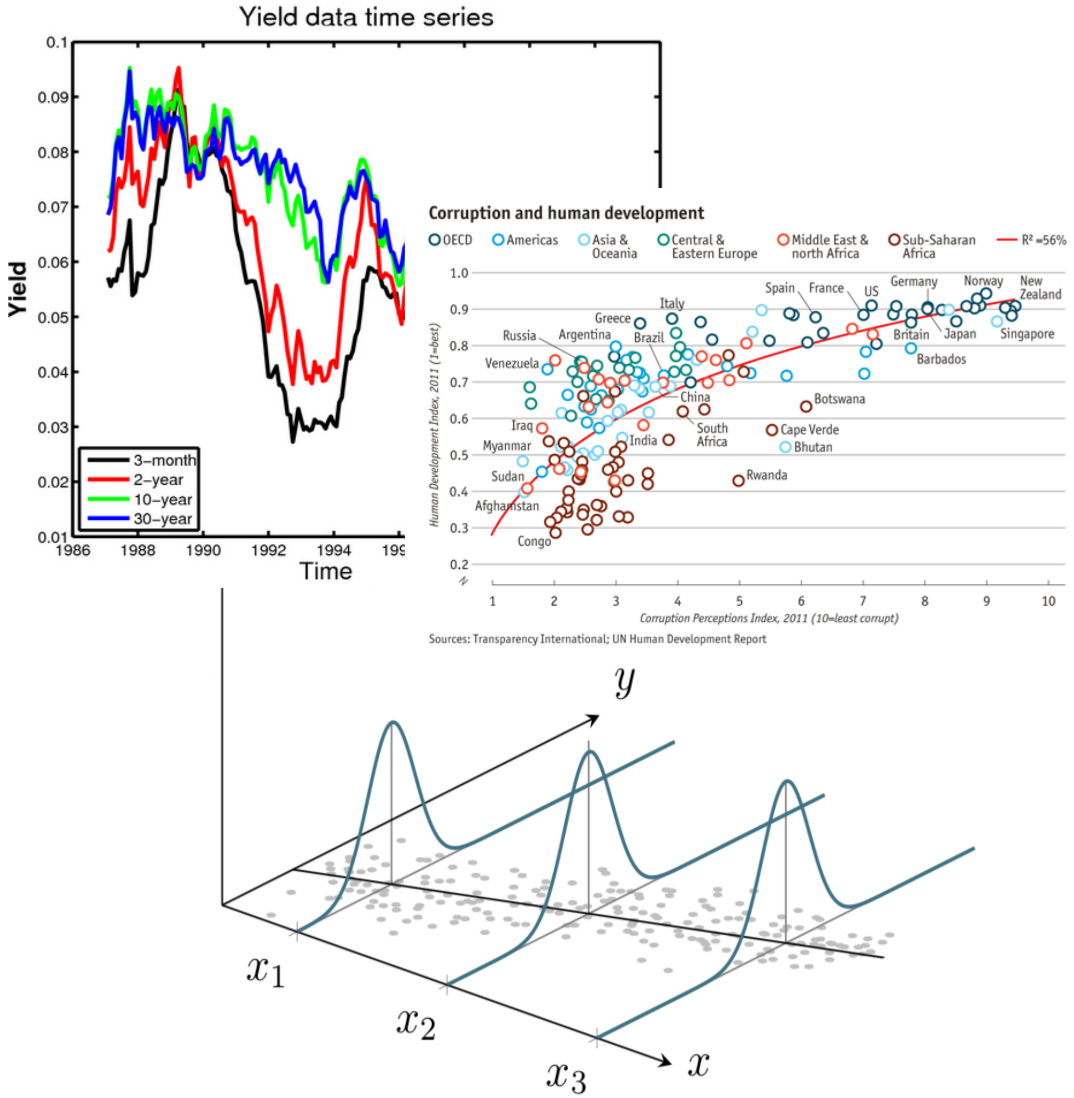
Chapter 14

Advanced Panel Data Methods



Advanced Panel Data Methods (1 of 15)

- **Fixed effects estimation**

Fixed effect, potentially correlated with explanatory variables

$$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \; i = 1, \ldots, N, t = 1, \ldots, T$$

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

Form time-averages for each individual

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

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)

Advanced Panel Data Methods (2 of 15)

- **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)

Advanced Panel Data Methods (3 of 15)

- **Discussion of fixed effects estimator**
  - Strict exogeneity in the original model has to be assumed.
  - The R-squared 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).

