Panel Data

Antonella Bandiera

March 28, 2023

Last class

- We like DD designs:
  - We can leverage time variation
  - And cross-sectional variation

Last class

- We like DD designs:
  - We can leverage time variation
  - And cross-sectional variation
- DD is very good as far as quasi-experiments go
- We can implement it easily via regression

Estimating DD via regression

We can just do:

$$\hat{\tau}_{DD} = (\bar{Y}(\text{treat,post}) - \bar{Y}(\text{treat,pre})) - (\bar{Y}(\text{untreat,post}) - \bar{Y}(\text{untreat,pre}))$$

Estimating DD via regression

We can just do:

$\hat{\tau}_{DD} = (\bar{Y}(\text{treat,post}) - \bar{Y}(\text{treat,pre})) - (\bar{Y}(\text{untreat,post}) - \bar{Y}(\text{untreat,pre}))$

Via regression:

$$Y_i = \alpha + \tau \text{Treat}_i \times \text{Post}_t + \beta \text{Treat}_i + \delta \text{Post}_t + \varepsilon_i$$

Estimating DD via regression

We can just do:

$\hat{\tau}_{DD} = (\bar{Y}(\text{treat,post}) - \bar{Y}(\text{treat,pre})) - (\bar{Y}(\text{untreat,post}) - \bar{Y}(\text{untreat,pre}))$

Via regression:

$$Y_i = \alpha + \tau \text{Treat}_i \times \text{Post}_t + \beta \text{Treat}_i + \delta \text{Post}_t + \varepsilon_i$$

Consider:

$\bar{Y}(\text{treat,post}) = \hat{\alpha} + \hat{\tau} + \hat{\beta} + \hat{\delta}$

$\bar{Y}(\text{treat,pre}) = \hat{\alpha} + \hat{\beta}$

$\rightarrow \bar{Y}(\text{treat,post}) - \bar{Y}(\text{treat,pre}) = \hat{\delta} + \hat{\tau}$

$\bar{Y}(\text{untreat,post}) = \hat{\alpha} + \hat{\delta}$

$\bar{Y}(\text{untreat,pre}) = \hat{\alpha}$

$\rightarrow \bar{Y}(\text{untreat,post}) - \bar{Y}(\text{untreat,pre}) = \hat{\delta}$

Identification assumption

The identifying assumption is:

In words: Parallel counterfactual trends

In math: $E[\varepsilon_{it} | \text{Treat}_i, \text{Post}_t, X_{it}] = 0$

In other words: Conditional on covariates, treatment is as good as randomly assigned; or treated and untreated units would be on similar trajectories if not for treatment

- We need a good story for why $\text{Treat}_i \times \text{Post}_t$ is quasi-random

Fixed effects models

DD is a specific case of the fixed effects model

- We can incorporate more units and time periods
- Different units can be treated at different times
- The effect of the treatment can vary over time
- We can assess the plausibility of the identifying assumption

Fixed effects models

Consider a basic panel data regression

$$Y_{it} = \beta X_{it} + \tau D_{it} + \varepsilon_{it}$$

Fixed effects models

Consider a basic panel data regression

$$Y_{it} = \beta X_{it} + \tau D_{it} + \varepsilon_{it}$$

We can split ε_{it} into three:

- α_i is an individual-specific time-invariant part
- δ_t is a common time-period specific part
- v_{it} is the remaining individual-by-time varying unobserved part

Fixed effects models

Consider a basic panel data regression

$$Y_{it} = \beta X_{it} + \tau D_{it} + \varepsilon_{it}$$

We can split ε_{it} into three:

- α_i is an individual-specific time-invariant part
- δ_t is a common time-period specific part
- v_{it} is the remaining individual-by-time varying unobserved part

Is there any way to remove at least some of these bits from the error term?

