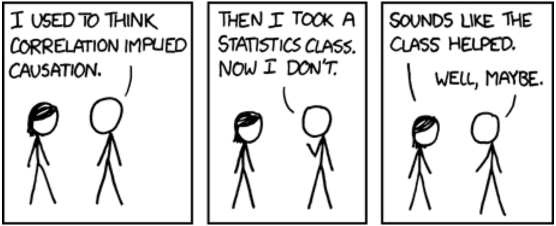
Advanced Applied Econometrics

Teacher: Felix Weinhardt



Topics covered

1. Review fixed effects regression models
2. Differences-in-differences basics: Card and Krueger (1994)
3. Regression differences-in-differences
4. Synthetic control: Abadie, Diamond and Hainmueller (2010)
5. Combining differences-in-differences with IV: Waldinger (2010)

Panel Methods

- Panel data: we observe the same units (individuals, firms, countries, schools, etc.) over several time periods
- Often our outcome variable depends on unobserved factors which are also correlated with our explanatory variable of interest
- If these omitted variables are constant over time, we can use panel data estimators to consistently estimate the effect of our explanatory variable
- Main estimators for panel data:
  - Pooled OLS
  - Fixed effects estimator
  - Random effects estimator

Panel setup

- Let y and $x \equiv (x_1, x_2, \ldots, x_k)$ be observable random variables and c be an unobservable random variable
- We are interested in the partial effects of variable x_j in the population regression function
$$E[y|x_1, x_2, \ldots, x_k, c]$$
- We observe a sample of $i = 1, 2, \ldots, N$ cross-sectional units for $t = 1, 2, \ldots, T$ time periods (a balanced panel)
    - For each unit i, we denote the observable variables for all time periods as $\{(y_{it}, x_{it}) : t = 1, 2, \ldots, T\}$
    - $x_{it} \equiv (x_{it1}, x_{it2}, \ldots, x_{itk})$ is a $1 \times K$ vector
- Typically assume that cross-sectional units are i.i.d. draws from the population: $\{y_i, x_i, c_i\}_{i=1}^N \sim i.i.d.$ (cross-sectional independence)
    - $y_i \equiv (y_{i1}, y_{i2}, \ldots, y_{iT})'$ and $x_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iT})$
    - Consider asymptotic properties with T fixed and $N \to \infty$

Panel setup

Single unit:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T\times 1} \qquad X_i = \begin{pmatrix} X_{i,1,1} & X_{i,1,2} & X_{i,1,j} & \dots & X_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,t,1} & X_{i,t,2} & X_{i,t,j} & \dots & X_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,T,1} & X_{i,T,2} & X_{i,T,j} & \dots & X_{i,T,K} \end{pmatrix}_{T\times K}$$

Panel with all units:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{pmatrix}_{NT\times 1} \qquad X = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{pmatrix}_{NT\times K}$$

Unobserved effects model: Farm output

- For a randomly drawn cross-sectional unit i, the model is given by

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \ \ t = 1, 2, \ldots, T$$

  - y_{it}: output of farm i in year t
  - x_{it} : $1 \times K$ vector of variable inputs for farm i in year t, such as labor, fertilizer, etc. plus an intercept
  - β : $K \times 1$ vector of marginal effects of variable inputs
  - c_i: sum of all time-invariant inputs known to farmer i (but unobserved for the researcher), e.g., soil quality, managerial ability, etc.
    - often called the unobserved effect, unobserved heterogeneity, etc
  - ε_{it}: time-varying unobserved inputs, such as rainfall, unknown to the farmer at the time the decision on the variable inputs x_{it} is made
    - often called the idiosyncratic error
- What happens when we regress y_{it} on x_{it}?

