Causal Data Analysis

Chapter 23: Methods for Panel Data

Álmos Telegdy

Corvinus University of Budapest

Multiple time periods

Diff-in-diff estimation: two time periods are enough (before/after treatment)

Multiple time periods: many pre- and post-treatment periods

- Pre-treatment:, PTA (sort of) testable
- Post-treatment: dynamics of the effect $\rightarrow$ one-off, continuous, inverse $\cap$
- Data structure: staggered treatment
  - E.g., FDI effects, replacement of managers

Time series data

1: Dependent variable in levels: $y_t^E = \alpha + \beta x_t$
- α equals average y, if $x = 0$
- β equals average y, if x increases/declines by one unit

2: Dependent variable in differences: $\Delta y_t^E = \alpha + \beta \Delta x_t$
- α = trend; what is the average change of y, if x is constant
- β: what is the average change of y relative to the trend, if x increases/declines by one unit

Causal analysis with time series data

Should we estimate levels or changes?

- Changes: the bias of the trend is controlled for (+ seasonal controls)

Identification hypothesis: Δx **is exogenous**

- Anything that affects x in t, must be independent of changes of y in t
- This a form of the PTA: in time periods when $\Delta x \neq 0$, the average change in Δy in absence of treatment should be equal to average Δy in the control group

Example: demand on the gasoline market

How does the demand for gas change at firm k, if the firm changes the price of gasoline?

- $\Delta ln(D_{k,t}) = \alpha + \beta \Delta ln(p_{k,t})$
- β estimates a causal effect if the change in p_k is not affected by anything, which can affect the demand for gas
  - Changes in p_k és D_k are correlated with the price setting of all other firms (e.g., similar costs)
  - This can be solved if other firms' prices are added as controls $\Rightarrow \beta$ measures the effect relative to the other prices

Lagged variables

What is the long-run effect of x? $\Rightarrow$ lagged variables

- Immediate effect: Δx_t
- Effect after one time-period: Δx_{t-1}
- ...
- Effect after K time-periods: Δx_{t-K}

$$\Delta y_t^E = \alpha + \beta_0 \Delta x_t + \beta_1 \Delta x_{t-1} + ... + \beta_K \Delta x_{t-K} \qquad (1)$$

- $\beta_0 + \beta_1 + ... + \beta_K$: the long-run effect of x on y

Cumulative effect

With simple manipulation we can estimate the cumulative effect directly

$$\Delta y_t^E = \alpha + \beta_{cumul}\Delta x_{t-K} + \delta_0\Delta(\Delta x_t) + ... + \delta_{K-1}\Delta(\Delta x_{t-(K-1)}) \quad (2)$$

- $\beta_{cumul} = \beta_0 + \beta_1 + ... + \beta_K$
- β_{cumul} estimates the average total change of y across K periods, if x increased/declined one unit
- If Δx exogenous, β_{cumul} estimates the long-run effect of Δx_t on Δy

Lead variables

Adding **lead values** of x to the regression tests...

1: PTA

$$\Delta y_t^E = \alpha + \beta \Delta x_t + \gamma_1 \Delta x_{t+1} + ... + \gamma_L \Delta x_{t+L} \qquad (3)$$

- Δy materializes l periods before $\Delta x_{t+l} \Rightarrow$ $\gamma_1 = \gamma_2 = ... = \gamma_l = 0$ suggest that changes in y are independent of changes in x in the last L periods
  - the evolution of y is the same L periods, regardless of the change of x

2: Existence of reverse causality

- If the change in x follows the change in $y \Rightarrow \gamma_1 > 0$

Example

Firms change the price of gasoline, if demand declines more than expected

- Interesting to know whether such behavior exists
- This will bias the estimated effect

If $\gamma_1 > 0$, such behavior probably exists

- Bias declines if we control for Δx_{t+1}

Limitations of time series data

We rarely use one subject's time series for estimation

$\Rightarrow$ Usually too short

$\Rightarrow$ Variability of x low

$\Rightarrow$ Data from the past far behind cannot be used for describing the immediate past

- Economic environment changes

Example: the effect of gasoline prices on demand

- 10 years, weekly data $\rightarrow$ 520 data points
- The price of firm k rarely changes independent of the other firms' prices
- If the time series were longer, we would compare the 90's with the current situation

