Panel Data

EC 421, Set 12

Luciana Etcheverry
November 26, 2019

Prologue

Schedule

Last Time

Instrumental variables and causality

Today

Panel data

Upcoming

- Assignment due **Tomorrow**
- Office hours:
  - Today: 3-4pm
  - Wednesday: 10-12pm
- Practice Final

Panel data

Panel data

Intro

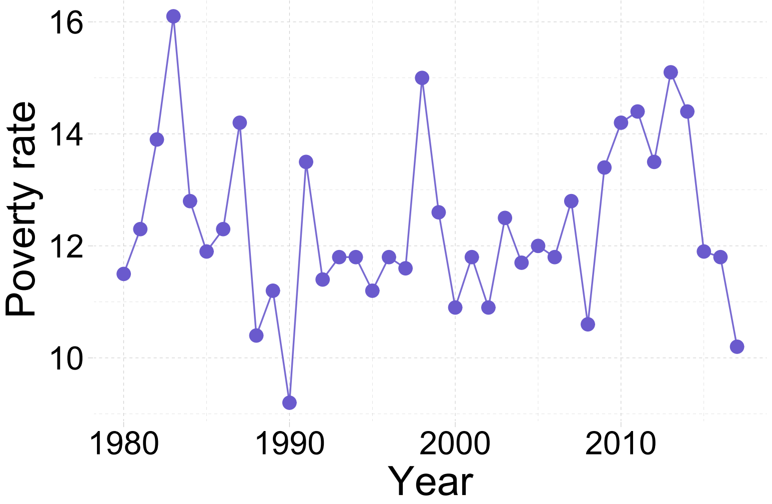
Panel data

Intro

We've considered two types of data (each with one dimension):

