The basic panel data model

The basic panel data model is given by:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \ldots + \beta_K x_{K,it} + a_i + u_{it}, \qquad (18)$$

with idiosyncratic errors u_{it} having following important property which is essential for unbiasedness of FE and FD estimation:

- $\mathrm{E}(u_{it}|\mathbf{X}_i, a_i) = 0$, which implies $\mathrm{E}(\Delta u_{it}|\mathbf{X}_i) = 0$

Exclude the constant β_0 to allow the a_is to be unconstrained.

The basic panel data model

An important distinction is made concerning dependence between a_i and the observed covariates:

- Fixed-effects model:
  a_i could be correlated with any covariates $x_{j,it}$ for some or all t for any $j = 1, \dots, K$; i.e. $\mathrm{Cov}(a_i, x_{j,it})$ might be different from 0;

- Random-effects model:
  a_i is uncorrelated with all covariates $x_{j,it}$ for all $t = 1, \dots, T$ for any $j = 1, \dots, K$; i.e. $\mathrm{Cov}(a_i, x_{j,it}) = 0$.

Estimation of fixed effect models - FE

Two-step procedure: averaging over time yields

$$\bar{y}_{i\cdot} = \beta_1 \bar{x}_{1,i\cdot} + \beta_2 \bar{x}_{2,i\cdot} + \ldots + \beta_K \bar{x}_{K,i\cdot} + a_i + \bar{u}_{i\cdot}.$$

Allows estimation of a_i, if $\beta_1, \ldots, \beta_K$ are known.

Substracting these time averages from (18) yields a regression model without a_i, which allows OLS estimation of $\beta_1, \ldots, \beta_K$:

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

where $\ddot{x}_{j,it} = x_{j,it} - \bar{x}_{j,i\cdot}$, and similarly for $\ddot{y}_{it}$ and $\ddot{u}_{it}$.

Estimation of fixed effect models - FE

Comments:

- Two step-procedure does not work, if any covariate x_{it} is time-invariant for all units i.
- Standard error of the fixed effects are underestimated, because $\beta_1, \ldots, \beta_K$ are held fixed at the estimated values.

Alternative - joint OLS estimation:

- Include a dummy variable for each unit in (18) and estimate a large regression model with parameter vector $(\beta_1, \beta_2, \ldots, \beta_K, a_1, \ldots, a_N)$.

Estimation of fixed effect models - FD

Consider first differences between t and $t-1$ as before and use OLS estimation:

$$\Delta y_{it} = \beta_1 \Delta x_{1,it} + \beta_2 \Delta x_{2,it} + \ldots + \beta_K \Delta x_{K,it} + \Delta u_{it}. \qquad (19)$$

Comments:

- Does not work, if any covariate x_{it} is time-invariant for all units i.
- No estimates for a_i are available.

Assumptions for FE estimation

Under following assumptions on u_{it}, FE estimation is BLUE and therefore more efficient than FD estimation:

- Homoskedasticity: $\mathrm{Var}\,(u_{it}|\mathbf{X}_i, a_i) = \sigma^2$ for all units $(i = 1, \ldots, N)$ and for all $t = 1, \ldots, T$;

- No serial correlation: the idiosyncratic errors are uncorrelated (conditional on all explanatory variables and the unobserved effect a_i): $\mathrm{Cov}(u_{it}, u_{is}|\mathbf{X}_i, a_i) = 0$, for all $s \neq t$, $s, t = 1, \ldots, T$.

If the idiosyncratic errors follow in addition a normal distribution, then t- and F-statistics have exact t- and F-distributions (p-values are correct).

Assumptions for FE estimation

The condition of no correlation among u_{it} implies that the differences Δu_{it} of the idiosyncratic errors in the FD equation (19) follow an MA(1)-Process,

$$\Delta u_{it} = -1 \cdot u_{i,t-1} + u_{it},$$

with $\theta = -1$ and negative lag 1 correlation equal to $\theta/(1+\theta^2) = -0.5$, and all remaining AC are equal to 0.

Explains loss of efficiency.

Assumptions for FD estimation

Under the following assumptions on the differences Δu_{it} of the idiosyncratic errors, FD estimation is BLUE and therefore more efficient than FE estimation:

- Homoskedasticity: $\mathrm{Var}\left(\Delta u_{it}|\mathbf{X}_i\right) = \sigma^2$ for all units $(i = 1, \ldots, N)$ and for all $t = 1, \ldots, T$;

- No serial correlation: $\mathrm{Cov}(\Delta u_{it}, \Delta u_{is}|\mathbf{X}_i) = 0$, for all $s \neq t$, $s, t = 1, \ldots, T$.

The last condition implies that the original idiosyncratic errors follow a random walk.

FD or FE estimation?

- Run FE estimation and check autocorrelation of residuals;
- if uncorrelated, use FE; if PAC(1) close to 1, run FD;
- Run FD estimation and check autocorrelation of residuals;
- if uncorrelated, use FD; if PAC(1) high negative value, run FE.

No clear decision, if partial autocorrelation function PAC(h) does not vanish for lags $h > 1$.