Fixed effects model

Enter the fixed effects model

- We want to control for α_i and δ_t: they may be correlated with D_i

Fixed effects model

Enter the fixed effects model

- We want to control for α_i and δ_t: they may be correlated with D_i

There are two ways of doing this:

1. Use dummy variables
2. De-meaning the data (equivalent)

Fixed effects model

Enter the fixed effects model

- We want to control for α_i and δ_t: they may be correlated with D_i

There are two ways of doing this:

1. Use dummy variables
2. De-meaning the data (equivalent)

Both are fixed-effects estimators

Fixed effects estimator 1: Dummy variables

Consider the following regression model with only individual (no time) effects

$$Y_{it} = \beta X_{it} + \tau D_{it} + \alpha_i + v_{it}$$

Fixed effects estimator 1: Dummy variables

Consider the following regression model with only individual (no time) effects

$$Y_{it} = \beta X_{it} + \tau D_{it} + \alpha_i + v_{it}$$

The dummy variable approach:

- We want to control for α_i

Fixed effects estimator 1: Dummy variables

Consider the following regression model with only individual (no time) effects

$$Y_{it} = \beta X_{it} + \tau D_{it} + \alpha_i + v_{it}$$

<u>The dummy variable approach:</u>

- We want to control for α_i
- What are those α_is?

Fixed effects estimator 1: Dummy variables

Consider the following regression model with only individual (no time) effects

$$Y_{it} = \beta X_{it} + \tau D_{it} + \alpha_i + v_{it}$$

The dummy variable approach:

- We want to control for α_i
- What are those α_is?
- Individual specific effects

Fixed effects estimator 1: Dummy variables

Consider the following regression model with only individual (no time) effects

$$Y_{it} = \beta X_{it} + \tau D_{it} + \alpha_i + v_{it}$$

The dummy variable approach:

- We want to control for α_i
- What are those α_is?
- Individual specific effects
- So we can add a control variable for each person
- We do this with dummy variables $I_i = 1$ for unit i, 0 for all $j \neq i$

$$Y_{it} = \beta X_{it} + \tau D_{it} + \sum_{i=1}^{N} 1[\text{unit} = 1] + v_{it}$$

Intuition: Example (From NHK)

- We want to know what the effect is of exercise on health
  - Many things can affect health, like genetics, dietary preferences
- We have the following data on exercise

Intuition: Example (From NHK)

- We want to know what the effect is of exercise on health
  - Many things can affect health, like genetics, dietary preferences
- We have the following data on exercise

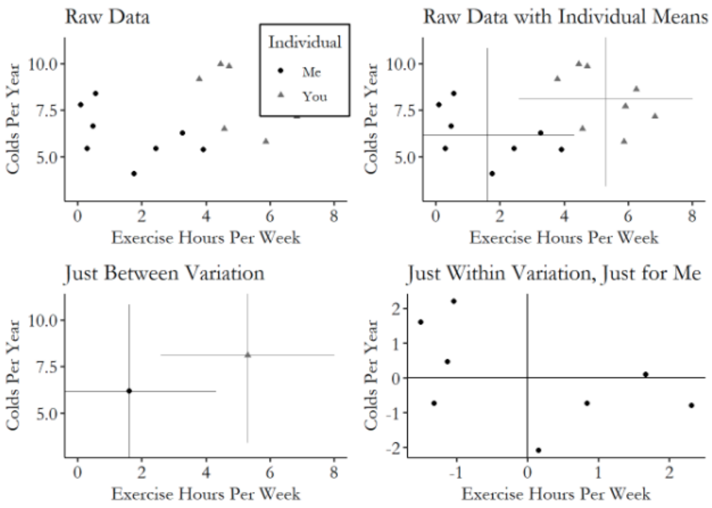
	Year	Excercise
You	2019	5
You	2020	7
Me	2019	4
Me	2020	3

Example (cont'd)