Pooled time series

Even if we are interested in one subject

- Estimate one regression with all time series in one database
- Measure separate α for each subject

$$\Delta y_{it}^{E} = \alpha_i + \beta \Delta x_{it} \tag{4}$$

- β equals the average change in y if x changes one unit and each i subject has its own trend

Use subjects that behave approximately similarly

Case study: the effect of import demand on industrial output

What is the effect of US import demand on the industrial output of Thailand? – causal question, although x not generated by an intervention

- US big consumer (many people)
- US big producer (global value chains)

$\Rightarrow$ Import demand of the US has an effect on the output of other countries

This is an important policy question about the effects of globalization

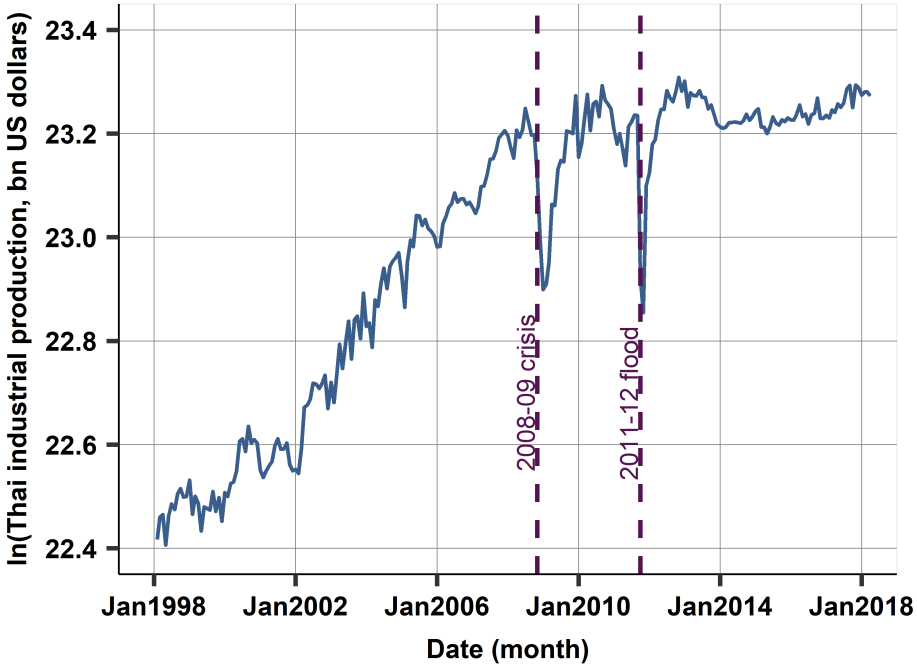
- What is the effect of globalization on industrial production?
- How much variability (= uncertainty) does globalization generate?

Data

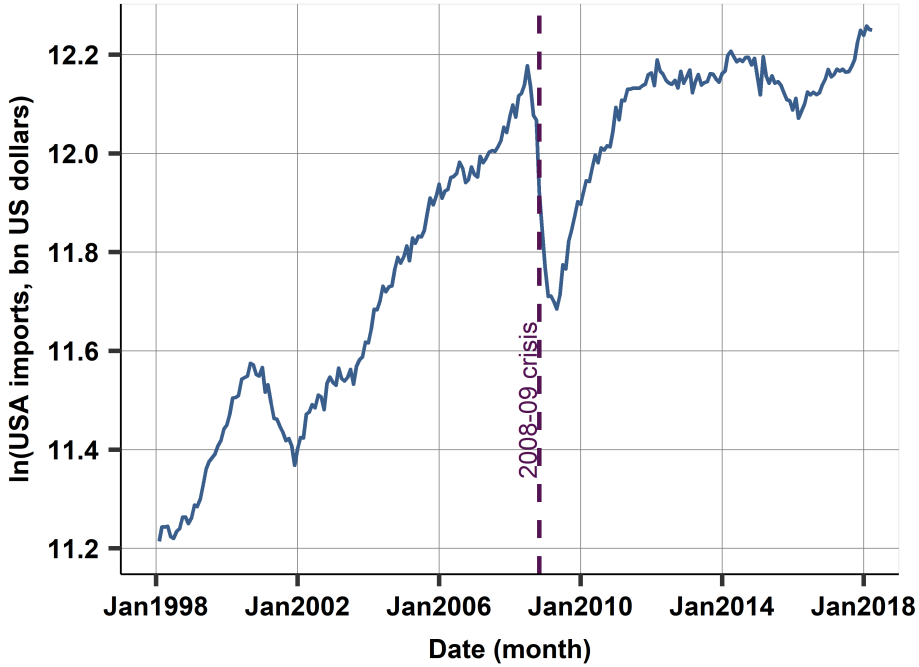
Asia-industry dataset

- Industrial production of 4 South-Asian countries
- US export
- 1998.02 – 2018.04, monthly data

