Pooling Cross Sections across Time

Lecture 2/1 EE426 - 2/2014

Chayanee Chawanote

Pooling Cross Sections Across Time

2 kinds of data sets:

- Independently pooled cross section: sampling randomly from a larger population at different points in time
  - Observations are not identically distributed.

- Panel data or longitudinal data: collecting the same individuals, families, firms, cities, states, etc, across time
  - Can't assume that observations are independently distributed across time: there are time-constant, unobserved attributes of the units being studied.

Year dummy variables

- The population may have different distributions in different time period. We need the intercept to differ across periods >> including year dummy variables
  - The coefficients on year dummy variables represent changes in the dependent variable for reasons that are not captured in the explanatory variables.
  - Compared to a based year
- We can interact a year dummy variable with key explanatory variables to see if the effect of that variable has changed over a certain time period.
  - Without interactions, the effect of each explanatory variables has remained constant
  - What happens if we interact all independent variables with a year dummy variable?

Year dummy variables

- Interactions:

  $\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85{*}educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85{*}female + u$

  - Intercept for 1985 = $\beta_0 + \delta_0$
  - Return to education in 1978, and in 1985 = β_1, and $\beta_1 + \delta_1$
  - The gender gap in log(wage) in 1985 = $\beta_5 + \delta_5$
  - Test that nothing has happened to the gender differential over this 7-year period? >> H_0: $\delta_5 = 0$

- Cautions:
  - Monetary term needs to adjust to real value since we compare across years
  - But if using logarithmic form, real or nominal wage only affects the coefficient on the year dummy, y85

Chow test for structural change across time

- Chow test: whether a multiple regression differs across 2 groups

$$F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{n - 2(k+1)}{k+1}$$

- Alternative Chow test for 2 periods: interacting each variable with a year dummy for one of the two years and testing for joint significance of the year dummy and all of the interaction terms
- If more than 2 periods?
  1. Estimate the restricted model by pooled regression allowing for different time intercepts >> give SSR_r
  2. Run a regression for each of T time periods >> get SSR_{ur}

     $SSR_{ur} = SSR_1 + SSR_2 + \ldots + SSR_T$
  3. If there are k explanatory variables (not including the intercept or time dummies), we are testing (T-1)k restrictions, and there are T + Tk parameters in the unrestricted model

$$F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \cdot \frac{n - T - Tk}{(T-1)k}$$

Policy analysis with pooled cross sections

- Evaluating the impact of a certain event or policy with the data collected before and after the occurrence of an event
- A natural experiment occurs when some exogenous event (a change in government policy) changes the environment in which individuals, families, firms operate.
  - control group (C): not affected by the policy change
  - treatment group (T): affected by the policy change
  - These two groups are not randomly assigned. To control for systematic differences between the two groups, we need 2 years of data: before and after the change
- Let dT = 1 if in the treatment group, 0 otherwise

  Let d2 = 1 if post-policy change time period, 0 otherwise

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + other\ factors$$

Policy analysis with pooled cross sections

- Without other factors in the regression, $\hat{\delta}_1$ will be the difference-in-differences estimator:

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C})$$

- Sometimes, $\hat{\delta}_1$ is called the "average treatment effect", measuring the effect of the treatment or policy on the average outcome of y

	Before	After	After - Before
Control	β_0	$\beta_0+\delta_0$	δ_0
Treatment	$\beta_0+\beta_1$	$\beta_0+\delta_0+\beta_1+\delta_1$	$\delta_0+\delta_1$
Treatment - Control	β_1	$\beta_1+\delta_1$	δ_1

Two-period panel data analysis

- Panel data can be used to address some kinds of omitted variable bias
- Think of the omitted variables as being fixed over time, then we can model 'a composite error'

$$v_{it} = a_i + u_{it}$$

- Suppose the population model is

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$$

- If a_i is correlated with x, OLS will be biased since we have a_i as a part of the composite error
- With panel data, we can difference-out the unobserved fixed effect, a_i, then apply OLS

Two-period panel data analysis

- View the unobserved factors affecting the dependent variable as: 1) constant; 2) vary over time

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$$

  - d2 = 1 when t = 2 and does not change across *i*
  - a_i: all unobserved, time-constant factors that affect y_{it} => unobserved effect, or fixed effect, or unobserved heterogeneity
  - u_{it}: idiosyncratic error or time varying-error, representing unobserved factors that change over time and affect y_{it}

- If we pool the 2-year data and use OLS (no a_i), it will result in heterogeneity bias, caused from omitted a time-constant variable.
- Reason for collecting panel data: allow for the unobserved effect, a_i, to be correlated with the explanatory variables. E.g., allow the unmeasured city factors in a_i that affect the crime rate to be correlated with the unemployment rate.