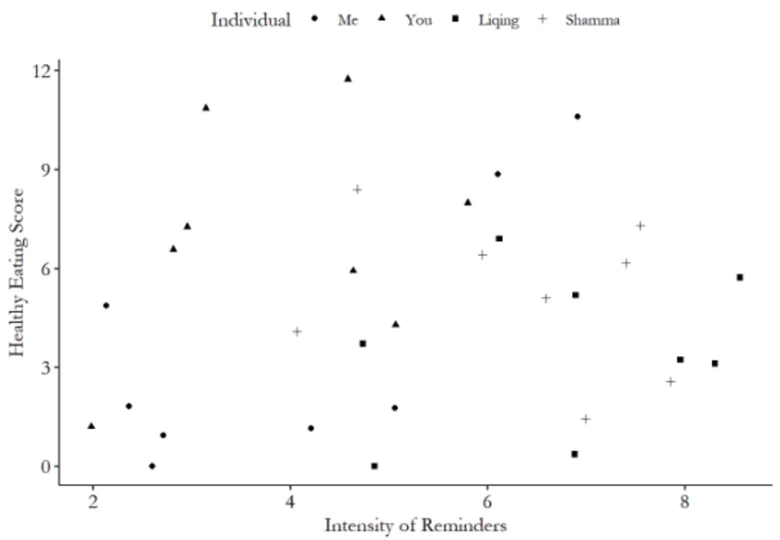
Fixed effects allow us to calculate within-variation

	Year	Excercise	Mean Excercise	Within Excercise
You	2019	5	6.0	-1.0
You	2020	7	6.0	1.0
Me	2019	4	3.5	0.5
Me	2020	3	3.5	-0.5

