynamics.
- Starting from the perspective of a single time series, the addition of cross-sectional information allows investigation of heterogeneity.
- In both cases, if parameters are common across units or over time, the additional data allows for more precise estimation.

The basic idea is to allow variables to have two indices, $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$. The simple linear model becomes

$$y_{it} = \alpha + x_{it}\beta + \epsilon_{it}$$

We could think of allowing the parameters to change over time and over cross sectional units. This would give

$$y_{it} = \alpha_{it} + x_{it}\beta_{it} + \epsilon_{it}$$

The problem here is that there are more parameters than observations, so the model is not identified. We need some restraint! The proper restrictions to use of course depend on the problem at hand, and a single model is unlikely to be appropriate for all situations. For example, one could have time and cross-sectional dummies, and slopes that are constant over time and across agents:

$$y_{it} = \alpha_i + \gamma_t + x_{it}\beta + \epsilon_{it}$$

There is a lot of room for playing around here. We also need to consider whether or not n and T are fixed or growing. We'll need at least one of them to be growing in order to do asymptotics.

To provide some focus, we'll consider common slope parameters, but agent-specific intercepts, which:

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it} \tag{1}$$

- I will refer to this as the ''simple linear panel model''. We assume that the regressors x_{it} are exogenous, with no correlation with the error term.
- This is the model most often encountered in the applied literature. It is like the original cross-sectional model, in that the $\beta's$ are constant over time for all i. However we're now allowing for the constant to vary across i (some individual heterogeneity).
- We can consider what happens as $n \to \infty$ but T is fixed. This would be relevant for microeconometric panels, (e.g., the PSID data) where a survey of a large number of individuals may be done for a limited number of time periods.

- Macroeconometric applications might look at longer time series for a small number of cross-sectional units (e.g., 40 years of quarterly data for 15 European countries). For that case, we could keep n fixed (seems appropriate when dealing with the EU countries), and do asymptotics as T increases, as is normal for time series.
- The asymptotic results depend on how we do this, of course.

Why bother using panel data, what are the benefits?

- The model

  $$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$$

  is a restricted version of

  $$y_{it} = \alpha_i + x_{it}\beta_i + \epsilon_{it}$$

  which could be estimated for each i in turn, using time series data. Why use the panel approach? Because the restrictions that $\beta_i = \beta_j = \ldots = \beta$, if true, lead to more efficient estimation. Estimation for each i in turn will be very uninformative if T is small. -Panel data allows us to estimate parameters that are not identified by cross sectional (time series) data. For example, if the model is

  $$y_{it} = \alpha_i + \gamma_t + x_{it}\beta + \epsilon_{it}$$

  and we have only cross sectional data, we cannot estimate the α_i. If we have only time series data on a single cross sectional unit $i = 1$, we cannot estimate the γ_t. Cross-sectional variation allows us to estimate parameters indexed by time, and time series variation allows us to estimate parameters indexed by cross-sectional unit. Parameters indexed by both i and t will require other forms of restrictions in order to be estimable.
- α_i can absorb any missing variables in the regression that don't change over time, and γ_t can absorb missing variables that don't change across i.

For example, suppose we have the model

$$y_{it} = \alpha + x_{it}\beta + z_i\gamma + \epsilon_{it} \tag{2}$$

where the variables in z_i are unobserved, but are constant over time. Assume that, as is usually the case, there is some correlation between the variables in x_{it} and z_i. That is to say, there is some ordinary collinearity of the regressors.

- If we have only one time period, then we have to estimate the model

  $$y_i = \alpha + x_i\beta + z_i\gamma + \epsilon_i$$

  using the observations $i = 1, 2, \ldots, n$. Because z_i is unobserved, we have to let it be absorbed in the error term. For convenience, and to keep the notation simple, assume that the mean of $z_i\gamma$ is zero (this does not affect the argument in any important way), so the model we can actually estimate is

  $$y_i = \alpha + x_i\beta + v_i$$

where $v_i = z_i\gamma + \epsilon_i$. This model has correlation between the regressors and the error, so the OLS estimates would be inconsistent. Furthermore, we don't have any natural instruments to estimate the model by IV.

However, suppose we have at least two time periods of data, and n cross-sectional observations. Then, we can let $z_i\gamma$ move into the constant, and we get the model

$$\begin{aligned} y_{it} &= \alpha + x_{it}\beta + z_i\gamma + \epsilon_{it} \\ y_{it} &= \alpha_i + x_{it}\beta + \epsilon_{it} \end{aligned}$$

where $\alpha_i = \alpha + z_i\gamma$. This is the simple linear panel data model.

```
- Notice that the problematic $z_{i}$ have now disappeared!
- It turns out that OLS estimation of this model will give consistent
estimates of the $\beta$ parameters, as the cross sectional size of the
sample, $n$ increases, as long as the regressors are exogenous. If it's
not clear how this can be estimated by OLS, then consider estimating it
using first differences: that model is pretty obviously consistently
estimable using OLS.
```

To begin with, assume that:

- the x_{it} are weakly exogenous variables (uncorrelated with ϵ_{it})
- the model is static: x_{it} does not contain lags of y_{it}.
- then the basic problem we have in the panel data model $y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$ is the presence of the α_i. These are individual-specific parameters. Or, possibly more accurately, they can be thought of as individual-specific variables that are not observed (latent variables). The reason for thinking of them as variables is because the agent may choose their values following some process, or may choose other variable taking these ones as given.

