# Econometrics 2

## Panel Data Models

Sergey Lychagin

CEU

Winter 2020

# Upcoming topics

- Panel data, first differences, fixed effects estimator
- Difference-in-Differences designs

# Omitted variables formula

If the *true* model is

$$Y_i = X_i'\beta + \varepsilon_i$$

and $E\left[X_i\varepsilon_i\right] \neq 0$, then we know that the population regression gives

$$\begin{aligned} \beta^{OLS} &= \beta + \pi \\ \pi &= E\left[X_i X_i'\right]^{-1} E\left[X_i\varepsilon_i\right] \end{aligned}$$

$\pi$ is the vector of regression coefficients when we regress $\varepsilon_i$ on $X_i$ (which is not practically possible)

# First Approach: observed covariates $Z$

True causal model

$$Y_i = X_i'\beta + \varepsilon_i$$

in which

- We know that $Y_i$ depends on observed $X_i$'s and unobserved $\varepsilon_i$
- Not all the observed covariates $Z$ belong in the model
- We know something about $\varepsilon_i$, e.g. $\varepsilon_i = u_i + w_i$, where $w_i$ — idiosyncratic shock (i.e., unrelated to $X$).

Possible Approaches

- Use the $Z's$ to "control for" $u_i$ (proxy)
- Use the $Z's$ to get the part $X_i$ that is unrelated to $u_i$ (IV)

# Second Approach: Panel data

We make assumptions on the nature of $\varepsilon_i$ and how it is related to $X_i$

Second approach: Panel data setting with units $i = 1, ..., N$ and periods $t = 1, ..., T$

$$\begin{aligned} Y_{it} &= X_{it}'\beta + \varepsilon_{it} \\ \varepsilon_{it} &= u_i + w_{it} \\ E\left[X_{it}\varepsilon_{it}\right] &= E\left[X_{it}u_i\right] + E\left[X_{it}w_{it}\right] \end{aligned}$$

It might be reasonable to assume $E\left[X_{it}w_{it}\right] = 0$, the idiosyncratic part of the error term is not correlated with observable $X$.

Example (*pre-determined covariates*): $Y_{it}$ firm $i$'s log of output at $t$, $X_{it}$ — capital, labor. If inputs are hired at $t-1$ and $w_{it}$ is unexpected, $E[X_{it}w_{it}] = 0$.

# Varieties of data

- Repeated cross section
  - We draw a sample of individuals, **independently for each** $t$. We do not track individuals over time
- Panel data
  - There is a population at $t = 1, .., T$,
  - We draw a sample of individuals
  - Every individual comes with a whole history of $(Y_{it}, x_{it})$
- Balanced panel: no holes; we have data for all $(i, t)$ pairs.
- Unbalanced panel: there are holes (due to random or non-random selection).

# Repeated Cross Sectional Data

Cross-sectional data

- Random sample of the population, independent observations
- Example: German Labor Force Survey (Mikrozensus), survey of 1% of German households: about 370,000 households with 870,000 individuals

Many cross-sectional surveys are repeated over time

# Repeated Cross Sectional Data

Example: returns to education over time

- cross section $t = 2000$

$$log(w_i) = \alpha + \beta educ_i + \varepsilon_i$$

- pooled cross sections $t = 2000, 2005$

$$log(w_{it}) = \alpha_0 + \alpha_1 y05 + \beta_0 edu_{it} + \beta_1 edu_{it} \times y05 + \varepsilon_{it}$$

- pooled cross sections $t = 2000, 2005, 2010$

$$\begin{aligned} log(w_{it}) &= \alpha_0 + \alpha_1 y05 + \alpha_2 y10 \\ &+ \beta_0 edu_{it} + \beta_1 edu_{it} * y05 + \beta_2 edu_{it} * y10 + \varepsilon_{it} \end{aligned}$$

As long as $E[X_{it}\varepsilon_{it}] = 0$, OLS gives the causal effect.

