Panel Data Estimators

William Brasic

The University of Arizona

Panel Data Set

Definition 1: Pooled Cross Section

A pooled cross-sectional data set contains data on *potentially different* agents observed at *multiple* points in time.

Panel Data Set

Definition 2: Panel Data Set

A panel data set contains data on the *same* agents observed at *multiple* points in time.

- E.g., tracking each country's GDP for the last five years.

Panel Data Set

Definition 3: Balanced Panel Data Set

A balanced panel data set is a data set where all individuals are observed in all time periods.

- Otherwise, we have an *unbalanced* panel data set.

Panel Data Set Advantages

Property 1: Panel Data Set Advantages

The advantages of a panel data set include:

- Tracking dynamics over time.
- More accurate regression models.
- Controlling for individual heterogeneity.
  - Potential solution to the endogeneity problem!

Panel Data Set Examples

Example 1: Panel Data Set Examples

- Studying the effect of policy changes on economic growth.
- Evaluating the effects of job training program on long-run income.
- Analyzing consumer behavior over time.

Panel Data Estimator Examples

Example 2: Panel Data Estimator Examples

- Pooled OLS: Typically used when only have a pooled cross-section. Estimation is the same as OLS.

Panel Data Estimator Examples

Example 2: Panel Data Estimator Examples

- Pooled OLS: Typically used when only have a pooled cross-section. Estimation is the same as OLS.
- Differenced Estimators: Use the difference in values of the variables over time to remove time-invariant individual effects.

Panel Data Estimator Examples

Example 2: Panel Data Estimator Examples

- Pooled OLS: Typically used when only have a pooled cross-section. Estimation is the same as OLS.
- Differenced Estimators: Use the difference in values of the variables over time to remove time-invariant individual effects.
- Fixed Effect Estimators: Can control for time-variant or individual-variant factors, or both.

Panel Data Estimator Examples

Example 2: Panel Data Estimator Examples

- Pooled OLS: Typically used when only have a pooled cross-section. Estimation is the same as OLS.
- Differenced Estimators: Use the difference in values of the variables over time to remove time-invariant individual effects.
- Fixed Effect Estimators: Can control for time-variant or individual-variant factors, or both.
- Difference-in-Differences (DiD) Estimators: Compare the differences in outcomes over time between a treatment group and a control group to estimate causal effects.

Pooled OLS

Definition 4: Pooled OLS w/ a Pooled Cross-Section

Pooled OLS is an estimation method typically used with a pooled cross-sectional data set that estimates β in the same way as traditional OLS does.

- It does not account for individual-specific effects or time-specific effects.
- Does allow us to include time indicator variables.

Pooled OLS

Definition 5: Pooled OLS Equation w/ a Pooled Cross-Section

The pooled OLS equation is given by

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \underbrace{c_i + u_{it}}_{\text{composite error term}} \\ &= \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + u_{it}. \end{aligned}$$

- Some of the x's could now be time indicator variables.
- Works well if we think $c_i = 0$ for each $i = 1, \ldots, n$ (unlikely).

Pooled OLS Time Indicators

Property 2: Time Indicator Variables

A critical aspect of pooled OLS is that we can estimate:

- How the dependent variable is impacted over time.
- How the effect of our covariates on our outcome evolves over time.

Pooled OLS Examples

Example 3: Time Controls

Suppose we have data in the even years from 1972 to 1976 on the number of kids born to woman, the age of the woman, and the woman's education. Our estimated regression is

$$\begin{aligned} kids_{it} = &-7.742 - 0.128 educ_{it} + 0.532 age_{it} \\ &- 0.0058 age_{it}^2 - 0.522 year_{1974} - 0.545 year_{1976}. \end{aligned}$$

- A woman in 1974 has an estimated 0.522 less children relative to 1972.
- One-hundred women in 1976 has an estimated 54 fewer children relative to one-hundred women in 1972.

Pooled OLS Examples

Example 4: Time Controls

Suppose we collect data on wage that is pooled across 1978 and 1985, and estimate the model of

$$\log(wage_{it}) = \beta_0 + \delta_0 year_{1985} + \beta_1 educ + \delta_1 year_{1985} \times educ + u_{it}.$$

- The return to education in 1985 is given by $\beta_1 + \delta_1$.

- Thus, δ_1 measures the change in the return to education over the seven year period.

