Chapter 14: Advanced Panel Data Methods

Introductory Econometrics: A Modern Approach

Outline

Fixed effects

- Intuition
- Assumptions
- Properties
- Fixed effects or first difference?

Random effects

- Assumption

Hausman test

Applying panel data methods to other data structures

Introduction

We will want to estimate the parameters of the model:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + a_i + u_{it}$$

We will see two aditional methods:

1) Fixed effects: Differences a_i away

2) Random effects: Puts a_i into the error term

Fixed Effects (FE)

$$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$$

- a_i fixed effect (does not change over time), potentially correlated with explanatory variables

Form time-averages for each individual

$$\bar{y}_i = \beta_1 \bar{x}_{i1} + \ldots + \beta_k \bar{x}_{ik} + \bar{a}_i + \bar{u}_i$$

$$[y_{it} - \bar{y}_i] = \beta_1 [x_{it1} - \bar{x}_{i1}] + \ldots + \beta_k [x_{itk} - \bar{x}_{ik}] + [u_{it} - \bar{u}_i]$$

- Because $a_i - \bar{a}_i = 0$ (the fixed effect is removed)

Estimate time-demeaned equation by OLS

- Uses time variation within cross-sectional units (= within estimator)

Fixed Effects

This is equivalent to a two steps process:

1) Average equation across each cross section i:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \ldots + \bar{u}_i$$

2) Subtract average from each observation:

$$y_{it} - \bar{y}_i = \beta_1 \left(x_{1it} - \bar{x}_{1i}\right) + \beta_2 \left(x_{2it} - \bar{x}_{2i}\right) + \ldots + u_{it} - \bar{u}_i$$

This process is called **time demeaning** or **within transformation**.

The resulting model above is usually written as:

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{1it} + \beta_2 \ddot{x}_{2it} + \ldots + \ddot{u}_{it}$$

Fixed effects using dummy variables

The fixed effects estimator is equivalent to introducing a dummy for each individual in the original regression and using pooled OLS:

$$y_{it} = a_1 ind1_{it} + a_2 ind2_{it} + \ldots + a_N indN_{it} + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + u_{it}$$

- where $ind1_{it}$ is 1 if observation stems from first individual and 0 otherwise

After fixed effects estimation, the fixed effects can be estimated as:

$$\hat{a}_i = \bar{y}_i - \widehat{\beta}_1 \bar{x}_{i1} - \ldots - \widehat{\beta}_k \bar{x}_{ik}$$

Fixed Effects Assumptions

The applicability of FE depends on a series of assumptions:

- **Strict exogeneity of all $\mathbf{x}$** : Same as in first differences (FD), the error term must not have an effect on future realizations of any x.
- **Error terms serially uncorrelated**: This is also a strong assumption, meaning that u_{it} has no influence on u_{it+1}. Remember that the error term includes unobserved variables that change over time, but also that their fixed component is included in a_i.
- **Error term has to be homoskedastic**.

Fixed Effects Properties

The FE estimator has similar properties to those of FD:

- All time constant variables are eliminated.
- Cross sectional variance is eliminated, need for sufficient variance in changes over time.
- Time constant variables can still be interacted with variables that change over time (e.g. year dummies).
- If a full set of year dummies is included, one cannot estimate the effect of any variable whose change across time is constant (e.g. experience, age).
- The estimator is very sensitive to measurement error.

Example: Effect of training grants on firm scrap rate

$$scrap_{it} = \beta_1 d88_{it} + \beta_2 d89_{it} + \beta_3 grant_{it} + \beta_4 grant_{it-1} + a_i + u_{it}$$

- Time-invariant reasons why one firm is more productive than another are controlled for.
- The important point is that these may be correlated with the other explanatory variables.

Fixed-effects estimation using the years 1987, 1988, and 1989:

$$\widehat{scrap}^*_{it} = -\underset{(.109)}{.080}\; d88^*_{it} - \underset{(.133)}{.247}\; d89^*_{it} - \underset{(.151)}{.252}\; grant^*_{it} - \underset{(.210)}{.422}\; grant^*_{it-1}$$

Stars denote time-demeaning

$$n = 162, R^2 = .201$$

Training grants significantly improve productivity (with a time lag)

Fixed Effects of First Difference?

The two estimators are very similar. The most important results are:

- When $T = 2$ the two estimators are exactly identical! But FD is computationally easier.
- When $\mathrm{T} > 2$ the following observations are important:
  - When the u_{it} are serially uncorrelated FE is preferred.
  - When the Δu_{it} are serially uncorrelated FD is preferred. $\Rightarrow$ Unfortunately, only one of these requirements can be fulfilled, but both can be simultaneously violated. $\Rightarrow$ In any case, methods do exist to correct for autocorrelation in the error term.
  - Bias induced by measurement error or a violation of the strict exogeneity assumption may decreases with T in FE, but not FD. This result depends on the assumption that the process u_{it} is weakly dependent.

Discussion of Fixed Effects estimator

- Strict exogeneity in the original model has to be assumed.
- The R^2 of the demeaned equation is inappropriate.
- The effect of time-invariant variables cannot be estimated.
- The effect of interactions with time-invariant variables can be estimated (e.g. the interaction of education with time dummies).
- If a full set of time dummies are included, the effect of variables whose change over time is constant cannot be estimated (e.g. experience).
- Degrees of freedom have to be adjusted because the N time averages are estimated in addition (resulting degrees of freedom = NT-N-k).

