1 Panel data: basics

As we saw in the first set of lecture notes, panel data provides a way of relaxing the assumption that the error term is uncorrelated with the explanatory variables. Remember the model:

\[ y_{it} = x_{it} \beta + \varepsilon_{it} \]  

(1)

where the subscript \( i \) refers to the observation unit (say a worker) and \( t \) to the time period. The standard assumption used in panel data analysis is that we can write the error as:

\[ \varepsilon_{it} = \theta_i + \nu_{it} \]  

(2)

where \( \theta_i \) is usually called a fixed or random effect that does not change over time. Even if \( \theta_i \) represents unobserved determinants of \( y_{it} \) that are correlated with \( x_{it} \), it is possible to consistently estimate \( \beta \) by first-differencing the data (or using other related “fixed effect” methods). So panel data enables us to replace the assumption that \( E(\varepsilon_{it} | x_{it}) = E(\theta_i + \nu_{it} | x_{it}) = 0 \) by the weaker assumption \( E(\nu_{it} | x_i) = 0 \).

Notice that there is a little bit of a catch here. In an \( iid \) sample, what we need for OLS to be consistent is that \( E(\varepsilon_{it} | x_{it}) = 0 \), where the \( x \) variable both has an \( i \) and \( t \) subscript. In the notation we used above, we have \( x_i \) instead, where \( x_i \) is defined to include all values of \( x_{it} \) for individual \( i \), i.e. \( x_i = [x_{i1}, x_{i2}, ..., x_{iT}] \), where \( T \) is the length of the panel (the number of observations for each unit \( i \)). So when we write \( E(\nu_{it} | x_i) = 0 \), we require all leads \( (x_{it+1}, ..., x_{iT}) \) and lags \( (x_{i1}, ..., x_{it-1}) \) of \( x_{it} \) to be unrelated to \( \nu_{it} \), while the condition \( E(\varepsilon_{it} | x_{it}) = 0 \) means that \( \varepsilon_{it} \) and \( x_{it} \) are only contemporaneously
unrelated. The assumption that \( E(\nu_{it}|x_i) = 0 \) could be violated, for example, in the case where \( x_{it} \) is “predetermined” in the sense that it may be related to earlier realization of \( \nu_{it} \) but not to the current realization of \( \nu_{it} \). Another example where the condition will fail is when \( x_{it} \) is a lagged value of the dependent variable, for example where \( x_{it} = y_{it-1} \).

The literature tends to distinguish between the case of a “random effect” where \( \theta_i \) is assumed not to be correlated with \( x_{it} \), and “fixed effects” where \( \theta_i \) and \( x_{it} \) are allowed to be correlated. Under the assumptions of the random effect model \( (E(\theta_i|x_i) = 0) \), OLS yields consistent estimates, but OLS standard errors are inconsistent because of the group effect problem (see lecture notes #3). This can be easily corrected, however, by “clustering” the standard errors at the individual level. Another issue is that OLS is not efficient. More efficient estimates can be obtained by doing GLS.

The more interesting case is when \( \theta_i \) is correlated with \( x \) \( (E(\theta_i|x_i) \neq 0) \). For a number of historical reasons, this tends to be called the “fixed effect” case, but what this really means is that we have correlated (with the \( x_{it} \) random effects. Let’s now see how we can consistently estimate \( \beta \) in the presence of fixed effects.

1.1 First-difference estimation

First-differencing is the easiest way of dealing with fixed effects. Since the lagged value of \( y_{it} \) is

\[
y_{it-1} = x_{it-1}\beta + \theta_i + \nu_{it-1}
\]

(3)

taking first differences yields:

\[
y_{it} - y_{it-1} = (x_{it} - x_{it-1})\beta + (\theta_i - \theta_i) + (\nu_{it} - \nu_{it-1})
\]

or

\[
\Delta y_{it} = \Delta x_{it}\beta + \Delta \nu_{it}
\]

(4)

(5)

OLS yields consistent estimates of \( \beta \) since \( \theta_i \) has been removed through first-differencing, while \( \Delta \nu_{it} \) is uncorrelated with \( \Delta x_{it} \) because of the assumption that \( E(\nu_{it}|x_i) = 0 \). Note, however, that first differencing tends to introduce spurious negative correlation across observations since \( cov(\Delta \nu_{it}, \Delta \nu_{it-1}) = cov(\nu_{it} - \nu_{it-1}, \nu_{it-1} - \nu_{it-2}) = -var(\nu_{it-1}) \).