Pooled OLS Advantages

Property 3: Advantages of Pooled OLS

- Simple to implement and interpret.
- Useful as a baseline model for comparison.

Pooled OLS Disadvantages

Property 4: Disadvantages of Pooled OLS

- If you have a panel data rather than the pooled cross-section, you aren't taking full advantage of your data set to eliminate potential sources of endogeneity.
- Can lead to biased and inconsistent estimates if individual-specific effects, c_i, are correlated with the explanatory variables, x_{it}.

Pooled OLS Disadvantage

Example 5: Pooled OLS Disadvantage

Suppose we have a panel data set on wages and education levels. Our model is

$$wage_{it} = \beta_0 + \beta_1 educ_{it} + a_i + u_{it}.$$

Innate ability is something that could be included in a_i that definitely impacts education and wages. Thus, by not accounting for a_i, we have an OVB issue!

- Pooled OLS doesn't help us fix issues like these.

Pooled OLS Alternatives

Question 1: Individual Specific Heterogeneity

How can we control for the individual unobserved *fixed* effect a_i?

Answer to Question 1

- First Differencing Estimator.
- Fixed Effects Estimator.

- We need a panel data set to control for the a_i (panel data is harder to collect).
- We're able to remove things from the unobservable making our estimators less biased.
- Panel data methods allow us to now control for things we couldn't with a cross-sectional data set such as innate ability, land quality, etc.

First Differencing

Definition 6: First Differencing

The process of first differencing involves transforming the data by subtracting the value of each variable at time t from its value at time $t-1$.

- This allows us to control for the individual specific error component that is fixed across time (a_i).
- By including time indicators after differencing, we can control for the unobserved time component, w_t, as well.

First Differencing

Definition 7: First Differencing Equation

If

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \underbrace{c_i + u_{it}}_{\text{composite error}},$$

then

$$y_{it} - y_{i(t-1)} = \left(\boldsymbol{x}_{it} - \boldsymbol{x}_{i(t-1)}\right)'\boldsymbol{\beta} + \left(u_{it} - u_{i(t-1)}\right)$$

is called the first-differenced equation.

- a_i is constant across time so it drops out!
- We now have $n(T-1)$ observations.

First Differencing Advantages

Property 5: Advantages of First Differencing

- Eliminates unobserved individual-specific effects (c_i), reducing bias.
- Controls for time-invariant characteristics of individuals.
- Simple and intuitive transformation.

First Differencing Disadvantages

Property 6: Disadvantages of First Differencing

- Loss of one period of data per cross section, reducing sample size.
- May not control for time-variant unobserved heterogeneity (w_t) effectively.
  - $w_t - w_{t-1}$ may be smaller, but likely won't be zero unless the time effect is constant across t.
- Biased in the present of error serial correlation $\text{Cov}(u_{it}, u_{i(t-1)}) \neq 0$.

Fixed Effects

Definition 8: Fixed Effects

Fixed effects (within) estimation is a method used in panel data analysis to control for unobserved heterogeneity via the process of *mean differencing* or *indicator variables*.

Individual Fixed Effects

Definition 9: Individual Fixed Effects

If

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \underbrace{c_i + u_{it}}_{\text{composite error}},$$

then

$$y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} = \left(\boldsymbol{x}_{it} - \overline{\boldsymbol{x}}_i\right)'\boldsymbol{\beta} + \left(u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it}\right)$$

is called the (individual) fixed effects model.

- We remove the unobserved effect that varies across individuals, but is fixed across time.

Time Fixed Effects

Definition 10: Time Fixed Effects

If

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \underbrace{w_t + u_{it}}_{\text{composite error}},$$

then

$$y_{it} - \frac{1}{n}\sum_{i=1}^{n} y_{it} = \left(\boldsymbol{x}_{it} - \overline{\boldsymbol{x}}_t\right)'\boldsymbol{\beta} + \left(u_{it} - \frac{1}{n}\sum_{i=1}^{n} u_{it}\right)$$

is called the (time) fixed effects model.

- We remove the unobserved effect that varies across time, but is fixed across individuals.

Two-Way Fixed Effects

Definition 11: Two-Way Fixed Effects

The two-way fixed effect model includes both individual *and* time fixed effects to control for unobserved heterogeneity across both dimensions.

- Most robust form of fixed effect estimation.
- Subtract mean across cross-section and mean across time from the original data.