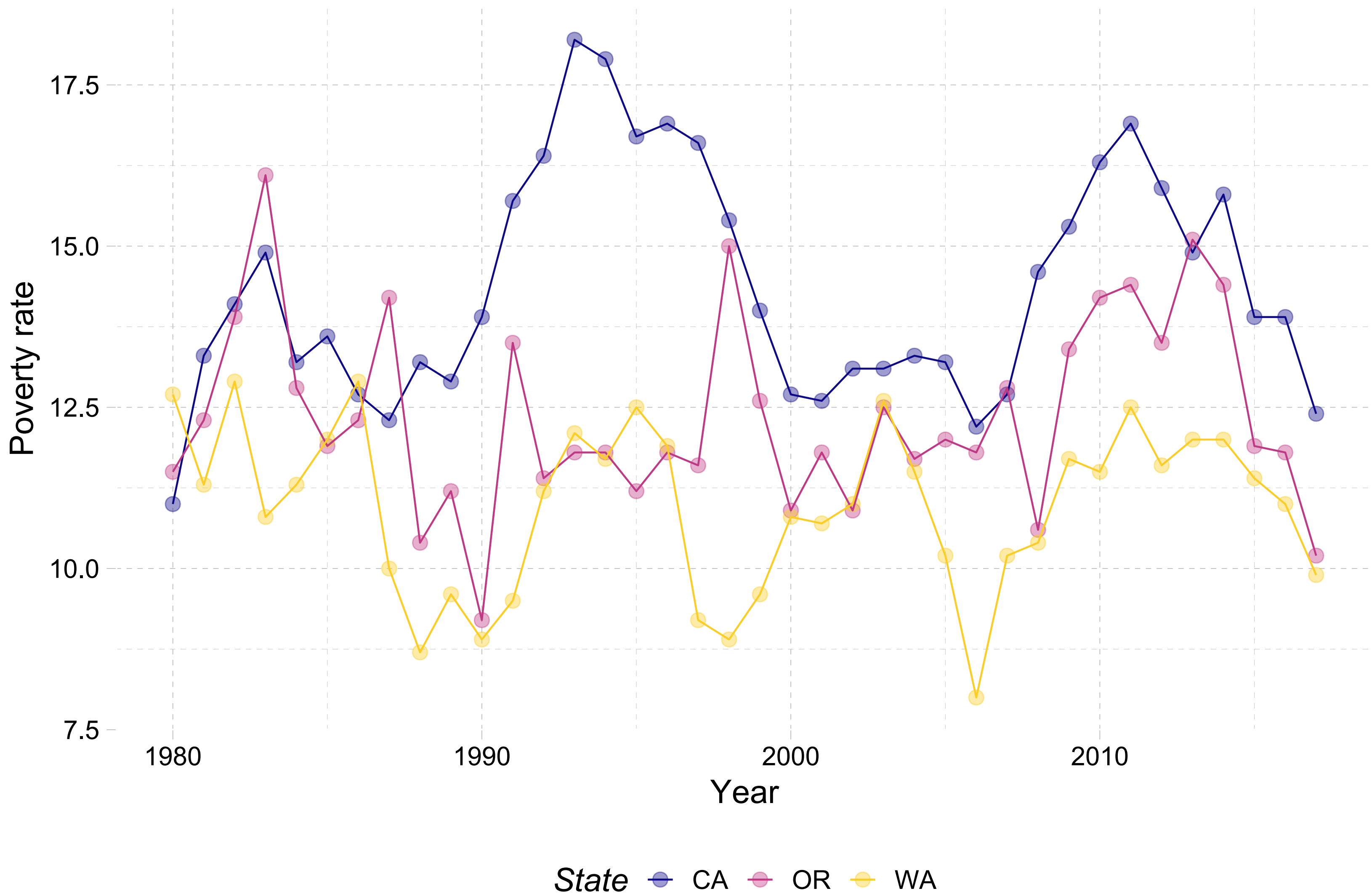
Cross-sectional data: individual i

Time-series data: time t



Panel data combine these data types/dimensions: individual i **at** time t.

Panel data combine these data types/dimensions: individual i **at** time t.



Panel data

Definition

With ***panel data***, we have

- **repeated observations** (t)
- on **multiple indiviuals** (i).

```
#>   state year poverty_rate min_wage
#> 1    CA 1990         13.9     4.25
#> 2    CA 2000         12.7     6.25
#> 3    CA 2010         16.3     8.00
#> 4    OR 1990          9.2     4.25
#> 5    OR 2000         10.9     6.50
#> 6    OR 2010         14.2     8.40
#> 7    WA 1990          8.9     4.25
#> 8    WA 2000         10.8     6.50
#> 9    WA 2010         11.5     8.55
```

Thus, our regression equation with a panel dataset looks like

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

for individual i in time t.

Panel data

Example

Minimum-wage laws involve many contentious/important policy questions.

- Do minimum-wage laws **increase well-being** for minimum-wage earners and their families?
- Do minimum-wage laws **increase unemployment**?
- Overall, do minimum-wage laws **decrease poverty**?

We want to know the causal effect of the minimum wage, *i.e.*, β_1 in

$$\left(\text{Poverty Rate}\right)_{it} = \beta_0 + \beta_1 \left(\text{Min. Wage}\right)_{it} + u_{it}$$

where i denotes state and t indexes year.

Panel data

Example

If we go ahead and run OLS in our panel, we find

OLS w/ outcome variable 'poverty rate'

Term	**Est.**	**S.E.**	**t stat.**	**p-Value**
Intercept	14.196	0.283	50.21	<0.0001
Min. Wage	**-0.203**	**0.051**	**-3.99**	**<0.0001**

which suggests that a one-dollar increase in the minimum wage significantly *reduces* poverty by approximately 0.203 percentage points.

Surprising?

Panel data

Example: Causality is still hard

To isolate the causal effect of minimum wage on poverty in

$$\left(\textbf{Poverty Rate}\right)_{it} = \beta_0 + \beta_1 \left(\textbf{Min. Wage}\right)_{it} + u_{it}$$

We still need exogeneity, *i.e.*, $\boldsymbol{E}[u_{it} \mid (\textbf{Min. Wage})] = 0$.

Exogeneity with *panel data:* Are there omitted factors that affect both a state's minimum wage *and* its poverty rate?

We are going to discuss two common panel-data strategies:

1. **Fixed effects**
2. **First differences**

Fixed effects

Fixed effects are binary indicator variables that *help* control for unobserved differences across individuals or time periods.

For example, we can include a **fixed effect for each individual state i** to control for unobserved, time-invariant differences between states:

$$(\textbf{Poverty Rate})_{it} = \beta_0 + \beta_1(\textbf{Min. Wage})_{it} + \text{State}_i + u_{it}$$

```
#>   state year poverty_rate min_wage fe_ca fe_or fe_wa
#> 1    CA 2000         12.7     6.25     1     0     0
#> 2    CA 2010         16.3     8.00     1     0     0
#> 3    OR 2000         10.9     6.50     0     1     0
#> 4    OR 2010         14.2     8.40     0     1     0
#> 5    WA 2000         10.8     6.50     0     0     1
#> 6    WA 2010         11.5     8.55     0     0     1
```

Panel data

Fixed effects

Notice that these individual fixed effects are just individual-specific intercepts—now each unit/individual gets her own intercept.