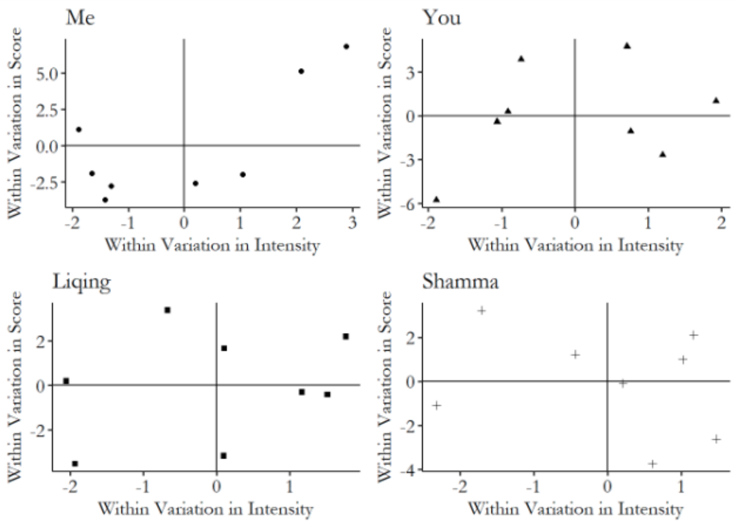
Example in action



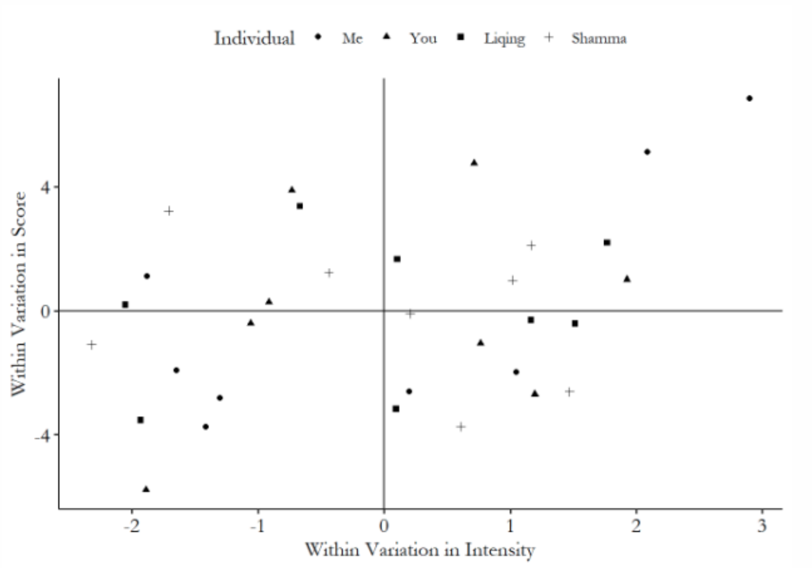
Isolating within-variation



Isolating within-variation



Isolating within-variation



We can extend these estimators

Consider the regression model with both individual and time effects

$$Y_{it} = \beta X_{it} + \tau D_{it} + \alpha_i + \delta_t + v_{it}$$

We can extend these estimators

Consider the regression model with both individual and time effects

$$Y_{it} = \beta X_{it} + \tau D_{it} + \alpha_i + \delta_t + v_{it}$$

$$Y_{it} = \beta X_{it} + \tau D_{it} + \sum_{i=1}^{N} 1[\text{unit=1}]_i + \sum_{t=1}^{T} 1[\text{time=t}]_t + v_{it}$$

Connecting FE to DD

Remember that we started with a simple DD model that we can implement via regression:

$$Y_i = \alpha + \tau \text{Treat}_i \times \text{Post}_t + \beta \text{Treat}_i + \delta \text{Post}_t + \varepsilon_i$$

Connecting FE to DD

Remember that we started with a simple DD model that we can implement via regression:

$$Y_i = \alpha + \tau \text{Treat}_i \times \text{Post}_t + \beta \text{Treat}_i + \delta \text{Post}_t + \varepsilon_i$$

In a different notation:

$$Y_{it} = \alpha + \tau D_{it} + \gamma \text{Treat}_i + \delta \text{Post}_t + \beta X_{it} + \varepsilon_{it}$$

Connecting FE to DD

Remember that we started with a simple DD model that we can implement via regression:

$$Y_i = \alpha + \tau \text{Treat}_i \times \text{Post}_t + \beta \text{Treat}_i + \delta \text{Post}_t + \varepsilon_i$$

In a different notation:

$$Y_{it} = \alpha + \tau D_{it} + \gamma \text{Treat}_i + \delta \text{Post}_t + \beta X_{it} + \varepsilon_{it}$$

Generalizing:

$$Y_{it} = +\tau D_{it} + \alpha_i + +\delta_t + \beta X_{it} + \varepsilon_{it}$$

- α_i individual FE which capture $Treat_i$;
- δ_t time FE, which captures $Post_t$

Extending DD to multiple treatment times

What happens if we have treated units who get treated at different times?

Our new general framework works

$$Y_{it} = +\tau D_{it} + \alpha_i + +\delta_t + \beta X_{it} + \varepsilon_{it}$$

Extending DD to multiple treatment times

What happens if we have treated units who get treated at different times?

Our new general framework works

$$Y_{it} = +\tau D_{it} + \alpha_i + +\delta_t + \beta X_{it} + \varepsilon_{it}$$

- D_{it} can turn from 0 to 1 at different times for different units
- Or can stay always 0 for some units
- Or stay always 1 for some units

Extending DD to multiple treatment times

What happens if we have treated units who get treated at different times?