Pooled OLS

- When we ignore the panel structure and regress y_{it} on x_{it} we get

$$y_{it} = x_{it}\beta + v_{it};\ \ t = 1, 2, \ldots, T$$

  with composite error $v_{it} \equiv c_i + \varepsilon_{it}$
- Main assumption to obtain consistent estimates for β is:
  - $E[v_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}] = E[v_{it}|x_{it}] = 0$ for $t = 1, 2, \ldots, T$
    - x_{it} are strictly exogenous: the composite error v_{it} in each time period is uncorrelated with the past, current and future regressors
    - But: labour input x_{it} likely depends on soil quality c_i and so we have omitted variable bias and $\widehat{\beta}$ is not consistent
  - No correlation between x_{it} and v_{it} implies no correlation between unobserved effect c_i and x_{it} for all t
    - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)
  - Additional problem: v_{it} are serially correlated for same i since c_i is present in each t and thus pooled OLS standard errors are invalid

Unobserved effects model: program evaluation

- Program evaluation model:

$$y_{it} = prog_{it}\beta + c_i + \varepsilon_{it}; t = 1, 2, \ldots, T$$

  - y_{it}: log wage of individual i in year t
  - $prog_{it}$: indicator coded 1 if individual i participants in training program at t and 0 otherwise
  - β: effect of program
  - c_i: sum of all time-invariant unobserved characteristics that affect wages, such as ability, etc.
- What happens when we regress y_{it} on $prog_{it}$? $\widehat{\beta}$ not consistent since $prog_{it}$ is likely correlated with c_i (e.g., ability)
- Always ask: is there a time-constant unobserved variable (c_i) that is correlated with the regressors? If yes, then pooled OLS is problematic

Fixed effect regression

- Our unobserved effects model is:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it};\, t = 1, 2, \ldots, T$$

- If we have data on multiple time periods, we can think of c_i as **fixed effects** or "nuisance parameters" to be estimated
- OLS estimation with fixed effects yields

$$(\widehat{\beta}, \widehat{c}_1, \ldots, \widehat{c}_N) = \underset{b, m_1, \ldots, m_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it} b - m_i)^2$$

this amounts to including N farm dummies in regression of y_{it} on x_{it}

Derivation: fixed effects regression

$$(\widehat{\beta}, \widehat{c}_1, \ldots, \widehat{c}_N) = \underset{b, m_1, \ldots, m_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it} b - m_i)^2$$

The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}'(y_{it} - x_{it}\widehat{\beta} - \widehat{c}_i) = 0$$

and

$$\sum_{t=1}^{T} (y_{it} - x_{it}\widehat{\beta} - \widehat{c}_i) = 0$$

for $i = 1, \ldots, N$.

Derivation: fixed effects regression

Therefore, for $i = 1, \dots, N$,

$$\widehat{c}_i = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - x_{it}\widehat{\beta}) = \bar{y}_i - \bar{x}_i\widehat{\beta},$$

where

$$\bar{x}_i \equiv \frac{1}{T}\sum_{t=1}^{T} x_{it}; \bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

Plug this result into the first FOC to obtain:

$$\widehat{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i)\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)'(y_{it} - \bar{y})\right)$$

$$\widehat{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}_{it}'\ddot{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}_{it}'\ddot{x}_{it}\right)$$

with time-demeaned variables $\ddot{x}_{it} \equiv x_{it} - \bar{x}, \ddot{y}_{it} \equiv y_{it} - \bar{y}_i$

Fixed effects regression

Running a regression with the time-demeaned variables $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ and $\ddot{x}_{it} \equiv x_{it} - \bar{x}$ is numerically equivalent to a regression of y_{it} on x_{it} and unit specific dummy variables.

Even better, the regression with the time demeaned variables is consistent for β even when $Cov[x_{it}, c_i] \neq 0$ because time-demeaning eliminates the unobserved effects

$$
\begin{array}{rcl}
y_{it} & = & x_{it}\beta + c_i + \varepsilon_{it} \\
\bar{y}_i & = & \bar{x}_i\beta + c_i + \bar{\varepsilon}_i \\
\hline
(y_{it} - \bar{y}_i) & = & (x_{it} - \bar{x})\beta + (c_i - c_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \\
\ddot{y}_{it} & = & \ddot{x}_{it}\beta + \ddot{\varepsilon}_{it}
\end{array}
$$

