For the today

- Pooling cross sections across time. Simple panel data methods.
- Chapter 13 (pp. 408 – 434).

For the next week

- Advanced Panel Data Methods.
- Chapter 14 (pp. 441 – 460).
Panel Data vs. Pooled Cross Sections

Up until now, we covered multiple regression analysis using:

- Pure cross-sectional data – each observation represents an individual, firm, etc.
- Pure time series data – each observation represents a separate period.
- In an economic applications, we often have data which have both these dimensions – we may have cross sections for different time periods.
- We will talk about 2 types of pooled data:
  - Independently pooled cross sections
  - Panel data (sometimes called longitudinal data)
- Panel data are not independently distributed across time as pooled cross sections!
- We will cover basics to introduce the methods.
Panel Data vs. Pooled Cross Sections

Pooled Cross Sections

- Population surveys - each period, Statistical Bureau independently samples the population
- \(\Rightarrow\) At each period, the sample is different.
- these are independent cross-sections, but we can take time into account.

Panel Data

- Each year, the European Community Household Panel surveys the same individuals on a range of questions: income, health, education, employment, etc.
- These are cross-sections with time order.
- We have observations for each individual with temporal ordering.
- Sample does not change.
What is the basic intuition?

Time Series vs. Panel Data

\[ y_{it} = \beta_0 + \beta_1 x_{it1} + u_{it} \quad u_{it} \sim N(0, \Sigma) \]

\[ t = 1, 2, \ldots, n \]

\[ i = 1, 2, \ldots, N \]

cross-sections

But let us introduce the intuition step by step first...
Pooling Independent Cross Sections Across Time

- If a random sample is drawn at each time period, resulting data are **independently pooled cross sections**.
- Reasons for pooling cross sections:
  - To increase sample size \( \Rightarrow \) more precise estimators.
  - To investigate the effect of time (simply add dummy variable)

We may assume that parameters remain constant

\[
wages_{it} = \beta_0 + \beta_1 educ_{it} + u_{it}
\]

Alternatively, we may want to investigate the effect of time

\[
wages_{it} = \beta_0t + \beta_1 educ_{it} + u_{it}
\]

Or, we may want to investigate whether relations change in time

\[
wages_{it} = \beta_0 + \beta_1 t educ_{it} + u_{it}
\]
Pooling Independent Cross Sections Across Time cont.

Let’s find out if an hourly wage pooled across the years 1978 and 1985 was dependent on an education and gender gap.

**Example: Changes in the Return to Education**

\[
\log(\text{wage}) = \beta_0 + \delta_0 y_{85} + \beta_1 \text{educ} + \delta_1 y_{85}\text{educ} + \beta_2 \text{female} + u
\]

- \(y_{85}\) is a dummy equal to 1 if observation is from 1985 and zero if it comes from 1978 (the number of observations is different + different people in the sample).
- Thus the intercept for 1978 is \(\beta_0\) and intercept for 1985 is \(\beta_0 + \delta_0\).
- The return to education in 1978 is \(\beta_1\) and the return to education in 1985 is \(\beta_1 + \delta_1\).
- \(\delta_1\) measures the 7-year change. We can test the null hypothesis that nothing has changed over the period \(H_0 : \delta_1 = 0\) to alternative that the effect has been reduced, \(H_0 : \delta_1 > 0\).
The Chow Test for Structural Change

We may want to determine if regression function differs across 2 groups

\[ y = \alpha_1 + \beta_1 x + u_1 \text{ vs. } y = \alpha_2 + \beta_2 x + u_2 \]

Clearly, \( \alpha_1 \neq \alpha_2 \text{ and } \beta_1 \neq \beta_2 \)

Resulting \( y = \alpha_0 + \beta_0 x + u_0 \) in the worst fit.
The Chow Test for Structural Change cont.

- We may want to test if there are 2 (or more) periods in a regression:
  
  \[ y_{it} = \beta_0 t + \beta_1 t x_{it} + u_{it} \]

- It may not be as obvious as from the example on previous slide.

- \( H_0 : \beta_{01} = \beta_{02}, \beta_{11} = \beta_{12}. \)

- Compute simple F test.

- Alternatively, use pooled cross-sections as in previous example with wages.
Example: What effect will building a garbage incinerator (in Czech “spalovna”) have on housing prices? (Wooldridge book, Ex.13.3.)

Consider only simple one-period case of 1981 data:

\[ \text{prices} = \beta_0 + \beta_1 \text{near} + u \]

where \text{near} is dummy variable (1 for houses near incinerator, 0 otherwise)

The hypothesis is that prices of houses near the incinerator fall when it is built:

\[ H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 < 0 \]

Unfortunately, this test does not really imply that building incinerator is causing the lower prices.

\( \hat{\beta}_1 \) is inconsistent as \( \text{cov(\text{near}, u)} \neq 0 \).
Let’s include also another year into the analysis so we can really study the impact.

\[ t = \{1978, 1981\} \text{ now:} \]

\[ \text{prices}_{it} = \beta_0 + \beta_1 \text{near}_{it} + \beta_3 D_{it}^{1981} + \beta_4 \text{near}_{it} D_{it}^{1981} u_{it} \]

The hypothesis is the same, that location of incinerator will decrease the prices, but:

\[ H_0 : \beta_4 = 0 \quad \text{vs.} \quad H_A : \beta_4 < 0 \]

\( \hat{\beta}_4 \) is called difference-in-differences estimator.
Let’s look at the logics of the **difference-in-differences** estimator.

