Panel Data I

Francisco Villamil

Applied Quantitative Methods II

MA in Social Sciences, Spring 2026

Today's goals

- Understand the structure and logic of panel data

Today's goals

- Understand the structure and logic of panel data
- See why cross-sectional OLS can be biased due to unobserved heterogeneity

Today's goals

- Understand the structure and logic of panel data
- See why cross-sectional OLS can be biased due to unobserved heterogeneity
- Learn the fixed effects (within) estimator and its key intuition

Today's goals

- Understand the structure and logic of panel data
- See why cross-sectional OLS can be biased due to unobserved heterogeneity
- Learn the fixed effects (within) estimator and its key intuition
- Add time fixed effects for two-way FE models

Today's goals

- Understand the structure and logic of panel data
- See why cross-sectional OLS can be biased due to unobserved heterogeneity
- Learn the fixed effects (within) estimator and its key intuition
- Add time fixed effects for two-way FE models
- Compare fixed effects and random effects; know when to use which

Today's goals

- Understand the structure and logic of panel data
- See why cross-sectional OLS can be biased due to unobserved heterogeneity
- Learn the fixed effects (within) estimator and its key intuition
- Add time fixed effects for two-way FE models
- Compare fixed effects and random effects; know when to use which
- Cluster standard errors correctly in panel settings

Roadmap

What is Panel Data?

The Problem: Unobserved Heterogeneity

Fixed Effects

Two-Way Fixed Effects

Random Effects and the FE/RE Choice

Clustered Standard Errors

Wrap-up

Panel data: the basic structure

Unit i	**Time** t	y_{it}	x_{it}
1	2010	0.42	12.1
1	2011	0.51	13.0
1	2012	0.48	12.7
2	2010	0.61	9.4
2	2011	0.59	9.8
2	2012	0.64	10.2

⋮

- N units, each observed at T time points
- Data indexed (i, t): unit i at time t

Panel data: examples

- **Cross-national**: GDP, democracy, conflict for 150+ countries $\times$ 50 years

Panel data: examples

- **Cross-national**: GDP, democracy, conflict for 150+ countries $\times$ 50 years
- **Survey panels**: same individuals surveyed in 2010, 2014, 2018

Panel data: examples

- **Cross-national**: GDP, democracy, conflict for 150+ countries $\times$ 50 years
- **Survey panels**: same individuals surveyed in 2010, 2014, 2018
  - $\rightarrow$ European Social Survey rotating panels

Panel data: examples

- **Cross-national**: GDP, democracy, conflict for 150+ countries $\times$ 50 years
- **Survey panels**: same individuals surveyed in 2010, 2014, 2018
  - $\rightarrow$ European Social Survey rotating panels
  - $\rightarrow$ British Household Panel Survey

Panel data: examples

- **Cross-national**: GDP, democracy, conflict for 150+ countries $\times$ 50 years
- **Survey panels**: same individuals surveyed in 2010, 2014, 2018
  - $\rightarrow$ European Social Survey rotating panels
  - $\rightarrow$ British Household Panel Survey
- **Sub-national**: US states $\times$ years; municipalities $\times$ election cycles

Panel data: examples

- **Cross-national**: GDP, democracy, conflict for 150+ countries $\times$ 50 years
- **Survey panels**: same individuals surveyed in 2010, 2014, 2018
  - $\rightarrow$ European Social Survey rotating panels
  - $\rightarrow$ British Household Panel Survey
- **Sub-national**: US states $\times$ years; municipalities $\times$ election cycles
- **Firms**: quarterly earnings reports for publicly traded companies

Panel data: examples

- **Cross-national**: GDP, democracy, conflict for 150+ countries $\times$ 50 years
- **Survey panels**: same individuals surveyed in 2010, 2014, 2018
  - $\rightarrow$ European Social Survey rotating panels
  - $\rightarrow$ British Household Panel Survey
- **Sub-national**: US states $\times$ years; municipalities $\times$ election cycles
- **Firms**: quarterly earnings reports for publicly traded companies
- **Running example**: US state-level presidential approval $\times$ years

Why panel data? Three advantages

- **More observations**: $N \times T$ rows instead of N — more statistical power

Why panel data? Three advantages

- **More observations**: $N \times T$ rows instead of N — more statistical power
- **Within-unit variation**: follow how y_{it} changes as x_{it} changes *for the same unit*

Why panel data? Three advantages

- **More observations**: $N \times T$ rows instead of N — more statistical power
- **Within-unit variation**: follow how y_{it} changes as x_{it} changes *for the same unit*
  - → Cleaner comparison than across different units