Thai industrial production and US export



Industrial production in Thailand, in logs, monthly time series



US total imports, in logs, monthly time series

Estimation equation

What's special about the data?

1. Trend before 2008
2. Special event in both time-series (global recession of 2008)
3. Trend declines after recession

$$\Delta ln(ipTHA)_t = \alpha + \beta_0 \Delta \ln(impUSA)_t + \beta_1 \Delta \ln(impUSA)_{t-1} + ... + \Delta\beta_4 \ln(impUSA)_{t-4} + \phi \ln(ipTHA)_{t-1} \quad (5)$$

- Variables in (log) difference so trends are taken care of
- Max 4 lags: change in US imports may affect Thai output for 4 months

Effect of US import on industrial production – 4 countries

Variables	(1) Thailand	(2) Malaysia	(3) Philippines	(4) Singapore	(5) pooled
USA imports log change, cumul. coeff.	0.400*	0.358**	0.556**	0.367	0.437**
	(0.190)	(0.112)	(0.185)	(0.289)	(0.103)
Industrial production log change, lag	-0.119	-0.460**	-0.242**	-0.376**	-0.315**
	(0.065)	(0.059)	(0.064)	(0.061)	(0.031)
Malaysia					0.000
					(0.004)
Philippines					-0.001
					(0.004)
Singapore					0.002
					(0.004)
Constant	0.002	0.004*	0.001	0.005	0.003
	(0.003)	(0.002)	(0.003)	(0.004)	(0.003)
Observations	238	238	238	238	952
R-squared	0.070	0.231	0.140	0.183	0.123

Results

- 1% increase of US import demand generates 0.4% increase in Thai industrial production (4 months cumulative effect)
- Standard errors quite large $\Rightarrow$ conf. int. $[0,02, 0,78]$
- Similar effect for the other countries
- Pulled data: effect = 0.437, more precise estimation (conf. int: $[0,24, 0,64]$)
- Do we trust this estimation?
  - Yes, if US import exogenous – uncorrelated with variables which also affect South-Asian production
  - Reverse causality unlikely – small countries
  - Potential endogeneity: innovation, financial shocks (e.g., recession)
- So the effect may partially be endogenous, but some of it probably causal

ATE with panel data

So far: one subject

From now on: multiple subjects $\Rightarrow$ **panel data** – cross-section and time-series at the same time

- Usually relatively short time series, lots of subjects $\Rightarrow$ cross-section features dominate over time-series features
- How can we use panel data to control for unobserved heterogeneity?
  1. Fixed effect method $\Rightarrow$ each subject gets its own constant
  2. First differences method $\Rightarrow$ variables in difference form

Fixed effects

Fixed effects: each subject gets a separate constant

$$y_{it}^{E} = \alpha_i + \beta x_{it} \tag{6}$$

$$\alpha_i = \alpha_1 D_1 + \alpha_2 D_2 + ... + \alpha_N D_N$$

Why is this good?

- It controls for time invariant variables
  - z endogenous variable (affects both x and y); if z does not change in time **within subjects**...
  - ...fixed effects regression takes care of its effect, even if we have no idea of the existence of z

Estimation equation of fixed effects

Regression equation:

$$y_{it} = \alpha_0 + \beta x_{it} + u_i$$

Error term:

$$u_i = n_i + \epsilon_{it}$$

- n_i – time invariant part of the error term
- ϵ_{it} – white noise

$$y_{it} = \alpha_0 + \beta x_{it} + n_i + \epsilon_{it}$$

Fixed effects estimation

Let's average the equation...

$$\bar{y}_i = \alpha_0 + \beta \bar{x}_i + n_i + \bar{\epsilon}_i$$

...and subtract it from the regression equation

$$y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

Notation:

- $y_{it} - \bar{y}_i = \tilde{y}_{it}$
- $x_{it} - \bar{x}_i = \tilde{x}_{it}$
- $\epsilon_{it} - \bar{\epsilon}_i = \tilde{\epsilon}_{it}$

Fixed effects equation

Demeaned regression

$$\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{\epsilon}_{it}$$

This regression doesn't contain time invariant endogeneity!!!

