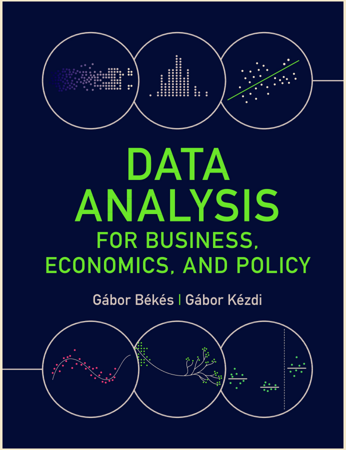
Békés-Kézdi: Data Analysis, Chapter 23: Methods for Panel Data



Data Analysis for Business, Economics, and Policy

Gábor Békés (Central European University)
Gábor Kézdi (University of Michigan)

Cambridge University Press, 2021

gabors-data-analysis.com

Central European University

Any comments or suggestions:
gabors.da.contact@gmail.com

Plan for today

- Talk about a key method for observational data
- Panel data methods = generalization of difference in differences in several ways
- Widely used in academia and real life
- Very useful to get closer to causality

Multiple Time Periods Can Be Helpful

- Diff-in-diffs estimates the effect at a single point in time.
- Issue 1: Immediate effect in one period
- Most real-life situations: delayed effect, variation of impact over time
  - Having a single endline time period is not enough to tell the full story.
- To estimate how an effect plays out in time, need more time periods.
- Issue 2: subjects may be treated at various points in time
- Need method(s) that generalize diff-in-diffs for multiple periods.

Plan for today and next week

- Time series
- Pooled time series
- Panel data and first difference
- Panel data and fixed effects
- Panel data with first difference and fixed effects

- Tweak panel data to design the control group
- Event studies with placebo controls

- Talk other useful ideas in causal inference

Generalization 1: Multiple time periods, comparison within subject

Estimating Effects Using Observational Time Series

- Generalization: multiple periods
- Estimating an effect from a single time series: within subject comparisons only.
- An average effect across time for the same subject.
  - we care about a single country / shop; the intervention happens at one place.

- Time series regressions
- specified in levels as well as changes.
  - y_t variable is measured at which t time period. Could have lags.
  - Δ denotes change: $\Delta y_t = y_t - y_{t-1}$

Estimating Effects Using Observational Time Series

- Time series regression specified in levels:

$$y_t^E = \alpha + \beta x_t \quad (1)$$

  - α is the average y when $x = 0$;
  - β shows how much larger y is, on average, when x is larger by one unit.

Estimating Effects Using Observational Time Series

- Time series regression specified in terms of changes in y and changes in x :

$$\Delta y_t^E = \alpha + \beta \Delta x_t \tag{2}$$

  - α: estimates the trend: the average change in y when x doesn't change.
  - β: how much y changes, on average when x increases (or decreases), by one unit; *in addition* to the trend.
    - as y_t changes by the trend anyway, so "how much more" is the question.
- Difference: avoid estimating spurious effects due to trends and random walks
  - Applied when x is binary or quantitative.

Estimating Effects Using Observational Time Series

- Causal effect? Yes, if variation in Δx_t is exogenous.
- time periods with different changes in x would have experienced the same change in y, had x changed the same way for them.
  - Yes, units are the time periods, as we have a single subject
- Whatever makes x change at time t should be independent of all other things that would make y change at time t.
  - Within-subject criterion: changes in x and y are for the same subject.
  - A version of PTA. In time periods when the treatment status changed ($\Delta x_t \neq 0$), y would have changed the same way, had the treatment status remained the same, as it changed in time periods when the treatment status did remain the same.
- Please read and think about the gasoline example on p652, p654, p655

Estimating Effects Using Observational Time Series

- Time series problems are the same as before
  - Trend - first difference takes care of it
  - Seasonality - if suspected, add seasonal dummies
  - Standard error - Newey-West SE
- Review Chapter 12, especially 12.4, 12.8

Capture dynamics and reverse effects with leads and lags

Lags to Estimate the Time Path of Effects

- Advantage of multiple time periods: estimate the time path of effects,
  - immediate effects,
  - effects in the near future,
  - long-run effects.
- Include appropriate lags of Δx_t.
  - Application of what we covered earlier

Lags to Estimate the Time Path of Effects

- With lags, we can estimate effects within the same time period (β_0 below), effects one time period later (β_1),etc.
- Time series regression that can estimate effects for up to K time periods has K lags of Δx:

$$\Delta y_t^E = \alpha + \beta_0 \Delta x_t + \beta_1 \Delta x_{t-1} + ... + \beta_K \Delta x_{t-K} \tag{3}$$

Lags to Estimate the Time Path of Effects

- One-lagged effect: A change in January affects the outcome change in February (and stay that way)
- A change in how demand grows affects how sales grow next month and beyond.
  - new growth rate, no reversal back

Lags to Estimate the Time Path of Effects

- Long-run effect on y = adding up the coefficients on all lags
- Or apply trick to get cumulative effect:

$$\Delta y_t^E = \alpha + \beta_{cumul} \Delta x_{t-K} + \delta_0 \Delta(\Delta x_t) + ... + \delta_{K-1} \Delta(\Delta x_{t-(K-1)}) \tag{4}$$

  - $\beta_{cumul} = \beta_0 + \beta_1 + ... + \beta_K$ above
  - β_{cumul} shows the total change in y within K time periods after a unit change in x, on average.