Fixed effects regression: main results

- Identification assumptions:
  1. $E[\varepsilon_{it}|x_{i1}, x + i2, \ldots, x_{iT}, c_i] = 0; t = 1, 2, \ldots, T$
     - regressors are strictly exogenous conditional on the unobserved effect
     - allows x_{it} to be arbitrarily related to c_i
  2. $rank\left(\sum_{t=1}^{T} E[\ddot{x}_{it}' \ddot{x}_{it}] \right) = K$
     - regressors vary over time for at least some i and not collinear
- Fixed effects estimator
  1. Demean and regress $\ddot{y}_{it}$ on $\ddot{x}_{it}$ (need to correct degrees of freedom)
  2. Regress y_{it} on x_{it} and unit dummies (dummy variable regression)
  3. Regress y_{it} on x_{it} with canned fixed effects routine
     - STATA: `xtreg y x, fe i(PanelID)`
- Properties (under assumptions 1-2):
  - $\widehat{\beta}_{FE}$ is consistent: $\underset{N\to\infty}{plim}\, \widehat{\beta}_{FE,N} = \beta$
  - $\widehat{\beta}_{FE}$ is unbiased conditional on $\mathbf{X}$

Fixed effects regression: main issues

- Inference:
  - Standard errors have to be "clustered" by panel unit (e.g., farm) to allow correlation in the ε_{it}'s for the same i.
    - STATA: `xtreg , fe i(PanelID) cluster( PanelID )`
  - Yields valid inference as long as number of clusters is reasonably large
- Typically we care about β, but unit fixed effects c_i could be of interest
  - $\widehat{c}_i$ from dummy variable regression is unbiased but not consistent for c_i (based on fixed T and $N \to \infty$)
  - `xtreg , fe` routine demeans the data before running the regression and therefore does not estimate $\widehat{c}_i$
    - intercept shows average $\widehat{c}_i$ across units
    - we can recover $\widehat{c}_i$ using $\widehat{c}_i = \bar{y}_i - \bar{x}_i\widehat{\beta}$
    - `predict c_i, u`

Example: Direct Democracy and Naturalizations

- Do minorities fare worse under direct democracy than under representative democracy?
- Hainmueller and Hangartner (2012) examine data on naturalization requests of immigrants in Switzerland, where municipalities vote on naturalization applications in:
  - referendums (direct democracy)
  - elected municipality councils (representative democracy)
- Annual panel data from 1,400 municipalities for the 1991-2009 period
  - y_{it} : naturalization rate $= \frac{no.naturalizations_{it}}{eligibleforeignpopulation_{i,t-1}}$
  - x_{it} : 1 if municipality used representative democracy, 0 if municipality used direct democracy in year t

Naturalization Panel Data

```
. des muniID muni_name year nat_rate repdem

              storage   display     value
variable name   type    format      label      variable label
-------------------------------------------------------------------------------------
muniID          float   %8.0g                  municipality code
muni_name       str43   %43s                   municipality name
year            float   %ty                    year
nat_rate        float   %9.0g                  naturalization rate (percent)
repdem          float   %9.0g                  1 representative democracy, 0 direct
```

Panel Data Long Format

```
. list muniID muni_name year nat_rate repdem in 31/40
```

	muniID	muni_name	year	nat_rate	repdem
31.	2	Affoltern A.A.	2002	4.638365	0
32.	2	Affoltern A.A.	2003	4.844814	0
33.	2	Affoltern A.A.	2004	5.621302	0
34.	2	Affoltern A.A.	2005	4.387827	0
35.	2	Affoltern A.A.	2006	8.115358	1
36.	2	Affoltern A.A.	2007	7.067371	1
37.	2	Affoltern A.A.	2008	8.977719	1
38.	2	Affoltern A.A.	2009	6.119704	1
39.	3	Bonstetten	1991	.8333334	0
40.	3	Bonstetten	1992	.8403362	0

