Panel data models

What are Panel Data?

Panel data (also known as longitudinal or cross-sectional time-series data) are repeated measures on entities over time. In other words, panel data have both a cross-sectional and a time series dimension, where (a relatively large number of) cross-section entities are observed during a (relatively short) period of time. The entities could be, e.g., countries, companies, universities or individuals. A **balanced panel** requires that all entities are present in all time periods. An **unbalanced panel** is a dataset where entities are observed a different number of times. A balanced panel is ideal but this is not always the case because of missing values, however most panel data regression models can be used for unbalanced datasets.

Panel data allow for richer models and estimation methods that cross-sectional data. The presence of multiple observations on the same entity across time allows the identification of causal effects under weaker assumptions compared to cross-sectional data. For instance, with panel data we know the time-ordering of events and thus we can investigate how an event (e.g. a policy treatment) changes the outcome (e.g. the performance of a specific entity, such as a company or a university). Furthermore, some panel data estimators (see below for a discussion) allow to control for variables that cannot be directly observed or measured (like cultural factors when entities are countries) but do not change across time (**time-fixed unobserved heterogeneity**). Finally, panel data allow to define more sophisticated specifications, by including lagged dependent variable among regressors (**dynamic panel data models**). For example, the level of investment in physical assets of a company at time *t* can depend on whether the company has already purchased physical assets at time *t-1*.

However, the presence of cross-sectional entities observed across multiple periods of time create some issues when estimating standard errors using standard Ordinary Least Squares (OLS) regression models. This is because each additional year of data is **not independent** of previous years. Some specific estimation methods have been therefore developed to deal with the specific nature of panel data. We focus here on Fixed Effects (FE) and Random Effects (RE) models. These models are typically used in specifications that do not include lagged dependent variables among regressors. In this latter case, strict exogeneity of the regressors no longer holds and more sophisticated estimation methods based on instrumental variables (e.g. GMM, see references for details) are more appropriate.

Fixed Effects

FE models are appropriate when researchers are interested in **analyzing the impact of variables that vary over time**, while taking into account for omitted variable bias due to the presence time-fixed unobserved heterogeneity. A simple FE model for data observed for entities (*i*) across time (*t*) can be described by the following equation:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}.$$

Where:

- Y_{it} is the dependent variable;
- X_{it} is the independent variable;
- β_1 is the coefficient for the independent variable. The interpretation of this coefficient would be: “for a given entity, as X varies across time by one unit, Y increases or decreases by β_1 units”;
- α_i is the unobserved entity-specific time-constant error term. It is possibly correlated with X_{it};
- u_{it} is the error term. This is assumed to be uncorrelated with X_{it}.

The key insight of FE models is that, if the unobserved variable α_i does not change over time, then any changes in the dependent variable must be due to influences other than these fixed characteristics. In other words, a FE model can consistently estimate β_1 for time-varying X_{it}, even if the unobserved variable α_i is correlated with X_{it}. FE estimators indeed eliminate the fixed characteristics α_i through mean-differencing or first-differencing the equation described above. This way **FE control for all time invariant differences between the entities**, so the estimated coefficients of the model cannot be biased because of omitted time-invariant characteristics.

A major limitation of FE is that it is not possible to estimate the effects of variables whose values do not change across time (because they are partialled out). Furthermore, it is worth pointing out that FE do not allow to control for omitted variables that are not time-invariant, so that a critical assumption is that the error term u_{it} must be uncorrelated with

X_{it}. Finally, if there is little variation across time, the standard errors from FE models tend to be large (FE are not efficient).

Random Effects

The rationale behind RE model is that, unlike the FE model, the variation across entities is assumed to be random and uncorrelated with the independent variables included in the model: *"…the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not" [Green, 2008, p.183]"*.