# Panel Data

- Repeated observations of the same unit
- $i = 1 \dots N$, can be countries, firms, regions
- $t = 1 \dots T$ denotes time
- **Example 1**: Financial statements of Hungarian firms
  - Collected by the State Tax Authority
  - All double-entry bookkeeping firms in Hungary (150-400 thousand, depending on year).
- **Example 2**:
  - $i$: family, and $t$: twin 1 or 2
- We consider applications where $N$ is large and $T$ small ("short panels")

# Outline

- Case 1: $T = 2$
- Case 2: $T > 2$
- Fixed effects estimation

## Case 1: $T = 2$

- Example: twin data, $i$ – family, $t$ – child.
- Fixed Effects Model:

$$Y_{it} = X_{it}\beta + u_i + w_{it}$$

- $u_i$ unobserved effect, fixed effect, or unobserved heteregeneity
- $w_{it}$ idiosyncratic error
- Estimate with OLS - *Pooled Model*:

$$\begin{aligned} Y_{it} &= \beta X_{it} + \varepsilon_{it} \\ \varepsilon_{it} &= u_i + w_{it}, \text{(composite error)} \end{aligned}$$

- If we estimate this model with OLS, we need $E(Xu) = 0$ to recover the causal effect.
- This implies the assumption that $u_i$ is uncorrelated with $X_{it}$.
- E.g., $Y_{it}$ – wage, $X_{it}$ – education, $u_i$ – genetic factors. Is $X_{it}$ unrelated to $u_i$?

# Panel Estimation

Individual effect $u_i$ causes trouble. Three solutions:

1. Control for it: include a dummy for each $i$ in the model.
   - LS dummy variable estimator
   - Large number of regressors: dimension of $X$ is $NT \times (k + N)$
   - computationally demanding
2. Get rid of it by first differencing. Then, apply OLS:

$$\begin{aligned} Y_{i2} - Y_{i1} &= (X_{i2} - X_{i1})\,\beta + (w_{i2} - w_{i1}) \\ \Delta Y_{it} &= \Delta X_{it}\beta + \Delta w_{it} \end{aligned}$$

3. Get rid of it by de-meaning:

$$\begin{aligned} Y_{it} - \overline{Y}_i &= \left(X_{it} - \overline{X}_i\right)\beta + (w_{it} - \overline{w}_i) \\ \widetilde{Y}_{it} &= \widetilde{X}_{it}\beta + \widetilde{w}_{it} \end{aligned}$$

If $T = 2$, the results are numerically identical.

# Estimation assumptions in the FD model

- $(X_{i2} - X_{i1})$ has to be uncorrelated with $(w_{i2} - w_{i1})$
- This holds if $w_{it}$ is uncorrelated with both $X_{i1}$ and $X_{i2}$: *Strict Exogeneity* (strong exogeneity in some textbooks)
- $(X_{i2} - X_{i1})$ must have some variation in $i$: e.g. not all twins have same level of schooling.

Example: Control for ability bias in return to education.

$$\begin{aligned} Y_{it} &= \alpha + X_{it}\beta + S_{it}\delta + A_i + \epsilon_{it} \\ \Delta Y_{it} &= \Delta X_{it}\beta + \Delta S_{it}\delta + \Delta\epsilon_{it} \end{aligned}$$

Usually there is no variation in $\Delta S_{it}$ in a sample of adult workers: cannot estimate $\delta$ with precision/at all.

# Case 2: $T > 2$

$$Y_{it} = X_{it}\beta + \delta_t + u_i + w_{it}$$

*First differenced* model with $(T-1)$ equations:

$$Y_{it} - Y_{it-1} = (X_{it} - X_{it-1})\beta + (\delta_t - \delta_{t-1}) + (w_{it} - w_{it-1})$$

- $\delta_t$ — time-varying intercepts.
- Condition to uncover the causal effect if we estimate the FD model with OLS: $E(\Delta X_{it} \Delta w_{it}) = 0$
- holds if $Cov(X_{it}, w_{is}) = 0$, for all $t, s$ : *Strict Exogeneity*

# Strict exogeneity assumption

- Example: dynamic model

$$\begin{aligned} Y_{it} &= Y_{it-1}\beta + u_i + w_{it} \\ Y_{it} - Y_{it-1} &= (Y_{it-1} - Y_{it-2})\beta + (w_{it} - w_{it-1}) \end{aligned}$$

  Here: $(w_{it} - w_{it-1})$ is correlated with $(Y_{it-1} - Y_{it-2})$ by construction

- Strict exogeneity also rules out cases where future explanatory variables react to changes in the idiosyncratic errors.