In this simple model, the autocorrelation in errors is equal to -.5 since \( var(\Delta \nu_{it}) = var(\nu_{it}) + var(\nu_{it-1}) \), which is twice (in absolute value) as large as the autocovariance \( (cov(\Delta \nu_{it}, \Delta \nu_{it-1})) \). A first consequence of the serial correlation is that first-difference
estimates are not efficient and that one could estimate $\beta$ more precisely with a GLS procedure. A second consequence is that the standard errors are incorrect. This problem can be fixed, however, by allowing for serial correlation in the errors. Clustering (on individuals $i$) should also correct for this since the formula used in lecture 3 allows for flexible forms of correlation across observations for the same individual $i$ (the group in this case).

### 1.2 Dummy variable estimation

Another estimation approach consists of including dummy variables for each individual $i$. For example, if we have 1000 individuals ($i = 1$ to 1000) and 10 observations for each individual ($t = 1$ to 10), we can construct 1000 dummy variables for each individual, and just stick these variables into the regression. OLS estimates from this very big regression with over one thousand explanatory variables would actually yield consistent estimates of $\beta$. The intuition is that whatever omitted person-specific factors end up in the fixed effect $\theta_i$, these factors will be entirely absorbed by the dummy variables.

Obviously, this is not a very convenient estimator to use in very large data sets as we may have to include thousands of explanatory variables, which eventually pushes the limits of even very powerful computers. Fortunately, the dummy variable model yields estimates of $\beta$ that are numerically equivalent to what is called the “within” estimator. Let’s now consider this alternative approach.

### 1.3 Within estimator

The idea of the within estimator is to solely use the variation in $y_{it}$ and $x_{it}$ within the same individual over time. This is formally done by subtracting the “within-individual” average of each variable in the main equation:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + \theta_i - \bar{\theta}_i + \nu_{it} - \bar{\nu}_i$$  \hspace{1cm} (6)

where $\bar{y}_i$ is the mean value of $y_{it}$ for individual $i$, etc. Since $\theta_i$ is constant over time for each individual $i$, it follows that $\theta_i = \bar{\theta}_i$ and that $\theta_i - \bar{\theta}_i = 0$. So this within or “difference from means” estimator is another way of removing the fixed effect $\theta_i$. Consistent estimates of $\beta$ can then be obtained by estimating the equation

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + \nu_{it} - \bar{\nu}_i$$  \hspace{1cm} (7)
by OLS.

The simplest way to show the equivalence between the within and the dummy variable estimator is to apply the Frisch-Waugh theorem. This theorem states that a regression of $y$ on a subset of the $x$ variables (say $x_1$) yields the same coefficients as in the full regression on all $x$’s ($x_1$ and $x_2$) provided that we first “partial out” the effect of $x_2$ (the other regressors) from $y$ and $x_1$ by running regressions of these variables on $x_2$ and keeping the residuals. A well known example is the case where $x_2$ is a time trend. One can then either detrend the data first and run a regression of the detrended value of $y$ on the detrended value of $x_1$, or include the trend directly in the regression.

Here we can think of the dummy variables as the $x_2$ variables. So the dummy variables estimator can be implemented by first running regressions of both $y_{it}$ and the regressors of interest ($x_{it}$) on the full set of dummies, and then keeping the residuals, which turn out to simply be the differences with respect to individual-specific means.

### 1.4 Within vs. first-differences

When there are only two observations by individual, the within and first-difference estimators turn out to be numerically equivalent. With longer panels, however, both estimators are consistent but not numerically equivalent. It can be shown that the within estimator is efficient while first-differences is not efficient. One reason why FD is not efficient is what we saw earlier, namely that errors are correlated and that GLS should be used instead of OLS.

Note that we could also use longer differences instead of first-differences. Some studies have shown that estimators based on longer differences had nicer properties, in particular that there were more robust to the problem of measurement error.