Interpretation: what is the change in y relative to its subject-level average value, if x increases one unit relative to its own average $\Rightarrow$ **within estimator**

Examples

Example for subject level variables constant in time

- **Individuals:** IQ, non-cognitive attributes, physical abilities...
- **Firms:** special market, average level of workers, corporate culture, ability to innovate...
- **Townships:** infrastructure, average income, number of innovative firms...

What we can define as fixed in time depends on the question and length of the analyzed period

Comparison of OLS and FE estimation

Effect of privatization on firm efficiency

- 2 periods, 2 firms (0: always state; 1: first state, later private)
- Productivity: $PR_0 = 5$, $PR_1 = 10$, time invariant

Comparison of OLS and FE estimation

Effect of privatization on firm efficiency

- 2 periods, 2 firms (0: always state; 1: first state, later private)
- Productivity: $PR_0 = 5$, $PR_1 = 10$, time invariant
- $\hat{\beta}_{OLS} =$

Comparison of OLS and FE estimation

Effect of privatization on firm efficiency

- 2 periods, 2 firms (0: always state; 1: first state, later private)
- Productivity: $PR_0 = 5$, $PR_1 = 10$, time invariant
- $\hat{\beta}_{OLS} = 3.33$
- $\hat{\beta}_{FE} =$

Comparison of OLS and FE estimation

Effect of privatization on firm efficiency

- 2 periods, 2 firms (0: always state; 1: first state, later private)
- Productivity: $PR_0 = 5$, $PR_1 = 10$, time invariant
- $\hat{\beta}_{OLS} = 3.33$
- $\hat{\beta}_{FE} = 0$

Why?

- The private owner simply chose the good firm $\Rightarrow$ pure selection
- OLS estimation: effect + selection
- FE estimation: effect
  - But you will see that FE can also be biased

Identification assumption

PTA: in absence of treatment, treated subjects have the same outcome as untreated ones

- Same as in Diff-in-diff estimation $\Rightarrow$ FE estimation is a diff-in-diff estimation
- Can be partially tested with the comparison of pre-trends

When is violated?

- Some variable not fixed in time, and correlated with x and y
  - Example 1: Firms with good prospects selected for privatization $\rightarrow$ pre-trends will be increasing
  - Example 2: Privatized firms get a big subsidy $\rightarrow$ pre-trends will be flat

Example: effect of income on healthy diet

Regression equation: $HF^E_{it} = \alpha_0 + \beta ln(\text{INC}_{it})$

Example: effect of income on healthy diet

Regression equation: $HF^E_{it} = \alpha_0 + \beta ln(\text{INC}_{it})$

Comparison of OLS and FE estimation

- What does β_{OLS} measure?

Example: effect of income on healthy diet

Regression equation: $HF^E_{it} = \alpha_0 + \beta ln(\text{INC}_{it})$

Comparison of OLS and FE estimation

- What does β_{OLS} measure?
  - If income is 1% higher, how much more healthy food is consumed?

Example: effect of income on healthy diet

Regression equation: $HF^E_{it} = \alpha_0 + \beta ln(\text{INC}_{it})$

Comparison of OLS and FE estimation

- What does β_{OLS} measure?
  - If income is 1% higher, how much more healthy food is consumed?
- What does β_{FE} measure?

Example: effect of income on healthy diet

Regression equation: $HF^E_{it} = \alpha_0 + \beta ln(\text{INC}_{it})$

Comparison of OLS and FE estimation

- What does β_{OLS} measure?
  - If income is 1% higher, how much more healthy food is consumed?
- What does β_{FE} measure?
  - If a person's income increases by 1% relative to its own average income, how will healthy food consumption change?