Lags to Estimate the Time Path of Effects

- Use either method
- β_{cumul} shows the total effect of Δx_t on Δy over the long run.
- Causal effect condition the same: when variation in Δx is exogenous.

Leads to Examine Pre-trends and Reverse Effects

- We can also include lead terms of Δx in the regression.
- It's analogous to pre-trends in diff-in-diffs regressions
- It helps capture reverse causality

$$\Delta y_t^E = \alpha + \beta \Delta x_t + \gamma_1 \Delta x_{t+1} + ... + \gamma_L \Delta x_{t+L} \tag{5}$$

Leads to Examine Pre-trends and Reverse Effects

- Include lead terms of Δx in the regression.
- Lead x = lagged y

- Similar role as looking at pre-trends
- Examine how y did change in the previous time period(s)
- The parallel trends assumption we need here: analogous to pre-trends in diff-in-diffs regressions
  - Formally adding a few periods. Few defined by data limitations.

Leads to Examine Pre-trends and Reverse Effects

- Specific case of endogenous change in x – reverse causality effect: y affecting x.
- With observations from multiple time periods - capture this reverse effect.
- IF it takes time.
- Result of reverse effect: a change in x would tend to follow a change in y.
- One time period, Δy_t is associated with Δx_{t+1},
  - coefficient capture that reverse effect

Leads to Examine Pre-trends and Reverse Effects

- Include lead terms of Δx in the regression. With L leads:

$$\Delta y_t^E = \alpha + \beta \Delta x_t + \gamma_1 \Delta x_{t+1} + ... + \gamma_L \Delta x_{t+L} \tag{6}$$

  - The lead terms are Δx_{t+1} through Δx_{t+L}.
  - γ_1 shows how y tends to change one time periods before x changes.
  - γ_L shows how y tends to change L time periods before x changes.
  - They show that because Δy_t is one time period before Δx_{t+1}, two time periods before Δx_{t+2}, etc.
  - $\gamma_1 = ...\gamma_L = 0$ would show that, regardless of how x changes, y tends to change the same way one through L time periods earlier.

Leads to Examine Pre-trends and Reverse Effects

- Causal model with a single series: combine leads and lags
- The lag terms help capture delayed effects.
- The lead terms help capture differences in pre-trends and reverse effects.
- A time series regression, in differences, with K lags and L leads, has the form

$$\Delta y_t^E = \alpha + \beta_0 \Delta x_t + \beta_1 \Delta x_{(t-1)} + ... + \beta_K \Delta x_{(t-K)} + \gamma_1 \Delta x_{(t+1)} + ... + \gamma_L \Delta x_{(t+L)} \quad (7)$$

Generalization 2: Multiple time periods, multiple subjects - pooled time series

Pooled Time Series to Estimate the Effect for One Unit

- Despite the advantages of estimating effects from time series, single time series are rarely used to estimate effects in practice.
- Time series are rarely long enough

Pooled Time Series to Estimate the Effect for One Unit

- Despite the advantages of estimating effects from time series, single time series are rarely used to estimate effects in practice.
- Time series are rarely long enough
- Even if long, are they relevant? Often, not.

- One solution: combine time series from several subjects i (cross-sectional units).
- Idea: time series of similar units are more representative than longer series of a single unit
- Use domain knowledge to select similar units

Pooled Time Series to Estimate the Effect for One Unit

- The simplest pooled time series regression estimates a single intercept and a single slope.
- Most often, though, we include separate intercepts for each i.
- Doing so allows for trends to be different across i.

$$\Delta y_{it}^{E} = \alpha_i + \beta \Delta x_{it} \tag{8}$$

  - Here β shows the average change in y, across time and units i, when x increases by one unit.
  - Conditional on i-specific trends: even if different subjects had different trends, this would not affect our estimate.

