

Panel-data analysis and difference-in-differences modeling

An applied Stata workshop

Chris Cheng

Senior Econometrician

StataCorp LLC

New York City

ASA Teaching Workshop

August 9, 2026

<https://www.stata.com/news/conferences/asa/>

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed
effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in- differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

What is panel data?

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † (math stats) repeated measures (*wide vs long* format)
- † (biostatistics) longitudinal data
- † (social sciences) panel data

How it looks

† A typical panel vs. cross-sectional data structure

country	year	consumption	gdp	irate
Canada	2010	1269.1912	1613.5428	.2
Canada	2011	1294.7359	1664.3191	.475
Canada	2012	1315.0192	1693.649	.48333333
Mexico	2010	815.78416	1057.8013	1.2125
Mexico	2011	842.78459	1096.5486	.95583333
Mexico	2012	863.83937	1136.4885	1.0816667
United States	2010	12695.979	14992.053	2.4000001
United States	2011	12812.144	15224.555	6.5
United States	2012	12932.334	15567.038	7

panel

country	consumption	gdp	irate
Canada	1357.525	1745.8304	.31666667
Mexico	905.63994	1189.6593	1.4738889
United States	13561.267	16307.542	5.0000001

cross-sectional

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Source: <http://databank.worldbank.org/data/Home.aspx>

- † Data management
- † Linear regression estimators *
- † Dynamic panel-data estimators
- † Nonlinear regression estimators *
- † Postestimation tools
- † Extended regression models
- † Causal inference models *

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

Panel-data linear models

Data-generating process (DGP)

† The data-generating process is given by

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + \eta_{it}$$

$$\eta_{it} \equiv \alpha_i + \varepsilon_{it}$$

$$i = 1, \dots, n$$

$$t = 1, \dots, T$$

† The random disturbance (η_{it}) includes two parts:

- ◇ α_i : the **unobservable** component is particular to each panel and is **time-invariant** (e.g., for individuals: ability, intelligence, work ethic). As in the regression case, the assumptions made about η_{it} , with particular emphasis on α_i , define the models we work with.
- ◇ ε_{it} : the idiosyncratic error term

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Model for aggregate consumption

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

$$\text{consumption}_{it} = \beta_0 + \beta_1 * \text{gdp}_{it} + \beta_2 * \text{irate}_{it} + \alpha_i + \varepsilon_{it}$$

- † World Bank public online data for
 - consumption**: final consumption expenditure (2010 US\$)
 - gdp**: gross domestic product (2010 US\$)
 - irate**: deposit interest rate
- † Example: 2010-2018 for 131 countries
- † Source: <http://databank.worldbank.org/data/Home.aspx>

Specifying the panel structure in Stata

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † Users can tell Stata that their data have a special structure for various types of datasets
- ✓ Repeated measures, panel, or longitudinal datasets – see **help xtset**
- † Time-series datasets – see **help tsset**
- † Survival time datasets – see **help stset**
- † Datasets arising from complex survey designs (called survey datasets) – see **help svyset**

Specifying the panel structure in Stata

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† Assuming that the second dimension corresponds to time series, we use the **xtset** command to specify the panel structure with

- ◇ Panel identifier variable (e.g., country)
- ◇ Time identifier variable (e.g., year)

```
. xtset country year
```

Panel variable: country (unbalanced)

Time variable: year, 2010 to 2018, but with a gap

Delta: 1 unit

Note: You can specify the panel structure using **xtset panelvar** if you want to ignore the time structure.

Theoretical framework (DGP revisit)

$$\begin{aligned}y_{it} &= \beta_0 + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + \eta_{it} \\ \eta_{it} &= \alpha_i + \varepsilon_{it}\end{aligned}$$

† As in the classical linear regression, all models are defined by two components:

1. The data-generating process (DGP)
2. The relationship between the random disturbance or idiosyncratic shock and the explanatory variables
3. How the relationship is structured defines your model: random- vs. fixed-effects

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † The regressors are unrelated to the ***unobserved time-invariant component*** α_i

$$E(\alpha_i | x_{it1}, \dots, x_{itk}) = E(\alpha_i)$$

- † Strict exogeneity, no lagged dependent variables

$$E(\varepsilon_{it} | x_{it1}, \dots, x_{itk}, \alpha_i) = 0$$

- † The previous two assumptions allow us to think about using a regression. But:

$$E(\varepsilon_i \varepsilon_i' | x_i, \alpha_i) = \sigma_\varepsilon^2 I_T$$

$$E(\varepsilon_{it}^2) = \sigma_\varepsilon^2$$

$$E(\varepsilon_{it} \varepsilon_{is}) = 0$$

$$V(\alpha_i^2) = E(\alpha_i^2 | x_i) = \sigma_\alpha^2$$

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Random-effects variance-covariance matrix

† For each individual i , we have:

$$\Omega = E(\eta_i \eta_i') = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 & \dots & \vdots \\ \vdots & \dots & \ddots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{pmatrix}$$

† This gives rise to an efficient estimator:

$$\begin{aligned} \Omega^{-1/2} y_i &= \Omega^{-1/2} x_i \beta + \Omega^{-1/2} \eta_i \\ \Omega^{-1/2} z_i &\equiv z_i^* \end{aligned}$$

† This implies that we have the following model:

$$\begin{aligned} y_i^* &= x_i^* \beta + \eta_i^* \\ E(\eta_i^* \eta_i^{*'}) &= I_T \end{aligned}$$

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Random-effects estimation with Stata

† Loading data and setup

```
. use gdp
. xtset country year
Panel variable: country (unbalanced)
Time variable: year, 2010 to 2018, but with a gap
Delta: 1 unit
. describe
Contains data from gdp.dta
Observations:      1,016
Variables:          10                      3 Aug 2026 12:17
```

Variable name	Storage type	Display format	Value label	Variable label
CountryName	str30	%30s		Country Name
year	float	%9.0g		
irate	double	%10.0g		Deposit interest rate
consumption	double	%10.0g		Consumption (Billions 2000 US\$)
gdp	double	%10.0g		GDP (Billions 2000 US\$)
country	long	%30.0g	country	Country Name
ln_cons	float	%9.0g		Log of consumption
ln_gdp	float	%9.0g		Log of gdp
ln_irate	float	%9.0g		Log of irate
region	long	%12.0g	region	Regional groups

Sorted by: country year

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Random-effects estimation with Stata

† The syntax:

```
xtreg depvar indepvars, [ re ]
```

† Fitting the model

```
. xtreg ln_cons ln_gdp ln_irate, re
```

Random-effects GLS regression

Group variable: country

R-squared:

Within = 0.8033

Between = 0.9859

Overall = 0.9847

Number of obs = 1,016

Number of groups = 131

Obs per group:

min = 1

avg = 7.8

max = 9

Wald chi2(2) = 13277.81

Prob > chi2 = 0.0000

corr(u_i, X) = 0 (assumed)

ln_cons	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
ln_gdp	.958856	.0084128	113.98	0.000	.9423672	.9753449
ln_irate	-.0039294	.0041147	-0.95	0.340	-.011994	.0041352
_cons	.760708	.2065915	3.68	0.000	.3557961	1.16562
sigma_u	.2339765					
sigma_e	.05205235					
rho	.95284182	(fraction of variance due to u_i)				

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † The Wald `chi2(df)` statistic is the overall relevance of the model
- † The three different R-squared statistics represent the variability of y explained by the model's predicted values. But there are three possible measures of y :

y_{it}	Overall
$\bar{y}_i$	Between
$y_{it} - \bar{y}_i$	Within

- † `corr(u_i, X)` refers to the correlation between the time-invariant component α_i , in this case called `u_i`, and the regressors. For the random effects, we assume it is zero.
- † $\text{sigma_u} = \sigma_\alpha$,
 $\text{sigma_e} = \sigma_\varepsilon$,
 $\text{rho} = \sigma_\alpha^2 (\sigma_\varepsilon^2 + \sigma_\alpha^2)^{-1}$

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† Random effects vs. pooled OLS

```
. xttest0
```

Breusch and Pagan Lagrangian multiplier test for random effects

```
ln_cons[country,t] = Xb + u[country] + e[country,t]
```

Estimated results:

	Var	SD = sqrt(Var)
ln_cons	4.027035	2.006747
e	.0027094	.0520524
u	.054745	.2339765

Test: Var(u) = 0

```
chibar2(01) = 3108.85  
Prob > chibar2 = 0.0000
```

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † The regressors are correlated with the ***unobserved time-invariant*** component α_i

$$\text{Cov}(\alpha_i, x_i) \neq 0$$

- † Strict exogeneity, no lagged dependent variables:

$$E(\varepsilon_{it} | x_{it1}, \dots, x_{itk}, \alpha_i) = 0$$

† In the model we have been discussing:

$$\ln(\text{consumption})_{it} = \beta_0 + \beta_1 \ln(\text{GDP})_{it} + \beta_2 \ln(\text{irate})_{it} + \alpha_i + \varepsilon_{it}$$

† It is difficult to maintain, for a particular model, that the unobserved individual component is independent of all regressors

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + \alpha_i + \varepsilon_{it} \quad (1)$$

† If we take the average over the T_i observations of each panel, we obtain

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \dots + \beta_k \bar{x}_{ik} + \alpha_i + \bar{\varepsilon}_i$$

where

$$\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it},$$

$$\bar{x}_{ij} = T^{-1} \sum_{t=1}^T x_{itj}$$

† We can now construct the following object:

$$y_{it} - \bar{y}_i = (\beta_0 - \beta_0) + \beta_1 (x_{it1} - \bar{x}_{i1}) + \dots + \beta_k (x_{itk} - \bar{x}_{ik}) + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

† We can then estimate the parameters of interest from equation (1):

$$\tilde{y}_i = \beta_1 \tilde{x}_{i1} + \dots + \beta_k \tilde{x}_{ik} + \tilde{\varepsilon}_i$$

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Within estimation

† The syntax:

xtreg depvar indepvars, fe

† Fitting the model

```
. xtreg ln_cons ln_gdp ln_irate, fe
```

Fixed-effects (within) regression

Group variable: country

R-squared:

Within = 0.8034

Between = 0.9858

Overall = 0.9845

Number of obs = 1,016

Number of groups = 131

Obs per group:

min = 1

avg = 7.8

max = 9

F(2, 883) = 1804.60

Prob > F = 0.0000

corr(u_i, Xb) = -0.0175

ln_cons	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
ln_gdp	.958713	.016523	58.02	0.000	.926284	.991142
ln_irate	-.0074047	.0042761	-1.73	0.084	-.0157972	.0009878
_cons	.7750608	.4063998	1.91	0.057	-.0225615	1.572683
sigma_u	.24585324					
sigma_e	.05205235					
rho	.95709727	(fraction of variance due to u_i)				

F test that all u_i=0: F(130, 883) = 152.63

Prob > F = 0.0000

Introduction

Panel-data linear
models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed
effects

Panel-data
nonlinear models

Binary outcome

Other nonlinear models

Difference-in-
differences models

Motivation

Homogeneous DID

Heterogeneous DID

Fixed effects vs. random effects

† Theory should be one of the main factors guiding your modeling decisions

Assumption	Fixed- vs. Random-effects			Implication
	Estimator	Consistent	Efficient	
$Cov(x_{it}, \alpha) = 0$	Within/F.D	Yes	No	$\hat{\beta}_{fe} \approx \hat{\beta}_{re}$
	Random-effects	Yes	Yes	$var(\hat{\beta}_{fe}) > var(\hat{\beta}_{re})$
$Cov(x_{it}, \alpha) \neq 0$	Within/F.D	Yes		$\hat{\beta}_{fe} \neq \hat{\beta}_{re}$
	Random-effects	No		

† However, you should present a statistical test to back up your claims

1. Hausman test for fixed effects vs random effects
2. Mundlak test for fixed effects vs random effects

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † The following object has a chi-squared distribution with degrees of freedom equal to the number of regressors:

$$H = \left(\hat{\beta}_{fe} - \hat{\beta}_{re} \right)' \left[\widehat{VCE}_{fe} - \widehat{VCE}_{re} \right]^{-1} \left(\hat{\beta}_{fe} - \hat{\beta}_{re} \right)$$

- † The test implicitly assumes that the random effects model is efficient, which in turn makes $\left[\widehat{VCE}_{fe} - \widehat{VCE}_{re} \right]$ positive definite.
- † The test does not accommodate heteroskedasticity and serial correlation.

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† Performing the test

```
. quietly xtreg ln_cons ln_gdp ln_irate, re
. estimates store eq_re
. quietly xtreg ln_cons ln_gdp ln_irate, fe
. estimates store eq_fe
. hausman eq_fe eq_re
```

	Coefficients		(b-B) Difference	sqrt(diag(V_b-V_B)) Std. err.
	(b) eq_fe	(B) eq_re		
ln_gdp	.958713	.958856	-.000143	.0142209
ln_irate	-.0074047	-.0039294	-.0034753	.0011638

b = Consistent under H0 and Ha; obtained from xtreg.

B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

```
chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 17.25
```

```
Prob > chi2 = 0.0002
```

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † As an alternative to the Hausman test, we can also implement the Mundlak approach, which is now commonly called a correlated random-effects model
- † The main idea is to model the correlation between the unobserved component and the explanatory variables

$$E(\alpha_i | x_{it}) = \theta_0 + \theta_1 \bar{x}_{i1} + \dots + \theta_k \bar{x}_{ik}$$

- † This implies that:

$$\begin{aligned} E(y_{it} | x_{it}) &= (\beta_0 + \theta_0) + \beta_1 x_{it1} + \dots + \beta_k x_{itk} \\ &\quad + \theta_1 \bar{x}_{i1} + \dots + \theta_k \bar{x}_{ik} \end{aligned}$$

$$E(y_{it} | x_{it}) = \gamma_0 + \gamma_1 x_{it} + \gamma_2 \bar{x}_i$$

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† If we have a random-effects model:

$$\theta_1 = \dots = \theta_k = 0$$

$$\gamma_2 = 0$$

† If not, the coefficients will have some meaning

† Therefore:

$$H_0 : \theta_1 = \dots = \theta_k = 0$$

$$H_0 : \gamma_2 = 0$$

Constructing the Mundlak test

† In **StataNow 18.5** –update 25jun2024–, we introduced a new option to `xtreg`. The syntax:

`xtreg depvar indepvars, cre`

```
. xtreg ln_cons ln_gdp ln_irate, cre
Correlated random-effects regression      Number of obs   =       1,016
Group variable: country                   Number of groups =        131
R-squared:                               Obs per group:
    Within = 0.8034                        min =           1
    Between = 0.9872                      avg =           7.8
    Overall = 0.9864                      max =           9
                                           Wald chi2(2)     =       3609.37
                                           Prob > chi2      =        0.0000

corr(xit_vars*b, xt_means*y) = 0.8552
```

ln_cons	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
xit_vars						
ln_gdp	.958713	.0165226	58.02	0.000	.9263292	.9910968
ln_irate	-.0074047	.004276	-1.73	0.083	-.0157855	.0009761
_cons	.6099668	.2413124	2.53	0.011	.1370031	1.08293
xt_means						
ln_gdp	.0031773	.0191767	0.17	0.868	-.0344083	.0407629
ln_irate	.0680654	.0184904	3.68	0.000	.0318248	.104306
sigma_u	.2339765					
sigma_e	.05205235					
rho	.95284182	(fraction of variance due to u_i)				

Mundlak test (xt_means = 0): $\chi^2(2) = 13.5984$

Prob > $\chi^2 = 0.0011$

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Performing the Mundlak test

† To perform the Mundlak test, we can issue

```
. estat mundlak  
  
Mundlak specification test  
H0: Covariates are uncorrelated with unobserved panel-level effects  
      chi2(2) = 13.60  
Prob > chi2 = 0.0011  
  
Notes: Fixed effects and correlated random effects are  
       consistent under H0 and Ha.  
       Random effects are efficient under H0.
```

† A significant test favors fixed-effects and correlated random-effects

† Available after

- ◇ xtreg, re
- ◇ xtreg, fe
- ◇ xtreg, cre

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Two-way fixed effects

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † What if we want to include fixed effects for another dimension, such as time/year?
- † We can add this dimension by introducing time/year dummy/indicator variables