Why panel data? Three advantages

- **More observations**: $N \times T$ rows instead of N — more statistical power
- **Within-unit variation**: follow how y_{it} changes as x_{it} changes *for the same unit*
  - $\rightarrow$ Cleaner comparison than across different units
- **Control for unobserved heterogeneity**: the big one

Why panel data? Three advantages

- **More observations**: $N \times T$ rows instead of N — more statistical power
- **Within-unit variation**: follow how y_{it} changes as x_{it} changes *for the same unit*
  - $\rightarrow$ Cleaner comparison than across different units
- **Control for unobserved heterogeneity**: the big one
  - $\rightarrow$ Units may differ in ways we cannot measure

Why panel data? Three advantages

- **More observations**: $N \times T$ rows instead of N — more statistical power
- **Within-unit variation**: follow how y_{it} changes as x_{it} changes *for the same unit*
  - → Cleaner comparison than across different units
- **Control for unobserved heterogeneity**: the big one
  - → Units may differ in ways we cannot measure
  - → Panel structure lets us "absorb" those differences

Roadmap

What is Panel Data?

The Problem: Unobserved Heterogeneity

Fixed Effects

Two-Way Fixed Effects

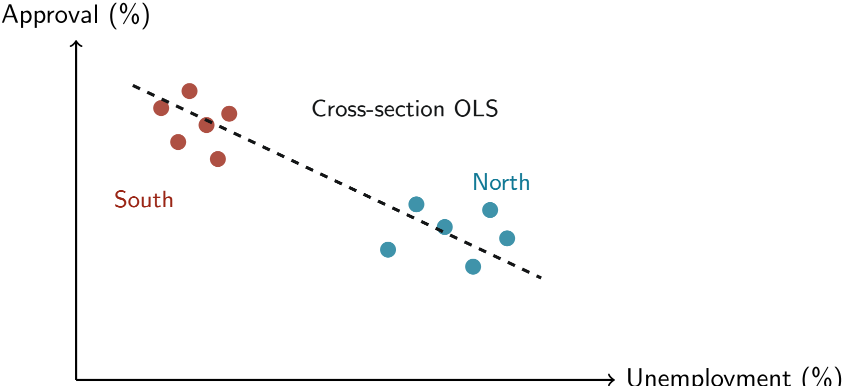
Random Effects and the FE/RE Choice

Clustered Standard Errors

Wrap-up

Motivating example: presidential approval

Does unemployment drive down presidential approval?



The cross-sectional slope is negative.

Does that mean unemployment *causes* lower approval?

What else might explain this pattern?

The problem: unit-level confounders

- Units differ in many **unobserved** ways:

The problem: unit-level confounders

- Units differ in many **unobserved** ways:
  - → Political culture, history, institutional quality

The problem: unit-level confounders

- Units differ in many **unobserved** ways:
  - → Political culture, history, institutional quality
  - → Personality (in individual panels)

The problem: unit-level confounders

- Units differ in many **unobserved** ways:
  - → Political culture, history, institutional quality
  - → Personality (in individual panels)
  - → Industrial structure, geography

The problem: unit-level confounders

- Units differ in many **unobserved** ways:
  - → Political culture, history, institutional quality
  - → Personality (in individual panels)
  - → Industrial structure, geography
- If these unobservables correlate with x_{it} **and** y_{it}: OLS is biased

The problem: unit-level confounders

- Units differ in many **unobserved** ways:
  - $\rightarrow$ Political culture, history, institutional quality
  - $\rightarrow$ Personality (in individual panels)
  - $\rightarrow$ Industrial structure, geography
- If these unobservables correlate with x_{it} **and** y_{it}: OLS is biased
- The model:

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$$

The problem: unit-level confounders

- Units differ in many **unobserved** ways:
  - → Political culture, history, institutional quality
  - → Personality (in individual panels)
  - → Industrial structure, geography
- If these unobservables correlate with x_{it} **and** y_{it}: OLS is biased
- The model:

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$$

- α_i = unit-specific intercept (the unobserved heterogeneity)

The problem: unit-level confounders

- Units differ in many **unobserved** ways:
  - → Political culture, history, institutional quality
  - → Personality (in individual panels)
  - → Industrial structure, geography

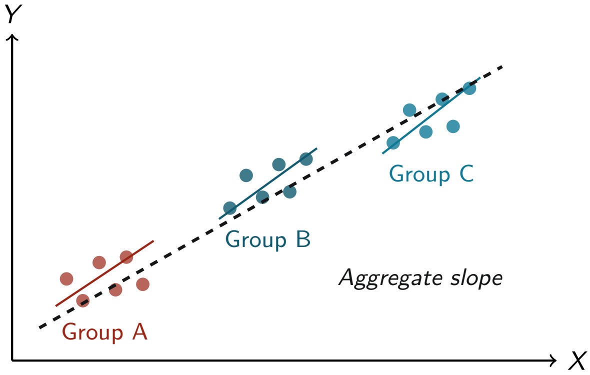
- If these unobservables correlate with x_{it} **and** y_{it}: OLS is biased

- The model:

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$$

- α_i = unit-specific intercept (the unobserved heterogeneity)
- Cross-section OLS ignores $\alpha_i \Rightarrow$ omitted variable bias

Simpson's paradox: the intuition



- Within each group: positive slope
- Cross-section OLS: also positive, but for the **wrong reason**
- The group-level differences dominate the estimate

Roadmap

What is Panel Data?