  Example: firm $i$ discovers its productivity improved (high $w_{it}$) $\Longrightarrow$ $i$ hires more workers tomorrow ($X_{it+1} \uparrow$).

# Remarks on FD.

- Always, always include time dummies. Example:
  - $Y_{it}$ is deflated revenue of firm $i$, $X_{it}$ is deflated capital stock (in logs).
  - Imperfect deflation: suppose $Y_{it}$ and $X_{ik}$ both have upward trends due to this — **coincidental trends!**
  - Coefficient on $\Delta X_{it}$ will pick this up even if capital does not have effect on output.
- Differencing may amplify measurement error. Example: coef. on capital often too low in production function estimates. Why?
  - Differencing wipes out cross-sectional variation in the data.
  - If measurement error is independent across $t$, it's variation stays the same.
  - $\Rightarrow$ Relative variation in error grows $\Rightarrow$ attenuation bias $\uparrow$.

# Fixed Effects Estimation

FE = OLS with individual dummies, but less computationally expensive.
Use OLS to estimate the *time demeaned* model:

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)\beta + (\delta_t - \bar{\delta}) + (w_{it} - \bar{w}_i)$$

also called within transformation or fixed-effects (FE) transformation

- All time constant variables drop out of the fixed effects model
- *Strict exogeneity* assumption necessary to get the causal effect
- Fixed effects and FD estimators are generally different for $t > 2$
- To compute correct standard errors use FE procedures (more on this later)

# Fixed effects estimation

Estimation of $\hat{u}_i$, $i = 1, ..., N$

$$\hat{u}_i = \bar{Y}_i - \bar{X}_i \hat{\beta} - \bar{\delta}$$

- We think of a situation where $T$ is fixed and $N$ increases to do statistical inference.
- For every $i$ that is added to the sample, we would get an additional estimate for $u_i$.
- In this model estimates $\hat{u}_i$ are not meaningful (inconsistent).

# Why does OLS work and how to compute std. errors?

FD and FE estimators are just OLS on transformed data. Can't one just use standard errors from the OLS output?

Not a good idea: OLS assumes $\Delta X_{it}$ and $\Delta X_{it+1}$ are independent. But they are not — transforming the data creates *serial correlation*.

# Recall theory behind OLS

Consider a model with one variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad E[\varepsilon_i | X_i] = 0$$

OLS

$$\widehat{\beta}_1 = \frac{\frac{1}{N}\sum_i (X_i - \bar{X}) Y_i}{\frac{1}{N}\sum_i (X_i - \bar{X})^2} = \beta_1 + \frac{\frac{1}{N}\sum_i (X_i - \bar{X})\varepsilon_i}{\frac{1}{N}\sum_i (X_i - \bar{X})^2}$$

CLT: $\frac{\frac{1}{N}\sum_i (X_i - \bar{X})\varepsilon_i - E[X\varepsilon]}{\sqrt{N}} \to N[0, V]$, where $V = E[(X - E[X])^2 \varepsilon^2]$
LLN: $\frac{1}{N}\sum_i (X_i - \bar{X})^2 \to \sigma_X^2$.

In the limit, standard error of $\widehat{\beta}_1$ is $\sigma_{\widehat{\beta}_1} = \sqrt{V}/\sigma_X^2$. We know neither $V$, nor $\sigma_X$, need to estimate both.

# Homoskedastic errors

Suppose $E[\varepsilon^2|X] = \sigma_\varepsilon^2$. Estimating $V$ and $\sigma_X$ in $\sigma_{\widehat{\beta}_1} = \sqrt{V}/\sigma_X^2$:

1. $\widehat{\sigma}_X = \sqrt{\frac{1}{N}\sum_i (X_i - \bar{X})}$
2. $V = E[(X - E[X])^2\varepsilon^2] = E[(X - E[X])^2]\sigma_\varepsilon^2$:

$$\widehat{V} = \frac{1}{N}\sum_i (X_i - \bar{X})^2 \widehat{\sigma}_\varepsilon^2, \quad \widehat{\sigma}_\varepsilon^2 = \frac{1}{N}\sum_i \widehat{\varepsilon}_i^2$$

This is what you get if you run

```
regress y x
```

without any extra options

# Heteroskedasticity-robust errors

Suppose $E[\varepsilon^2|X] \neq const$:

1. $\widehat{\sigma}_X^2$ — same as before
2. Estimating $V = E[(X - E[X])^2\varepsilon^2]$

$$\widehat{V} = \frac{1}{N}\sum_i (X_i - \bar{X})^2 \widehat{\varepsilon}_i^2$$

This is what you get if you run

```
regress y x, robust
```

# Dealing with serial correlation

So, what's the deal with serial correlation?