Two-way fixed effects

† Adding yearly indicators

```
. xtreg ln_cons ln_gdp ln_irate i.year, fe
Fixed-effects (within) regression
Group variable: country
R-squared:
    Within   = 0.8183
    Between  = 0.9862
    Overall  = 0.9848

Number of obs   =    1,016
Number of groups =     131
Obs per group:
    min = 1
    avg = 7.8
    max = 9
F(10, 875)      =    394.11
Prob > F         =    0.0000

corr(u_i, Xb) = 0.8149
```

ln_cons	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
ln_gdp	.792342	.0257757	30.74	0.000	.7417526	.8429313
ln_irate	-.001492	.004219	-0.35	0.724	-.0097725	.0067884
year						
2011	.007819	.0065987	1.18	0.236	-.0051322	.0207702
2012	.0197851	.0068678	2.88	0.004	.0063058	.0332645
2013	.0364415	.0072375	5.04	0.000	.0222366	.0506465
2014	.0437999	.0076573	5.72	0.000	.0287712	.0588287
2015	.0523564	.0081018	6.46	0.000	.0364551	.0682577
2016	.0537609	.0086048	6.25	0.000	.0368725	.0706494
2017	.0619798	.0091493	6.77	0.000	.0440226	.079937
2018	.0693091	.010371	6.68	0.000	.0489541	.089664
_cons	4.811732	.62864	7.65	0.000	3.577913	6.04555
sigma_u	.42631054					
sigma_e	.05027135					
rho	.98628514	(fraction of variance due to u_i)				

F test that all $u_i=0$: $F(130, 875) = 162.48$

Prob > F = 0.0000

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects**

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

Two-way fixed effects

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † In fact, we estimated year fixed effects
- † But what if there are many year levels, or we want to include yet another dimension?
- † In **StataNow 18.5** -update 30apr2024-, high-dimensional fixed effects are incorporated into Stata with the `areg` and `xtreg, fe` commands

<https://www.stata.com/new-in-stata/high-dimensional-fixed-effects/>
- † We can include another dimension now using the `absorb` option

Two-way fixed effects

† The syntax:

xtreg depvar indepvars, fe absorb(abs_vars)

```
. xtreg ln_cons ln_gdp ln_irate, fe absorb(year) nolog
note: term year has 2018 levels in a dataset of size 1016
Alternating projection maximum absolute difference = 6.064e-09
Fixed-effects (within) regression      Number of obs   =      1,016
Group variable: country                Number of groups =      131
R-squared:                             Obs per group:
    Within = 0.8183                      min =          1
    Between = 0.9860                     avg =          7.8
    Overall = 0.9848                     max =          9
                                         F(2, 875)       =      472.60
                                         Prob > F         =      0.0000

corr(u_i, Xb) = 0.8152
```

Absorbed variable	Levels
year	9

ln_cons	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
ln_gdp	.792342	.0257757	30.74	0.000	.7417526	.8429313
ln_irate	-.001492	.004219	-0.35	0.724	-.0097725	.0067884
_cons	4.847955	.6321408	7.67	0.000	3.607265	6.088644
sigma_u	.42631054					
sigma_e	.05027135					
rho	.98628514	(fraction of variance due to u_i)				

F test that all u_i=0: F(130, 875) = 164.19

Prob > F = 0.0000

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † The results are the same as when we include year indicator variables, but the year fixed effects do not need to be estimated.
- † The `absorb()` option accepts multiple variables, so you can include additional fixed effects to be absorbed.

Panel-data nonlinear models

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

Panel-data models for binary latent dependent variables

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

$$\begin{aligned}y_{it}^* &= x_{it}'\beta + z_i'\gamma + \alpha_i + \varepsilon_{it} \\ &= w_{it}'\delta + \alpha_i + \varepsilon_{it}\end{aligned}$$

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

† As in the linear panel models

- ◇ If α_i is related to x_{it} or z_i , it is known as a fixed-effects model
- ◇ If α_i is unrelated to x_{it} and z_i , it is known as a random-effects model

- † For fixed-effects models:
 - ◇ We require a method to handle α_i (feasible for logit models).
 - ◇ The Mundlak approach provides a useful alternative in this scenario.
 - ◇ Partial effects cannot be estimated under this specification unless we assume the fixed effect is zero.
- † For random-effects models:
 - ◇ Integrate out the unobserved component.
 - ◇ We can estimate partial effects (marginal effects).
- † We can test random vs. fixed effects using model diagnostics.

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† The syntax:

```
xtprobit depvar indepvars, [ re ]
```

```
xtlogit depvar indepvars, [ re ] | fe
```

† We can fit a random-effects probit model as well as both random-effects and fixed-effects logit models.

Panel-data probit models

† Loading the data

```
. use nlswork, clear
(National Longitudinal Survey of Young Women, 14-24 years old in 1968)

. describe

Contains data from nlswork.dta
Observations:      28,534      National Longitudinal Survey of
                                Young Women, 14-24 years old in
                                1968
Variables:          8         24 Jul 2026 11:24
                                (_dta has notes)
```

Variable name	Storage type	Display format	Value label	Variable label
idcode	int	%8.0g		NLS ID
year	byte	%8.0g		Interview year
age	byte	%8.0g		Age in current year
race	byte	%8.0g	racelbl	Race
collgrad	byte	%8.0g		1 if college graduate
south	byte	%8.0g		1 if south
union	byte	%8.0g		1 if union
tenure	float	%9.0g		Job tenure, in years

Sorted by: idcode year

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Panel-data probit models

† Let's see what determines workers' union membership by fitting random-effects probit models

```
. xtprobit union tenure age i.collgrad i.south, re nolog
Random-effects probit regression               Number of obs   = 19,007
Group variable: idcode                       Number of groups  = 4,134
Random effects u_i ~ Gaussian                Obs per group:
                                             min   =      1
                                             avg   =     4.6
                                             max   =     12
Integration method: mvaghermite              Integration pts. = 12
Wald chi2(4)                                = 287.07
Prob > chi2                                  = 0.0000
Log likelihood = -7661.754
```

union	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
tenure	.0568389	.0051368	11.07	0.000	.046771	.0669068
age	-.008689	.003115	-2.79	0.005	-.0147942	-.0025837
1.collgrad	.4186621	.074082	5.65	0.000	.273464	.5638602
1.south	-.6029729	.0556627	-10.83	0.000	-.7120699	-.493876
_cons	-1.182167	.0994035	-11.89	0.000	-1.376995	-.9873401
/lnsig2u	.8334059	.053721			.7281148	.938697
sigma_u	1.516952	.040746			1.439157	1.598952
rho	.6970746	.0113438			.6743914	.7188364

LR test of rho=0: chibar2(01) = 4556.75

Prob >= chibar2 = 0.000

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

† We can fit the same model using the logit link

```
. xtlogit union tenure age i.collgrad i.south, re nolog
Random-effects logistic regression      Number of obs   = 19,007
Group variable: idcode                 Number of groups  = 4,134
Random effects u_i ~ Gaussian          Obs per group:
                                     min =      1
                                     avg =     4.6
                                     max =     12
Integration method: mvaghermite        Integration pts.  = 12
Wald chi2(4)                          = 294.10
Prob > chi2                           = 0.0000
Log likelihood = -7649.6368
```

union	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
tenure	.1026306	.0091387	11.23	0.000	.0847191	.1205422
age	-.015351	.005548	-2.77	0.006	-.0262248	-.0044771
i.collgrad	.744387	.1297776	5.74	0.000	.4900277	.9987464
i.south	-1.086953	.0986403	-11.02	0.000	-1.280284	-.8936214
_cons	-2.094546	.1767535	-11.85	0.000	-2.440976	-1.748115
/lnsig2u	1.966757	.0545097			1.85992	2.073594
sigma_u	2.673474	.0728651			2.534408	2.82017
rho	.684798	.0117659			.6612956	.7073911

LR test of rho=0: chibar2(01) = 4580.48

Prob >= chibar2 = 0.000

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

† The logit estimator allows us to condition out fixed effects

```
. xtlogit union tenure age i.collgrad i.south, fe nolog
note: multiple positive outcomes within groups encountered.
note: 2,899 groups (11,485 obs) omitted because of all positive or
      all negative outcomes.
note: 1.collgrad omitted because of no within-group variance.
Conditional fixed-effects logistic regression      Number of obs   = 7,522
Group variable: idcode                          Number of groups = 1,235
                                                Obs per group:
                                                min   =      2
                                                avg   =     6.1
                                                max   =     12
                                                LR chi2(3)      = 80.72
                                                Prob > chi2     = 0.0000
```

Log likelihood = -2776.0934

union	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
tenure	.0623928	.0109854	5.68	0.000	.0408619	.0839237
age	-.0029985	.0061811	-0.49	0.628	-.0151132	.0091162
1.collgrad	0 (omitted)					
1.south	-1.035083	.1753132	-5.90	0.000	-1.37869	-.691475

† We cannot estimate the effects of time-invariant variables, such as college graduate status.

