Econometrics

Week 5

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Fall 2022

Recommended Reading

For today

- Pooling cross-sections across time. Simple panel data methods
- Chapter 13

For next week

- Advanced Panel Data Methods
- Chapter 14

Panel Data vs. Pooled Cross-Sections

Until now, we have covered multiple regression analysis using:

- Pure cross-sectional data - each observation represents an individual, firm, country, etc.
- Pure time series data - each observation represents a separate time period.

In economic applications we often observe data which have both these dimensions - we may observe cross-sections from several time periods.

- We will talk about 2 types of such data:
  - **Independently pooled cross-sections**.
  - **Panel data** (also called **longitudinal data**).
- We will discuss statistical characteristics of such data...
- ... and introduce some useful estimation methods

What is the basic intuition?

Cross-section vs. Panel / Pooled Cross-section Data

$$y_{it} = \beta_0 + \beta_1 x_{1it} + u_{it} \qquad u_{it} \sim N(0, \sigma^2)$$

$$i = 1, 2, \ldots, N$$

$$t = 1, 2, \ldots, T$$

But let as introduce the intuition step by step first...

Panel Data vs. Pooled Cross-Sections

Independently Pooled (Repeated) Cross-Sections

- These are cross-sections drawn from the same population *independently* each year
- At each period, the sample is different!
- Each quarter, statistical offices of EU Member States independently sample the population for a Labor Force Survey.

Panel Data

- We have panel data when we observe a single cross-sectional sample for several time periods
- Sample does not change!
- We have observations for each individual/unit with temporal ordering.
- Since 1979, the National Longitudinal Survey of Youth surveys a sample of Americans born in 1957 - 1964.

Pooling Independent Cross-Sections Across Time

- If a random sample is drawn at each time period, resulting data are **independently pooled cross-sections.**
- Such data *can be analyzed as cross-sections*, but also allow for more interesting analysis
- Reasons for pooling cross-sections:
  - To increase sample size $\Rightarrow$ more precise estimates.
  - To investigate how a relationship changes in time.

Pooling Independent Cross-Sections Across Time

Let us consider a simple regression model estimated on cross-sections collected in year 1 ($t = 1$) and year 2 ($t = 2$):

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

We may assume that parameters remain constant:

$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$ for $t = 1$
$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$ for $t = 2$

We may want to investigate whether the mean changes in time

$y_{it} = \beta_{01} + \beta_1 x_{it} + u_{it}$ for $t = 1$
$y_{it} = \beta_{02} + \beta_1 x_{it} + u_{it}$ for $t = 2$

We may want to investigate whether relations change in time

$y_{it} = \beta_0 + \beta_{11} x_{it} + u_{it}$ for $t = 1$
$y_{it} = \beta_0 + \beta_{12} x_{it} + u_{it}$ for $t = 1$

Pooling Independent Cross-Sections Across Time

Let us define a dummy variable distinguishing between time periods:

$$t2_{it} = \begin{cases} 0, & \text{if } t = 1 \\ 1, & \text{if } t = 2 \end{cases}$$

We may assume that parameters remain constant:

$$y_{it} = \beta_0 + \beta_1 \cdot x_{it} + u_{it}$$

We may want to investigate whether the mean changes in time

$$y_{it} = \beta_0 + \delta_0 \cdot t2_{it} + \beta_1 \cdot x_{it} + u_{it}$$

We may want to investigate whether relations change in time

$$y_{it} = \beta_0 + \beta_1 \cdot x_{it} + \delta_1 \cdot x_{it} t2_{it} + u_{it}$$

Pooling Independent Cross-Sections Across Time

Let us use a pooled dataset (cross-section from 2019 and cross-section from 2022) to find out if the gender wage gap changed between 2019 and 2022.

Example: Changes in the Gender Wage Gap

$$log(wage_{it}) = \beta_0 + \delta_0 y2022_t + \beta_1 female_i + \delta_1 y2022_t * female_i + \beta_2 college_{it} + u_{it}$$

- $y2022$ is a dummy equal to 1 if observation is from 2022 and zero if it comes from 2019.
  - The intercept for 2019 is β_0
  - The intercept for 2022 is $\beta_0 + \delta_0$
  - The gender wage gap in 2019 is β_1
  - The gender wage gap in 2022 is $\beta_1 + \delta_1$
  - δ_1 measures the 15-year change in the gender wage gap.
- We can test the null hypothesis that the gender gap has not changed $H_0 : \delta_1 = 0$ against the alternative that the gap has been reduced, $H_A : \delta_1 > 0$