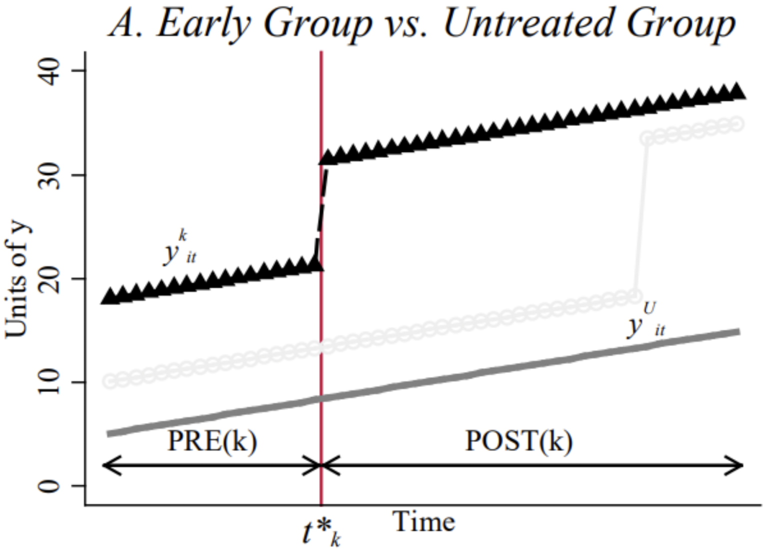
Our new general framework works

$$Y_{it} = +\tau D_{it} + \alpha_i + +\delta_t + \beta X_{it} + \varepsilon_{it}$$

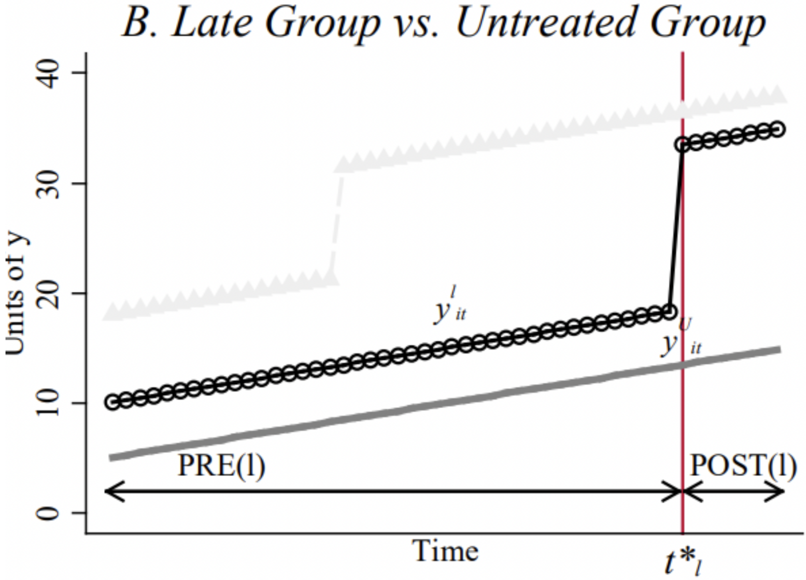
- D_{it} can turn from 0 to 1 at different times for different units
- Or can stay always 0 for some units
- Or stay always 1 for some units

It is important to note that this will get you a **weighted average** of different types of comparisons. This may not be what you want.

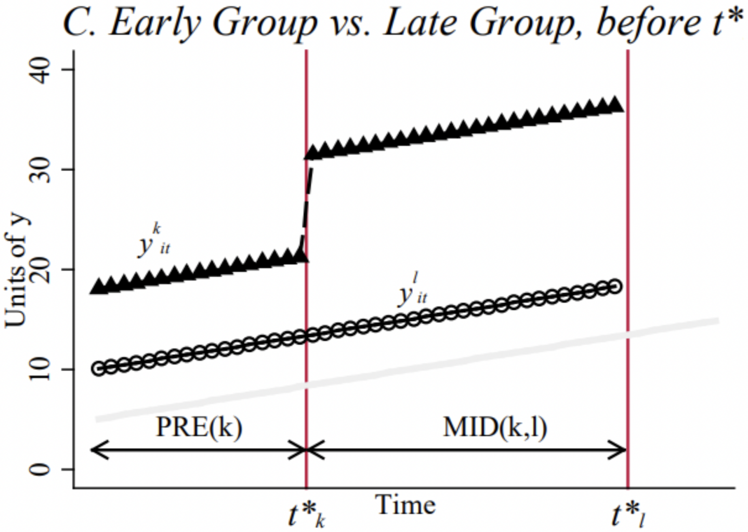
What does multiple-treatment-timing FE get us?