Two-Way Fixed Effects

Example 6: Two-Way Fixed Effects

Suppose we are examining the impact of a new policy on the productivity of firms across different regions over a ten year period. Our model would be:

$$Productivity_{it} = \beta_0 + \beta_1 Policy_{it} + \boldsymbol{x}'_{it}\boldsymbol{\delta} + \underbrace{c_i + w_t + u_{it}}_{\text{composite error}}.$$

- Factors specific to each region (contained in c_i) could include climate and infrastructure quality.
- Factors specific to each time period (contained in w_t) could include recessions or changes in technology.

- Two-way fixed effects would remove these sources of endogeneity!

Fixed Effects

Property 7: Fixed Effects

Including individual specific indicator variables will produce practically identical results to the demeaning process described above.

- We typically like the demeaning process because if we have a panel data set with thousands of categories, the indicator variable approach produces a computational burden.

Fixed Effects or First Differences?

Question 2: Fixed Effects or First Differences?

Do we like fixed effects or first differences better?

Fixed Effects or First Differences?

Question 2: Fixed Effects or First Differences?

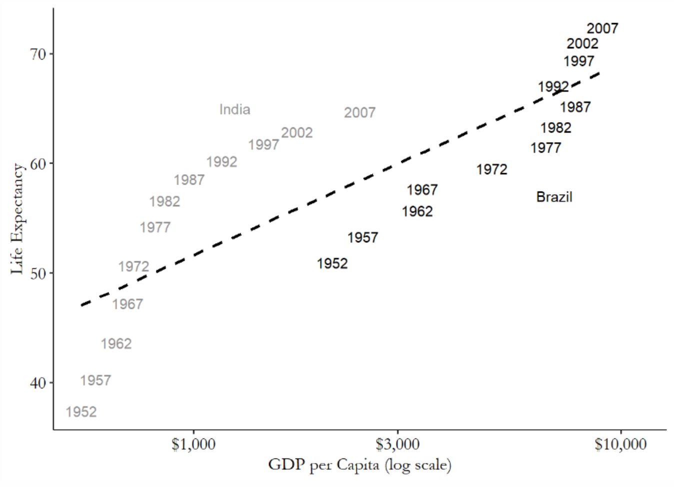
Do we like fixed effects or first differences better?

Answer to Question 2

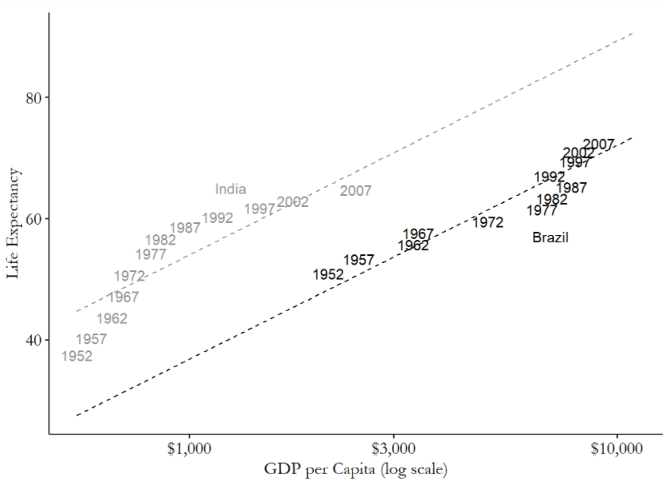
It's good to estimate both models and report their results. However, in general, economists like fixed effects because:

- There is no loss of data.
- We can effectively control for the unobserved time-variant effect better.

Without Fixed Effects



With Fixed Effects



- What if I wanted to include differing slopes across countries?

Difference-in-Differences (DiD)

Definition 12: Difference-in-Differences (DiD)

Difference-in-Differences (DiD) is a pooled cross-section/panel data estimator typically used to evaluate the effects of a policy.