Pooled OLS

```
. reg nat_rate repdem , cl(muniID)

Linear regression                                      Number of obs =    4655
                                                       F(  1,   244) =  130.04
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0748
                                                       Root MSE      =    3.98

                                (Std. Err. adjusted for 245 clusters in muniID)
```

nat_rate	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
repdem	2.503318	.2195202	11.40	0.000	2.070921	2.935714
_cons	2.222683	.10088	22.03	0.000	2.023976	2.421389

Decompose within and between variation

```
. tsset  muniID year ,  yearly
       panel variable:  muniID (strongly balanced)
        time variable:  year, 1991 to 2009
                delta:  1 year

. xtsum  nat_rate
```

Variable		Mean	Std. Dev.	Min	Max	Observations
nat_rate	overall	2.938992	4.137305	0	24.13793	N = 4655
	between		1.622939	0	7.567746	n = 245
	within		3.807039	-3.711323	24.80134	T = 19

Time-demeaning for fixed effects: $y_{it} \rightarrow \ddot{y}_{it}$

```
. * get municipality means
. egen means_nat_rate = mean(nat_rate) , by(muniID)

. * compute deviations from means
. gen  dm_nat_rate = nat_rate - means_nat_rate

. list muniID muni_name year nat_rate means_nat_rate dm_nat_rate in 20/40 ,ab(20)
```

	muniID	muni_name	year	nat_rate	means_nat_rate	dm_nat_rate
20.	2	Affoltern A.A.	1991	.2173913	3.595932	-3.37854
21.	2	Affoltern A.A.	1992	.9473684	3.595932	-2.648563
22.	2	Affoltern A.A.	1993	1.04712	3.595932	-2.548811
23.	2	Affoltern A.A.	1994	.8342023	3.595932	-2.761729
24.	2	Affoltern A.A.	1995	2.002002	3.595932	-1.59393
25.	2	Affoltern A.A.	1996	1.7769	3.595932	-1.819031
26.	2	Affoltern A.A.	1997	1.862745	3.595932	-1.733186
27.	2	Affoltern A.A.	1998	2.054155	3.595932	-1.541776
28.	2	Affoltern A.A.	1999	2.402135	3.595932	-1.193796

Fixed effects regression with demeaned data

```
. egen means_repdem   = mean(repdem) , by(muniID)

. gen  dm_repdem   = repdem   - means_repdem

.
. * regression with demeaned data
. reg dm_nat_rate dm_repdem , cl(muniID)

Linear regression                                      Number of obs =    4655
                                                       F(  1,   244) =  265.18
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1052
                                                       Root MSE      =  3.6017

                                (Std. Err. adjusted for 245 clusters in muniID)
------------------------------------------------------------------------------
             |               Robust
 dm_nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   dm_repdem |     3.0228   .1856244    16.28   0.000     2.657169    3.388431
       _cons |   6.65e-10   5.81e-09     0.11   0.909    -1.08e-08    1.21e-08
------------------------------------------------------------------------------
```

Fixed effects regression with canned routine

```
. xtreg nat_rate repdem , fe cl(muniID) i(muniID)

Fixed-effects (within) regression               Number of obs      =      4655
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.1052                         Obs per group: min =        19
       between = 0.0005                                        avg =      19.0
       overall = 0.0748                                        max =        19

                                                F(1,244)           =    265.18
corr(u_i, Xb)  = -0.1373                        Prob > F           =    0.0000

                              (Std. Err. adjusted for 245 clusters in muniID)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      repdem |     3.0228   .1856244    16.28   0.000     2.657169    3.388431
       _cons |   2.074036   .0531153    39.05   0.000     1.969413    2.178659
-------------+----------------------------------------------------------------
     sigma_u |  1.7129711
     sigma_e |    3.69998
         rho |  .17650677   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Fixed effects regression with dummies

```
. reg nat_rate repdem i.muniID, cl(muniID)

Linear regression                               Number of obs =     4655
                                                F(  0,    244) =        .
                                                Prob > F      =        .
                                                R-squared     =   0.2423
                                                Root MSE      =      3.7