Pooled Time Series to Estimate the Effect for One Unit

- We can add leads and lags as before
- We had two ways to tackle serial correlation: Newey-West SE and adding lagged y_t. Here it's the lagged y_t
- Data table with pooled time series, N units, each with T_i observations.
  - There is no specific, ideal N, it's typically 5-20, depends on domain, could be more.
  - Ideally, each unit has same time series, but can work with them even if not —> end of lecture

Panel Regression: taking stock 1

- Pooled time series helps to estimate an effect for one unit
- Often in first difference - trend
- Using pooled time series, N is small, T is large
- Add lagged Δy_t to the right hand side of equation to tackle serial correlation

Case study – Import Demand and Industrial Production

- Interested in understanding how external demand affects production
- Thai industrial production and US total imports: individual time series
  - Industrial production in Thailand, in logs, monthly time series
  - US total imports, in logs, monthly time series
- Source: asia-industry dataset. N=243.
  - Monthly data, seasonally adjusted, February 1998–April 2018.

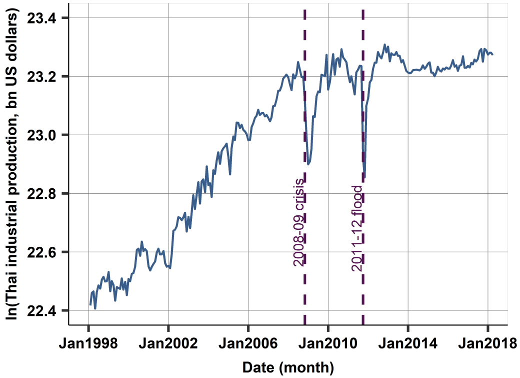
Case study – Thai industrial production and US total imports

- Question: how the import demand of the USA affects industrial production in Thailand.
- Causal question, but no explicit intervention.
- what happens in a mid-sized open economy when something changes externally - major trading partner.
- Mechanism: global supply chains, Thailand sells to USA directly, and indirectly (often through China).
- We care about coefficient not just if there is an effect – policy

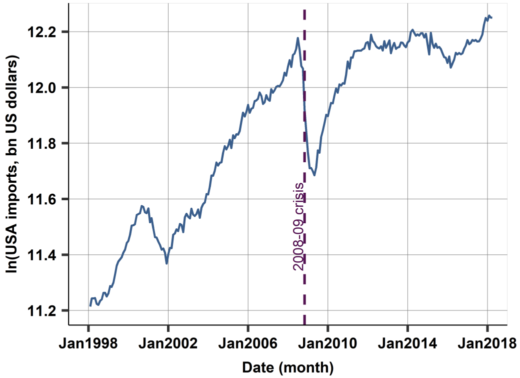
Case study – Thai industrial production and US total imports

- Thai industrial production and US total imports: individual time series
  - Industrial production in Thailand, in logs, monthly time series
  - US total imports, in logs, monthly time series
- Source: World Bank WDI – asia-industry dataset. N=243.
  - Monthly data, seasonally adjusted, February 1998–April 2018.

Case study – Thai industrial production and US total imports



Thailand IP, in logs, Feb 1998–April 2018, monthly



US total imports, in logs, monthly

Case study – Thai industrial production and US total imports [REV]

- There is a trend, an extreme event (2009 great crisis), care about relative change
- First difference. Log values.
- Lags=4 -a one-time change in U.S. imports can have an effect on how Thai industrial production changes through four months.
- No leads - expect no reverse causality
- TS regression estimate the effect of U.S. import demand on Thai industrial production (IP):

$$\Delta(\ln(ipTHA)_t) = \alpha + \beta_0\Delta(\ln(impUSA)_t) + \beta_1\Delta(\ln(impUSA)_{t-1}) + \ldots + \beta_4\Delta(\ln(impUSA)_{t-4}) + \phi\Delta(\ln(ipTHA)_{t-1}) \tag{9}$$

Case study – Import Demand and Industrial Production

- US imports and industrial production in Thailand and three other countries
- Dependent variable is change of log industrial production in each country;
- Explanatory variable cumulative effect of the change in log US imports, four lags.
- Add lagged dependent variable to capture serial correlation
- Monthly time series, seasonally adjusted, February 1998–April 2018. N=243 - for all units

Case study - US imports and IP in Thailand + 3 other countries

Variables	(1) Thailand	(2) Malaysia	(3) Philippines	(4) Singapore	(5) Pooled
USA imports log change, cumulative 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

TS regression; dep.var= change of log industrial production in country; log US imports change: 4 lags.
Monthly, SA, Feb 1998–April 2018. N=243. Standard error estimates in parentheses. ** p <0.01, * p <0.05.

Case study - US imports and IP in Thailand + 3 other countries

- Estimate is 0.44, 95% confidence interval is [0.24,0.64].
- Causality: we have good reasons to take estimate as causal effect
  - First difference takes care of level, trend.
  - Unlikely reverse causality (but may add leads Try at home.)
- What can go wrong?

Case study - US imports and IP in Thailand + 3 other countries

- Estimate is 0.44, 95% confidence interval is [0.24,0.64].
- Causality: we have good reasons to take estimate as causal effect
  - First difference takes care of level, trend.
  - Unlikely reverse causality (but may add leads Try at home.)
- What can go wrong?
- A confounder affecting the change in output and demand
- Examples?