Difference-in-Differences (DiD) Model

Definition 12: Difference-in-Differences (DiD) Model

The basic DiD model is often written as:

$$y_{it} = \beta_0 + \beta_1 \text{post}_t + \beta_2 \text{treatment}_i + \beta_3 (\text{treatment}_i \times \text{post}_t) + \epsilon_{it}$$

where

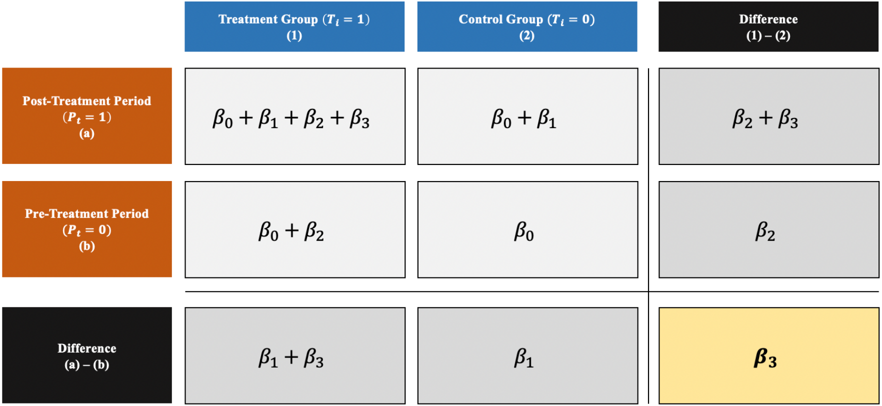
- y_{it}: The outcome variable for unit i at time t.
- post_t: A binary indicator that equals 1 if time t is after the treatment has been implemented, 0 otherwise.
- treatment_i: A binary indicator that equals 1 if unit i is in the treatment group, 0 otherwise.
- $\text{treatment}_i \times \text{post}_t$: The interaction term, which captures the effect of the treatment after it has been implemented.

Difference-in-Differences (DiD) Model

Property 8: Difference-in-Differences (DiD) Interpretation

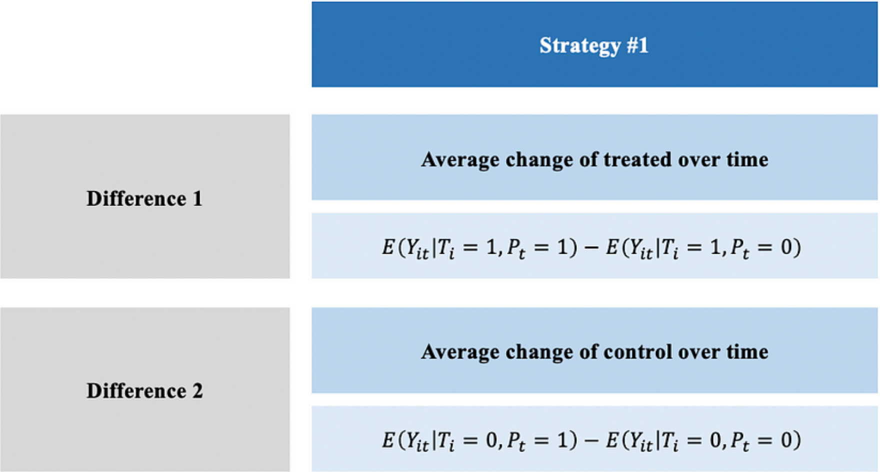
- β_1 represents the change in the average outcome over time for the control group (i.e., the time effect).
- β_2 represents the difference in the average outcome between the treated and untreated groups before the treatment is implemented.
- β_3 represents the treatment effect, which is the difference in the change in the average outcome between the treatment and control groups over time.

Difference-in-Differences (DiD)



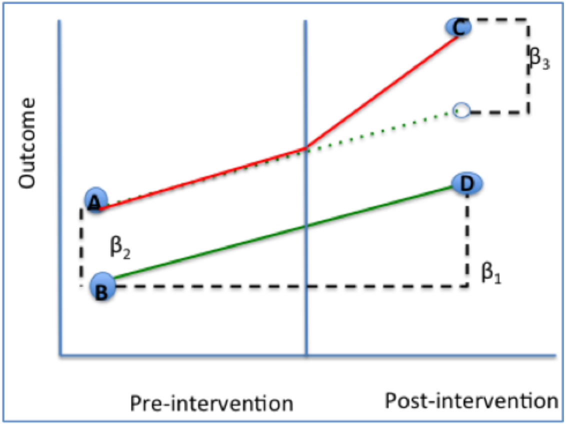
	Treatment Group ($T_i = 1$) **(1)**	**Control Group ($T_i = 0$)** **(2)**	**Difference** **(1) – (2)**
Post-Treatment Period **($P_t = 1$)** **(a)**	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$	$\beta_2 + \beta_3$
Pre-Treatment Period **($P_t = 0$)** **(b)**	$\beta_0 + \beta_2$	β_0	β_2
Difference **(a) – (b)**	$\beta_1 + \beta_3$	β_1	$\boldsymbol{\beta_3}$