                              (Std. Err. adjusted for 245 clusters in muniID)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      repdem |     3.0228   .1906916    15.85   0.000     2.647188    3.398412
             |
      muniID |
          2  |   1.367365   5.17e-14  2.6e+13   0.000     1.367365    1.367365
          3  |   1.292252   5.17e-14  2.5e+13   0.000     1.292252    1.292252
          9  |   1.284652   5.17e-14  2.5e+13   0.000     1.284652    1.284652
         10  |   1.271783   5.17e-14  2.5e+13   0.000     1.271783    1.271783
         13  |   .3265469   5.17e-14  6.3e+12   0.000     .3265469    .3265469
```

Applying fixed effects

- We can use fixed effects for other data structures to restrict comparisons to within unit variation
  - Matched pairs
    - Twin fixed effects to control for unobserved effects of family background
  - Cluster fixed effects in hierarchical data
    - School fixed effects to control for unobserved effects of school

Problems that even fixed effects do not solve

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it};\ t = 1, 2, \ldots, T$$

- Where y_{it} is murder rate and x_{it} is police spending per capita
- What happens when we regress y on x and city fixed effects?
  - $\widehat{\beta}_{FE}$ inconsistent unless strict exogeneity conditional on c_i holds
    - $E[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}, c_i] = 0; t = 1, 2, \ldots, T$
    - implies ε_{it} uncorrelated with past, current and future regressors
- Most common violations
  1. Time-varying omitted variables
     - Economic boom leads to more police spending and less murders
     - Can include time-varying controls, but avoid post-treatment bias (i.e., collider)
  2. Simultaneity
     - if city adjusts police based on past murder rate, then spending$_{t+1}$ is correlated with ε_t (since higher ε_t leads to higher murder rate at t)
     - strictly exogenous x cannot react to what happens to y in the past or the future!
- Fixed effects do not obviate need for good research design!

Random Effects

- Reconsider our unobserved effects model:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \quad t = 1, 2, \ldots, T$$

  - Cannot use the fixed effects regression to estimate effects of time-constant regressors in x_{it} (eg., soil quality, farm location, etc.)
    - Since fixed effect estimator allows c_i to be correlated with x_{it}, we cannot distinguish the effects of time-invariant regressors from the time-invariant unobserved effect c_i
  - Need orthogonality assumption: $Cov[x_{it}, c_i] = 0; \quad t = 1, \ldots, T$
    - Strong assumption: Unobserved effects c_i are uncorrelated with each explanatory variable in x_{it} in each time period
    - For example if we include soil quality in x_{it} we have to assume it is uncorrelated with all other time-invariant inputs

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \dots, T$$

1. $E[\varepsilon_{it}|x_i, c_i] = 0;\ \ t = 1, 2, \dots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
3. rank $E[X_i'\Omega X_i] = K$: no collinearity among regressors
    - $\Omega = E[v_i v_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that Ω takes a special form:
    - $E[\varepsilon_i\varepsilon_i'|x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
    - $E[c_i^2|x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \ldots, T$$

1. $E[\varepsilon_{it}|x_i, c_i] = 0;\ t = 1, 2, \ldots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
3. rank $E[X_i'\Omega X_i] = K$: no collinearity among regressors
   - $\Omega = E[v_i v_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that Ω takes a special form:
   - $E[\varepsilon_i \varepsilon_i' | x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
   - $E[c_i^2|x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic

$$\text{Assumption 4 implies } \Omega = E[v_i v_i'|x_i] = \begin{pmatrix} \sigma_c^2 + \sigma_\varepsilon^2 & \sigma_c^2 & \ldots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_\varepsilon^2 & \ldots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2 + \sigma_\varepsilon^2 \end{pmatrix}_{T \times T}$$

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \ldots, T$$

1. $E[\varepsilon_{it}|x_i, c_i] = 0;\ \ t = 1, 2, \ldots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
3. rank $E[X_i'\Omega X_i] = K$: no collinearity among regressors
   - $\Omega = E[v_i v_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that Ω takes a special form:
   - $E[\varepsilon_i \varepsilon_i'|x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
   - $E[c_i^2|x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic

- Given assumptions 1-3, pooled OLS is consistent, since composite error v_{it} is uncorrelated with x_{it} for all t
- However, pooled OLS ignores the serial correlation in v_{it}

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \ldots, T$$

1. $E[\varepsilon_{it}|x_i, c_i] = 0$; $t = 1, 2, \ldots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
3. rank $E[X_i'\Omega X_i] = K$: no collinearity among regressors
   - $\Omega = E[v_i v_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that Ω takes a special form:
   - $E[\varepsilon_i \varepsilon_i' | x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
   - $E[c_i^2|x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic

- Random effects estimator $\widehat{\beta}_{RE}$ exploits this serial correlation in a generalized least squares (GLS) framework
  - $\widehat{\beta}_{RE}$ is consistent under assumption 1-3: $\underset{N\to\infty}{plim}\,\widehat{\beta}_{RE,N} = \beta$
  - $\widehat{\beta}_{RE}$ is asymptotically efficient given assumption 4 (in the class of estimators consistent under $E[v_i|x_i] = 0$)

Random effects estimator

- Consider the transformation parameter

$$\lambda = 1 - \left(\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_c^2}\right)^{\frac{1}{2}} \; with \; 0 \leq \lambda \leq 1$$

  - $\sigma_\varepsilon^2 = Var[\varepsilon_{it}]$: variance of idiosyncratic error
  - $\sigma_c^2 = Var(c_i)$: Variance of unobserved effect
- $\widehat{\beta}_{RE}$ is equivalent to pooled OLS on:

$$\begin{aligned} y_{it} - \bar{y}_i &= (x_{it} - \lambda \bar{x}_i)\beta + (v_{it} - \lambda \bar{v}_i), \forall i, t \\ \tilde{y}_{it} &= \tilde{x}_{it}\beta + \tilde{v}_{it} \end{aligned}$$

  - As $\lambda \to 1$, $\widehat{\beta}_{RE} \to \widehat{\beta}_{FE}$
  - As $\lambda \to 0$, $\widehat{\beta}_{RE} \to \widehat{\beta}_{Pooled\ OLS}$
    - $\lambda \to 1$ as $T \to \infty$ or if variance of c_i is large relative to variance of ε_{it}
- λ can be estimated from data $\widehat{\lambda} = 1 - (\widehat{\sigma}_\varepsilon^2/(\widehat{\sigma}_\varepsilon^2 + T\widehat{\varepsilon}_c^2))^{\frac{1}{2}}$
- Usually wise to cluster the standard errors since assumption 4 is strong

Random effects regression