Generalization 3: Multiple time periods and subjects - xt panel data, FE model

Panel Regression

- Pooled time series from a few subjects to estimate the expected effect of a causal variable x on outcome y. Policy question was for one of the subjects.

- Change of question: the average effect of x on y across many subjects.
- Same kind of question to diff-in-diffs, but multiple periods
- So we'll have: N units, over T periods
- Will look at different models, approaches

Panel Regression with Fixed Effects

- Typically N is large, T is relatively small
- –> time series aspects less important
- Start with levels (y_t and not Δy_t)

Panel Regression with Fixed Effects

- So, setup: multi-period panel data
- First model is the fixed-effects regression (FE regression).
- In FE regressions we have y and x (in levels)
- Fixed effects are separate intercepts for different cross-sectional units.
- The simplest linear panel regression with cross-section fixed effects:

$$y_{it}^E = \alpha_i + \beta x_{it} \tag{10}$$

  - The fixed effects are denoted by α_i.
  - Intercept varies for different cross-sectional units.

Panel Regression with Fixed Effects

- Like pooled time series in levels, but
- ... many units,
- ...a short time series

- We look for average relationship

Panel Regression with Fixed Effects

- Why do we include the fixed effects?
  - Separate intercepts for each xsec unit instead of a common intercept?
- IF subjects tend to have higher y on average due to some unobserved confounder that affects x or y in the same way at all times.
- THEN, fixed effects help avoid/mitigate bias. Including fixed effects = conditioning on all variables that don't change through time.
- Fixed effects condition on confounders that do not change in time (time-invariant)

Panel Regression with Fixed Effects - mean differencing

- Technical detour: fixed effects is like mean differencing.
- Inclusion of the cross-sectional fixed effects acts as a transformation of the y and x variables into differences from their cross-sectional means: $y_{it} - \bar{y}_i$ and $x_{it} - \bar{x}_i$,
- where $\bar{y}_{it}$ and $\bar{x}_i$ are average values of y and x across all time periods within cross-sectional unit i.

- β in the model $y_{it}^E = \alpha_i + \beta x_{it}$ is exactly the same as the β in the model $(y_{it} - \bar{y}_i)^E = \alpha + \beta(x_{it} - \bar{x}_i)$.

Panel Regression with Fixed Effects - coefficients

- In the FE regression, β shows how much larger y is, on average, compared to its mean within the cross-sectional unit, where and when x is higher by one unit compared to its mean within the cross-sectional unit.
- Compare two observations that are different in terms of the value of x compared to its i-specific mean. On average, y is larger, compared to its i-specific mean, by β, for the observation with the larger x value.

Panel Regression with Fixed Effects - coefficients

- In the FE regression, β shows how much larger y is, on average, compared to its mean within the cross-sectional unit, where and when x is higher by one unit compared to its mean within the cross-sectional unit.

- That's a within-subject comparison, and it's not affected by whether one subject has larger average y.
- That's why it's not affected by whether an unobserved confounder affects the average y values of the different subjects.

Panel Regression with Fixed Effects - coefficients: where / when

- In the FE regression, β shows how much larger y is, on average, compared to its mean within the cross-sectional unit, where and when x is higher by one unit compared to its mean within the cross-sectional unit.
- "where and when" - β approximates the average pattern of association across both time and space. Linear specification = an approximation to the average pattern of association :
  - different cross-sectional units (cross-sectional heterogeneity of the association),
  - and/or across different time periods (changing patterns of association).
  - + as always, the pattern itself may be different for different values of x (nonlinearity),

Panel Regression with Fixed Effects : Fruits and income

- how much more fruit and vegetables people eat (compared their 10-year average) when their income is higher than their average 10-year income.

- this effect is not confounded by anything that is stable over time (such as gender, genes)
- Fixed effects soak up all time-invariant variation
- Other potential confounders that affect both how income in year t is different from average and how fruit consumption is different to average.
- Example?

Panel Regression with Fixed Effects

- Technical note 1. R-squared has two versions
- within R-squared - based on the transformed model, ie comparing mean differenced y and x
- R-squared of full model - the FE model as is, with binary variables for all cross-sectional units
- Preference for within R-squared - more meaningful re model.
  - Key: make sure to report which one.
- Technical note 2. Don't publish constant - fixed effects include it.

Aggregate Trend in panel data

- Aggregate trend is a global trend that affects all unit the same way
- such as global business cycle
- varies across time periods but not units
- With xt panel data, we can can condition on an aggregate trend, whatever form it has, including nonlinear trends or even ups and downs.