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

Other nonlinear panel-data models

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

`xtcloglog` Random-effects and population-averaged cloglog models

`xtologit` Random-effects ordered logistic model

`xtoprobit` Random-effects ordered probit model

`xtmlogit` Fixed-effects and random-effects multinomial logit models

`xtpoisson` Fixed-effects, random-effects, and population-averaged Poisson models

`xtnbreg` Fixed-effects, random-effects, & population-averaged negative binomial models

`xtstreg` Random-effects parametric survival models

† And more

Introduction

Panel-data linear
models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed
effects

Panel-data
nonlinear models

Binary outcome

Other nonlinear models

**Difference-in-
differences models**

Motivation

Homogeneous DID

Heterogeneous DID

Difference-in-differences models

What is DID?

- † Difference-in-differences (DID) compares outcomes along two dimensions:
 - ◇ time: pre vs. post
 - ◇ groups: treated vs. control
- † The DID estimator is the difference in outcome changes between the two groups.

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Why not just pre vs. post?

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † Example: Do higher cigarette taxes reduce smoking?
- † Comparing smoking rates before and after the tax increase among treated smokers is not enough:
 - ◇ smoking may be declining over time for everyone
- † A control group not exposed to the tax increase captures this common time trend.
- † DID isolates the effect of the tax by comparing changes between treated and control groups.

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

Motivation

- Homogeneous DID
- Heterogeneous DID

- † We observe outcomes before and after the intervention for:
 - ◇ a treated group
 - ◇ a control group
- † First difference: before vs. after within each group.
- † Second difference: the difference between these two changes.
- † This second difference is the DID treatment effect.

Parallel trends assumption

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

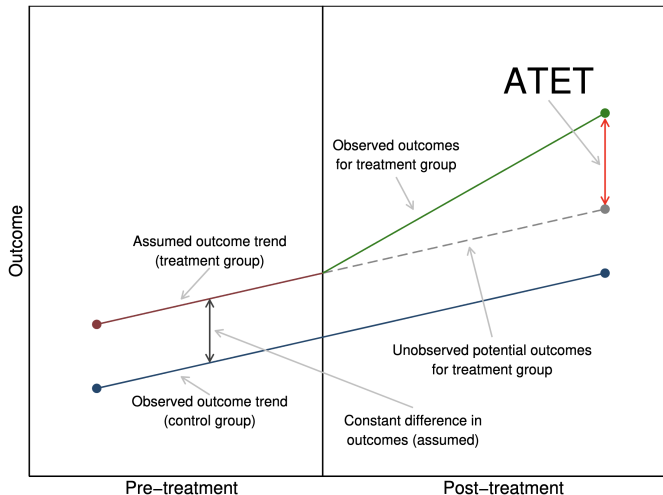
Motivation

Homogeneous DID

Heterogeneous DID

- † A key identifying assumption in DID is *parallel trends*.
- † Prior to the intervention:
 - ◇ treated and control groups follow the same trend
- † In the absence of treatment:
 - ◇ the trends for both groups would have remained the same
- † If this assumption is violated, the DID treatment effect is not identified.
- † DID estimates the average treatment effect on the treated (ATET).

DID estimates ATET



Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

A simple DID model: 2×2 design

	Pre	Post
Treated	$Y_{T,pre}$	$Y_{T,post}$
Control	$Y_{C,pre}$	$Y_{C,post}$

† To illustrate the basic idea of DID, we start with a simple 2×2 example: one treated unit, one control unit, observed once before and once after the intervention.

```
. use ncigs, clear  
. list id treat post ncigs state in 1/5
```

	id	treat	post	ncigs	state
1.	1	0	0	1.278782	24
2.	2	1	0	4.385671	26
3.	3	1	1	-1.121003	57
4.	4	1	1	1.076579	53
5.	5	1	1	.9774689	35

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

A simple DID model: 2×2 design

- † We generalize to a larger dataset with many individuals in both the treatment and control groups.
- † We fit our first DID model using a linear regression with both the treatment and pre/post indicator as well as their interaction.

```
. regress ncigs i.treat##i.post
```

Source	SS	df	MS	Number of obs	=	1,500
				F(3, 1496)	=	7.76
Model	39.5893808	3	13.1964603	Prob > F	=	0.0000
Residual	2545.26062	1,496	1.70137742	R-squared	=	0.0153
				Adj R-squared	=	0.0133
Total	2584.85	1,499	1.72438292	Root MSE	=	1.3044

ncigs	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
1.treat	-.0136317	.0942025	-0.14	0.885	-.1984148	.1711513
1.post	-.0417691	.094846	-0.44	0.660	-.2278145	.1442763
treat#post						
1 1	-.3400387	.1347554	-2.52	0.012	-.6043682	-.0757091
_cons	1.212943	.0662193	18.32	0.000	1.08305	1.342835

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Cell means and DID

$$DID = (Y_{T,post} - Y_{T,pre}) - (Y_{C,post} - Y_{C,pre})$$

```
. margins treat#post, post coeflegend
```

Adjusted predictions

Number of obs = 1,500

Model VCE: OLS

Expression: Linear prediction, predict()

	Margin	Legend
treat#post		
0 0	1.212943	$_b[0bn.treat\#0bn.post]$ ← $\hat{Y}_{C,pre}$
0 1	1.171174	$_b[0bn.treat\#1.post]$ ← $\hat{Y}_{C,post}$
1 0	1.199311	$_b[1.treat\#0bn.post]$ ← $\hat{Y}_{T,pre}$
1 1	.8175033	$_b[1.treat\#1.post]$ ← $\hat{Y}_{T,post}$

```
. display (_b[1.treat#1.post] - _b[1.treat#0bn.post]) ///  
>      - (_b[0bn.treat#1.post] - _b[0bn.treat#0bn.post])  
-.34003867
```

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Stata's `didregress` command

- † Often, data contain multiple pre- and post-treatment observations, but the treatment is administered at a single point in time.
- † The data may be either repeated cross-sectional or panel data.
- † Stata commands for DID (introduced in Stata 17):
 - ◇ `didregress` — repeated cross-sectional
 - ◇ `xtdidregress` — panel data
- † Various methods are available for standard errors, including bias correction and wild cluster bootstrap.
- † Postestimation commands allow tests and diagnostics, including evaluating the parallel trends assumption.

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† Repeated cross-sectional data

```
didregress (y x) (treatment), group() time()
```

† Panel data

```
xtdidregress (y x) (treatment), group() time()
```

See `help didregress` for the full syntax and options.