The Problem: Unobserved Heterogeneity

Fixed Effects

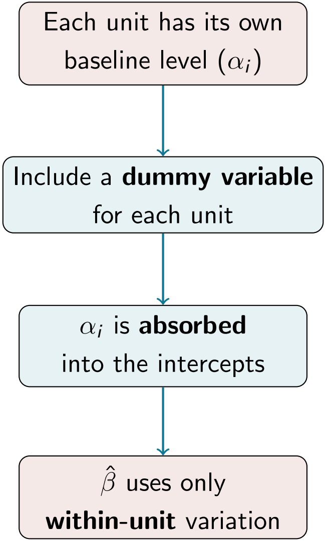
Two-Way Fixed Effects

Random Effects and the FE/RE Choice

Clustered Standard Errors

Wrap-up

Fixed effects: the key idea



The within (demeaning) estimator

Starting from $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$, subtract unit means:

$$\underbrace{y_{it} - \bar{y}_i}_{\tilde{y}_{it}} = \beta \underbrace{(x_{it} - \bar{x}_i)}_{\tilde{x}_{it}} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\tilde{\varepsilon}_{it}}$$

- α_i **cancels out** — the unit effect is gone

The within (demeaning) estimator

Starting from $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$, subtract unit means:

$$\underbrace{y_{it} - \bar{y}_i}_{\tilde{y}_{it}} = \beta \underbrace{(x_{it} - \bar{x}_i)}_{\tilde{x}_{it}} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\tilde{\varepsilon}_{it}}$$

- α_i **cancels out** — the unit effect is gone
- Regressing $\tilde{y}_{it}$ on $\tilde{x}_{it}$ gives the FE estimator

The within (demeaning) estimator

Starting from $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$, subtract unit means:

$$\underbrace{y_{it} - \bar{y}_i}_{\tilde{y}_{it}} = \beta \underbrace{(x_{it} - \bar{x}_i)}_{\tilde{x}_{it}} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\tilde{\varepsilon}_{it}}$$

- α_i **cancels out** — the unit effect is gone
- Regressing $\tilde{y}_{it}$ on $\tilde{x}_{it}$ gives the FE estimator
- Uses only variation *within* each unit over time

The within (demeaning) estimator

Starting from $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$, subtract unit means:

$$\underbrace{y_{it} - \bar{y}_i}_{\tilde{y}_{it}} = \beta \underbrace{(x_{it} - \bar{x}_i)}_{\tilde{x}_{it}} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\tilde{\varepsilon}_{it}}$$

- α_i **cancels out** — the unit effect is gone
- Regressing $\tilde{y}_{it}$ on $\tilde{x}_{it}$ gives the FE estimator
- Uses only variation *within* each unit over time
- Works because α_i is constant: $\bar{\alpha}_i = \alpha_i$

FE = dummies for each unit

- Mathematically equivalent to including unit dummies:

$$y_{it} = \sum_{i=1}^{N} \alpha_i D_i + \beta x_{it} + \varepsilon_{it}$$

where $D_i = 1$ if observation belongs to unit i

FE = dummies for each unit

- Mathematically equivalent to including unit dummies:

$$y_{it} = \sum_{i=1}^{N} \alpha_i D_i + \beta x_{it} + \varepsilon_{it}$$

  where $D_i = 1$ if observation belongs to unit i

- "Least squares dummy variable" (LSDV) estimator

FE = dummies for each unit

- Mathematically equivalent to including unit dummies:

$$y_{it} = \sum_{i=1}^{N} \alpha_i D_i + \beta x_{it} + \varepsilon_{it}$$

  where $D_i = 1$ if observation belongs to unit i

- "Least squares dummy variable" (LSDV) estimator
- Same $\hat{\beta}$, different computational approach

FE = dummies for each unit

- Mathematically equivalent to including unit dummies:

$$y_{it} = \sum_{i=1}^{N} \alpha_i D_i + \beta x_{it} + \varepsilon_{it}$$

  where $D_i = 1$ if observation belongs to unit i

- "Least squares dummy variable" (LSDV) estimator
- Same $\hat{\beta}$, different computational approach