Unbalanced panels

An unbalanced panel is when not all cross-sectional units have the same number of observations.

- Dropping units with only one time period does not cause bias or inconsistency.

Fixed effects (FE) or First Differencing (FD) with unbalanced panels

- FE will preserve more data than FD when we have unbalanced panels, since FD requires that each observation have data available for both t and t-1.

--

- For example, consider a scenario in which we have seven years of data, but data is missing for all even numbered years. Thus, we observe t=1,3,5,7.
    - FE will use time periods 1,3,5,7
    - FD will lose all observations.

Random Effects (RE)

The crucial assumption in RE is that all the x_{it} are uncorrelated with a_i :

$$\mathrm{cov}\left(x_{jti}, a_i\right) = 0 \qquad \forall \quad j$$

If this is the case a_i can simply be put into the error term:

$$\nu_{it} = a_i + u_{it}$$

Since a_i is constant across t the composite error term ν_{it} is serially correlated across time!

This model has to be estimated by Generalized Least Squares (GLS)

Random Effects (RE) models

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}$$

- The individual effect a_i is assumed to be "random" (i.e. completely unrelated to explanatory variables)

Random effects assumption:

$$\text{Cov}\left(x_{itj}, a_i\right) = 0, j = 1, 2, \ldots, k$$

The composite error a_i + u_{it} is uncorrelated with the explanatory variables but it is serially correlated for observations coming from the same i:

For example, in a wage equation, for a given individual the same unobserved ability appears in the error term of each period. Error terms are thus correlated across periods for this individual.

$$\text{Cov}(a_i + u_{it}, a_i + u_{is}) = \text{Cov}(a_i, a_i) = \sigma_a^2$$

Example: Wage equation using panel data

$$\widehat{\log}(wage_{it}) = \underset{(.011)}{.092}\ educ_{it} - \underset{(.048)}{.139}\ black_{it} + \underset{(.043)}{.022}\ hispan_{it}$$

$$+ \underset{(.015)}{.106}\ exper_{it} - \underset{(.0007)}{.0047}\ exper_{it}^2 + \underset{(.017)}{.064}\ married_{it}$$

$$+ \underset{(.018)}{.106}\ union_{it} + time\ dummies$$

Random effects is used because many of the variables are time-invariant. But is the random effects assumption realistic?

Random effects or fixed effects?

- In economics, unobserved individual effects are seldomly uncorrelated with explanatory variables so that fixed effects is more convincing.

Hausman test

One can use FE and RE, and then formally test for statistically significant differences in the coefficients on the time varying explanatory variables

- If the Hausman test statistic is *not significantly different* from zero, then both the FE and RE estimators are consistent. RE estimator should be used because it is more efficient.
- If the Hausman test statistic is *significantly different* from zero, then only the FE estimator is consistent and should be used.

```r
library(plm)
data(jtrain)
# Fixed effects within estimator using plm package
model_fe <- plm(formula = lscrap ~ tothrs + d88 + d89 + grant + grant_1,
              data = jtrain,
              index = c("fcode", "year"), # c(group index, time index)
              model = "within", effect = "individual")

# Fixed effects within estimator using feols package
model_fe_2<-feols( lscrap ~ tothrs + d88 + d89 + grant + grant_1 | year + fcode , data = jtrain)

# summary
msummary(list(model_fe,model_fe_2), stars = TRUE)
```

	Model 1	Model 2
tothrs	−0.005	−0.005
	(0.003)	(0.002)
d88	−0.075	
	(0.121)	
d89	−0.218	
	(0.156)	
grant	−0.118	−0.118
	(0.181)	(0.068)
grant_1	−0.410+	−0.410+
	(0.228)	(0.125)
Num.Obs.	146	146
R2	0.229	0.922
R2 Adj.	−0.216	0.877
R2 Within		0.061
R2 Pseudo		

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```r
# Random effects estimator
# model_re <- update(model_ols, model = "random", random.method = "walhus")
model_re <- plm(formula = lscrap ~ tothrs + d88 + d89 + grant + grant_1,
                data = jtrain,
                index = c("fcode", "year"), # c(group index, time index)
                model = "random", random.method = "walhus")
summary(model_re)
```

	Model 1
(Intercept)	0.664**
	(0.217)
tothrs	−0.005+
	(0.003)
d88	−0.092
	(0.119)
d89	−0.248
	(0.152)
grant	−0.074
	(0.175)
grant_1	−0.355
	(0.220)
Num.Obs.	146
R2	0.162
R2 Adj.	0.133
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Hausman test

```
# The Hausman test is used to decide whether to use fixed effects or random effects.
# H0: FE coefficients are not significantly different from the RE coefficients
# Ha: FE coefficients are significantly different from the RE coefficients

# Hausman test for fixed versus random effects
phtest(model_fe, model_re)
```

Hausman test

```
##
##      Hausman Test
##
## data:  lscrap ~ tothrs + d88 + d89 + grant + grant_1
## chisq = 1.225, df = 5, p-value = 0.9425
## alternative hypothesis: one model is inconsistent
```