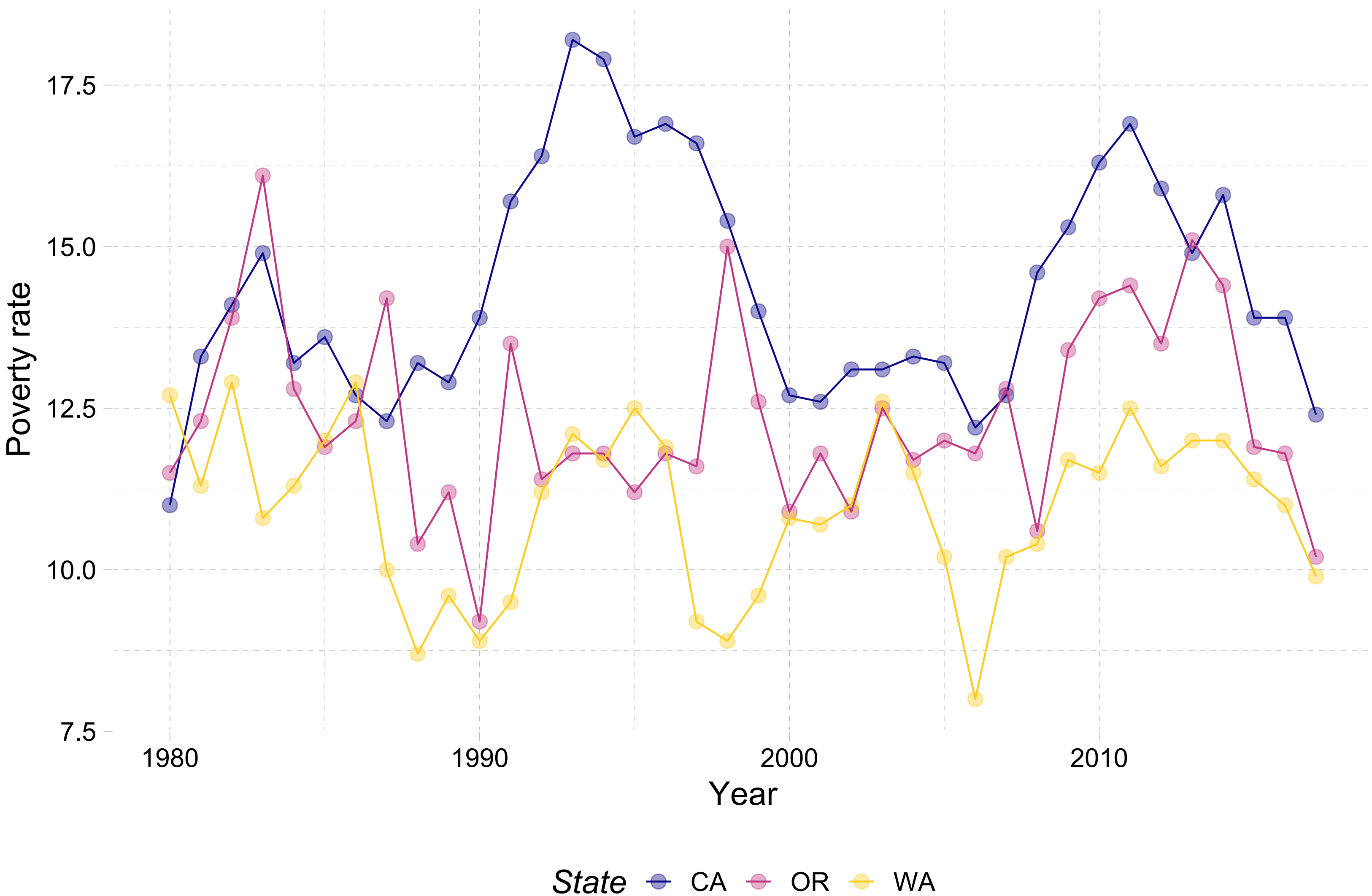
Q: What are these individual-level fixed effects (FEs) doing?

A$_1$: They remove each individual's mean, *i.e.*, $y_{it} - \overline{y}_i$ and $x_{it} - \overline{x}_i$.