Technical considerations

- Not possible to study variables that are fixed in time (at least not directly)
- If the variability of x is small, the estimation will lack precision
- Two types of R^2
  - How much of the variation of y is explained by the demeaned regression (*Within*-R^2)?
  - How much of the variation of y is explained by a regression, where each subject has its own constant?
- We usually do not show the constant in the FE regression
  - As many constants as subjects

Aggregate trend

If there is an aggregate trend, the regression will be biased

- E.g., household income increases in time (economic development); healthy food consumption increases (attitudes change) $\Rightarrow$ bias

Solution: controls for aggregate trend

$$y_{it}^{E} = \alpha_i + \theta_t + \beta x_{it} \tag{7}$$

θ_t: time period controls (e.g., year dummies)

Time period dummies: flexible functional form

Example: 6 periods

$$y_{it}^{E} = \alpha_i + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6 + \beta x_{it} \tag{8}$$

Dummies are better controls than a trend

- 2006 – 2014 (Great financial crisis): coefficients increase – decline – increase – decline – increase
- 2014 – 2019 (boom): coefficients increase
- 2020 – 2024 (Covid): coefficients decline – increase – decline

Estimation of standard errors

Standard errors are biased

1. Subjects vary along many dimensions $\Rightarrow$ error terms are heteroscedastic
2. There is autocorrelation between observations of the same subject $\Rightarrow$ error terms are not independent from each other

Solution: **clustered standard errors**

- Corrects for autocorrelation within subjects
- Corrects for heteroscedasticity
- If these are not present in the data, clustering won't bias standard errors

Case study: effect of vaccination against measles on child survival

- Highly contagious disease: fever, coughing, rashes
- No treatment, only vaccination (at ages 1 and 5)
- Combined with malnutrition can be fatal

Causal variable: **rate of vaccination**

- Shows the will of the government regarding vaccination (is it available or compulsory)

Outcome variable: **child survival**

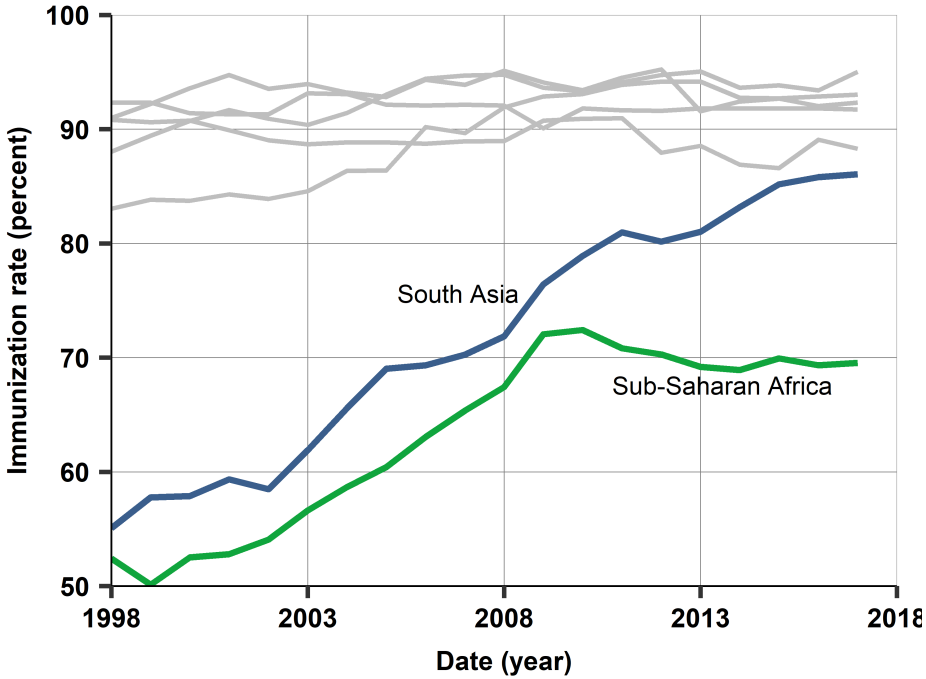
- Not the whole effect of the vaccine, but an important part