If there are reasons to believe that differences across entities have some influence on the dependent variable then the RE model is a more appropriate than the FE model. An advantage of RE is that it is possible to **estimate the effects of time-invariant variables**. A basic equation for RE model is as follows:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}.$$

Where:

- Y_{it} is the dependent variable;
- X_{it} is the independent variable. It must be exogenous;
- β_1 is the coefficient for the independent variable;
- β_0 is the intercept. Contrary to FE, it can be estimated;
- α_i is the unobserved entity-specific time-constant error term. Contrary to FE, it is assumed to uncorrelated with X_{it};
- u_{it} is the error term. Like FE, this is assumed to be uncorrelated with X_{it}.

RE assume that all error terms (α_i and u_{it}) are not correlated with the independent variable X_{it}, which allows for time-invariant variables to play a role as explanatory variables. In RE it is therefore important to specify all entity-specific characteristics that may influence the dependent variable. The problem with this is that some variables may not be available therefore leading to **omitted variable bias** in the model.

The choice between Fixed Effects and Random Effects

To sum up, under the assumption of no omitted variables (or omitted variables that are uncorrelated with the explanatory variables that are in the model), a RE model is probably more appropriate. It will produce unbiased estimates of the coefficients, use all the data available, and produce the smallest standard errors. More likely, however, is that omitted variables will produce at least some bias in the estimates. To decide between FE or RE it is possible to run a Hausman test where the null hypothesis is that the preferred model is RE vs. the alternative FE (see Green, 2008, chapter 9). It basically tests whether errors are correlated with the regressors (the null hypothesis is they are not). Under the null hypothesis of no correlation, there should be no difference between the two estimators, implying that both FE and RE are consistent, but FE is inefficient. So RE should be preferred to FE.

References

Comprehensive panel texts
Baltagi, B.H. (1995, 2001), Econometric Analysis of Panel Data, 1st and 2nd editions, New York, John Wiley.
Hsiao, C. (1986, 2003), Analysis of Panel Data, 1st and 2nd editions, Cambridge, UK, Cambridge University Press.

More selective advanced panel texts
Arellano, M. (2003), Panel Data Econometrics, Oxford, Oxford University Press.
Lee, M.-J. (2002), Panel Data Econometrics: Methods-of-Moments and Limited Dependent Variables, San Diego, Academic Press.

Texts with several chapters on panel data
Cameron, A.C. and P.K. Trivedi (2005), Microeconometrics: Methods and Applications, New York, Cambridge University Press.
Greene, W.H. (2008), Econometric Analysis, sixth edition, Upper Saddle River, NJ, Prentice-Hall.
Wooldridge, J.M. (2002), Econometric Analysis of Cross Section and Panel Data, Cambridge, MA, MIT Press.

Appendix

1 Some useful commands in STATA

Before using panel data commands in STATA it is required to set the software to handle panel data by using the command *xtset* (In this case “ID” represents the entities and “year” represents the time variable):

```
xtset ID year
```

Here some command for panel summary:

xtdescribe: extent to which panel is unbalanced;

xtsum: separate within (over time) and between (over individuals) variation;

xtdata: scatterplots for within and between variation;

xtline: time series plot for each individual on one chart;

xttab: tabulations within and between for discrete data e.g. binary;

xttrans: transition frequencies for discrete data.

The command to run FE is:

```
xtreg y x1 x2, fe
```

The command to run RE is :

```
xtreg y x1 x2, re
```

Procedure to perform the Hausman test:

1) Run a fixed effects model and save the estimates:

```
xtreg y x1, fe

estimates store fixed
```

2) Then run a random model and save the estimates:

```
xtreg y x1, re

estimates store random
```

3) Then perform the test (If the null hypothesis is rejected (p-value<0.05, ie significant) use FE):

```
hausman fixed random
```

2 Example: Understanding the impact of VC financing on firm growth

Sample

VICO dataset available on the RISIS platform. VICO contains geographical, industry and accounting information on companies founded starting from 1/1/1988, which have received at least one venture capital or angel investment starting from 1/1/1998, operating in seven European countries (Belgium, Finland, France, Germany, Italy, Spain, and the United Kingdom) and Israel. It refers to 17,863 companies, 7,834 distinct investors, of which 6,182 Venture Capitalists and 1,511 Business Angels for a total number of observations concerning investment-level data (i.e. all the company-investor-round dyads) equals to 52,657.