Note: Introduced in Stata 17

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Treatment as time-varying variable

- † Both commands require the treatment variable to be specified as a time-varying variable.

```
. generate policy = treat*post  
. list id treat post policy in 1/10, sep(0)
```

	id	treat	post	policy
1.	1	0	0	0
2.	2	1	0	0
3.	3	1	1	1
4.	4	1	1	1
5.	5	1	1	1
6.	6	0	0	0
7.	7	0	0	0
8.	8	0	1	0
9.	9	0	0	0
10.	10	0	0	0

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

† We use the following (two-way fixed-effects) model:

$$y_{ist} = \gamma_s + \gamma_t + \mathbf{z}_{ist}\beta + D_{st}\delta + \epsilon_{ist}$$

where

y_{ist}	is the outcome of individual i in group s at time t
γ_s	group fixed effects
γ_t	time fixed effects
$\mathbf{z}_{ist}$	covariates
β	coefficients for the covariates
D_{st}	time-varying treatment status at the group and time levels
δ	ATET
ϵ_{ist}	error term

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID**
- Heterogeneous DID

An example with hospital new procedure data

- † Hospital admissions data in which a new procedure has been implemented.

```
. use hospdd, clear
(Artificial hospital admission procedure data)
```

```
. describe
```

Contains data from hospdd.dta

```
Observations:      7,368      Artificial hospital admission
                                procedure data
Variables:          5          3 Aug 2026 12:41
```

Variable name	Storage type	Display format	Value label	Variable label
hospital	byte	%9.0g		Hospital ID
frequency	byte	%9.0g	size	Hospital visit frequency
month	byte	%8.0g	mnth	Month
procedure	byte	%9.0g	pol	Admission procedure
satis	float	%9.0g		Patient satisfaction score

Sorted by: hospital

```
. xtset hospital
```

Panel variable: hospital (unbalanced)

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Snippet of the data

† Let's look at a treated group and a control group.

```
. list hospital month procedure satis if inlist(hospital,1,27), sepby(hospital) noobs
```

	hospital	month	proced-e	satis
Treated group	1	January	Old	2.98438
	1	February	Old	4.319318
	1	March	Old	2.851151
	1	April	New	4.486683
	1	May	New	4.444119
	1	June	New	4.65553
	1	July	New	4.417949
Control group	27	January	Old	2.21451
	27	February	Old	2.097121
	27	March	Old	3.120116
	27	April	Old	2.13712
	27	May	Old	2.962273
	27	June	Old	3.781428
	27	July	Old	2.760494

Annotations:

- Green arrow (Jan-Mar): old procedure in effect
- Blue arrow (Apr-Jul): New procedure adopted
- Orange arrow (Apr-Jul): new procedure in effect

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID**
- Heterogeneous DID

Using xtdidregress

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID
Heterogeneous DID

```
. xtdidregress (satis) (procedure), group(hospital) time(month)
```

Treatment and time information

Time variable: month

Control: procedure = 0

Treatment: procedure = 1

	Control	Treatment
Group		
hospital	28	18
Time		
Minimum	1	4
Maximum	1	4

Difference-in-differences regression

Number of obs = 7,368

Data type: Longitudinal

(Std. err. adjusted for 46 clusters in hospital)

satis	Robust					
	Coefficient	std. err.	t	P> t	[95% conf. interval]	
ATET						
procedure						
(New vs Old)	.8479879	.0320138	26.49	0.000	.7835088	.9124669

Note: ATET estimate adjusted for panel effects and time effects.

† The previous `xtdidregress` specification is equivalent to

```
xtreg satis i.procedure, fe absorb(month) vce(cluster hospital)
```

† The `fe` option absorbs the hospital fixed effects.

† The `absorb(month)` option absorbs the month (time) fixed effects.

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † `aequations` – display auxiliary-equation results;
- † `nointeract` – exclude `group()` and `time()` interactions (useful when multiple group variables, like DDD models);
- † `nogteffects` – do not include group and time effects in the model

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† Useful postestimation commands available after `didregress` and `xtdidregress`:

† `estat trendplots`: graphical diagnostics for parallel trends

† `estat ptrends`: parallel-trends test

† `estat granger`: Granger causality test

† `estat grangerplot`: time-specific treatment effects

Parallel trends assumption

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

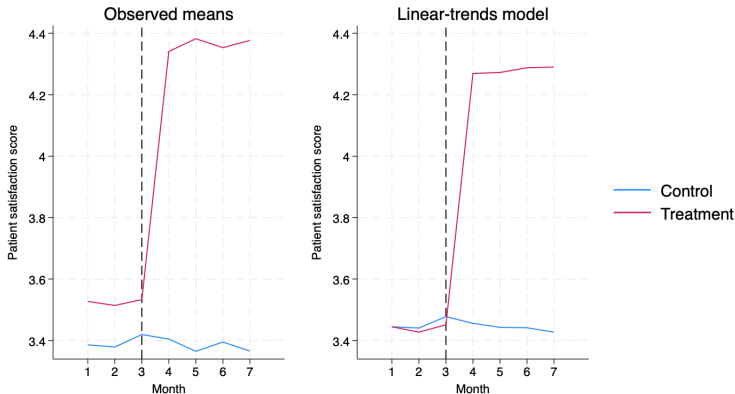
Heterogeneous DID

- † The ATET is identified only if the parallel trends assumption holds.
- † This means treatment and control groups would have followed the same trends in the absence of the intervention.
- † The assumption cannot be directly tested because the counterfactual is unknown.
- † A common indirect check is to compare pre-treatment trends between groups.
- † If pre-treatment trends are similar, it is plausible that they would have remained parallel without the intervention.

Graphical representation - parallel-trends

```
. estat trendplots
```

Graphical diagnostics for parallel trends



Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID**
- Heterogeneous DID

- † The parallel-trends assumption can be tested using a linear-trends model

```
. estat ptrends  
  
Parallel-trends test (pretreatment time period)  
H0: Linear trends are parallel  
  
F(1, 45) = 0.56  
Prob > F = 0.4602
```

- † Use the verbose option to display the underlying linear-trends specification

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID**
- Heterogeneous DID

Linear-trends specification

- † The underlying linear-trends model extends our original DID specification:

$$y_{ist} = \gamma_s + \gamma_t + \mathbf{z}_{ist}\beta + D_{st}\delta + tw_id_{t,0}\zeta_1 + tw_id_{t,1}\zeta_2 + \varepsilon_{ist}$$

- ◇ $d_{t,0} = 1(d_t = 0)$ indicates pretreatment periods
- ◇ $d_{t,1} = 1(d_t = 1)$ indicates posttreatment periods
- ◇ $w_i = 1$ for ever-treated units and 0 for never-treated units
- ◇ ζ_1 captures differences in slopes between treated and control groups in **pretreatment** periods
- ◇ ζ_2 captures differences in slopes between treated and control groups in **posttreatment** periods

- † Parallel trends in **pretreatment** periods require $\zeta_1 = 0$
- ◇ `estat ptrends` tests this restriction using a Wald test

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- ◇ Estimate treatment effects using panel data or repeated cross-sections
- ◇ Treatment adoption may occur at different points in time
- ◇ Staggered treatment assignment: once treated, always treated
- † **Groups (cohorts):** units that begin treatment in the same period; different groups adopt treatment at different times
- † **Objective:** estimate treatment effects for treated groups (cohorts), allowing for heterogeneity across cohorts and over time

AKC registrations and movies

† **Treatment:** dogs starring in a movie as protagonist (movie)

† **Outcome:** AKC registrations (registered)

```
. use akc, clear
(Fictional dog breed and AKC registration data)
. describe

Contains data from akc.dta
Observations:      1,410      Fictional dog breed and AKC
                               registration data
Variables:          5         24 Jul 2026 11:24
```

Variable name	Storage type	Display format	Value label	Variable label
year	int	%10.0g		Year
breed	int	%34.0g	Breed	Dog breed
movie	byte	%9.0g		Was a movie protagonist
best	byte	%9.0g		Won best in show in past 10 years
registered	int	%9.0g		Number of AKC registrations

```
Sorted by: breed
. xtset breed year

Panel variable: breed (strongly balanced)
Time variable: year, 2031 to 2040
Delta: 1 unit
```

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Which dogs are treated when?

breed	year	movie
Akita	2031	0
Akita	2032	0
Akita	2033	0
Akita	2034	0
Akita	2035	0
Akita	2036	0
Akita	2037	0
Akita	2038	0
Akita	2039	0
Akita	2040	0

} never treated

breed	year	movie
Bulldog	2031	0
Bulldog	2032	0
Bulldog	2033	0
Bulldog	2034	0
Bulldog	2035	0
Bulldog	2036	1
Bulldog	2037	1
Bulldog	2038	1
Bulldog	2039	1
Bulldog	2040	1

← 2036 cohort

breed	year	movie
Boxer	2031	0
Boxer	2032	0
Boxer	2033	0
Boxer	2034	1
Boxer	2035	1
Boxer	2036	1
Boxer	2037	1
Boxer	2038	1
Boxer	2039	1
Boxer	2040	1

← 2034 cohort

breed	year	movie
Pointer	2031	0
Pointer	2032	0
Pointer	2033	0
Pointer	2034	0
Pointer	2035	0
Pointer	2036	0
Pointer	2037	1
Pointer	2038	1
Pointer	2039	1
Pointer	2040	1