Aggregate Trend in panel data

- To condition on aggregate trends, we need to include time dummies: binary variables for each time period.
- Sometimes called time fixed effects

$$y_{it}^E = \alpha_i + \theta_t + \beta x_{it} \tag{11}$$

- β shows how much larger y is, on average, compared to its mean within the cross-sectional units and its mean within the time period, where and when x is higher by one unit compared to its mean within the cross-sectional unit and its mean within the time period.

Aggregate Trend in panel data

$$y_{it}^E = \alpha_i + \theta_t + \beta x_{it} \tag{12}$$

- β shows how much larger y is, on average, compared to its mean within the cross-sectional units and its mean within the time period, where and when x is higher by one unit compared to its mean within the cross-sectional unit and its mean within the time period.
- β shows how much larger y is, on average, compared to its *long run mean* and its *aggregate trend*, where and when x is higher by one unit compared to its *long run mean* and its *aggregate trend*.

Clustered Standard Errors

- Instead of heteroskedasticity robust SE (cross-section) or Newey West SE (time series), we'll use a new type called clustered standard error.
- Standard errors clustered at the level of cross-sectional units
- Idea: adjust standard errors to capture that observations in time series are not independent (serial correlation is likely)
- Clustered standard errors are robust in two aspects. They are fine in the presence of any kind of serial correlation, and they are also fine without any serial correlation.
- They are also fine in the presence of heteroskedasticity as well as homoskedasticity
- Thus, with panel models, we always use clustered SE.
- .... we need a not small (>30) number of units

Panel Regression: taking stock 2

- Fixed effect regression with dummies for aggregate trend - causality?
- The cross-sectional FE regression can get us closer to causal effect of x on y
- Conditioning on confounders that don't change;
- Condition on aggregate trends of any shape.
- Use clustered standard error

Case study B: Immunization against Measles and Saving Children

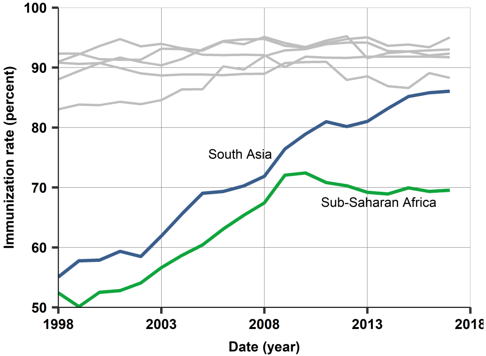
Case study – Immunization against Measles and Saving Children

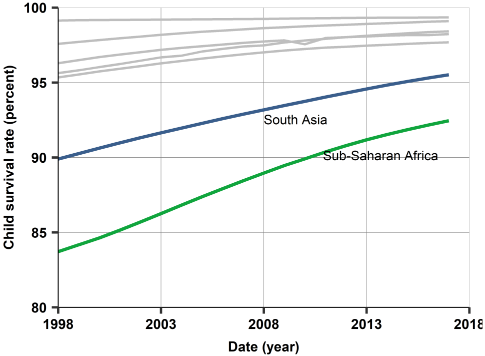
- A case study about vaccines.
- We picked in 2018

Case study – Immunization against Measles and Saving Children

- Immunization against measles and child survival rate in seven regions of the world
  - Immunization rate
  - Child survival rate
  - Immunization rate: percentage of children of age 12 to 23 months who received vaccination against measles.
  - Child survival rate: 100% minus the percentage of children of age 0 to 5 years who died in the given year.
- Source: worldbank-immunization dataset.
- Annual data, 1998–2017, aggregated to seven geographical regions.
- Many, but not all countries, N=172

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





Immunization rate Child survival rate
Source: worldbank-immunization dataset. Annual data, 1998–2017, aggregated to seven geographical regions. N=140

Case study - the effect of measles immunization on child survival. FE regressions

- The effect of measles immunization on child survival.
- FE regressions
- Within R-squared presented for FE regressions.
- Source: worldbank-immunization dataset;
- balanced yearly panel, years 1998–2017 in 172 countries.

Case study - the effect of measles immunization on child survival. FE regressions

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 countries	172	172

Within R-squared presented for FE regressions. Appropriate standard error estimates in parentheses. ** p <0.01, * p <0.05. Source: `worldbank-immunization dataset`; balanced yearly panel, years 1998–2017 in 172 countries.

Case study - the effect of measles immunization on child survival. FE regressions

- The slope parameter estimate on immunization is 0.077 without conditioning on any confounders
- drops to 0.038 when we condition on GDP per capita and population
- When we compare years with the same GDP and population, in years when the immunization rate is higher by 10 percentage points than its average rate within a country, child survival tends to be 0.38 percentage points higher than its average within the country, conditional on aggregate trends in the world.
- We can expect it to be 0.16 to 0.6 percentage points higher in the general pattern represented by our data.
  - 100 percent - 3.8 percent, but not a realistic improvement, 10% makes more sense

Case study - the effect of measles immunization on child survival.