### 1.5 Random effect estimation

In the presence of autocorrelation in a simple time-series setting, the GLS estimates can be obtained by estimating OLS on the quasi-differenced data where all variables (left and right hand side) $z_t$ are replaced by the quasi-difference $z_t - \rho z_{t-1}$, where $\rho$ is the (estimated) autocorrelation coefficient. By analogy (see Wooldridge or Greene for details), GLS estimates in the random effect setting are obtained by running OLS on quasi-differences relative to the mean, where the implied model is

$$y_{it} - \lambda \bar{y}_i = (x_{it} - \lambda \bar{x}_i) \beta + (1 - \lambda) \theta_i + \nu_{it} - \lambda \eta_i$$  \hspace{1cm} (8)
and where it can be shown that \( \lambda = 1 - \left[ \sigma^2_v / (\sigma^2_v + T\sigma^2_\theta) \right]^{1/2} \). On the one hand, like OLS, random effect estimates are not consistent when the error component \( \theta_i \) is correlated with \( x \). On the other hand, if \( \theta_i \) is not correlated with \( x \), then random effect estimates are more efficient than fixed effects. Furthermore, it is still possible to estimate the effect of time invariant regressors (see below) with random effects but not with fixed effects. Given these potential advantages of random effects when we do not need fixed effects, it is advisable to perform a Hausman specification test for random against fixed effects before settling on either procedure.

2 Other Issues

2.1 Time-invariant regressors

One important limitation of fixed-effect estimation (general term for either first-difference or within estimation) is that we cannot identify the effects of time-invariant variables, i.e. variables that vary across individuals but not across time. The reason is that these variables, just like the fixed effect, are removed by taking first differences or differences relative to the mean. For example, if the subset \( x_2 \) of the \( x \) variables is time invariant, we have:

\[
y_{it} - y_{it-1} = (x_{1it} - x_{1it-1})\beta_1 + (x_{2i} - x_{2i})\beta_2 + (\theta_i - \theta_i) + (\nu_{it} - \nu_{it-1}) = (x_{1it} - x_{1it-1})\beta_1 + (\nu_{it} - \nu_{it-1})
\]

(9)

The term related to time-invariant regressors, \( (x_{2i} - x_{2i})\beta_2 \), drops out from the equation, which means that it is not possible to estimate \( \beta_2 \). Some estimators (Hausman and Taylor) have been suggested to estimate the effect of time invariant regressors in the presence of fixed effects, but they involve additional assumptions that are difficult to test. When we are interested in the effect of time-invariant regressors, it is thus imperative to check whether we really need fixed effects. This can be done using a variant of the Hausman test that is easy to use in Stata (see below).

2.2 Year/time effects

Given the time dimension in panel data, it is well advised to make sure appropriate controls are introduced for possible time trends, just like in traditional time series analysis. For example, we can introduce a linear or quadratic trend in the model. With panel
data, it is also possible to include a full set of time dummies that control for time effects in the most unrestricted way. It is highly advisable to check how robust the main results are to the inclusion of richer and richer set of controls for time effects.

3 Stata commands

First differences are typically computed manually. Make sure that the data are sorted by individual and time period, then simply compute first-differences. For example, if we are running a regression of $y_1$ on $x_1$ where individuals are identified by a variable “indiv” and time by “year”, we can do the following:

```
sort indiv year

by indiv: gen dy1=y1-y1[_n-1]

by indiv: gen dx1=x1-x1[_n-1]

reg dy1 dx1, vce(cluster indiv)
```

where the cluster option is used to account for serial correlation in the errors. To compare this to OLS, just run

```
reg y1 x1, vce(cluster indiv)
```

The main command in Stata to perform fixed effect estimation using the “within” method is `xtreg`. In the case we just saw the command would be:

```
xtreg y1 x1, fe i(indiv)
```

The same model with a quadratic time trend is obtained by doing

```
gen year2=year^2

xtreg y1 x1 year year2, fe i(indiv)
```
Year dummies are obtained automatically by doing:

```
tab year, gen(yr)
```

This creates a set of dummy variable where yr1=1 for the first year, yr2=1 for the second year, etc. For example, if we have ten years of data, we can include nine of the dummies (since there is a constant in the model):

```
xtreg y1 x1 yr2-yr10, fe i(indiv)
```

The Hausman specification test is obtained by comparing random effect and fixed effect estimates. Random effect estimates need to be stored before running fixed effect estimation. For the simplest model above, the commands are:

```
xtreg y1 x1, re i(indiv)
```

```
estimates store random_effects
```

```
xtreg y1 x1, fe i(indiv)
```

```
hausman . random_effects
```

Note that in the “hausman” procedure in Stata, we always have to list the consistent and then the efficient estimator. The “.” just means we are using the latest estimates (fixed effects) as the consistent estimate. An equivalent way of performing the test is:

```
xtreg y1 x1, fe i(indiv)
```

```
estimates store fixed_effects
```

```
xtreg y1 x1, re i(indiv)
```

```
hausman fixed_effects .
```