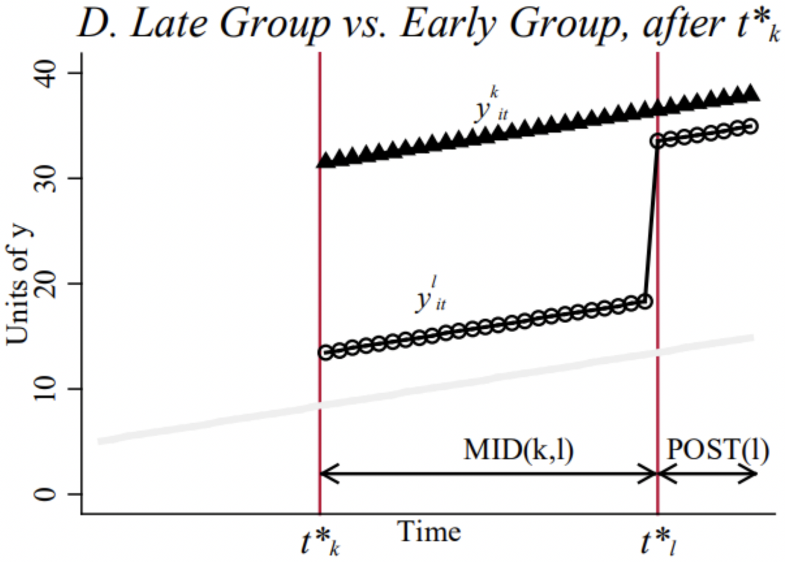
What does multiple-treatment-timing FE get us?



What does multiple-treatment-timing FE get us?



What does multiple-treatment-timing FE get us?



How to get it right?

Two alternatives:

1. Weighted balance test to make sure this is not a problem in your study (Not gonna focus on this during this class)
2. Artificially treat all units at the same time

The event study design

An event study is a more general FE design

Our standard FE regression model:

$$Y_{it} = +\tau D_{it} + \alpha_i + +\delta_t + \beta X_{it} + \varepsilon_{it}$$

- This imposes the constraint that $\tau_t = \tau$ for all t
- And all i, but let's not worry about that now

The event study design

An event study is a more general FE design

Our standard FE regression model:

$$Y_{it} = +\tau D_{it} + \alpha_i + +\delta_t + \beta X_{it} + \varepsilon_{it}$$

- This imposes the constraint that $\tau_t = \tau$ for all t
- And all i, but let's not worry about that now

A more general model will allow for differential effects over time:

$$Y_{it} = \sum_{r=0}^{R} \tau_r D_i \times 1[\text{periods post-treatment} = r]_{it} + \alpha_i + \delta_t + \beta X_i t + \varepsilon_{it}$$

$1[\text{periods post treatment} = \text{r}]_{it} = 1$ when we re r periods after treatment, 0 otherwise

- The τ_rs pick up the average treatment effect r periods after

The event study design

In the general version of this model, we also include pre-treatment effects

- We don't need the same number of pre- and post-treatment periods
- We need to exclude one category, typically the first prior to the treatment

This is good because:

- The treatment starts for everybody "at the same time"
- We can still use fixed effects
- We get a partial test of the identifying assumption

The event study design

In the general version of this model, we also include pre-treatment effects

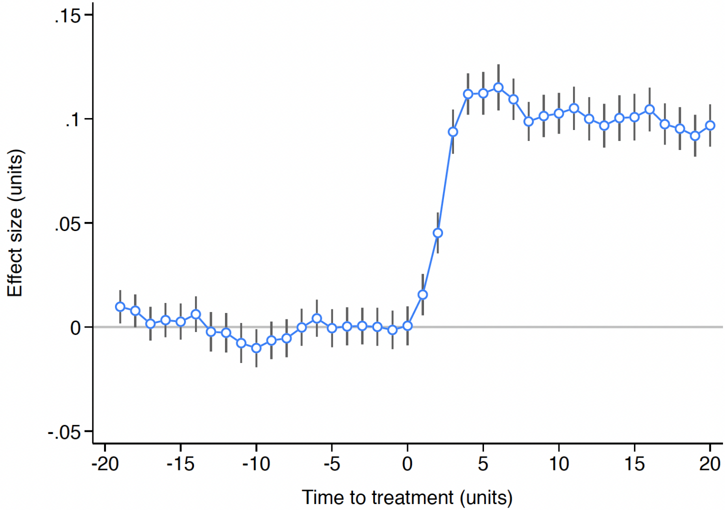
- We don't need the same number of pre- and post-treatment periods
- We need to exclude one category, typically the first prior to the treatment