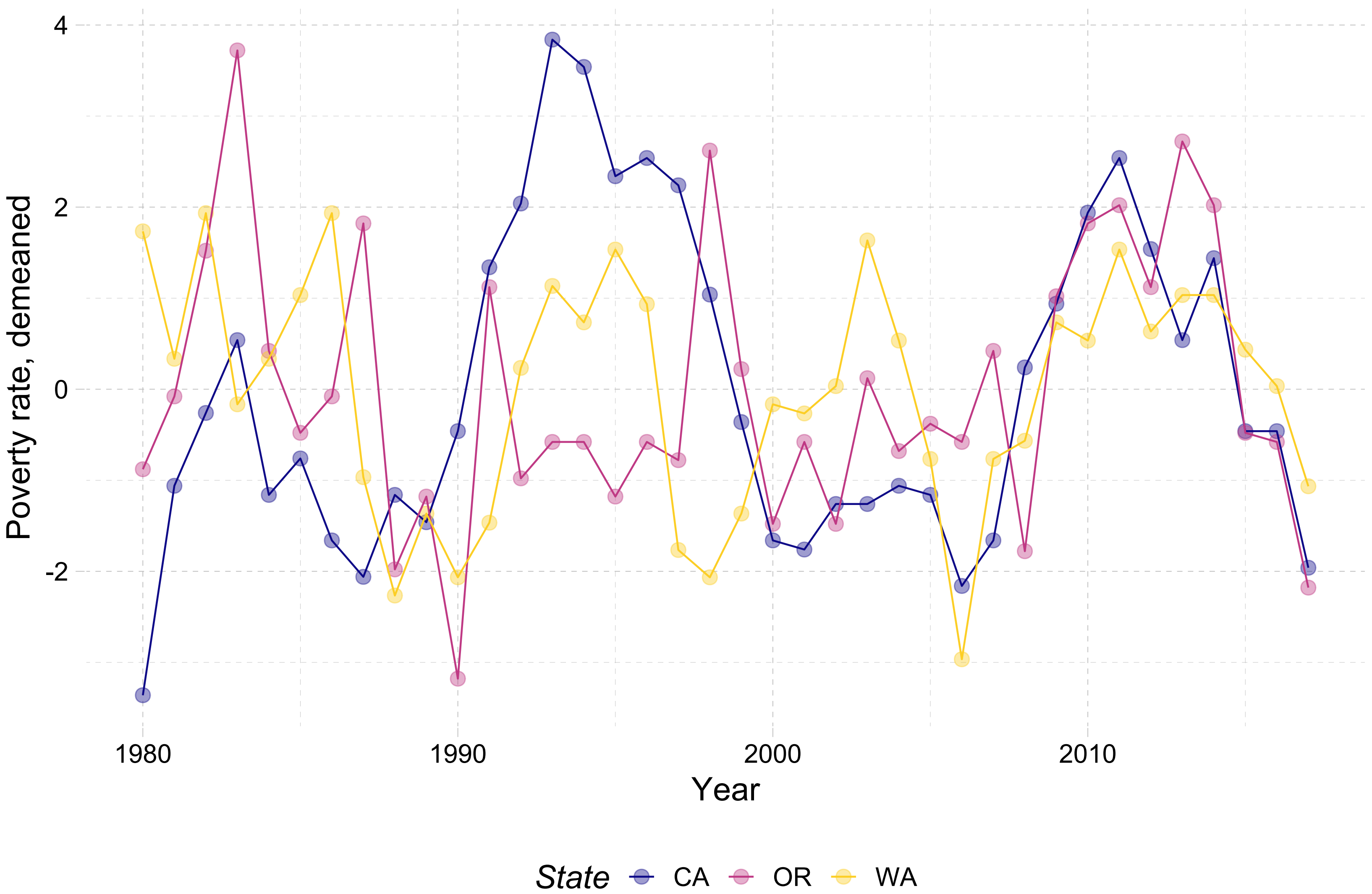
A$_2$: They control for unobserved, time-invariant differences between units.†

† By *time-invariance differences* we mean differences between individuals that do not change over time.

In the raw data (no fixed effects/demeaning), individuals differ in levels.



Individual-fixed effects remove individuals' means.



Fixed effects

Fixed effects are one method econometricians try to "match" individuals to generate a valid control group for our treated individuals.