- LLN and CLT work if $(X_i, Y_i)$ is independent of $(X_j, Y_j)$.
- When we use FD we run $\Delta Y_{it}$ on $\Delta X_{it}$. But $\Delta Y_{it}$ is correlated with $\Delta Y_{it-1}$!

Would OLS on transformed data still work? Standard errors?

# Dealing with serial correlation: clustering

Suppose we run FD (balanced panel, zero time effects):

$$\Delta Y_{it} = \beta_1 \Delta X_{it} + \Delta w_{it}, \quad E[w_{it}|X_{is}] = 0 \text{ for all } s, t$$

$$\widehat{\beta}_1 = \beta_1 + \frac{\frac{1}{NT}\sum_i \sum_t \Delta X_{it} \Delta w_{it}}{\frac{1}{NT}\sum_i \sum_t (\Delta X_{it})^2}$$

Cluster observations by cross-sectional unit. For instance, in the numerator:

$$\frac{1}{N}\sum_i \underbrace{\left[\frac{1}{T}\sum_t \underbrace{\Delta X_{it}\Delta w_{it}}_{\text{not i.i.d.}}\right]}_{i.i.d.}$$

Quantities in brackets are i.i.d.: trajectories of $(X_i, Y_i)$ are sampled randomly. Terms in the inner $\sum$ are correlated with one another.

# Dealing with serial correlation: clustering

So we can use CLT, but apply it to clusters rather than observations.
CLT:

$$\frac{1}{N}\sum_i \left[\frac{1}{T}\sum_t \Delta X_{it}\Delta w_{it}\right] \to N[0, V], \quad V = Var\left[\frac{1}{T}\sum_t \Delta X_{it}\Delta w_{it}\right]$$

Estimating $V$:

$$\widehat{V} = \frac{1}{N}\sum_i \left[\frac{1}{T}\sum_t \Delta X_{it}\widehat{\Delta w}_{it}\right]^2$$

Standard error for $\widehat{\beta}_1$: $\widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\widehat{V}}/\widehat{\sigma}^2_{\Delta X}$ This is what you get if you run

```
regress y x, vce(cluster i)
```

# Stata tips

- Set up your panel data before use:

  ```
  xtset i t
  ```

  where `i` – cross-sectional id, `t` – time id.
- First difference of $X$, $\Delta X_{it}$: `D.X`
- First lag, $X_{it-1}$: `L.X`
- Second lag, $X_{it-2}$: `L2.X`

FD estimator w/o time dummies:

```
regress D.Y D.X, vce(cluster i)
```

FE estimator:

```
xtreg Y X, fe vce(cluster i)
```

Don't forget to use the "fe" option!

# Clusters and sample size

Standard tests and confidence intervals are precise when the sample is large enough for the CLT to kick in. Otherwise $\widehat{\beta} - \beta \not\approx$ normal r.v.

## Typical faulty reasoning

*"I run GDP on interest rate for the 15 post-Soviet countries. A sample of 15 observations is clearly too small. However, I use a 20-year panel, which gives me 300 observations. I heard that OLS is fairly reliable if the number of observations is 100 or higher"*

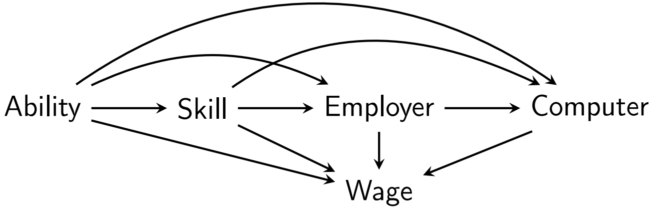
Remember: in the panel setting CLT is applied to **countries** (outer sum) rather than country-year pairs (inner sum):

$$\frac{1}{N}\sum_{i}\left[\frac{1}{T}\sum_{t}\Delta X_{it}\Delta w_{it}\right] \to N[0, V]$$

Relevant sample size is still $N = 15$.