This is good because:

- The treatment starts for everybody "at the same time"
- We can still use fixed effects
- We get a partial test of the identifying assumption

- We are looking for pre-treatment τ_s to be centered on 0 and not trending

Visually



.15
.1
.05
0
-.05
Effect size (units)
-20
-15
-10
-5
0
5
10
15
20
Time to treatment (units)

Cumulative effects

We can estimate cumulative effects with a distributed lag model

$$Y_{it} = \sum_{s=0}^{S} \tau_s D_{i,t-s} + \alpha_i + \delta_t + \beta X_{it} + \varepsilon_{it}$$

where $D_{i,t-s}$ is an indicator equal to the treatment status in period $t-s$

- τ_0 gives the effect in the treatment period
- τ_1 gives the marginal effect of treatment 1 period later ...
- τ_s give the marginl effect of treatment S periods later

The cumulative effect q periods after treatment is

$$T_q = \sum_{s=0}^{q} \tau_s$$

Example: Cyclones

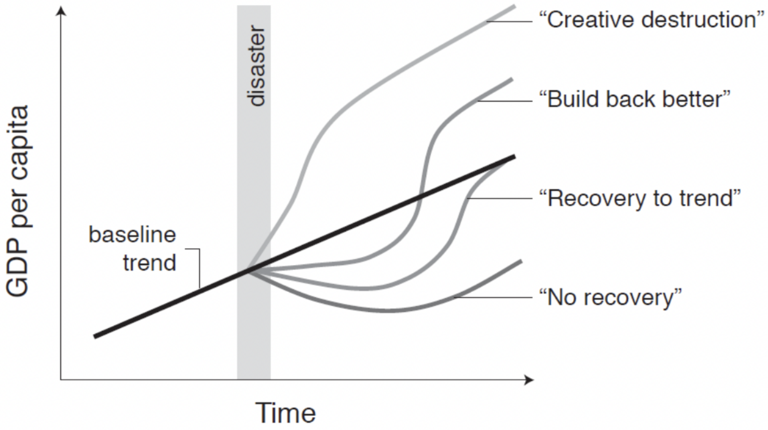
Policy issue:

- Climate change is projected to increase the intensity and number of cyclones
- What do cyclones actually do?
- And how long do these effects last?

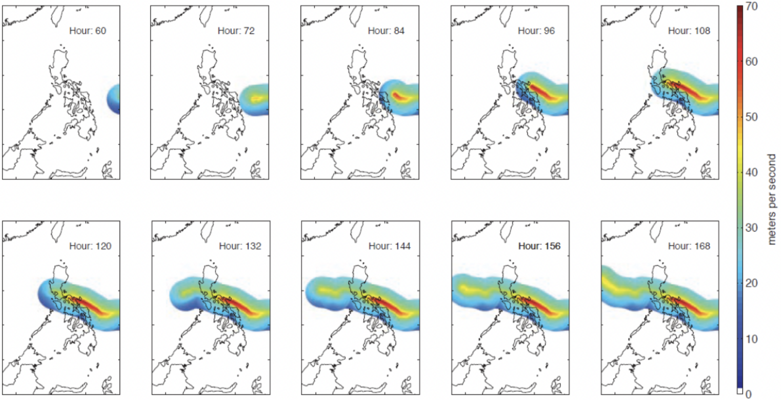
Approach:

- Construct a model of every hurricane 1950-2008
- Combine this with data on economic growth around the world
- Nobody randomized hurricanes
- But conditional on location and time FE, they are arguably exogenous. Why?
- Use a distibuted lag model to compute cumulative effects