- Clustered SE versus biased simple SE in a FE panel regression

- FE regressions with different SE estimates
- Clustered SE versus biased simple SE in a FE panel regression
- Measles immunization and child survival, FE panel regression estimates.
  - Within R-squared presented for FE regressions.
- Source: worldbank-immunization dataset;
- balanced yearly panel, years 1998–2017 in 172 countries.

Case study - the effect of measles immunization on child survival.

FE regressions with different Simple and Clustered SE estimates.

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 countries	172	172

Within R-squared presented for FE regressions. Standard error estimates in parentheses. ** p<0.01, * p<0.05.
Source: `worldbank-immunization dataset`; balanced yearly panel, years 1998–2017 in 172 countries.

Generalization 4: Allowing for flexible dynamics in effect, Panel FD model

Panel Regression: taking stock 3

- So far 1: Pooled time series
  - N= small, T = long
  - First difference
  - Focus on time dimension, dynamics
- So far 1: FE Panel Regression
  - N= large, T = small (may be large)
  - Level
  - Focus on within-unit comparison
- Now combine key ideas
  - Large N, focus on within unit variation
  - Allowing for dynamics

Panel Regression in First Differences

- Setup is the same: xt panel data, many cross-sectional units,
- panel regression in first differences or FD panel regression.
- FD = changes -> $\Delta y_{it} = y_{it} - y_{i(t-1)}$.
- FD panel regression with a common intercept across all i.

$$\Delta y_{it}^{E} = \alpha + \beta \Delta x_{it} \tag{13}$$

- Looks like a pooled a cross-section with first difference.
- But N is large
- We have a single intercept, α

Panel Regression in First Differences

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

- β shows the difference in the average change of y for units that experience a change in x during the same period.
- Comparing different cross-sectional units for the same time, or comparing different time periods for the same unit, β shows how much more y changes, on average, where and when x increases by one unit.

FD Panel Regression : what's new

- FD panel regression is alternative to FE panel regression
  - Similar setting
  - We can capture dynamics
- Related to pooled time series. But here N is large, T is small(er)

Lags and Leads in FD Panel Regressions

- Often, we want to estimate not only immediate effects but longer run effects, too.
- Multiple time periods allow us to capture the time path of the effects by including lags of Δx in the regression.
- Same idea as with pooled time series

- Regression in FD with K lags:

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

Lags and Leads in FD Panel Regressions

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

- α shows the trend in y: average change in y when x does/did not change
- β_0 is the contemporaneous slope and it shows how much more y changes, on average, for observations with a change in x in the same time period
  - ... but no change in all K preceding time periods.
- β_k is the slope on lag number k ($k = 1, 2, \ldots, K$) – how much more y changes, on average, for observations with a one unit higher increase in x in the k-th preceding time period
  - ... but with the same change of x in the current time period as well as all K preceding time periods except for the kth.

Lags and Leads in FD Panel Regressions

- cumulative effect or long-run effect of the change of x = sum of the immediate effect and all lagged effects:

$$\beta_{cumul} = \beta_0 + \beta_1 + ... + \beta_K \tag{15}$$

- Same as before, can add difference to get the cumulative effect directly out

Lags and Leads in FD Panel Regressions ***

- The trick in a regression with K lags is to include the K^{th} lag and the differences of the previous lags: the cumulative coefficient is then the one on the K^{th} lag.
- When variables are in first differences, the K^{th} lag is in difference, and the previous lags are differences of the difference.
- The formula for panel regression is

$$\begin{aligned}\Delta y_{it}^{E} &= \alpha + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + \dots + \beta_K \Delta x_{i(t-K)} \\ &= \alpha + \beta_{cumul} \Delta x_{i(t-K)} + \gamma_0 \Delta(\Delta x_{it}) + \gamma_1 \Delta(\Delta x_{i(t-1)}) + \\ &+ \dots + \gamma_{K-1} \Delta(\Delta x_{i(t-K+1)}) \\ &with \quad \beta_{cumul} = \beta_0 + \beta_1 + \dots + \beta_K\end{aligned} \tag{16}$$

Lags and Leads in FD Panel Regressions

- We can also add lead terms to an FD regression to examine pre-trends and capture reverse effects, just like with single time series.
- An FD panel regression with K lags and L leads looks like this:

$$\Delta y_{it}^{E} = \alpha + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + \ldots + \beta_K \Delta x_{i(t-K)} + \gamma_1 \Delta x_{i(t+1)} + \ldots + \gamma_L \Delta x_{i(t+L)} \quad (17)$$

  - The γ coefficients on the lead terms are zero if, prior to time periods when x may change, y tends to change the same way regardless of whether and how much x actually changes.