The Chow Test for Structural Change

- Do you remember the Chow test used to test whether the regression model differs between groups?
- The same test can be used to test whether the model differs across time periods!
- Test if the slope coefficient is constant over time:

$$y_{it} = \beta_0 + \beta_{1t} x_{it} + u_{it} \quad T = 2$$

$$H_0 : \beta_{11} = \beta_{12} \quad H_A : \beta_{11} \neq \beta_{12}$$

- Run the regression for each time period separately and obtain SSR_{ur} as the sum of each regression SSR: $SSR_{ur} = SSR_1 + SSR_2$
- Run the pooled regression with one common slope coefficient and obtain SSR_r.
- Compute simple F test: $\frac{SSR_r - SSR_{ur}}{SSR_{ur}} \cdot \frac{(n-T-Tk)}{(T-1)k} \sim F((T-1)k; n-T-Tk)$
- Unfortunately the test is not robust to heteroskedasticity! Better to use the approach from previous slide.

Policy Analysis with Pooled cross-sections

- Example: Did elections speed up spread of the covid-19 pandemics?
- In early October 2020, when the 2nd wave (actually the 1st real wave in the Czech Republic) of covid-19 was taking off, Senate elections were taking place in the Czech Republic.
- Did these elections cause an increase in new covid cases?
- This example is based on the research article: 'Do elections accelerate the COVID-19 pandemic?'.

Policy Analysis with Pooled cross-sections - elections and covid-19 spread

- Consider a simple regression using district-level data from the period before elections (September 2020) and after elections (October 2020):

$$cases_{it} = \beta_0 + \beta_1 October_t + \mathbf{\Gamma} X_{it} + u_{it}$$

  where $cases_{it}$ is number of new covid cases per 100thousand citizens in district i in month t, and X_{it} is a vector of district-level characteristics such as population density, share of college graduates, or unemployment rate.
- The hypothesis we want to test is whether elections cause an increase in new covid cases.:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 > 0$$

- Unfortunately, rejecting the null **does not** really imply that elections did speed-up spread of the virus. Why?
- Spurious regression: pandemic might have been speeding up anyway, even without elections.

Policy Analysis with Pooled cross-sections - elections and covid-19 spread

Every two years 1/3 of Senate mandates are renewed $\Rightarrow$ elections were taking place in 1/3 of Czech voting-districts.

- We can compare covid spread in the affected group (i.e. in districts were elections took place) with covid spread in an unaffected group (i.e. in districts where elections did not take place).
- The assumption is that the behavior of both groups would be evolving, on average, the same in the absence of elections.
- In other words, if there were no elections, covid cases would change at the same rate in voting and not voting districts.
- Then, any difference in the number of new covid cases between October 2020 and September 2020 between the voting and not voting districts would be because of the elections.
- Such approach is called **difference-in-differences**.

Policy Analysis with Pooled cross-sections - elections and covid-19 spread

- How to introduce the control group to the regression?
- Before the elections (September 2020), data from all election districts:
  $cases_i = \beta_{0,Sept} + \beta_{2,Sept} elections_i + \mathbf{\Gamma} X_{it} + u_i$
- After the elections (October 2020), data from all election districts: $cases_i = \beta_{0,Oct} + \beta_{2,Oct} elections_i + \mathbf{\Gamma} X_{it} + u_i$
- The change in β_2 from September to October can be interpreted as the effect of the elections.
- How to put this into a single regression? $cases_{it} = \beta_0 + \beta_1 October_t + \beta_2 elections_i + \beta_3 October_t \cdot elections_i + \Gamma X_{it} + u_{it}$
- Which coefficient captures the effect of the reform?

Introduction of the Waiting Period

Control variables X disregarded for simplicity:

$$cases_{it} = \beta_0 + \beta_1 October_t + \beta_2 elections_i + \beta_3 October_t \cdot elections_i + u_{it}$$

- $E[cases|elections, September] =$
- $E[cases|noelections, September] =$
- The difference:

$$E[\Delta cases_{September}] =$$

- $E[cases|elections, October] =$
- $E[cases|noelections, October] =$
- The difference:

$$E[\Delta cases_{October}] =$$

- The difference:
- β_3 reflects the effect of elections.

Panel Data

- For a cross-section of individuals, schools, firms, cities, etc., we have several periods of data.
- Data are not independent, as in pooled cross-sections, the same individuals are observed in each time period.
- This means we might face similar problems as with time series data! e.g. autocorrelation
- This also means we can take advantage of panel structure of the data and use it to solve some kinds of omitted variable bias.
- To see this, let us write a model capturing the panel structure of the data

General Model for Panel Data

Unobserved Effects Model (Fixed Effects Model)

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \underbrace{a_i + u_{it}}_{\nu_{it}}$$

- ν_{it} is the **composite error**. It consists of
  - Time-invariant, individual specific, **unobserved effect** a_i
  - Time and individual specific **idiosyncratic error** u_{it}
- a_i is also referred to as **unobserved heterogeneity**, or individual heterogeneity, or **fixed effect**, because it is fixed over time.