Data

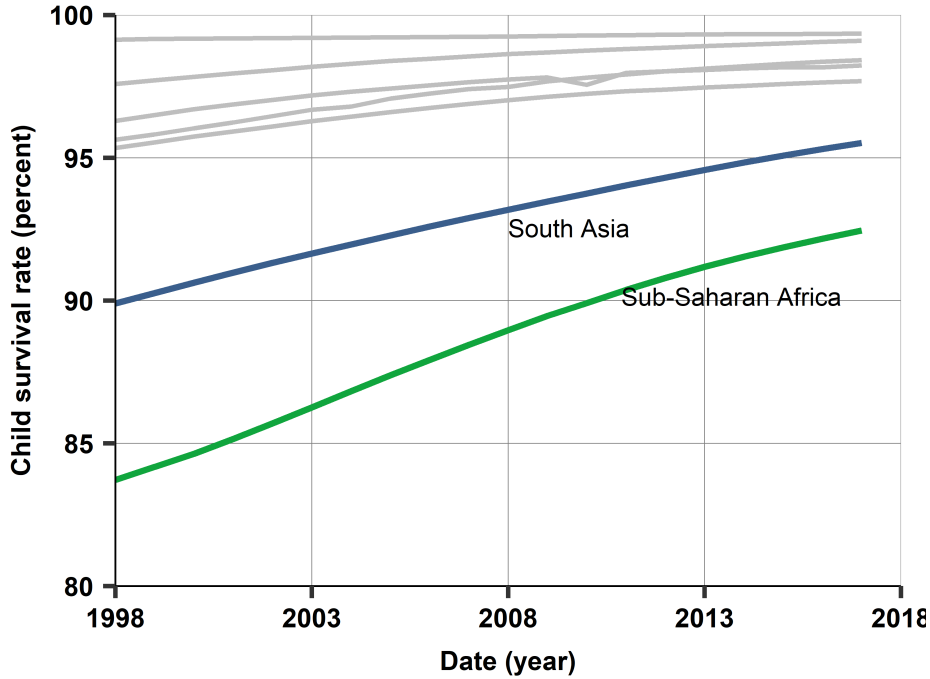
World Development Indicators (World Bank)

- Subjects: country (almost all)
- Period: year (1998-2017)
- 172 countries $\times$ 20 years = 3 440 observations
- Variables
  - Survival rate $\Rightarrow$ rate of those children who reach 5 years
  - Immunization rate $\Rightarrow$ the rate of 12-23 month old children who received the vaccine
  - Population (mill.)
  - GDP/Cap

Case study - Immunization against measles and child survival rate in seven regions of the world



Immunization rate



Child survival rate

Source: World Bank-immunization dataset. Annual data, 1998–2017, aggregated to seven geographical regions. N=140.

Descriptive statistics

Vaccine against measles

- Low in both South-Asia and Sub-Saharan Africa in 1988 ($\sim$55%)
- Constantly increases in South-Asia, gets close to other countries
- Increases in Sub-Saharan Africa by 2009, stagnates thereafter ($\sim$70%)

Child survival under age 5

- Constant increase
- Over 90% in all countries by the end of the period

Effects of vaccination on child survival

Two effects

1: Direct $\Rightarrow$ vaccinated children do not catch the disease
   - This effect increases if children are not malnourished (this increases survival in general)

2: Indirect $\Rightarrow$ vaccinated children do not spread the contagion (spillover effect)

How to estimate?

What would be the best estimation?

- Random assignment
- Direct effect + spillovers $\Rightarrow$ subjects should be groups (e.g., villages), not individuals
- There is a time lag between the vaccine and its effect
- Causal variable continuous, we are interested in changes
  - Controls: villages, where the vaccination rate changed only little
  - Treated: villages, where the vaccination rate changed a lot

How to estimate?

What would be the best estimation?

- Random assignment
- Direct effect + spillovers $\Rightarrow$ subjects should be groups (e.g., villages), not individuals
- There is a time lag between the vaccine and its effect
- Causal variable continuous, we are interested in changes
  - Controls: villages, where the vaccination rate changed only little
  - Treated: villages, where the vaccination rate changed a lot

This is impossible – no on would agree to randomize vaccination

Available estimation method

Observational data

- Country-level data ✓
- Long time series ✓
- Time fixed effects $\Rightarrow$ controls for trends
- Country fixed effects $\Rightarrow$ controls for everything constant at the country level
  - Habits, infrastructure...
- Control variables
  - Population
  - Grade of development

Possible bias: anything, that *changes at the country level*, and affects both the vaccination rate and child survival