# Application: Return to computer use

Stylized fact: workers who use computers earn more. Interpretation is hard:




- Data: French Labor Force Survey, rotating panel of 9000 individuals 1991-1993. $T = 3$, $N = 9000$.
- Computer use supplement in 1993.
- 905 individuals change status from non-users to computer users in the sample.

$$ln(w_{it}) = \rho Comp_{it} + X_{it}\beta + \delta_t + F_{j(i,t)} + A_i + \epsilon_{it}$$

Table 4
The Effect of New Technologies on Pay: Cross-Section Results, $N$ = 9,345

| Dependent Variable, ln(Monthly Wage) | Cross Section: Krueger (1993) (1) | (2) | Firm-Fixed Effects (3) | Di Nardo and Pischke (1997) (4) | Krueger (1993) (5) |
|---|---|---|---|---|---|
| Uses a computer (yes = 1) | .1824 (.0076) | .0700 (.0128) | .0809 (.0143) | .0979 (.0089) | .1612 (.0088) |
| Uses a robot (yes = 1) | ... | .0197 (.0306) | −.0171 (.0344) | .0249 (.0162) | ... |
| Uses fax or minitel (yes = 1) | ... | .0804 (.0084) | .1359 (.0091) | ... | ... |
| Uses fax (yes = 1) | ... | ... | ... | .1204 (.0093) | ... |
| Uses minitel (yes = 1) | | | | .0470 (.0091) | |
| Uses video or laser (yes = 1) | | .0197 (.0084) | .0436 (.0095) | .0711 (.0090) | |
| Experience with computer | | −.0013 (.0033) | −.0010 (.0038) | ... | |
| Experience with computer$^2$ | | .0001 (.0002) | .0002 (.0002) | ... | |
| Experience with robot | | .0057 (.0090) | .0131 (.0102) | ... | |
| Experience with robot$^2$ | | −.0004 (.0004) | −.0009 (.0005) | ... | |
| Tenure | .0106 (.0012) | .0119 (.0011) | .0112 (.0012) | .0110 (.0012) | .0161 (.0014) |
| Tenure$^2$ | −.0006 (.0004) | −.0014 (.0003) | −.0009 (.0004) | −.0008 (.0003) | −.0015 (.0004) |
| Experience | .0163 (.0015) | .0108 (.0013) | .0155 (.0014) | .0155 (.0015) | .0148 (.0016) |
| Experience$^2$ | −.0022 (.0003) | −.0017 (.0002) | −.0024 (.0003) | −.0021 (.0003) | .0022 (.0003) |
| $R^2$ | .5479 | .6350 | .6384 | .5634 | .4954 |

SOURCE.—Enquête Emploi, 1991–93. Columns 1–4 use 1993 data; col. 5 uses 1991 data.
NOTE.—Standard errors in parentheses. Models 1, 4, and 5 also include years of education (and square), a part-time effect, a sex effect, a married effect, a married female effect, a region effect (= 1 for Ile de France), and size of firms and government agencies (five indicators). Models 2 and 3 also include regional effect (= 1 for Ile de France), part-time effect, size of firm effects, government agencies effects, sex effect, eight education effects, short-term contract effect (interacted with computer use), five occupation effects, and 14 sector effects. Finally, model 3 also includes 1,016 firm effects.

**Table 5**
**The Effect of New Technologies on Pay: Longitudinal Results**

| Dependent Variable, ln(Monthly Wage) | Individual-Fixed Effects (5) | Individual-Fixed Effects with Firm-Fixed Effects (6) |
|---|---|---|
| Uses a computer (yes = 1) | .0105 | .0112 |
| | (.0082) | (.0084) |
| Uses a robot (yes = 1) | .0423 | .0368 |
| | (.0222) | (.0225) |
| Experience with computer | .0047 | .0034 |
| | (.0045) | (.0046) |
| Experience with computer$^2$ | −.0006 | −.0006 |
| | (.0003) | (.0003) |
| Experience with robot | −.0132 | −.0132 |
| | (.0116) | (.0116) |
| Experience with robot$^2$ | .0002 | .0001 |
| | (.0008) | (.0008) |
| Tenure | .0039 | .0035 |
| | (.0010) | (.0011) |
| Tenure$^2$ | −.0007 | −.0006 |
| | (.0003) | (.0003) |
| Experience | .0468 | .0502 |
| | (.0047) | (.0048) |
| Experience$^2$ | −.0025 | −.0026 |
| | (.0006) | (.0007) |
| $R^2$ | .9126 | .9160 |