- **Key implication**: cannot estimate effect of **time-invariant** variables

FE = dummies for each unit

- Mathematically equivalent to including unit dummies:

$$y_{it} = \sum_{i=1}^{N} \alpha_i D_i + \beta x_{it} + \varepsilon_{it}$$

  where $D_i = 1$ if observation belongs to unit i

- "Least squares dummy variable" (LSDV) estimator
- Same $\hat{\beta}$, different computational approach
- **Key implication**: cannot estimate effect of **time-invariant** variables
  - → If z_i does not vary over time, it is collinear with D_i

FE = dummies for each unit

- Mathematically equivalent to including unit dummies:

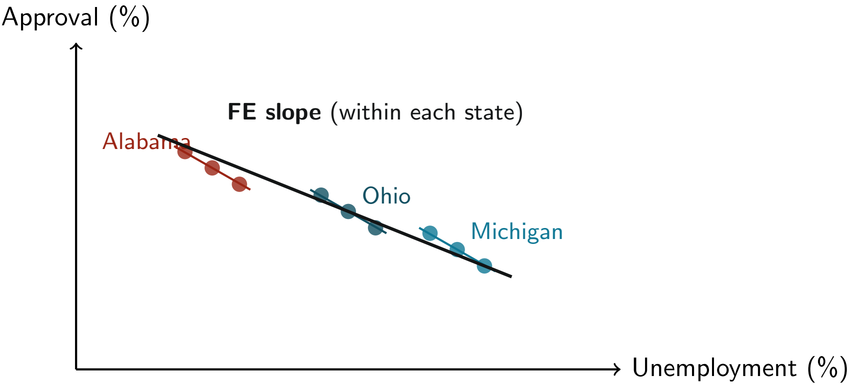
$$y_{it} = \sum_{i=1}^{N} \alpha_i D_i + \beta x_{it} + \varepsilon_{it}$$

  where $D_i = 1$ if observation belongs to unit i

- "Least squares dummy variable" (LSDV) estimator
- Same $\hat{\beta}$, different computational approach

- **Key implication**: cannot estimate effect of **time-invariant** variables
  - → If z_i does not vary over time, it is collinear with D_i
  - → Example: "South" dummy, gender, country of birth

Presidential approval: what FE does



- Each state has its own intercept; the slope is shared
- FE estimates: as *this state's* unemployment rises, its approval falls

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
```

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
feols(approval ~ unemp | state, data = df)
```

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
feols(approval ~ unemp | state, data = df)
```

- Alternative: `plm` package

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
feols(approval ~ unemp | state, data = df)
```

- Alternative: `plm` package

```
library(plm)
```

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
feols(approval ~ unemp | state, data = df)
```

- Alternative: `plm` package

```
library(plm)
pdata.frame(df, index = c("state","year"))
```

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
feols(approval ~ unemp | state, data = df)
```

- Alternative: `plm` package

```
library(plm)
pdata.frame(df, index = c("state","year"))
plm(approval ~ unemp, data = pdf, model = "within")
```

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
feols(approval ~ unemp | state, data = df)
```

- Alternative: `plm` package

```
library(plm)
pdata.frame(df, index = c("state","year"))
plm(approval ~ unemp, data = pdf, model = "within")
```

- The | in `feols()` separates regressors from fixed effects

Fixed effects in R

- Preferred: `fixest` package (fast, flexible, clustered SEs built in)

```
library(fixest)
feols(approval ~ unemp | state, data = df)
```

- Alternative: `plm` package