- \( E[prices|\text{near}, 1981] = \beta_0 + \beta_1 + \beta_3 + \beta_4 \)
- \( E[prices|\text{near}, 1978] = \beta_0 + \beta_1 \)
- The difference:
  \[
  E[\Delta prices_{\text{near}}] = E[prices_{1981} - prices_{1987}|\text{near}] = \beta_3 + \beta_4
  \]
- \( E[prices|\text{far}, 1981] = \beta_0 + \beta_3 \)
- \( E[prices|\text{far}, 1978] = \beta_0 \)
- The difference:
  \[
  E[\Delta prices_{\text{far}}] = E[prices_{1981} - prices_{1987}|\text{far}] = \beta_3
  \]
- The difference: \( E[\Delta prices_{\text{near}} - \Delta prices_{\text{far}}] = \beta_4 \)
- \( \beta_4 \) reflects the policy effect if the incinerator is built with no “different inflation”.
Two-period Panel Data

- For a cross-section of individuals, schools, firms, cities, etc., we have two time periods of data.
- Data are not independent as in pooled cross-sections, so they can be more helpful.
- It can be used to address some kinds of omitted variable bias.

Fixed Effects Model (Unobserved Effects Model)

\[ y_{it} = \beta_0 + \beta x_{it} + a_i + u_{it} \]

- \( a_i \) is time-invariant, individual specific, \textit{unobserved effect} on the level of \( y_{it} \).
- \( a_i \) is referred to as \textbf{fixed effect} – fixed over time.
- \( a_i \) is referred to as \textbf{unobserved heterogeneity}, or individual heterogeneity.
- \( u_{it} \) is \textit{idiosyncratic error}. 

Two-period Panel Data cont.

- We can rewrite the model as:

\[ y_{it} = \beta_0 + \beta x_{it} + \nu_{it}, \]

where \( \nu_{it} = a_i + u_{it} \)

- \( \nu_{it} \) is also called composite error.

- We can simply estimate this model by pooled OLS.

- But it will be biased and inconsistent if \( a_i \) and \( x_{it} \) are correlated: \( \text{cov}(a_i, x_{it}) \neq 0 \Rightarrow \text{heterogeneity bias} \).

- In real-world applications, the main reason for collecting the panel data is to allow for the unobserved effect \( a_i \) to be correlated with explanatory variables.

- I.e. we want to explain crime, and allow unmeasured city factors \( a_i \) affecting the crime to be correlated with i.e. unemployment rate.

- Simple solution follows...
Two-period Panel Data cont.

- This is simple to solve as $a_i$ is constant over time.
- **Solution:** first - differenced estimator

Let’s write

\begin{align*}
    y_{i2} &= \beta_0 + \beta x_{i2} + a_i + u_{i2}, \quad (t = 2) \\
    y_{i1} &= \beta_0 + \beta x_{i1} + a_i + u_{i1}, \quad (t = 1)
\end{align*}

Subtracting second equation from the first one gives:

$$
\Delta y_i = \beta \Delta x_i + \Delta u_i
$$

- Here, $a_i$ is “differenced away”.
- And we have standard cross-sectional equation.
- If $\text{cov}(\Delta u_i, \Delta x_i) = 0$, $\hat{\beta}_{FD}$ is consistent (strict exogeneity).
Differencing two years is a powerful way to control unobserved effects.

If we use standard cross-sections instead, it may suffer from omitted variables.

But, it is often very difficult to collect panel data (i.e., for individuals), as we have to have the same sample.

Moreover, $a_i$ can greatly reduce the variation in the explanatory variables.

Still, this is solution only when we have 2 data periods (more general next lecture).

Organization of Panel Data is crucial (more during the seminars).
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=3.

\[ y_{it} = \delta_1 + \delta_2 d_{2t} + \delta_3 d_{3t} + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \]

- The total number of observations is 3N.
- The key assumption is that idiosyncratic errors are uncorrelated with explanatory variables: \( \text{cov}(x_{itj}, u_{is}) = 0 \) for all \( t, s \) and \( j \Rightarrow \text{strict exogeneity} \).
- 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.
Assumptions for Pooled OLS Using First Differences

Assumptions revisited

**Assumption FD1**
For each $i$, the model is
\[ y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \quad t = 1, \ldots, T, \]
where parameters $\beta_j$ are to be estimated and $a_i$ is the unobserved effect.

**Assumption FD2**
We have a random sample from the cross section.

**Assumption FD3**
Let $X_i$ 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:
\[ E(u_{it}|X_i, a_i) = 0. \]
Assumptions for Pooled OLS Using First Differences cont.

An important implication of Ass. FD3 is that $E(\Delta u_{it}|X_i) = 0$, $t = 2, \ldots, T$. Once we control for $a_i$, there is no correlation between the $x_{isj}$ and remaining error $u_{it}$ for all $s$ and $t$. $x_{itj}$ is strictly exogenous conditional on the unobserved effect.

Assumption FD4: No Perfect Collinearity

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

Assumption FD5: Homoskedasticity

The variance of the differenced error, conditional on all explanatory variables, is constant: $Var(\Delta u_{it}|X_i) = \sigma^2$, for all $t = 2, \ldots, T$. 
Assumptions for Pooled OLS Using First Differences cont.

**Assumption FD6: No Serial Correlation**

For all $t \neq s$, the differences in the idiosyncratic errors are uncorrelated (conditional on all explanatory variables):

$$\text{Cov}(\Delta u_{it}, \Delta u_{is}|X_i) = 0, \ t \neq s.$$ 

- Under the Ass. FD1 – FD4, the first-difference estimators are unbiased.
- Under the Ass. FD1 – FD6, the first-difference estimators are BLUE.

**Assumption FD7: Normality**

Conditional on $X_i$, the $\Delta u_{it}$ are independent and identically distributed normal random variables.

- Last assumptions assures us that FD estimator is normally distributed, $t$ and $F$ statistics from the pooled OLS on the differenced data have exact $t$ and $F$ distributions.
Thank you

Thank you very much for your attention!