← 2037 cohort

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID**

† Estimation

- ◇ `xthdidregress` for panel data
- ◇ Four estimators: RA, IPW, and AIPW (Callaway and Sant'Anna, 2021); TWFE (Wooldridge, 2021)

† Postestimation

- ◇ `estat atetplot`: visualize ATETs
- ◇ `estat aggregation`: aggregate ATETs across dimensions
- ◇ `estat ptrends`: pre-treatment parallel-trends tests
- ◇ `estat sci`: simultaneous CIs for RA, IPW, and AIPW

Note: Introduced in Stata 18

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † We estimate the standard two-way fixed-effects (TWFE) model:

$$y_{ist} = \gamma_s + \gamma_t + \mathbf{z}_{ist}\beta + D_{st}\delta + \epsilon_{ist}$$

where δ is the average treatment effect on the treated (ATET) and TWFE is a weighted average of “good” and “bad” comparisons (Goodman-Bacon 2021):

- † Newly treated vs control group (good)
- † Newly treated vs not-yet-treated group (good)
- † Newly treated vs already treated group (bad)

Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

Next step: decomposing TWFE

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

† Decompose TWFE treatment effects (Goodman-Bacon 2021)

◇ `estat bdecomp` after `didregress` or `xtdidregress`

† Wooldridge (2021) extended TWFE

† Callaway & Sant'Anna (2021) approach will be discussed afterward:

◇ Estimators: RA, IPW, and AIPW

Bacon decomposition

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

```
. quietly xtdidregress (registered) (movie), group(breed) time(year)
```

```
. estat bdecomp, summaryonly
```

DID treatment-effect decomposition

```
ATET = 2129.655                                Number of obs    = 1,410
                                                Number of groups  = 141
                                                Number of cohorts = 4
```

ATET decomposition summary	ATET component	Weight
Treated vs never treated	2166.021	0.970810
Treated earlier vs later	936.70729	0.013157
Treated later vs earlier	906.60588	0.016033

Note: Number of cohorts includes never treated.

Note: The ATET reported by xtdidregress is a weighted average of the ATET components. If any component is substantially different from the ATET reported by xtdidregress and the weight is large, consider accounting for treatment-effect heterogeneity by using xthdidregress.

Note: Introduced in Stata 18

Next step: extended TWFE

- † Wooldridge (2021) shows that fixed-effects estimation can accommodate heterogeneous treatment effects by extending TWFE.

$$Y_{it} = \gamma_g + \gamma_t + \sum_{g \in G} \sum_{s=g}^T \delta_{gs} \mathbb{1}(G_i = g, t = s) + \epsilon_{ist}$$

- † γ_g and γ_t are group/cohort and time fixed effects
- † $\delta_{gs} \equiv ATET(g, s)$
- † With covariates z , we add full interactions with γ_g , γ_t , and $\mathbb{1}(G_i = g, t = s)$

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Extended TWFE: estimation and implementation

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † The estimator uses the full sample and does not partition the data
- † Treatment effects vary by cohort and time within a unified regression
- † If covariates are included, they enter fully interacted
- † This corresponds to an extended two-way fixed-effects estimator
- ‡ Implemented in Stata as:

```
xthdidregress twfe (y x1 x2) (treat), group()
```

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

```
. xthdidregress twfe (registered) (movie), group(breed)
note: variable _did_cohort, containing cohort indicators formed by treatment variable
      movie and group variable breed, was added to the dataset using the estimation
      sample.
```

Treatment and time information

Time variable: year

Time interval: 2031 to 2040

Control: _did_cohort = 0

Treatment: _did_cohort > 0

	_did_cohort
Number of cohorts	4
Number of obs	
Never treated	1190
2034	40
2036	30
2037	150

- TWFE estimation output (1 of 2)
- Estimation setup and model specification
- A $_did_cohort$ (G_{ig}) variable is created

Wooldridge (2021) – TWFE

Heterogeneous-treatment-effects regression

Number of obs = 1,410

Number of panels = 141

Estimator: Two-way fixed effects

Panel variable: breed

Treatment level: breed

Control group: Never treated

Heterogeneity: Cohort and time

Over: Cohort

(Std. err. adjusted for 141 clusters in breed)

- TWFE estimation output (2 of 2)
- Treatment effects by cohort and time

Cohort	Robust					
	ATET	std. err.	t	P> t	[95% conf. interval]	
2034						
year						
2034	502.2111	172.6038	2.91	0.004	160.9641	843.458
2035	847.1034	200.9662	4.22	0.000	449.7825	1244.424
2036	1114.676	169.8532	6.56	0.000	778.8675	1450.485
2037	1756.973	284.131	6.18	0.000	1195.231	2318.715
2038	2269.362	173.2464	13.10	0.000	1926.844	2611.879
2039	2399.868	547.9829	4.38	0.000	1316.477	3483.26
2040	2824.66	519.5897	5.44	0.000	1797.403	3851.917
2036						
year						
2036	1328.58	89.55409	14.84	0.000	1151.527	1505.633
2037	1333.46	282.0488	4.73	0.000	775.8347	1891.086
2038	1956.766	334.0546	5.86	0.000	1296.322	2617.21
2039	2864.855	575.904	4.97	0.000	1726.262	4003.449
2040	2504.897	584.4678	4.29	0.000	1349.373	3660.421
2037						
year						
2037	1752.219	214.2624	8.18	0.000	1328.611	2175.828
2038	2104.325	214.8665	9.79	0.000	1679.522	2529.127
2039	2522.948	301.1487	8.38	0.000	1927.56	3118.335
2040	2883.19	310.2944	9.29	0.000	2269.721	3496.658

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Other heterogeneity types

- † By default, `xthdidregress` estimates heterogeneous treatment effects by cohort and time using the option `hettype(timecohort)`.
- † Alternatively, we can restrict heterogeneity to a single dimension:
 - ◇ `hettype(time)`: ATETs vary only over time
 - ◇ `hettype(cohort)`: ATETs vary only across cohorts

```
xthdidregress twfe (registered) (movie), group(breed)
hettype(time)
```

```
xthdidregress twfe (registered) (movie), group(breed)
hettype(cohort)
```

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Beyond TWFE: three estimators

Callaway and Sant'Anna (2021)

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

† Regression adjustment (RA)

```
xthdidregress ra (y [x]) (treat), group()
```

† Inverse-probability weighting (IPW)

```
xthdidregress ipw (y) (treat [x]), group()
```

† Augmented inverse-probability weighting (AIPW)

```
xthdidregress aipw (y [x]) (treat [x]), group()
```

Regression adjustment (RA)

- † **Idea:** Adjust outcome changes using a regression model estimated on a valid control group.

$$ATET(g, t) = E \left\{ \frac{G_g}{E(G_g)} [Y_t - Y_{g-1} - m_{gt}(X)] \right\}$$

- ◇ $Y_t - Y_{g-1} \leftarrow$ **treated differences**
- ◇ $m_{gt}(X) = E(Y_t - Y_{g-1} \mid X, C = 1) \leftarrow$ **untreated differences**
- ◇ $C = 1$ denotes the never-treated group ($G_g = \infty$)
- ◇ Outcome dynamics are learned from untreated units and applied to cohort g

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Inverse probability weighting (IPW)

† **Idea:** Reweight outcome changes so treated and control units are comparable.

$$ATE(g, t) = E \left\{ \left(\frac{G_g}{E(G_g)} - \frac{\frac{p_g(Z)}{1-p_g(Z)}}{E \left[\frac{p_g(Z)}{1-p_g(Z)} \right]} \right) [Y_t - Y_{g-1}] \right\}$$

- ◇ $p_g(X) = P(G_g = 1 \mid Z, G_g + C = 1)$
- ◇ Propensity scores are estimated within the retained sample
- ◇ Implementation mirrors RA, replacing outcome modeling with reweighting

Augmented inverse probability weighting (AIPW)

- † **Idea:** Combine outcome modeling and reweighting for robustness.

$$ATET(g, t) = E \left\{ \left(\frac{G_g}{E(G_g)} - \frac{\frac{p_g(Z)}{1-p_g(Z)}}{E \left[\frac{p_g(Z)}{1-p_g(Z)} \right]} \right) [Y_t - Y_{g-1} - m_{gt}(X)] \right\}$$

