# Panel Data

Panel data is an important area in applied econometrics, simply because much of the available data has this structure. Also, it provides an example where things we've already studied (GLS, endogeneity, GMM, Hausman test) come into play. There has been much work in this area, and the intention is not to give a complete overview, but rather to highlight the issues and see how the tools we have studied can be applied.

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.
- Another reason is that 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 $\alpha_i$. If we have only time series data on a single cross sectional unit $i = 1$, we cannot estimate the $\gamma_t$. 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.
- A **very important reason** is that $\alpha_i$ can absorb any missing variables in the regression that don't change over time, and $\gamma_t$ 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

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

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 $\epsilon_{it}$)
- 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 $\alpha_i$. 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 $\alpha_i$ 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, $\alpha_i$ 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 $\alpha_i$.

For example, if $y$ is the quantity demanded of a luxury good, then a high value of $\alpha_i$ 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 $\alpha_i$ only when income is also high, otherwise, the budget constraint would be violated. If income is one of the variables in $x_{it}$, then $\alpha_i$ and $x_{it}$ are not independent.

Another example: consider returns to education, modeling wage as a function of education. $\alpha_i$ 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 $\alpha_i$ 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 $\eta_{it}$ are independent or at least uncorrelated, if the distribution of $x_{it}$ is constant with respect to the realization of $\alpha_i$. 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 $\alpha$ 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 $\alpha_i$ are uncorrelated with the $x_{it}$.