Difference-in-Differences (DiD)

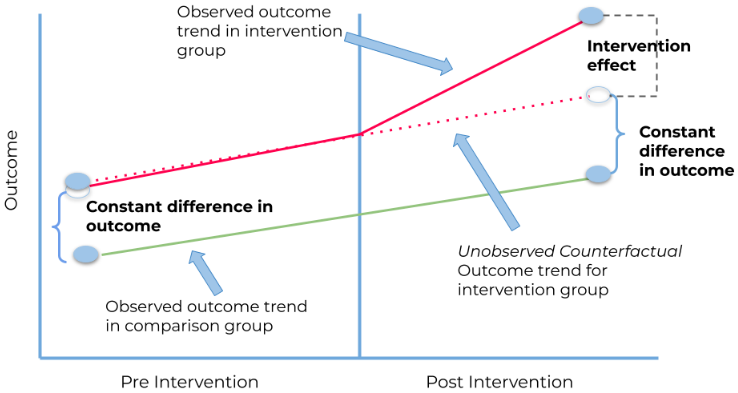


	Strategy #1
Difference 1	Average change of treated over time $E(Y_{it}\|T_i = 1, P_t = 1) - E(Y_{it}\|T_i = 1, P_t = 0)$
Difference 2	Average change of control over time $E(Y_{it}\|T_i = 0, P_t = 1) - E(Y_{it}\|T_i = 0, P_t = 0)$

Difference-in-Differences (DiD)



Difference-in-Differences (DiD)


Observed outcome trend in intervention group
Intervention effect
Constant difference in outcome
Outcome
Constant difference in outcome
Unobserved Counterfactual Outcome trend for intervention group
Observed outcome trend in comparison group
Pre Intervention
Post Intervention

Difference-in-Differences (DiD)

Example 7: Difference-in-Differences

Suppose we are analyzing the impact of a new healthcare policy on the health outcomes of individuals across two different states, where one state implements the policy and the other does not. The DiD model would be:

$$\begin{aligned} HealthOutcome_{it} &= \beta_0 + \beta_1 Policy_i + \beta_2 Post_t \\ &\quad + \beta_3 (Policy_i \times Post_t) + \boldsymbol{x}_{it}' \boldsymbol{\delta} + u_{it}. \end{aligned}$$

Difference-in-Differences (DiD)

Example 7: Difference-in-Differences

Suppose we are analyzing the impact of a new healthcare policy on the health outcomes of individuals across two different states, where one state implements the policy and the other does not. The DiD model would be:

$$\begin{aligned} HealthOutcome_{it} &= \beta_0 + \beta_1 Policy_i + \beta_2 Post_t \\ &\quad + \beta_3 (Policy_i \times Post_t) + \boldsymbol{x}'_{it}\boldsymbol{\delta} + u_{it}. \end{aligned}$$

- $Policy_i$ is an indicator variable that equals 1 if the individual is in the treatment state, and 0 otherwise.
- $Post_t$ is an indicator variable that equals 1 in the period after the policy implementation, and 0 otherwise.
- The interaction term $(Policy_i \times Post_t)$ captures the differential effect of the policy over time.

Interpreting DiD Estimates

Property 9: Interpreting DiD Estimates

The coefficient on the interaction term in the DiD model provides the estimate of the causal effect of the policy on the outcome variable, assuming that the parallel trends assumption holds.

- The parallel trends assumption implies that in the absence of the treatment, the difference in outcomes between the treatment and control groups would have remained constant over time.

DiD Assumptions

Property 10: DiD Assumptions

The validity of the DiD estimator relies on key assumptions, particularly:

- **Parallel Trends Assumption**: The treatment and control groups would have followed the same trend over time in the absence of the treatment.
- **No Simultaneous Treatment Effects**: No other events or policies differentially affect the treatment and control groups at the same time as the policy being studied.

Advantages of DiD

Question 3: Advantages of DiD

What are the advantages of using a Difference-in-Differences approach?

Advantages of DiD

Question 3: Advantages of DiD

What are the advantages of using a Difference-in-Differences approach?

Answer to Question 3

The Difference-in-Differences approach is advantageous because:

- It allows for controlling for individual and time fixed effects that could bias estimates.
- It is simple to implement and interpret.
- It does not require the availability of panel data for every individual in the sample.
  - However, the panel data approach is more robust.

Thank You!