- ◇ $m_{gt}(X)$ is the **augmented term** in addition to IPW
- ◇ X and Z may differ across the propensity score and outcome models
- ◇ Consistent estimation requires correct specification of either:
 - ▶ $p_g(Z)$ or $m_{gt}(X)$

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

- † $ATE(g, t)$ is constructed using only two groups:
 - ◇ Treated cohort g
 - ◇ Never-treated units ($C = 1$)
- † Outcomes are compared at two points in time: $g - 1$ and t
- † This forms a sequence of 2×2 comparisons
- † The procedure is repeated for all cohorts g and post-treatment periods t
- † Alternative comparisons (e.g., not-yet-treated units) are also possible
- † Identification assumptions must hold for each 2×2 comparison

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

1. Keep observations with time t or $g - 1$
2. Keep units in cohort g or never treated ($C = 1$)
3. Run parametric model
 - a. Construct outcome differences: $\Delta Y = Y_t - Y_{g-1}$
 - b. Regress ΔY on X using $C = 1$ observations and predict $\hat{m}_{gt}(X)$
 - c. Run a logit regression of G_g on Z and predict $\hat{p}_g(Z)$
4. Plug in the nuisance functions $\hat{m}_{gt}(X)$ and $\hat{p}_g(Z)$, or both into RA, IPW, or AIPW.
5. Repeat for all (g, t) pairs

RA, IPW, and AIPW: comparison

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † RA: model outcomes
 - † IPW: model treatment assignment
 - † AIPW: model both (**doubly robust**)
-
- ‡ All rely on valid 2×2 comparisons
 - ‡ All estimate $ATET(g, t)$ under parallel trends

An example using RA

```
. xthdidregress ra (registered best) (movie), group(breed)
note: variable _did_cohort, containing cohort indicators formed by treatment variable
      movie and group variable breed, was added to the dataset using the estimation
      sample.
```

Computing ATET for each cohort and time:

Cohort 2034 (9): done

Cohort 2036 (9): done

Cohort 2037 (9): done

Treatment and time information

Time variable: year

Time interval: 2031 to 2040

Control: _did_cohort = 0

Treatment: _did_cohort > 0

	_did_cohort
Number of cohorts	4
Number of obs	
Never treated	1190
2034	40
2036	30
2037	150

- RE estimation output (1 of 2)
- ATETs computed over cohort and time

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

An example using RA

```
Heterogeneous-treatment-effects regression      Number of obs   = 1,410
                                                Number of panels = 141

Estimator:      Regression adjustment
Panel variable:  breed
Treatment level: breed
Control group:   Never treated
                (Std. err. adjusted for 141 clusters in breed)
```

Cohort		Robust		z	P> z	[95% conf. interval]	
		ATET	std. err.				
2034	year						
	2032	-254.8927	266.1024	-0.96	0.338	-776.4439	266.6584
	2033	-257.5329	217.9389	-1.18	0.237	-684.6852	169.6194
	2034	701.1318	127.0935	5.52	0.000	452.0331	950.2304
	2035	1099.044	282.0704	3.90	0.000	546.196	1651.892
	2036	1367.632	225.8702	6.05	0.000	924.9343	1810.329
	2037	2008.294	237.2396	8.47	0.000	1543.313	2473.275
	2038	2472.624	278.2949	8.88	0.000	1927.176	3018.072
	2039	2689.615	504.3324	5.33	0.000	1701.142	3678.088
	2040	3110.97	568.916	5.47	0.000	1995.915	4226.025
2036	year						
	2032	216.0259	122.9107	1.76	0.079	-24.87472	456.9265
	2033	-172.5154	372.0776	-0.46	0.643	-901.7741	556.7433
	2034	-218.0495	504.5267	-0.43	0.666	-1206.904	770.8045
	2035	621.033	156.1306	3.98	0.000	315.0227	927.0434
	2036	999.0781	180.1055	5.55	0.000	646.0779	1352.078
	2037	1003.333	250.5916	4.00	0.000	512.1829	1494.484
	2038	1556.669	451.6914	3.45	0.001	671.3697	2441.967
	2039	2590.674	662.6979	3.91	0.000	1291.81	3889.538
	2040	2225.712	486.9917	4.57	0.000	1271.225	3180.198
2037	year						
	2032	-114.582	160.0972	-0.72	0.474	-428.3668	199.2028
	2033	-127.9856	183.3941	-0.70	0.485	-487.4315	231.4603
	2034	33.40901	168.0312	0.20	0.842	-295.9262	362.7442
	2035	130.3496	166.2261	0.78	0.433	-195.4477	456.1468
	2036	-10.48288	167.5059	-0.06	0.950	-338.7884	317.8226
	2037	1717.016	268.5692	6.39	0.000	1190.65	2243.383
	2038	2086.798	278.0215	7.51	0.000	1541.886	2631.71
	2039	2473.611	268.186	9.22	0.000	1947.976	2999.246
	2040	2835.117	378.6699	7.49	0.000	2092.938	3577.296

Note: ATET computed using covariates.

- RA estimation output (2 of 2)
- Pretreatment ATETs also computed

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

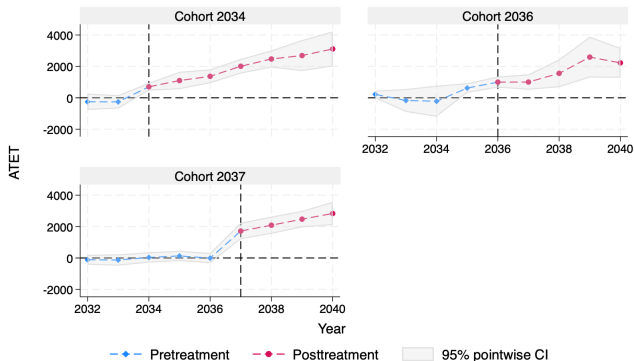
Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

Postestimation: visualizing ATETs

† Estimated ATETs can be visualized using the postestimation command

```
. estat atetplot
```



Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID**

Summarize ATET over time and/or cohort

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

- † How do the ATETs vary with the length of exposure to the treatment? (event study)
- † How do the ATETs vary with cohorts? (does starting treatment earlier matter?)
- † How do the ATETs vary with time? (Good year vs. lousy year)
- † Overall ATETs across time and cohorts

We can express the aggregations as a weighted mean of all ATETs

$$\theta = \sum_{g \in G} \sum_{t=2}^T \underbrace{w(g, t)}_{\text{weight}} ATET(g, t)$$

Types of aggregation

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

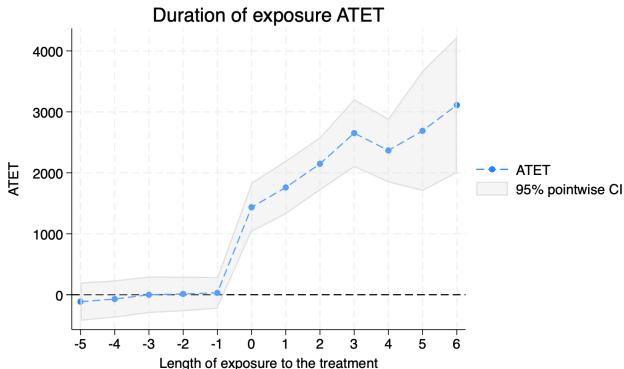
- † **Overall aggregation**: aggregate all the treatment effects across cohorts and posttreatment periods.
- † **Dynamic aggregation**: aggregate the treatment effects given a fixed length of exposure to the treatment. Commonly referred to as event studies.
- † **Cohort aggregation**: aggregate the treatment effects within a specific cohort across the posttreatment periods.
- † **Time aggregation**: aggregate the treatment effects at a specific time across the cohorts.