```
library(plm)
pdata.frame(df, index = c("state","year"))
plm(approval ~ unemp, data = pdf, model = "within")
```

- The | in `feols()` separates regressors from fixed effects
- Tables with `modelsummary()` work seamlessly with `feols` objects

Roadmap

What is Panel Data?

The Problem: Unobserved Heterogeneity

Fixed Effects

Two-Way Fixed Effects

Random Effects and the FE/RE Choice

Clustered Standard Errors

Wrap-up

FE controls for everything time-invariant.

But what if something happens in 2008 that affects *all* states simultaneously?

A new threat: time trends and common shocks

- Unit FE removes time-invariant unit characteristics

A new threat: time trends and common shocks

- Unit FE removes time-invariant unit characteristics
- But what about events that affect **all** units at the same time?

A new threat: time trends and common shocks

- Unit FE removes time-invariant unit characteristics
- But what about events that affect **all** units at the same time?
  - → A global recession hits every state simultaneously

A new threat: time trends and common shocks

- Unit FE removes time-invariant unit characteristics
- But what about events that affect **all** units at the same time?
  - → A global recession hits every state simultaneously
  - → A presidential scandal lowers approval everywhere

A new threat: time trends and common shocks

- Unit FE removes time-invariant unit characteristics
- But what about events that affect **all** units at the same time?
  - → A global recession hits every state simultaneously
  - → A presidential scandal lowers approval everywhere
  - → A pandemic affects all countries in 2020

A new threat: time trends and common shocks

- Unit FE removes time-invariant unit characteristics
- But what about events that affect **all** units at the same time?
  - → A global recession hits every state simultaneously
  - → A presidential scandal lowers approval everywhere
  - → A pandemic affects all countries in 2020
- If these shocks also correlate with x_{it}: new bias

A new threat: time trends and common shocks

- Unit FE removes time-invariant unit characteristics
- But what about events that affect **all** units at the same time?
  - → A global recession hits every state simultaneously
  - → A presidential scandal lowers approval everywhere
  - → A pandemic affects all countries in 2020
- If these shocks also correlate with x_{it}: new bias
- Solution: add **time fixed effects** γ_t

Two-way fixed effects model

$$y_{it} = \alpha_i + \gamma_t + \beta x_{it} + \varepsilon_{it}$$

- α_i: unit FE — absorbs all time-invariant unit characteristics

Two-way fixed effects model

$$y_{it} = \alpha_i + \gamma_t + \beta x_{it} + \varepsilon_{it}$$

- α_i: unit FE — absorbs all time-invariant unit characteristics
- γ_t: time FE — absorbs all unit-invariant time shocks

Two-way fixed effects model

$$y_{it} = \alpha_i + \gamma_t + \beta x_{it} + \varepsilon_{it}$$

- α_i: unit FE — absorbs all time-invariant unit characteristics
- γ_t: time FE — absorbs all unit-invariant time shocks
- $\hat{\beta}$: identified from variation **within units**, **across time**, **net of common trends**

Two-way fixed effects model

$$y_{it} = \alpha_i + \gamma_t + \beta x_{it} + \varepsilon_{it}$$

- α_i: unit FE — absorbs all time-invariant unit characteristics
- γ_t: time FE — absorbs all unit-invariant time shocks
- $\hat{\beta}$: identified from variation **within units**, **across time**, **net of common trends**
- In R:

Two-way fixed effects model

$$y_{it} = \alpha_i + \gamma_t + \beta x_{it} + \varepsilon_{it}$$

- α_i: unit FE — absorbs all time-invariant unit characteristics
- γ_t: time FE — absorbs all unit-invariant time shocks
- $\hat{\beta}$: identified from variation **within units**, **across time**, **net of common trends**
- In R:

```
feols(approval ~ unemp | state + year, data = df)
```

What TWFE absorbs

Type of variation	**Unit FE**	**Two-way FE**
Time-invariant unit differences	absorbed	absorbed
Unit-invariant time shocks	not absorbed	absorbed
Within-unit, across-time variation	used for $\hat{\beta}$	used for $\hat{\beta}$

- TWFE is conservative: only uses within-unit, net-of-time-trends variation

What TWFE absorbs

Type of variation	Unit FE	Two-way FE
Time-invariant unit differences	absorbed	absorbed
Unit-invariant time shocks	not absorbed	absorbed
Within-unit, across-time variation	used for $\hat{\beta}$	used for $\hat{\beta}$

- TWFE is conservative: only uses within-unit, net-of-time-trends variation
- Leaves less variation to identify $\hat{\beta}$ $\Rightarrow$ larger standard errors

What TWFE absorbs

Type of variation	Unit FE	Two-way FE
Time-invariant unit differences	absorbed	absorbed
Unit-invariant time shocks	not absorbed	absorbed
Within-unit, across-time variation	used for $\hat{\beta}$	used for $\hat{\beta}$

- TWFE is conservative: only uses within-unit, net-of-time-trends variation
- Leaves less variation to identify $\hat{\beta} \Rightarrow$ larger standard errors
- But more credible: fewer threats to identification

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)
  - → Predictors: `Apct` (attractive), `Enrollment`, `Required`

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)
  - → Predictors: `Apct` (attractive), `Enrollment`, `Required`
- OLS ignores that instructors differ systematically:

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)
  - → Predictors: `Apct` (attractive), `Enrollment`, `Required`
- OLS ignores that instructors differ systematically:
  - → Friendliness, teaching experience, subject difficulty

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)
  - → Predictors: `Apct` (attractive), `Enrollment`, `Required`
- OLS ignores that instructors differ systematically:
  - → Friendliness, teaching experience, subject difficulty
- Unit FE (instructor):

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)
  - → Predictors: `Apct` (attractive), `Enrollment`, `Required`
- OLS ignores that instructors differ systematically:
  - → Friendliness, teaching experience, subject difficulty
- Unit FE (instructor):