Note that some variables are time variant and some time invariant (don't have the t-index)

Panel Data

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \underbrace{a_i + u_{it}}_{\nu_{it}}$$

- It is tempting to estimate this model by pooled OLS, but...
- It will be inefficient if errors are serially correlated
- It will be biased and inconsistent if u_{it} and x_{it} are correlated $\Rightarrow$ this is an endogeneity bias that can be met also in cross-sectional or time-series models
- It will be biased and inconsistent if a_i and x_{it} are correlated: $Cov(a_i, x_{it}) \neq 0 \Rightarrow$ **heterogeneity bias**.
- In real-world applications, the main reason for collecting panel data is to deal with the unobserved effect a_i that is correlated with explanatory variables (example)
- Simple solution follows...

First-differenced estimator

- ...Simple, because a_i is constant over time.

First-differenced estimator

- ...Simple, because a_i is constant over time.

First-differenced estimator (FD)

Let us start with two-period panel data.

$$\begin{aligned} y_{i2} &= (\beta_0 + \delta_0) + \beta_1 x_{i2} + \beta_2 z_i + a_i + u_{i2}, \quad (t = 2) \\ y_{i1} &= \beta_0 + \beta_1 x_{i1} + \beta_2 z_i + a_i + u_{i1}, \quad (t = 1) \end{aligned}$$

Subtracting second equation from the first one gives:

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- Here, a_i is *"differenced away"*.
- Note that as a side effect z_i is also "differenced away"
- Can we estimate this equation by OLS and get a reliable estimate of β_1?

First-differenced estimator

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \underbrace{a_i + u_{it}}_{\nu_{it}}$$

$$\downarrow$$

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- Differencing is a powerful way to deal with time constant unobserved effects
- However, first-differencing can greatly reduce variation in the explanatory variables, and
- First-differencing removes observed time-constant variables from the regression.
- OLS estimates of parameters in the first-differenced equation are unbiased as long as the following assumptions are satisfied:

Assumptions for Pooled OLS Using First Differences

Assumption FD1

For each observation i, the model is
$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \qquad t = 1, \ldots, T,$
where parameters β_j are to be estimated and a_i is the unobserved fixed effect.

Assumption FD2

Each period we observe the same random sample.

Assumption FD3

Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables.

Assumptions for Pooled OLS Using First Differences

Assumption FD4

Let $\mathbf{X}_i$ denote x_{itj}, $t = 1, \ldots, T$, $j = 1, \ldots, k$. For each t, the expected value of the idiosyncratic error given the explanatory variables in *all* time periods and the unobserved effect is zero: $E(u_{it}|\mathbf{X}_i, a_i) = 0$.

An important implication of FD4 is that $E(\Delta u_{it}|\mathbf{X}_i) = 0$, $t = 2, \ldots, T$. Once we control for a_i, there is no correlation between the x_{isj} and the remaining error u_{it} for all s and t. x_{itj} is strictly exogenous conditional on the unobserved effect.

- Under assumptions FD1 - FD4, the first-difference estimator is unbiased.

Assumptions for Pooled OLS Using First Differences

Assumption FD5

The variance of the differenced error, conditional on all explanatory variables, is constant: $Var(\Delta u_{it}|\mathbf{X}_i) = \sigma^2$, for all $t = 2, \ldots, T$.

Assumption FD6

For all $t \neq s$, the differences in the idiosyncratic errors are uncorrelated (conditional on all explanatory variables): $Cov(\Delta u_{it}, \Delta u_{is}|\mathbf{X}_i) = 0$, $t \neq s$.

- Under assumptions FD1 - FD6, the first-difference estimator is BLUE.

Assumptions for Pooled OLS Using First Differences

Assumption FD7

Conditional on $\mathbf{X}_i$, the Δu_{it} are independent and identically distributed normal random variables.

- This last assumptions assures that FD estimator is normally distributed, t and F statistics from the pooled OLS on the differenced data have exact t and F distributions.

Differencing with More than Two Periods

- We can extend FD to more than two periods.
- We simply difference adjacent periods.

A general fixed effects model for N individuals and t=1,2,3

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$$

- The total number of observations is 3N.
- The key assumption is that idiosyncratic errors are uncorrelated with explanatory variables: $Cov(x_{it}, u_{is}) = 0$ for all t, s and $i \Rightarrow$ **strict exogeneity**.
- How to estimate? Simply difference equation for $t = 1$ from $t = 2$ and $t = 2$ from $t = 3$.
- It will result in 2 equations which can be estimated by pooled OLS consistently under the CLM assumptions.
- We can simply further extend to T periods.
- Correlation and heteroskedasticity are treated in the same way as in time series data.

Thank you

Thank you very much for your attention!