Dynamic aggregation (event-study plot)

† Let $e = t - g$ be the length of exposure to the treatment

$$\theta(e) = \sum_{g \in G} \underbrace{1(g + e \leq T) Pr(G = g | g + e \leq T)}_{\text{proportions used to estimate ATET}(g, g+e)} ATET(g, g + e)$$

```
. estat aggregation, dynamic graph
```



Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID

Introduction

Panel-data linear models

DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID

† Postestimation

- ◇ `estat atetplot`: visualize ATETs
- ◇ `estat aggregation`: aggregate ATETs across dimensions
 - ▶ `sci` option to compute and graph simultaneous CIs
- ◇ `estat ptrends`: pre-treatment parallel-trends tests
- ◇ `estat sci`: simultaneous CIs for RA, IPW, and AIPW

Approaches to heterogeneous treatment

† Solution 1: Callaway and Sant'Anna (2021)

- ◇ Construct valid comparison groups
- ◇ Decompose the problem into multiple 2×2 comparisons
- ◇ Estimate $ATET(g, t)$, where g denotes cohort and t denotes time

† Solution 2: Wooldridge (2021)

- ◇ Include treatment effect heterogeneity directly in the regression
- ◇ Introduce appropriate interactions between cohort and time
- ◇ Estimate $ATET(g, t)$; aggregation to $ATET(g)$ or $ATET(t)$ is also possible

† Other approaches:

- ◇ Good surveys include: Roth et al. (2022) and de Chaisemartin and D'Haultfœuille (2023)

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

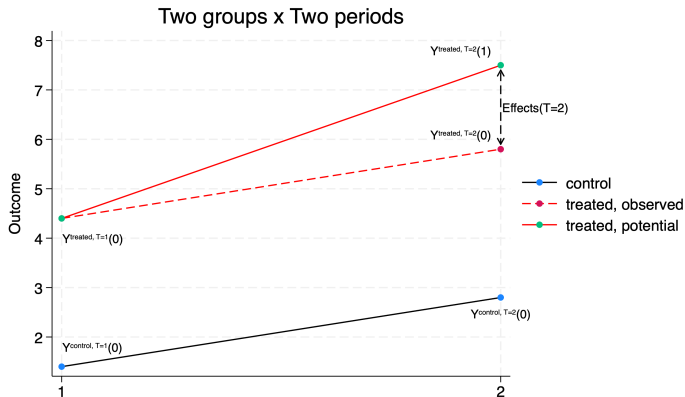
Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Appendix – Canonical DID



Introduction

Panel-data linear models

- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID**

Appendix – Homogeneous effects across time

Introduction

Panel-data linear models

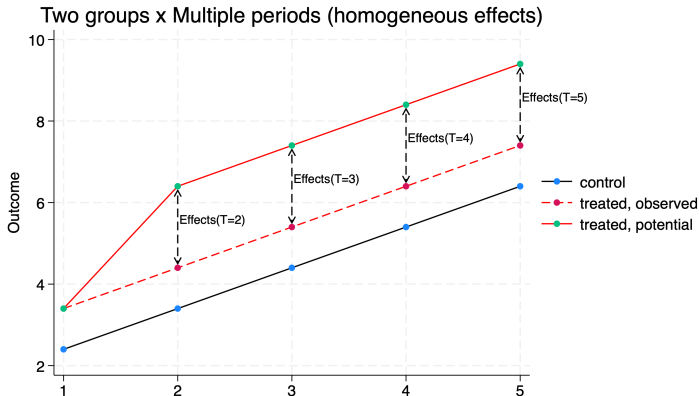
- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID**



Appendix – Heterogeneous effects across time

Introduction

Panel-data linear models

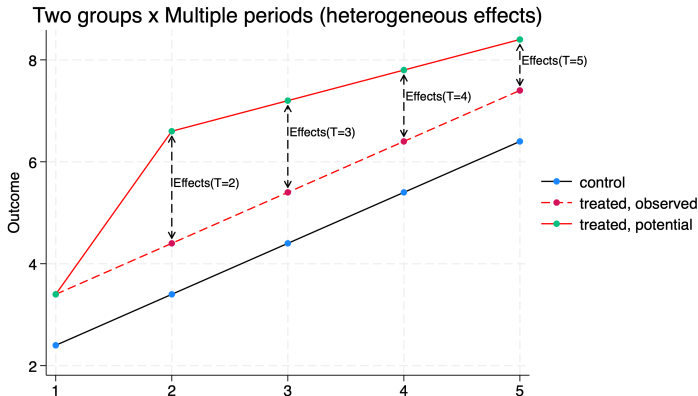
- DGP
- Random effects
- Fixed effects
- FE vs RE
- Correlated random effects
- High-dimensional fixed effects

Panel-data nonlinear models

- Binary outcome
- Other nonlinear models

Difference-in-differences models

- Motivation
- Homogeneous DID
- Heterogeneous DID**



Appendix – Homogeneous effects across groups

Introduction

Panel-data linear models

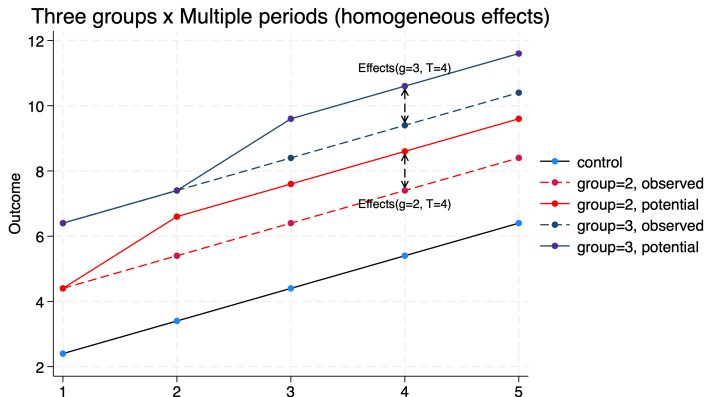
DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID



Appendix – Heterogeneous effects across groups and time

Introduction

Panel-data linear models

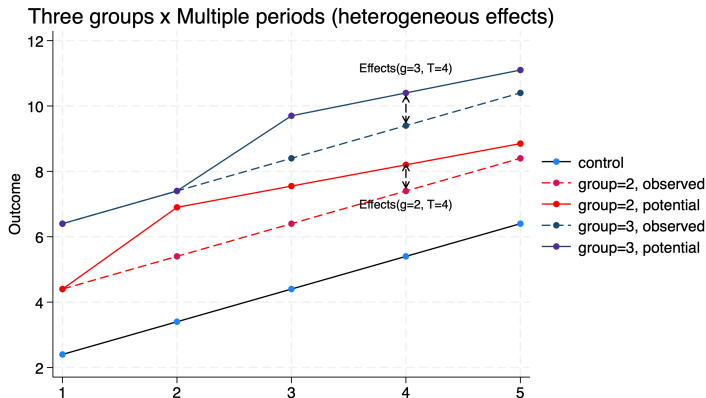
DGP
Random effects
Fixed effects
FE vs RE
Correlated random effects
High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome
Other nonlinear models

Difference-in-differences models

Motivation
Homogeneous DID
Heterogeneous DID



[Back to Main](#)

Reference

- † Baltagi, B. H. 2009. *A Companion to Econometric Analysis of Panel Data*. Chichester, UK: Wiley.
- † Callaway, B., and P. H. C. Sant'Anna. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics* 225: 200–230.
<https://doi.org/10.1016/j.jeconom.2020.12.001>.
- † de Chaisemartin, C., and X. D'Haultfœuille. 2023. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal* 26: C1–C30. <https://doi.org/10.1093/ectj/utac017>.
- † Goodman-Bacon, A. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* 225: 254–277.
<https://doi.org/10.1016/j.jeconom.2021.03.014>.
- † Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe. 2022. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. arXiv:2201.01194 [econ.EM], <https://doi.org/10.48550/arXiv.2201.01194>.
- † Wooldridge, J. M. 2019. Correlated random effects models with unbalanced panels. *Journal of Econometrics* 211: 137–150.
<https://doi.org/10.1016/j.jeconom.2018.12.010>.
- † Wooldridge, J. M. 2020. *Introductory Econometrics: A Modern Approach*. 7th ed. Boston: Cengage.
- † Wooldridge, J. M. 2021. Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. Working paper, Department of Economics, Michigan State University, East Lansing, MI.
<https://doi.org/10.2139/ssrn.3906345>.

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Introduction

Panel-data linear models

DGP

Random effects

Fixed effects

FE vs RE

Correlated random effects

High-dimensional fixed effects

Panel-data nonlinear models

Binary outcome

Other nonlinear models

Difference-in-differences models

Motivation

Homogeneous DID

Heterogeneous DID

Thank you!