Two-period panel data analysis

- First-differenced (FD) estimator:

$$\Delta y_{it} = \delta_0 + \beta_1 \Delta x_{it} + \Delta u_{it}$$

  - using 2 periods data results in a cross-sectional regression of the differenced data.

- The differencing used to eliminate a_i can greatly reduce the variation in the explanatory variables.
- Potential problem occurs when the key explanatory variables do not vary much over time.

Differencing with more than 2 periods

- Suppose we have N individuals and T = 3 (total no. of obs = 3N). A general fixed effects model is

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + a_i + u_{it}$$

- If the unobserved effect a_i is correlated with any of explanatory variables, then using pooled OLS results in biased and inconsistent estimates
- Key assumption: the idiosyncratic errors are uncorrelated with the explanatory variable in each time period:

$$Cov(x_{jit}, u_{it}) = 0 \quad \text{for all t, s and j}$$

  - This rules out cases where future explanatory variables react to current changes in the idiosyncratic errors.
  - If we have omitted an important time-varying variable, then the key assumption is violated.

Differencing with more than 2 periods

- If a_i is correlated with x_{jit}, then x_{jit} will be correlated with the composite error, $v_{it} = a_i + u_{it}$. We can eliminate a_i by differencing adjacent periods:

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{1it} + \ldots + \beta_k \Delta x_{kit} + \Delta u_{it} \quad \text{for t} = 2, 3$$

- Then, apply OLS with a requirement that Δu_{it} is uncorrelated with Δx_{jit} for all j and $t = 2, 3$
- There will be no intercept. It is better to estimate the 1st-differenced equation with an intercept and a single time period dummy
- For balanced panel (same T for each of N), after first differencing, the equation is

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \ldots + \alpha_T dT_T + \beta_1 \Delta x_{1it} + \ldots + \beta_k \Delta x_{kit} + \Delta u_{it} \quad \text{for t} = 2, 3, \ldots, \text{T}$$

  - We have T-1 time periods, total number of obs is N(T-1)

Assumptions for pooled OLS using First Differences

- FD.1: For each I, the model is

$$y_{it} = \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + a_i + u_{it}, \quad t = 1, \ldots, T$$

- FD.2: A random sample from the cross section
- FD.3: Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables
- FD.4: For each t, the expected value of the idiosyncratic error given the explanatory variables in all time periods and the unobserved effect is zero: $E(u_{it}|\boldsymbol{X}_i, a_i) = 0$, $\boldsymbol{X}_i$ contains x_{jit}, $t = 1, \ldots, T$; $j = 1, \ldots, k$
  - x_{jit} are strictly exogenous conditional on the unobserved effect.
  - implication of FD.4: $E(\Delta u_{it}|\boldsymbol{X}_i) = 0$, t = 2,…,T
  - FD estimator is also consistent with a fixed T and as N → ∞

Assumptions for pooled OLS using First Differences

- Next 2 assumptions ensure that the standard errors and test statistics resulting from pooled OLS on the FD are valid
- FD.5: The variance of the differenced errors, conditional on all explanatory variables, is constant (homoskedasticity)

$$\text{Var}(\Delta u_{it} | \boldsymbol{X}_i) = \sigma^2, \text{ t} = 2,\ldots,\text{T}$$

- FD.6: For all t ≠ s, the differences in the idiosyncratic errors are uncorrelated (serially uncorrelated)

$$\text{Cov}(\Delta u_{it}, \Delta u_{is} | \boldsymbol{X}_i) = 0, \text{ t} \neq \text{s}$$

  - FD.1-6: FD estimator β_j is the best linear unbiased estimator (conditional on the explanatory variables).
- FD.7: Conditional on $\boldsymbol{X}_i$, the Δu_{it} are independently and identically distributed normal random variables
  - to let t and F statistics from FD have exact t and F distributions