Goal

What about the growth of VC backed companies before and after receipt of a first round of VC?

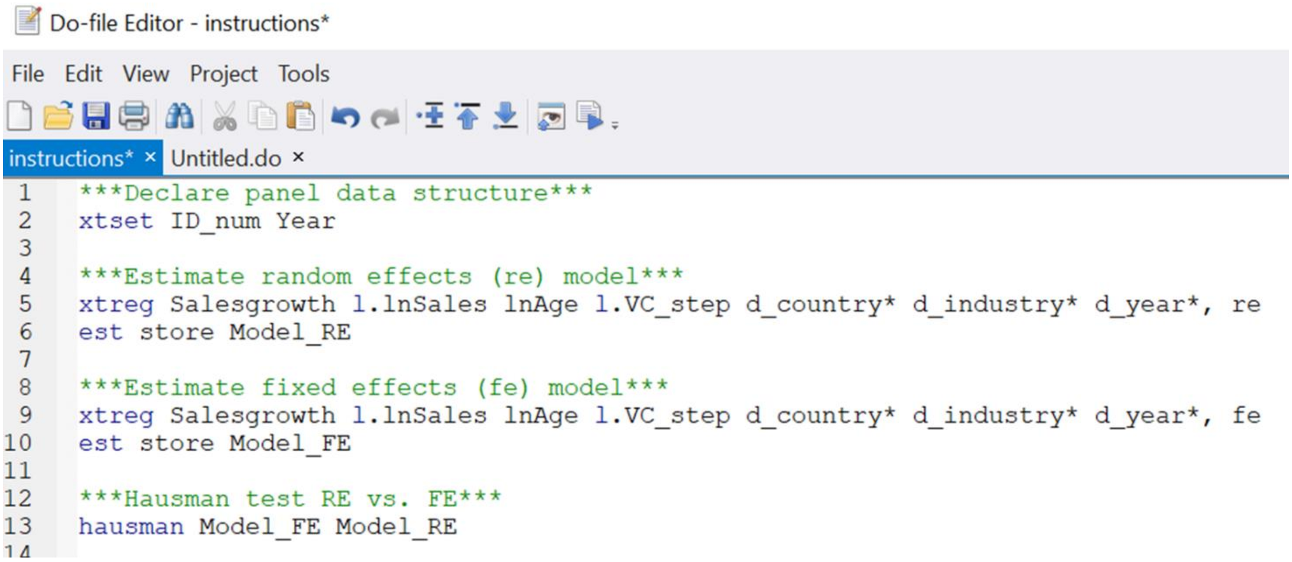
Econometric specification

$$LogSalesGrowth_{i,t} = \alpha + \beta_1\, LogSales_{i,t-1} + \beta_2\, VC_{i,t-1} + \text{Controls} + \varepsilon_{i,t}.$$

Where:

- $LogSalesGrowth_{i,t}$ is the logarithmic sales growth between time $t-1$ and t (i.e., $Log\ Sales_{i,t} - Log\ Sales_{i,t-1}$);
- $LogSales_{i,t-1}$ is the logarithm of the firm size at time t–1;
- $VC_{i,t-1}$ is a step dummy variable that indicate the VC status (i.e., the variable switches from 0 to 1 in the year following the first VC round).

STATA instructions



```
***Declare panel data structure***
xtset ID_num Year

***Estimate random effects (re) model***
xtreg Salesgrowth l.lnSales lnAge l.VC_step d_country* d_industry* d_year*, re
est store Model_RE

***Estimate fixed effects (fe) model***
xtreg Salesgrowth l.lnSales lnAge l.VC_step d_country* d_industry* d_year*, fe
est store Model_FE

***Hausman test RE vs. FE***
hausman Model_FE Model_RE
```