```
feols(Eval ~ Apct + Enrollment + Required
```

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)
  - → Predictors: `Apct` (attractive), `Enrollment`, `Required`
- OLS ignores that instructors differ systematically:
  - → Friendliness, teaching experience, subject difficulty
- Unit FE (instructor):

```
feols(Eval ~ Apct + Enrollment + Required
  | instructor, data = evals)
```

Teaching evaluations: worked example

- Dataset: instructors evaluated across multiple courses
  - → Outcome: `Eval` (student evaluation score)
  - → Predictors: `Apct` (attractive), `Enrollment`, `Required`
- OLS ignores that instructors differ systematically:
  - → Friendliness, teaching experience, subject difficulty
- Unit FE (instructor):

```
feols(Eval ~ Apct + Enrollment + Required
  | instructor, data = evals)
```

- $\hat{\beta}_{\text{Apct}}$: compares same instructor's courses, not different instructors

Roadmap

What is Panel Data?

The Problem: Unobserved Heterogeneity

Fixed Effects

Two-Way Fixed Effects

Random Effects and the FE/RE Choice

Clustered Standard Errors

Wrap-up

Random effects: a different assumption

$$y_{it} = \alpha + \beta x_{it} + \underbrace{\eta_i}_{\text{unit random effect}} + \varepsilon_{it}$$

- $\eta_i \sim N(0, \sigma^2_\eta)$: unit-specific deviation, treated as **random**

Random effects: a different assumption

$$y_{it} = \alpha + \beta x_{it} + \underbrace{\eta_i}_{\text{unit random effect}} + \varepsilon_{it}$$

- $\eta_i \sim N(0, \sigma^2_\eta)$: unit-specific deviation, treated as **random**
- **Key assumption**: $\eta_i \perp x_{it}$ (unit effects uncorrelated with regressors)

Random effects: a different assumption

$$y_{it} = \alpha + \beta x_{it} + \underbrace{\eta_i}_{\text{unit random effect}} + \varepsilon_{it}$$

- $\eta_i \sim N(0, \sigma_\eta^2)$: unit-specific deviation, treated as **random**
- **Key assumption**: $\eta_i \perp x_{it}$ (unit effects uncorrelated with regressors)
- If this holds: RE is more **efficient** than FE

Random effects: a different assumption

$$y_{it} = \alpha + \beta x_{it} + \underbrace{\eta_i}_{\text{unit random effect}} + \varepsilon_{it}$$

- $\eta_i \sim N(0, \sigma_\eta^2)$: unit-specific deviation, treated as **random**
- **Key assumption**: $\eta_i \perp x_{it}$ (unit effects uncorrelated with regressors)
- If this holds: RE is more **efficient** than FE
- If this fails: RE is **biased**; FE is consistent

Random effects: a different assumption

$$y_{it} = \alpha + \beta x_{it} + \underbrace{\eta_i}_{\text{unit random effect}} + \varepsilon_{it}$$

- $\eta_i \sim N(0, \sigma_\eta^2)$: unit-specific deviation, treated as **random**
- **Key assumption**: $\eta_i \perp x_{it}$ (unit effects uncorrelated with regressors)
- If this holds: RE is more **efficient** than FE
- If this fails: RE is **biased**; FE is consistent
- Unlike FE: can estimate effects of **time-invariant** variables

FE vs. RE: the tradeoff

	Fixed Effects	**Random Effects**
Assumption	α_i correlated with X?	$\eta_i \perp X$
Consistency	Always (if $T \to \infty$)	Only if $\eta_i \perp X$
Efficiency	Less efficient	More efficient
Time-invariant vars	Cannot estimate	Can estimate

- If you are unsure: **use FE**

FE vs. RE: the tradeoff

	Fixed Effects	**Random Effects**
Assumption	α_i correlated with X?	$\eta_i \perp X$
Consistency	Always (if $T \to \infty$)	Only if $\eta_i \perp X$
Efficiency	Less efficient	More efficient
Time-invariant vars	Cannot estimate	Can estimate

- If you are unsure: **use FE**
- More conservative, more credible

FE vs. RE: the tradeoff

	Fixed Effects	**Random Effects**
Assumption	α_i correlated with X?	$\eta_i \perp X$
Consistency	Always (if $T \to \infty$)	Only if $\eta_i \perp X$
Efficiency	Less efficient	More efficient
Time-invariant vars	Cannot estimate	Can estimate

- If you are unsure: **use FE**
- More conservative, more credible
- RE requires an untestable assumption; FE does not

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)
- H_1: correlation exists (only FE is consistent)

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)
- H_1: correlation exists (only FE is consistent)
- If $p < 0.05$: reject $H_0 \Rightarrow$ use FE

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)
- H_1: correlation exists (only FE is consistent)
- If $p < 0.05$: reject $H_0 \Rightarrow$ use FE
- If $p > 0.05$: fail to reject $\Rightarrow$ RE or FE both OK

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)
- H_1: correlation exists (only FE is consistent)
- If $p < 0.05$: reject H_0 $\Rightarrow$ use FE
- If $p > 0.05$: fail to reject $\Rightarrow$ RE or FE both OK
- In R (`plm` package):

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)
- H_1: correlation exists (only FE is consistent)
- If $p < 0.05$: reject $H_0 \Rightarrow$ use FE
- If $p > 0.05$: fail to reject $\Rightarrow$ RE or FE both OK
- In R (`plm` package):