Define $\alpha = E(\alpha_i)$, so $E(\alpha_i - \alpha) = 0$, where the expectation is with respect to the density that describes the distribution of the α_i in the population. Our model $y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$ may be written

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}\beta + \epsilon_{it} \\ &= \alpha + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ &= \alpha + x_{it}\beta + \eta_{it} \end{aligned}$$

Note that $E(\eta_{it}) = 0$. A way of thinking about the data generating process is this:

- First, α_i is drawn, from the population density
- then T values of x_{it} are drawn from $f_X(z|\alpha_i)$.
- the important point is that the distribution of x may vary depending on the realization of α_i.

For example, if y is the quantity demanded of a luxury good, then a high value of α_i means that agent i will buy a large quantity, on average. This may be possible only when the agent's income is also high. Thus, it may be possible to draw high values of α_i only when income is also high, otherwise, the budget constraint would be violated. If income is one of the variables in x_{it}, then α_i and x_{it} are not independent.

Another example: consider returns to education, modeling wage as a function of education. α_i could be an individual specific measure of ability. Ability could affect wages, but it could also affect the number of years of education. When education is a regressor and ability is a component of the error, we may expect an endogeneity problem.

Thus, there may be correlation between α_i and x_{it}, in which case $E(x_{it}\eta_{it}) \neq 0$ in the above equation.

- This means that OLS estimation of the model would lead to biased and inconsistent estimates.
- However, it is possible (but unlikely for economic data) that x_{it} and η_{it} are independent or at least uncorrelated, if the distribution of x_{it} is constant with respect to the realization of α_i. In this case OLS estimation would be consistent.

Fixed effects: when $E(x_{it}\eta_{it}) \neq 0$, the model is called the "fixed effects model"

Random effects: when $E(x_{it}\eta_{it}) = 0$, the model is called the "random effects model"

I find this to be pretty poor nomenclature, because the issue is not whether ''effects'' are fixed or random (they are always random, unconditional on i). The issue is whether or not the ''effects'' are correlated with the other regressors. In economics, it seems likely that the unobserved variable α is probably correlated with the observed regressors, x (this is simply the presence of collinearity between observed and unobserved variables, and collinearity is usually the rule rather than the exception). So, we expect that the ''fixed effects'' model is probably the relevant one unless special circumstances imply that the α_i are uncorrelated with the x_{it}.

Example: Agricultural Production

Suppose

$$y_{it} = \beta x_{it} + \eta_{it} + v_{it}$$

represents the Cobb-Douglas production function of an agricultural product. The index i denotes farms and t time periods.

- y_{it} = log output
- x_{it} = log input (labor)
- η_i = input that remains constant over time (soil quality)
- v_{it} = stochastic input which is outside of the farmer's control (rainfall)

Suppose η_i is known by the farmer but not by the econometrician.

- If farmers maximize expected profits there will be cross-sectional correlation between labour and soil quality.
- Therefore, the population coefficient in a simple regression of y_{i1} on x_{i1} will differ from β.

If η were observed by the econometrician,

- the coefficient on x in a multiple cross-sectional regression of y_{i1} on x_{i1} and η_i will coincide with β.

Now suppose that data on y_{i2} and x_{i2} for a second period become available. Moreover, suppose that rainfall in the second period is unpredictable from rainfall in the first period (permanent differences in rainfall are part of η_i), so that rainfall is independent of a farm's labour demand in the two periods. Thus, even in the absence of data on ηi the availability of panel data affords the identification of the technological parameter β.

Fixed Effects: the within estimator

The within estimator involves subtracting the time series average from each cross sectional unit.

$$
\begin{aligned}
\overline{x}_i &= \frac{1}{T}\sum_{t=1}^{T} x_{it} \\
\overline{\epsilon_i} &= \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} \\
\overline{y}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} = \alpha_i + \frac{1}{T}\sum_{t=1}^{T} x_{it}\beta + \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} \\
\overline{y}_i &= \alpha_i + \overline{x}_i\beta + \overline{\epsilon_i} \qquad (3)
\end{aligned}
$$

The transformed model is

$$
\begin{aligned}
y_{it} - \overline{y}_i &= \alpha_i + x_{it}\beta + \epsilon_{it} - \alpha_i - \overline{x}_i\beta - \overline{\epsilon_i} \qquad (4) \\
y^*_{it} &= x^*_{it}\beta + \epsilon^*_{it}
\end{aligned}
$$

where $x^*_{it} = x_{it} - \overline{x}_i$ and $\epsilon^*_{it} = \epsilon_{it} - \bar{\epsilon}_i$. In this model, it is clear that x^*_{it} and ϵ^*_{it} are uncorrelated, as long as the original regressors x_{it} are exogenous with respect to the original error ϵ_{it} ($E(x_{it}\epsilon_{is}) = 0,\ \forall t, s$). In this case, OLS will give consistent estimates of the parameters of this model, β.