Pooled estimation

OLS estimation of the regression coefficients in the basic panel data model (18) using a (pooled) regression model which ignores a_i:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \ldots + \beta_K x_{K,it} + \tilde{u_{it}}. \qquad (20)$$

- Fixed-effects - OLS may be biased: possible correlation between a_i and any covariates $x_{j,it}$ implies that $\mathrm{E}(u_{it}|\mathbf{X}_i)$ might be different from 0.

- Random-effects - OLS unbiased: $\mathrm{Cov}(a_i, x_{j,it}) = 0$ implies $\mathrm{E}(u_{it}|\mathbf{X}_i) = 0$ for all i, t, j. However, standard errors may be wrong, because the composite error $\tilde{u_{it}} = a_i + u_{it}$ is serially correlated across time.

Assumptions for Random effects Estimation

The basic assumption

$$\mathrm{E}(a_i|\mathbf{X}_i) = \beta_0,$$

rules out correlation between the unobserved effect a_i and the explanatory variables and implies $\mathrm{Cov}(a_i, x_{j,it}) = 0$.

What assumptions are needed to obtain correct standard errors? Same assumption as for FE estimation, namely

- Homoskedasticity: $\mathrm{Var}\left(u_{it}|\mathbf{X}_i, a_i\right) = \sigma_u^2$ for all i, t;
- No serial correlation: $\mathrm{Cov}(u_{it}, u_{is}|\mathbf{X}_i, a_i) = 0$, for all $s \neq t$.

Assumptions for Random effects Estimation

Additional assumption: the variance of the unobserved effect a_i given all explanatory variables is constant,

$$\mathrm{Var}\,(a_i|\mathbf{X}_i) = \sigma_a^2.$$

In this case, the correlation between the composite errors $\tilde{u_{it}}$ and $\tilde{u_{is}}$, $t \neq s$, is given by:

$$\mathrm{Corr}(u_{it}, u_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}.$$

Compound symmetry: all off-diagonal elements of the correlation matrix of $(u_{i1}, \ldots, u_{iT})$ are the same!

Assumptions for Random effects Estimation

For each unit, the random effects model could be written as a system of equations with correlated errors, or in matrix notation

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{u}_i, \qquad \mathrm{Cov}(\boldsymbol{u}_i) = \mathbf{R},$$

where $\mathbf{R}$ has compound symmetry (as has the inverse matrix).

GLS (Generalized least square):

- transform this system to a new system with uncorrelated errors and estimate $\beta_1, \beta_2, \ldots, \beta_K$ from this system;
- yields correct standard errors for $\beta_1, \beta_2, \ldots, \beta_K$; two-stage estimation of the unobserved effects $a_1, \ldots, a_N$.

Assumptions for Random effects Estimation

GLS estimation corresponds to pooled OLS estimator of $\beta_1, \ldots, \beta_K$ in the following, quasi-demeaned equation,

$$y_{it} - \lambda \bar{y}_{i\cdot} = \beta_1 (x_{j,it} - \lambda \bar{x}_{j,i\cdot}) + \ldots + \beta_K (x_{j,it} - \lambda \bar{x}_{j,i\cdot}) + \ddot{u}_{it}.$$

λ depends on σ_a^2 and σ_u^2:

$$\lambda = 1 - \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}.$$

Assumptions for Random effects Estimation

- $\lambda = 1$: FE estimation; $\lambda = 0$: pooled OLS estimation.

- If λ is close to 1, GLS is close to FE estimation (e.g. large T, large variance σ_a^2 of unobserved effects relative to error variance σ_u^2).

- If λ is close to 0, GLS is close to pooled OLS estimation (e.g. small variance σ_a^2 of unobserved effects relative to error variance σ_u^2).

Assumptions for Random effects Estimation

Additional assumption: the unobserved effect a_i given all explanatory variables follows a normal distribution,

$$a_i \sim \text{Normal}\left(\beta_0, \sigma_a^2\right).$$

Use ML (maximum likelihood) estimation for the system

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{u}_i, \qquad \boldsymbol{u}_i \sim \text{Normal}\left(\mathbf{0}, \mathbf{R}\right).$$

Allows joint estimation of $\beta_1, \ldots, \beta_K, \sigma_u^2$ and the parameters β_0 and σ_a^2 of the distribution of unobserved heterogeneity.

Case Study: Investment equation

Small balanced sample: $N = 10$ manufacturing firms, $T = 20$ years (Baltagi, 2001).

Investment equation:

$$I_{it} = \alpha + \beta_1 F_{it} + \beta_2 C_{it} + u_{it},$$

with I_{it} real gross investment of a firm i in year t, F_i real value of the firm (shares outstanding), C_{it} real value of the capital stock.

Eviews Workfile: `baltagi-gruenfeld-panel.wf1`

Further Issues

Baltagi, B. H. (2001): Econometric Analysis of Panel Data (2 ed.). Chichester: Wiley.

- Testing various specifications and model selection: quite involved.
- It is possible to correct for heteroskedasticity of the idiosyncratic errors, e.g. unit specific variances $\mathrm{Var}\,(u_{it}|\mathbf{X}_i) = \sigma_i^2$.
- It is possible to correct for serial correlation of the idiosyncratic errors, i.e. $\mathrm{Cov}(u_{it}, u_{is}|\mathbf{X}_i) \neq 0$, for all $s \neq t$.
- It is possible to allow for heterogeneity across units and across time for selected regression coefficients β_j.