Estimation equation

1: Without controls

$$\text{Survive}_{it}^{E} = \alpha_i + \theta_t + \beta \text{VACC}_{it} \tag{9}$$

2: With controls

$$\text{Survive}_{it}^{E} = \alpha_i + \theta_t + \beta \text{VACC}_{it} + \gamma_1 Pop_{it} + \gamma_2 Dev_{it} \tag{10}$$

Both regressions control for country and year fixed effects

Results

Variables	(1) Survival rate	(2) Survival rate
Immunization rate	0.077**	0.038**
	(0.010)	(0.011)
ln GDP per capita		1.593**
		(0.399)
ln population		12.049**
		(1.648)
Year dummies	Yes	Yes
Observations	3,440	3,440
R-squared	0.717	0.848
Number of clusters	172	172

Clustered and unclustered standard errors

Variables	(1) Clustered SE	(2) Simple SE
Immunization rate	0.038**	0.038**
	(0.011)	(0.002)
ln GDP per capita	1.593**	1.593**
	(0.399)	(0.071)
ln population	12.049**	12.049**
	(1.648)	(0.227)
Observations	3,440	3,440
R-squared	0.848	0.848
Number of c	172	172

Panel regressions on first differences (FD)

We can first difference the data

- Simplest FD model: x the only explanatory variable

$$\Delta y_{it}^{E} = \alpha + \beta \Delta x_{it} \tag{11}$$

Interpretation of coefficients

- α: trend – average change in y when x constant
- β: Diff-in-diff effect – average change in y when x changes by one unit
  - We compare the same subjects in different time periods
  - We compare different subjects within the same period

Example

Binary treatment at different times, three subjects

1. Untreated – treated
2. Untreated – untreated
3. Treated – treated

Interpretation of β

- Measured only when $\Delta x \neq 0$
- Number of treated subjects is important
  - Untreated – treated $\Rightarrow$ treated group
  - Untreated – untreated, treated – treated $\Rightarrow$ control group

Differences and fixed controls

$$y_{it} = \alpha_0 + \beta x_{it} + n_i + \epsilon_{it} \tag{12}$$

Subtract from the equation the same equation lagged one year $(t-1)$

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta \epsilon_{it} \tag{13}$$

FD also controls for fixed effects!!!

Lagged variables

Linear panel regression with K lags

$$\Delta y_{it}^{E} = \alpha + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + ... + \beta_K \Delta x_{i(t-K)} \qquad (14)$$

- $\alpha =$ the trend of y: the average change in y if x does not change and did not change in the last K periods
- $\beta_0 =$ instantaneous effect: the average change in y if x changes in the same period by one unit, but x is constant in the previous K periods
- $\beta_k =$ delayed effect: the average change in y if x changed one unit k periods before, but x was constant in the other previous K periods

Cumulative effect

Same as with time series

Long-term effect: equals the coefficient of the K^{th} lag, if the other lags in difference form are included as controls

$$\begin{aligned}\Delta y_{it}^{E} &= \alpha + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + ... + \beta_K \Delta x_{i(t-K)} \\ &= \alpha + \beta_{cumul} \Delta x_{i(t-K)} + \gamma_0 \Delta(\Delta x_{it}) + \gamma_1 \Delta(\Delta x_{i(t-1)}) + \\ &+ ... + \gamma_{K-1} \Delta(\Delta x_{i(t-K+1)}) \\ \beta_{cumul} &= \beta_0 + \beta_1 + ... + \beta_K \end{aligned} \tag{15}$$

Lead variables, PTA

We can check the PTA with the leads of Δx

$$\Delta y_{it}^{E} = \alpha + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + ... + \beta_K \Delta x_{i(t-K)} + \gamma_1 \Delta x_{i(t+1)} + ... + \gamma_L \Delta x_{i(t+L)} \quad (16)$$

- If the γ coefficients were equal to 0, the average change of y would be the same in the absence of treatment regardless of the future effects of Δx (0 or 1)

Aggregate trends

α controls for a linear trend $\Rightarrow$ replace it with time dummies (θ_t), so we can control for any kind of aggregate trend

$$\Delta y_{it}^{E} = \theta_t + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + ... + \beta_K \Delta x_{i(t-K)} \qquad (17)$$