```
fe_mod = plm(y ~ x, data = pdf, model = "within")
```

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)
- H_1: correlation exists (only FE is consistent)
- If $p < 0.05$: reject $H_0 \Rightarrow$ use FE
- If $p > 0.05$: fail to reject $\Rightarrow$ RE or FE both OK
- In R (plm package):

```
fe_mod = plm(y ~ x, data = pdf, model = "within")
re_mod = plm(y ~ x, data = pdf, model = "random")
```

Hausman test: FE vs. RE

- **Hausman test** (1978): formally test whether $\eta_i \perp x_{it}$
- H_0: no correlation between unit effects and regressors (RE is consistent)
- H_1: correlation exists (only FE is consistent)
- If $p < 0.05$: reject $H_0 \Rightarrow$ use FE
- If $p > 0.05$: fail to reject $\Rightarrow$ RE or FE both OK
- In R (`plm` package):

```
fe_mod = plm(y ~ x, data = pdf, model = "within")
re_mod = plm(y ~ x, data = pdf, model = "random")
phtest(fe_mod, re_mod)
```

Roadmap

What is Panel Data?

The Problem: Unobserved Heterogeneity

Fixed Effects

Two-Way Fixed Effects

Random Effects and the FE/RE Choice

Clustered Standard Errors

Wrap-up

Why panel data violates iid

- OLS standard errors assume errors are **independent** across observations

Why panel data violates iid

- OLS standard errors assume errors are **independent** across observations
- In panel data: observations from the *same unit* are correlated over time

Why panel data violates iid

- OLS standard errors assume errors are **independent** across observations
- In panel data: observations from the *same unit* are correlated over time
  - → Alabama in 2010 and Alabama in 2011 are not independent

Why panel data violates iid

- OLS standard errors assume errors are **independent** across observations
- In panel data: observations from the *same unit* are correlated over time
  - → Alabama in 2010 and Alabama in 2011 are not independent
  - → Same unit experiences same shocks, trends, institutions

Why panel data violates iid

- OLS standard errors assume errors are **independent** across observations
- In panel data: observations from the *same unit* are correlated over time
  - → Alabama in 2010 and Alabama in 2011 are not independent
  - → Same unit experiences same shocks, trends, institutions
- Ignoring this: standard errors are **too small**

Why panel data violates iid

- OLS standard errors assume errors are **independent** across observations
- In panel data: observations from the *same unit* are correlated over time
  - → Alabama in 2010 and Alabama in 2011 are not independent
  - → Same unit experiences same shocks, trends, institutions
- Ignoring this: standard errors are **too small**
- Too-small SEs ⇒ inflated *t*-statistics ⇒ false positives

Why panel data violates iid

- OLS standard errors assume errors are **independent** across observations
- In panel data: observations from the *same unit* are correlated over time
  - → Alabama in 2010 and Alabama in 2011 are not independent
  - → Same unit experiences same shocks, trends, institutions
- Ignoring this: standard errors are **too small**
- Too-small SEs ⇒ inflated t-statistics ⇒ false positives
- Solution: **cluster standard errors by unit**

Clustering in practice

- `feols()` clusters by the FE variable automatically:

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
```

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

- Explicit clustering:

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

- Explicit clustering:

```
feols(y ~ x | state, data = df, cluster = ~state)
```

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

- Explicit clustering:

```
feols(y ~ x | state, data = df, cluster = ~state)
```

- In `modelsummary()` tables:

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

- Explicit clustering:

```
feols(y ~ x | state, data = df, cluster = ~state)
```

- In `modelsummary()` tables:

```
modelsummary(m, vcov = ~state)
```

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

- Explicit clustering:

```
feols(y ~ x | state, data = df, cluster = ~state)
```

- In `modelsummary()` tables:

```
modelsummary(m, vcov = ~state)
```

- Rule of thumb: cluster at the level of treatment assignment

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

- Explicit clustering:

```
feols(y ~ x | state, data = df, cluster = ~state)
```