Advanced Panel Data Methods (4 of 15)

- **Interpretation of fixed effects as dummy variable regression**
  - The fixed effects estimator is equivalent to introducing a dummy for each individual in the original regression and using pooled OLS:

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

For example, = 1 if the observation stems from individual N, = 0 otherwise

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

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - \ldots - \hat{\beta}_k \bar{x}_{ik}, \; i = 1, \ldots, N$$

Estimated individual effect for individual i

Advanced Panel Data Methods (5 of 15)

- **Fixed effects or first differencing?**
  - Remember that first differencing can also be used if $T > 2$.
  - In the case $T = 2$, fixed effects and first differencing are identical.
  - For $T > 2$, fixed effects is more efficient if classical assumptions hold.
  - First differencing may be better in the case of severe serial correlation in the errors, for example if the errors follow a random walk.
  - If T is very large (and N not so large), the panel has a pronounced time series character and problems such as strong dependence arise.
  - In these cases, it is probably better to use first differencing.
  - Otherwise, it is a good idea to compute both and check robustness.

Advanced Panel Data Methods (6 of 15)

- **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.

Advanced Panel Data Methods (7 of 15)

- **Random effects (RE) models**

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

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

<u>Random effects assumption</u>: $Cov(x_{itj}, a_i) = 0, j = 1,2, \dots, k$

- 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:

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

Under the assumption that idiosyncratic errors are serially uncorrelated

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.

Advanced Panel Data Methods (8 of 15)

- **Estimation in the random effects model**
  - Under the random effects assumptions explanatory variables are exogenous so that pooled OLS provides consistent estimates.
  - If OLS is used, standard errors have to be adjusted for the fact that errors are correlated over time for given i (= clustered standard errors).
  - But, because of the serial correlation, OLS is not efficient.
  - One can transform the model so that it satisfies the GM-assumptions:

$$[y_{it} - \lambda \bar{y}_i] = \beta_1 [x_{it1} - \lambda \bar{x}_{i1}] + \ldots + \beta_k [x_{itk} - \lambda \bar{x}_{ik}]$$ ← Quasi-demeaned data

$$+ [a_i - \lambda \bar{a}_i + u_{it} - \lambda \bar{u}_i]$$ ← Error can be shown to satisfy GM-assumptions

Advanced Panel Data Methods (9 of 15)

- **Estimation in the random effects model (cont.)**

$$\lambda = 1 - \left[\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)\right]^{1/2}, \ 0 \leq \lambda \leq 1$$

  - The quasi-demeaning parameter is unknown but it can be estimated.
  - FGLS using the estimated λ is called random effects estimation.
  - If the random effect is relatively unimportant compared to the idosyncratic error, FGLS will be close to pooled OLS (because λ goes to 0).
  - If the random effect is relatively important compared to the idiosyn-cratic term, FGLS will be similar to fixed effects (because λ goes to 1).
  - Random effects estimation can be used to estimate the effect of time-invariant variables.

Advanced Panel Data Methods (10 of 15)

- **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.