Hypotheses


disaster
"Creative destruction"
"Build back better"
"Recovery to trend"
"No recovery"
baseline
trend
GDP per capita
Time

Hurricane data



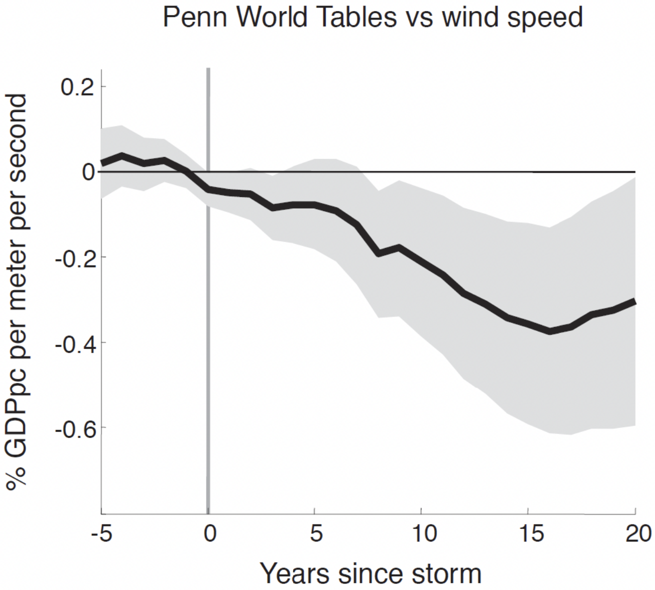
Estimating the effects of hurricanes on growth

The authors run a version of:

$$Y_{it} = \sum_{s=0}^{S} \tau_s D_{i,t-s} + \alpha_i + \delta_t + \beta X_{it} + \varepsilon_{it}$$

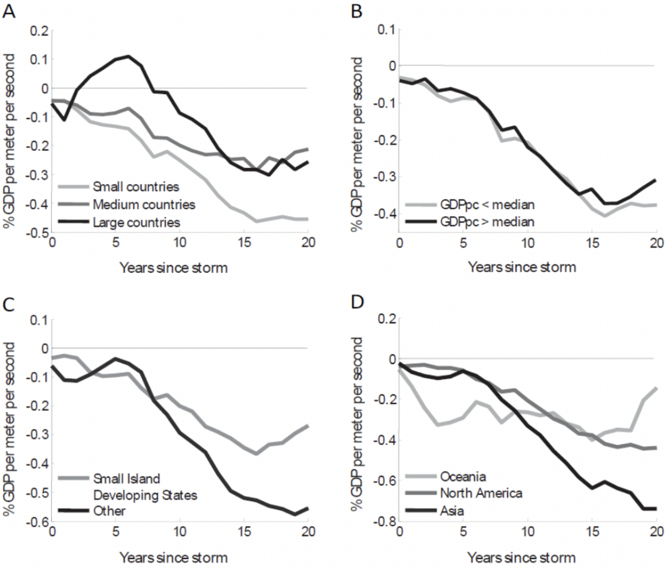
where - $Y_{it} = ln(GDP_{i,t}) - ln(GDPi, t-1)$ is the change in economic growthbetween $t-1$ and t - $D_{i,t}$ is the hurricane - Then they compute the cumulative effects

Results



Penn World Tables vs wind speed
% GDPpc per meter per second
0.2
0
-0.2
-0.4
-0.6
-5
0
5
10
15
20
Years since storm

Different locations



Another example

Policy issue

- Large gender gaps in education persist in developing countries
- We want to increase girl's participation in schooling
- How do we accomplish this?

Approach

- Program to provide girls with bikes
- Estimate the impact of cycling program
- No randomization of bikes
- But we can do a DD, even a DDD

Triple differences

- The same setting can be extended to a triple DD (DDD)
- Example: https://www.youtube.com/watch?v=6nG63ISt_Ek&t=14s