- In `modelsummary()` tables:

```
modelsummary(m, vcov = ~state)
```

- Rule of thumb: cluster at the level of treatment assignment
  - → If state gets the treatment, cluster by state

Clustering in practice

- `feols()` clusters by the FE variable automatically:

```
feols(y ~ x | state, data = df)
# SEs already clustered by state
```

- Explicit clustering:

```
feols(y ~ x | state, data = df, cluster = ~state)
```

- In `modelsummary()` tables:

```
modelsummary(m, vcov = ~state)
```

- Rule of thumb: cluster at the level of treatment assignment
  - → If state gets the treatment, cluster by state
  - → If individual gets the treatment, cluster by individual

Comparing specifications: a template

	(1) OLS	(2) Unit FE	(3) TWFE
Unemployment	-0.82^{***}	-0.51^{**}	-0.43^{**}
	(0.15)	(0.18)	(0.16)
State FE	No	Yes	Yes
Year FE	No	No	Yes
Clustered SE	No	Yes	Yes
N	1000	1000	1000

- Coefficient shrinks as we add FEs: the OLS estimate was partly confounded

Comparing specifications: a template

	(1) OLS	**(2) Unit FE**	**(3) TWFE**
Unemployment	-0.82^{***}	-0.51^{**}	-0.43^{**}
	(0.15)	(0.18)	(0.16)
State FE	No	Yes	Yes
Year FE	No	No	Yes
Clustered SE	No	Yes	Yes
N	1000	1000	1000

- Coefficient shrinks as we add FEs: the OLS estimate was partly confounded
- Report all three specifications for transparency

Comparing specifications: a template

	(1) OLS	(2) Unit FE	(3) TWFE
Unemployment	-0.82^{***}	-0.51^{**}	-0.43^{**}
	(0.15)	(0.18)	(0.16)
State FE	No	Yes	Yes
Year FE	No	No	Yes
Clustered SE	No	Yes	Yes
N	1000	1000	1000

- Coefficient shrinks as we add FEs: the OLS estimate was partly confounded
- Report all three specifications for transparency
- Preferred specification: (3)

Roadmap

What is Panel Data?

The Problem: Unobserved Heterogeneity

Fixed Effects

Two-Way Fixed Effects

Random Effects and the FE/RE Choice

Clustered Standard Errors

Wrap-up

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)
- Unobserved heterogeneity α_i biases cross-sectional OLS

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)
- Unobserved heterogeneity α_i biases cross-sectional OLS
  - $\rightarrow$ If α_i correlates with x_{it}: omitted variable bias

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)
- Unobserved heterogeneity α_i biases cross-sectional OLS
  - $\rightarrow$ If α_i correlates with x_{it}: omitted variable bias
- **Fixed effects** (within estimator): demeans data by unit, eliminates α_i

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)
- Unobserved heterogeneity α_i biases cross-sectional OLS
  - $\rightarrow$ If α_i correlates with x_{it}: omitted variable bias
- **Fixed effects** (within estimator): demeans data by unit, eliminates α_i
  - $\rightarrow$ Cannot estimate time-invariant variables

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)
- Unobserved heterogeneity α_i biases cross-sectional OLS
  - $\rightarrow$ If α_i correlates with x_{it}: omitted variable bias
- **Fixed effects** (within estimator): demeans data by unit, eliminates α_i
  - $\rightarrow$ Cannot estimate time-invariant variables
- **Two-way FE**: add time dummies to remove common shocks

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)
- Unobserved heterogeneity α_i biases cross-sectional OLS
  - $\rightarrow$ If α_i correlates with x_{it}: omitted variable bias
- **Fixed effects** (within estimator): demeans data by unit, eliminates α_i
  - $\rightarrow$ Cannot estimate time-invariant variables
- **Two-way FE**: add time dummies to remove common shocks
- **FE vs. RE**: use FE when unit effects may correlate with X; use Hausman test

Summary: key takeaways

- Panel data: N units $\times$ T time periods; indexed (i, t)
- Unobserved heterogeneity α_i biases cross-sectional OLS
  - $\rightarrow$ If α_i correlates with x_{it}: omitted variable bias
- **Fixed effects** (within estimator): demeans data by unit, eliminates α_i
  - $\rightarrow$ Cannot estimate time-invariant variables
- **Two-way FE**: add time dummies to remove common shocks
- **FE vs. RE**: use FE when unit effects may correlate with X; use Hausman test
- Always **cluster standard errors** by unit

For next session

- Complete Assignment 5
- Read the assigned paper using panel FE
- Next session: Panel Data II
  - → Difference-in-Differences (DiD)
  - → Event studies
  - → Staggered treatment timing
  - → Recent advances in DiD (Callaway–Sant'Anna, etc.)

Questions?