Advanced Panel Data Methods (11 of 15)

- **Correlated Random Effects (CRE)**
  - When using CRE to choose between FE and RE, we must include any time-constant variables that appear in RE estimation:

$$y_{it} = \alpha_1 + \alpha_2 d2_t + \cdots + \alpha_T dT_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk}$$
$$+\gamma_1 \bar{x}_{i1} + \cdots + \gamma_k \bar{x}_{ik} + \delta_1 z_{i1} + \cdots + \delta_m z_{im} + r_i + u_{it}$$

- Estimating this equation by RE (or even just pooled OLS) yields:

$$\hat{\beta}_{CRE,j} = \hat{\beta}_{FE,j};\ j = 1, \dots, k \longleftarrow$$ Time varying estimates will be the same as in FE
$$\hat{\alpha}_{CRE,t} = \hat{\alpha}_{FE,t};\ t = 1, \dots, T$$

$$H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0 \longleftarrow$$ Under the null, RE is sufficient. If we reject the null, then FE is preferred.

  - An advantage of CRE is that it allows for estimation of the effects of time-constant explanatory variables, not possible using FE.

Advanced Panel Data Methods (12 of 15)

- **General policy analysis with panel data**
  - The two-period, before-after setting is a special case of a more general policy analysis framework when T ≥ 2.

$$y_{it} = \eta_1 + \alpha_2 d2_t + \cdots + \alpha_T dT_t + \beta w_{it} + \boldsymbol{x_{it}\psi} + a_i + u_{it}$$

w_{it} is the binary policy variable and β estimates the average treatment effect of the policy

  - To allow w_{it} to be systematically related to the unobserved fixed effect a_i, we estimate the regression with either FD or FE, using cluster-robust standard errors.
  - We can also include lags of the policy intervention: w_{it-1}, w_{it-2},…

Advanced Panel Data Methods (13 of 15)

- **Testing for feedback from the error term to the policy variable**
- We need to be careful if the policy variable w_{it} it reacts to past shocks.
  - Example: y_{it} is the poverty rate and w_{it} is some measure of government assistance.
  - A large shock to the poverty rate in year t could prompt an increase in government assistance the following year.

- If we have at least three time periods, we can test for feedback

$$y_{it} = \eta_1 + \alpha_2 d2_t + \cdots + \alpha_{T-1} dT - 1_t + \beta w_{it} + \delta w_{it+1} + \boldsymbol{x_{it}\psi} + a_i + u_{it}$$

Estimate with FE and compute a cluster robust t-statistic for $\hat{\delta}$

  - This is known as a “falsification test.”
    - If the forward policy variable is statistically significant, there is potential feedback from the error term to the policy variable.

Advanced Panel Data Methods

- **The heterogeneous trend model**
- What if time trends are unique across individuals?

$$y_{it} = \eta_1 + \alpha_2 d2_t + \cdots + \alpha_T dT_t + \beta w_{it} + \boldsymbol{x}_{it}\boldsymbol{\psi} + a_i + g_i t + u_{it}$$

The new term $g_i t$ is a unit-specific time trend.

- This allows the policy intervention to not only be correlated with level differences among units (captured by a_i), but also by trend differences.
- We can estimate this model by taking first differences:

$$\Delta y_{it} = \alpha_2 \Delta d2_t + \cdots + \alpha_T \Delta dT_t + \beta \Delta w_{it} + \boldsymbol{\Delta x}_{it}\boldsymbol{\psi} + g_i + \Delta u_{it}$$

Estimate by FE., though we need to ensure we have T ≥ 3

Advanced Panel Data Methods (15 of 15)

- **Applying panel data methods to other data structures**
  - Panel data methods can be used in other contexts where constant unobserved effects have to be removed.
- Example: Wage equations for twins

Unobserved genetic and family characteristics that do not vary across twins

$$\log(wage_{i1}) = \beta_0 + \beta_1 educ_{i1} + \ldots + a_i + u_{i1}$$

Equation for twin 1 in family i

$$\log(wage_{i2}) = \beta_0 + \beta_1 educ_{i2} + \ldots + a_i + u_{i2}$$

Equation for twin 2 in family i

$$\Rightarrow \quad \Delta \log(wage_i) = \beta_1 \Delta educ_i + \ldots + \Delta u_i$$

Estimate differenced equation by OLS