SOURCE.—Enquête Emploi, 1991–93.
NOTE.—$N$ = 27,893 (both models). Standard errors in parentheses. Models 5 and 6 include an indicator for year 1991, size of firm effects, government agencies and firms effects, short-term contract effect (interacted with computer use), five occupation effects, 14 sector effects, experience (and square), seniority (and square), and 9,344 individual effects. Model 6 also includes 1,045 firm effects of which 494 are identified.

# FE vs FD?

- FD discards many observations if the dataset has many holes.
- FE may be less sensitive to timing misspecification than FD
  - Consider $\log WAGE = \beta_0 + \beta_1 YEARS\_EDUC + \varepsilon$
  - On years when $YEARS\_EDUC$ grows, $WAGE$ is zero or close to zero (school/college) and vice versa.
  - Plot FD, demeaned data.
- FD: Strict exogeneity sufficient, but not necessary.

Try both.

# Notes on FD and FE

Unbalanced panel? Beware of sample selection

- Selection on observables is not an issue, as usual.
- Selection on individual intercept is okay (e.g., lower ability workers having more holes in the data).
- Selection on time-varying unobservables is still a problem.

The effect is identified off switchers (e.g., no computer $\rightarrow$ computer and vice versa).

- No way to tease out the effect of time-invariant variables (e.g., gender, race).
- Selection on treatment effects! In theory, switchers are almost indifferent between treatment and no treatment.

Think hard what makes people switch. Why is this unrelated to the effect of interest?

# Random effects estimator?

Some of you have heard about the random effects estimator.

- More efficient (tighter confidence intervals) than FE.
- But requires stronger assumptions.
- A dealbreaker: $u_i$ cannot correlate with $X_{it}$. E.g., computer use is independent of ability (!)
- Hence, RE is never used for causal inference.

# Dynamic models?

A model with lags of outcome on the RHS. For instance, think of firm-level employment ($i$ – firm, $t$ –year):

$$\log EMPLOYMENT_{it} = \rho \log EMPLOYMENT_{it-1} + \beta X_{it} + u_i + w_{it}$$

$X_{it}$ — gov't grants, access to credit. Lag: adjusting employment is subject to frictions.

Cannot use FD or FE: strict exogeneity is violated —$w_{it}$ is correlated with the RHS at $t+1$.

# Dynamic models? Arellano-Bond estimators.

Take differences:

$$\Delta \log EMPLOYMENT_{it} = \rho \Delta \log EMPLOYMENT_{it-1} + \beta \Delta X_{it} + \Delta w_{it}$$

Fixed effect $u_i$ is gone. Next step — use lagged levels of $X_{it}$ to instrument for $\log EMPLOYMENT_{it-1}$ and $\Delta X_{it}$. Intuition?

- $X_{it}$ fluctuates around some "steady state". If $X_{it-1}$ is too high, $\Delta X_{it}$ is likely to be negative and vice versa.
- $\log EMPLOYMENT_{it-1}$ is partly driven by $X_{it-1}, X_{it-2}, \ldots$. Generous gov't support in the past $\rightarrow$ more workers at $t-1$.

Need contemporaneous and past exogeneity only: "future shocks do not affect past choices"

$$E[w_{it} | X_{it}, X_{it-1}, \ldots] = 0.$$

$X_{it+1}$ can respond to $w_{it}$, strict exogeneity is not necessary.

# Arellano-Bond estimators. Notes.

- Many ways to implement — what lags to include? Instrument with past levels or past differences or both?
- Degrees of freedom $\uparrow \Rightarrow$ trust in estimates $\downarrow$. Fishing for significance, convenient results?
- Non-transparent identification — most applications rely on handwaving rather than carefully exploited natural experiments.
- But sometimes this is the best you can do (quick and dirty policy research).