Toward this goal, we include **fixed effects for each time period t**, to (attempt to) control for shocks that affected all observations.

$$\text{(Poverty Rate)}_{it} = \beta_0 + \beta_1 (\text{Min. Wage})_{it} + \text{State}_i + \text{Year}_t + u_{it}$$

```
#>   state year poverty_rate min_wage fe_ca fe_or fe_wa fe_2000 fe_2010
#> 1    CA 2000         12.7     6.25     1     0     0       1       0
#> 2    CA 2010         16.3     8.00     1     0     0       0       1
#> 3    OR 2000         10.9     6.50     0     1     0       1       0
#> 4    OR 2010         14.2     8.40     0     1     0       0       1
#> 5    WA 2000         10.8     6.50     0     0     1       1       0
#> 6    WA 2010         11.5     8.55     0     0     1       0       1
```

Panel data

Fixed-effects estimation in `R`

`R` makes estimation with fixed-effects really easy.

As always, you have options.

We're going to use the `felm()` function from the `lfe` package.

General notation:

```
felm(y ~ x1 + x2 + … | fe1 + fe2 …, data = some_data)
```

Our example:

```
felm(poverty_rate ~ min_wage | state + year, data = panel_df)
```

Fixed-effects estimation in R

```
felm(poverty_rate ~ min_wage | state + year, data = panel_df)
```

Fixed effects w/ outcome variable 'poverty rate'

Term	Est.	S.E.	t stat.	p-Value
Min. Wage	**0.374**	**0.109**	**3.43**	**0.0006**

```
lm(poverty_rate ~ min_wage, data = panel_df)
```

OLS w/ outcome variable 'poverty rate'

Term	Est.	S.E.	t stat.	p-Value
Intercept	14.196	0.283	50.21	<0.0001
Min. Wage	**-0.203**	**0.051**	**-3.99**	**<0.0001**

Fixed-effects estimation in R

Q: Which set of estimates should we believe?

A: The set that you believe meets our exogeneity requirement.

First differences

Another route—related to our time-series studies—uses ***first differences***.

The ***first difference*** for variable y is the difference between individual i's current value of y (*i.e.*, $y_{i,t}$) and his previous (lagged) value of y (*i.e.*, $y_{i,t-1}$).

We write the first difference as

$$\Delta y_{it} = y_{i,t} - y_{i,t-1}$$

First differences

From our example, write the model for t and $t-1$

$$\begin{aligned}
(\text{Poverty Rate})_{i,t} &= \beta_0 + \beta_1 (\text{Min. Wage})_{i,t} + u_{i,t} && \text{(t)}\\
(\text{Poverty Rate})_{i,t-1} &= \beta_0 + \beta_1 (\text{Min. Wage})_{i,t-1} + u_{i,t-1} && \text{(t-1)}
\end{aligned}$$

taking the difference between (t) and $(t-1)$ gives

$$\begin{aligned}
&(\text{Poverty Rate})_{i,t} - (\text{Poverty Rate})_{i,t-1} = \\
&\qquad \beta_0 - \beta_0 + \beta_1 (\text{Min. Wage})_{i,t} - \beta_1 (\text{Min. Wage})_{i,t-1} + u_{i,t} - u_{i,t-1}
\end{aligned}$$

which implies

$$\Delta (\text{Poverty Rate})_{i,t} = \beta_1 \Delta (\text{Min. Wage})_{i,t} + \Delta u_{i,t}$$

First differences

Estimating our model via first differences gives us the results

First diff. w/ outcome variable 'poverty rate'

Term	Est.	S.E.	t stat.	p-Value
Intercept	-0.064	0.047	-1.34	0.1811
Min. Wage	**0.221**	**0.157**	**1.41**	**0.1584**

Fixed effects w/ outcome variable 'poverty rate'

Term	Est.	S.E.	t stat.	p-Value
Min. Wage	**0.374**	**0.109**	**3.43**	**0.0006**

OLS w/ outcome variable 'poverty rate'

Term	Est.	S.E.	t stat.	p-Value
Intercept	14.196	0.283	50.21	<0.0001
Min. Wage	**-0.203**	**0.051**	**-3.99**	**<0.0001**

Q: Conclusions?

A: Models (and their requirements) can ***really*** affect your results

Q: Comparing FE and FD: Which one should you choose?

A:

- If there is no serial correlation in the error term, we know FE is BLUE, so that is the clear choice. Alternatively, suppose the residual follows a random walk. Now, FD is much better because it eliminates the autocorrelation. In between these two extremes, it is difficult to know which is more efficient.
- When the time span is long, the assumption that individual's FE is not time-varying becomes an untenable assumption. In this case, both FE and FD will yield biased estimates and the estimates from the two methods will tend to diverge.

Table of contents