What about the α_i?

Can they be consistently estimated?

An estimator is

$$
\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}\left(y_{it} - x_{it}\hat{\beta}\right)
$$

It's fairly obvious that this is a consistent estimator if $T \to \infty$. For a short panel with fixed T, this estimator is not consistent. Nevertheless, the variation in the $\hat{\alpha}_i$ can be fairly informative about the heterogeneity.

An equivalent approach is to estimate the model

$$
y_{it} = \sum_{j=1}^{n} d_{j,it}\alpha_j + x_{it}\beta + \epsilon_{it}
$$

by OLS. The d_j, $j = 1, 2, \ldots, n$ are n dummy variables that take on the value 1 if $j = i$, zero otherwise. They are indicators of the cross sectional unit of the observation. For example, with 3 cross sectional units and 3 time periods, and a single x regressor, the model in matrix form would look like

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & x_{11} \\ 1 & 0 & 0 & x_{12} \\ 1 & 0 & 0 & x_{13} \\ 0 & 1 & 0 & x_{21} \\ 0 & 1 & 0 & x_{22} \\ 0 & 1 & 0 & x_{23} \\ 0 & 0 & 1 & x_{31} \\ 0 & 0 & 1 & x_{32} \\ 0 & 0 & 1 & x_{33} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{bmatrix}$$

Estimating this model directly by OLS gives numerically exactly the same results as the OLS version of the ''within'' estimator, and you get the $\hat{\alpha}_i$ automatically. See Cameron and Trivedi, section 21.6.4 for details. An interesting and important result known as the Frisch-Waugh-Lovell Theorem can be used to show that the two means of estimation give identical results.

Example in which Panel Data Does Not Work: Returns to Education

“Structural” returns to education are important in the assessment of educational policies. It has been widely believed in the literature that cross-sectional regression estimates of the returns could not be trusted because of omitted “ability” potentially correlated with education attainment. In the earlier notation:

- y_{it} = Log wage (or earnings).
- x_{it} = Years of full-time education.
- η_i = Unobserved ability.
- β = Returns to education. The problem in this example is that x_{it} typically lacks time variation. So a regression in firstdifferences will not be able to identify β in this case. In this context data on siblings and cross-sectional instrumental variables have proved more useful for identifying returns to schooling free of ability bias than panel data.

This example illustrates a more general problem. Information about β in the regression in firstdifferences will depend on the ratio of the variances of Δv and Δx. In the earnings−education equation, we are in the extreme situation where Var(Δx)=0, but if Var(Δx) is small regressions in changes may contain very little information about parameters of interest even if the cross-sectional sample size is very large.

Random Effects

α_i is assumed to be random.

The original model is

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}$$

This can be written as

$$\begin{aligned} y_{it} &= \alpha + x_{it}\beta + (\alpha_i - \alpha + \epsilon_{it}) \\ y_{it} &= \alpha + x_{it}\beta + \eta_{it} \end{aligned} \tag{5}$$

Under random effects, the α_i are assumed not to be correlated with the x_{it}, so $E(\eta_{it}) = 0$, and $E(x_{it}\eta_{it}) = 0$. As such, the OLS estimator of this model is consistent. We can recover estimates of the α_i as discussed above.

It is to be noted that the error η_{it} is almost certainly heteroscedastic and autocorrelated, so OLS will not be efficient, and inferences based on OLS need to be done taking this into account. One could attempt to use GLS, or panel-robust covariance matrix estimation, or both, as above.

There are other estimators when we have random effects, a well-known example being the ''between'' estimator, which operates on the time averages of the cross sectional units. There is no advantage to doing this, as the overall estimator is already consistent, and averaging looses information (efficiency loss). One would still need to deal with cross sectional heteroscedasticity when using the between estimator, so there is no gain in simplicity, either.

It is to be emphasized that ''random effects'' is not a plausible assumption with most economic data, so use of this estimator is discouraged, even if your statistical package offers it as an option. Think carefully about whether the assumption is warranted before trusting the results of this estimator.

Hausman Test

Suppose you're doubting about whether fixed or random effects are present.

- If we have correlation between x_{it} and α_i (fixed effects), then the ''within'' estimator will be consistent, but the random effects estimator of the previous section will not.
- A Hausman test statistic can be computed, using the difference between the two estimators.
  - The null hypothesis is that the effects are uncorrelated with the regressors in x_{it} (''random effects'') so that both estimators are consistent under the null.
  - When the test rejects, we conclude that fixed effects are present, so the ''within'' estimator should be used.
  - Now, what happens if the test does not reject? One could optimistically turn to the random effects model, but it's probably more realistic to conclude that the test may have low power. Failure to reject does not mean that the null hypothesis is true. After all, estimation of the covariance matrices needed to compute the Hausman test is a non-trivial issue, and is a source of considerable noise in the test statistic (noise=low power).