Lags and Leads in FD Panel Regressions

- Causality if assumption holds re pre-intervention trends too: what happened to y before an intervention, or more generally, before a change in the causal variable x
- Adding leads directly to the model is a bit better than inspecting.
- But still about the past - it's still an assumption

Aggregate Trend in FD Models

- As for FE models, we can add time dummies to capture non-linear trend
- FD regression with K lags and time dummies (time FE) is the following:

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

- θ_t = coefficients of the time dummies
  - = time-specific intercepts = time fixed effects.

Individual Trends in FD Models

- Time dummies capture an aggregate trend in a completely flexible way
- Cross-sectional units in the data may have their own trends, too.
  - Here we don't have the opportunity to estimate flexible trends, because we have only one observation for each time period for each unit.

- Can capture individual linear trends: allow the intercept to be different across cross-sectional units.
  - trend = average change per unit
  - as with pooled time series

Individual Trends in FD Models

- FD regression with K lags, time dummies, and individual-specific intercepts:

$$\Delta y_{it}^{E} = \alpha_i + \theta_t + \beta_0 \Delta x_{it} + \beta_1 \Delta x_{i(t-1)} + \ldots + \beta_K \Delta x_{i(t-K)} \tag{19}$$

- α_i: the average change in y in cross-sectional unit i across all time periods
  - measured as a deviation from the flexibly estimated aggregate trend θ_t,
  - and when x does not change (and didn't change for the past K time periods).

Panel Regression: taking stock 4

- Model in first difference takes care of level differences (as we look at differences by design)
- Add aggregate trend (as dummies) and individual linear trends (as unit specific intercept)

- So this model takes care of confounders that are
  - correlated with levels of x_t and y_t
  - correlated the global trends affecting x and y the same way
  - makes linear trends in unit specific x and y correlate

Case study – Immunization against Measles and Saving Children

- The immediate and lagged effect of measles immunization on child survival
- FD panel regression estimates
- Cumulative effect estimates calculated via transformation.
- Clustered standard error
- balanced yearly panel, years 1998–2017 in 172 countries.

Variables	(1) $\Delta surv$	(2) $\Delta surv$	(3) $\Delta surv$	(4) $\Delta surv$
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)
R-squared	0.013	0.078	0.078	0.093
Observations	3,268	2,408	2,408	1,892

Case study – Immunization against Measles and Saving Children

- The effect of measles immunization on child survival. FD panel regression estimates with year dummies, confounders, and country-specific trends
- FD panel regressions with 5 lags of all right-hand-side variables.
  - Cumulative coefficient on the change of immunization over the 5 lags.
  - Clustered standard error estimates in parentheses.
- Adding leads - 3 periods

Case study - The effect of measles immunization on child survival

The effect of measles immunization on child survival - FD model estimates

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

FD panel regressions with 5 lags of all right-hand-side variables. Confounders: GDP per cap, population. Cumulative coefficient w 5 lags. Clustered SE estimates in parentheses. ** p<0.01, * p<0.05. Source: `worldbank-immunization dataset`; balanced yearly panel, years 1998–2017 in 172 countries.

Detour: good tables

- Focus on key causal variable
- Note but not publish values we don't care about (like global time trend dummies)
- Detailed footnote
- N of obs, key stats (here R-squared)
- Opted for long title - could be a shorter one

Case study - The effect of measles immunization on child survival

- Baseline result 0.05
- Year dummies + confounders: 0.030 - confounders clearly important
- Adding individual linear time trend: 0.011 - small but precisely measured

- A 10 percent increase in the immunization rate tends to be followed by a 0.1 percentage point increase in the child survival rate within five years in the data relative to its country-specific trend
- Corresponding expected increase in child survival is 0.05 to 0.17 percentage points in the general pattern represented by the data.

Case study - The effect of measles immunization on child survival

- Causal effect?
- We can't be certain. It's observational data.
  - Country-specific trends: can't be certain that this ensures that the parallel trends assumption,
- We did a great deal of efforts to condition on all kinds of confounders.
  - FD model with lags - takes out level differences and accounts for dynamics
  - Key confounders added: GDP per capita and population + individual linear trends
  - PTA - make a very good effort: Adding leads or confounders like population, gdp makes no difference.
- Good approximation to what the true effect: A 10 percent increase in the immunization rate leads to a 0.1 percentage point increase in the child survival rate within five years.

Case study - The effect of measles immunization on child survival

- Ok, so consider a nutrition supplement policy where a nurse visits people.
- this would help both survival and also makes families easier to get to vaccination.
- FE, or FD models take care of differences on average, ie like development
- FD + FE models take care the variation in speed of how such policies are implemented / carried out on site.

- Broader set of confounders: cross country differences in long-term efforts to make the health system better, with most of its elements getting improved in parallel.