Individual trends

We can also control for individual trends (α_i)

$$\Delta y_{it}^{E} = \alpha_i + \theta_t + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + ... + \beta_K \Delta x_{i(t-K)} \quad (18)$$

Individual trends

We can also control for individual trends (α_i)

$$\Delta y_{it}^{E} = \alpha_i + \theta_t + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + ... + \beta_K \Delta x_{i(t-K)} \quad (18)$$

- α_i: what is the average change of y relative to an aggregate trend (θ_t), if x did not change in the last K periods
- β: what is the change of y relative to the aggregate and individual trend, if x changed by one unit

Case study: FD with lags and leads

Variables	$\Delta surv$	$\Delta surv$	$\Delta surv$	$\Delta surv$
Δimm	0.009**	0.010**		
	(0.002)	(0.002)		
Δimm lag 1		0.010**		
		(0.002)		
Δimm lag 2		0.011**		
		(0.002)		
Δimm lag 3		0.009**		
		(0.002)		
Δimm lag 4		0.007**		
		(0.002)		
Δimm lag 5		0.006**		
		(0.002)		
Δimm lead 1				0.008**
				(0.002)
Δimm lead 2				0.007**
				(0.002)
Δimm lead 3				0.005
				(0.003)
Δimm cumul			0.053**	0.054**
			(0.010)	(0.008)
Constant	0.188**	0.136**	0.136**	0.125**
	(0.024)	(0.018)	(0.018)	(0.018)
Observations	3,268	2,408	2,408	1,892
R-squared	0.013	0.078	0.078	0.093

Interpretation of the results

- Immediate effect: 0.009**
- Long-term effects
  - constant for 3 years, declines thereafter
  - Cumulative effect: 0.053**
- Trend before treatment already positive: child survival increases already before vaccination rate increases

FD regressions with controls, aggregated and individual trends

Variables	(1) $\Delta surv$	(2) $\Delta surv$	(3) $\Delta surv$
Δimm cumulative ,	0.052**	0.030**	0.011**
	(0.010)	(0.009)	(0.003)
Year dummies	Yes	Yes	Yes
Confounder variables	No	Yes	Yes
Country-specific trends	No	No	Yes
Observations	2,408	2,408	2,408
R-squared	0.088	0.212	0.331

Interpretation of the results

- Year effects: treatment effect does not change (0,052**)
- Control variables (GDP/Cap, population): treatment effect declines (0.030**)
- Country-specific trends: treatment effect further declines (0.011**), conf. int. $[0,005, 0,17]$
  - 10% increase in immunization induces a 0.1% increase in child survival
  - Coefficients of lead variables are small and statistically insignificant $\Rightarrow$ no evidence for violation of PTF

Immunization against measles does save children's life

FE or FD?

Similar in principle $\Rightarrow$ both methods control for fixed effects.
BUT...

- FD: directly estimates how changes in x affect y
- FD: easy to model how the effect evolves in time
- FD: easy to control for individual trends
- FE: easy to estimate long-run effects
  - No need for variable transformation
  - No need to think about lags
  - Nobs do not change because of lags
- FE: less vulnerable to missing data
  - Lose only one observation, not two

Conclusion: FE vs. FD

When to use which method?

- FD easier, if evolution in time is important and individual trends are important
- FE easier, if cumulative effect is important and losing data is an issue

In firm analysis usually FE is used

Unbalanced panel

Some subjects are observed along fewer time periods.
Two problems

1. Less accurate estimation
2. It is possible that missing data are not random $\Rightarrow$ **can lead to biased results**

Example: immunization against measles

- Missing data mostly from countries with no modern data collection methods
  - These countries are less developed...
  - ...or there is a crisis in the country (e.g., war)
- Missing years are likely to be different from the average year $\Rightarrow$ leads to bias

Missing data: solutions

- Leave out years altogether where data are scant
  - Good solution, if those years are not important (e.g., similar to year with good data)
  - In some years data may be so bad we cannot use in analysis
- Exclude subject with missing data
  - Good solution, if these subjects are close to average
  - Look into differences between excluded and included groups
  - Exclude data based on some attribute (e.g., development)
- Use the whole data
  - Good solution, if missing data are random
  - Even if missing data is part of the mechanism
    - Example: firm exit