- And so identification comes from different functional form of changes around individual unit trends

Case study - The effect of measles immunization on child survival

- Adding all these confounders - could it be too much?
- Is it possible we partial out some of exogenous variation in x?
- Yes. Individual linear trends - if measles vaccination is linear
- Unlikely
- But 0.01 may be a lower bound, while 0.03 an upper bound.
- While of course, other confounders may lurk
- Analytical choice how to present.

Working with panel data, making decisions

Dealing with Unbalanced Panels

- Missing observations: missing at random or not
- If missing at random - okay to keep. Maybe FE models will be better.
- If not
  - Reduce T - focus only on more recent years when coverage is high
  - Reduce N - drop unit (countries) where coverage is low
- Sample design (filtering out observation) means we have a different sample, and may not be representative to what we started with.
- Many analytical choice, but must make notes

Panel Regressions and Causality

- FE regressions and FD regressions can estimate the effect of x on y without the bias due to confounders that don't change over time.
- Confounders that change through time need to be observed and included in the FE or FD regression.
- Conditioning on individual trends is feasible with FD regressions
  - Can do something similar in FE, but (even more) complicated

- Panel model allow us conditioning on a great deal of confounding factors
- But, as always, there can be omitted variables - so never certain.

First Differences or Fixed Effects?

- Have seen many models, which one to choose?
- FE and FD regressions are similar because both condition on confounders that affect the level of y and x and don't change through time.
  - FE regressions do that by comparing values of y and x to their cross-sectional means.
  - FD regressions do something similar by comparing values of y and x to their values in the previous time period.
- Confounders that affect the change in y or x still matter for both FE and FD regressions, whether the confounders themselves change through time or not

First Differences or Fixed Effects?

- FD main advantage 1: capture serial correlation by first differencing
  - important if time series properties key
- FD main advantage 2: capture transparent dynamics
- As long as we keep adding lags. But that means smaller and smaller panel for estimation.
  - FD takes care of linear trend automatically, but as we add anyway, no big deal
- FD main advantage 3: can easily capture individual linear trends

First Differences or Fixed Effects?

- FE main advantage 1: simple method of estimating longer run effects, easier to use
    - estimate of the average of short-term and longterm effects.
    - When the long-term effects kick in fast, that's a good approximation of the long-term effects themselves
- FE main advantage 2: Works when missing values in panel (see next bit)

First Differences or Fixed Effects?

- In many cases, both FD and FE can work.
- If similar result, use either, and show robustness of method
- If different - should investigate
  - Time path of effect
  - Missing values, too short series for lags
  - Very strong serial correlation

One more option: Long difference

- We have seen FE and FD.
- One other model is the long difference: considering the difference between the end and beginning of our panel.
- Intervention happens sometime during a long period
- Technically: a difference in differences regression
  - before and after are further apart
  - To capture long run difference

- α Expected long-term change in y when x does not change
- β: Compare two units with different changes in x across the long time horizon. y is expected to increase by β more more where or when x increases by one more unit.

One more option: Long difference

- LD is useful first step, simple to carry out, interpret.
- Especially when we do not really know when and how interventions happened.
- least likely to condition on confounders
- PTA is like: Compare countries that experienced different changes in immunization between 1998-2018.
- Had they experienced the same change in immunization, child survival would have changed the same way, on average.
- Least likely to be believed.
- Hence: in most cases, FD or FE is preferred.

Don't do it at home: POLS

- "Pooled OLS".

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

- Has time dummies
- NO fixed effects for cross-sectional units.
- It is an average of cross-sectional OLS regressions.
  - Looks like a panel model, but in fact, it's not.
- It's better to pick a single year for cross section.
  - Maybe check for a few years.

Don't do it at home: RE

- Random Effects (RE) model
- A mixture of fixed effects and POLS
- In rare case, a model leads to it. In PhD only
- Otherwise, just don't do it.

Panel Regression: taking stock 4

- Panel data: first difference or fixed effects
  - Each with some technical issues to pay attention to.
- Often similar results, but not always
  - Worth investigating
- The strictest way is first difference model with cross-sectional unit FE. Protects against individual trends and time-invariant confounders.
- Some other models
  - LD is okay to do, less informative re causality
  - POLS, RE - to avoid.

Summary: Panel Regression

- Data with multiple time periods can help uncover short- and long-run effects and examine pretrends.
- When interested in the effects on a single cross-sectional unit, we may analyze a single time series or pool several time series of similar units.
- With panel data having multiple time periods, several modeling options
- use an FD regression to uncover the development of the effect over time, and an FD or an FE regression to uncover the long-run effect
- Watch out for interpretation - hard

- Overall big picture: using panel data methods can take us much closer to a causal interpretation.

Panel Regression: Some issues

- Panel data methods – some deeper understanding of possible issues in 2020s
- Ask your instructor
- Plus, online updates coming in 2025