```
. xtreg nat_rate repdem , re cl(muniID) i(muniID)

Random-effects GLS regression                   Number of obs      =      4655
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.1052                         Obs per group: min =        19
       between = 0.0005                                        avg =      19.0
       overall = 0.0748                                        max =        19

                                                Wald chi2(1)       =    227.99
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

                               (Std. Err. adjusted for 245 clusters in muniID)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      repdem |   2.859397   .1893742    15.10   0.000      2.48823    3.230564
       _cons |   2.120793   .0972959    21.80   0.000     1.930096    2.311489
-------------+----------------------------------------------------------------
     sigma_u |  1.3866768
     sigma_e |    3.69998
         rho |   .1231606   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Summary: Fixed effects, random effects, Pooled OLS

- Main assumptions
  1. Regressors are strictly exogenous conditional on the time-invariant unobserved effects
  2. Regressors are uncorrelated with the time-invariant unobserved effects
- Results
  - Fixed effects estimator is consistent given assumption 1, but rules out time-invariant regressors
  - Random effects estimators and pooled OLS are consistent under assumptions 1-2, and allow for time-invariant regressors
  - Given homoskedasticity assumptions (random effects assumption 4), the random effects estimator is asymptotically efficient
- Assumption 2 is strong so fixed effects are typically more credible
  - Often the main reason for using panel data is to rule out all time-invariant unobserved confounders!

Hausman test

	$\widehat{\beta}_{RE}$	$\widehat{\beta}_{FE}$
$H_0 : Cov[x_{it}, c_i] = 0$	Consistent and efficient	Consistent
$H_1 : Cov[x_{it}, c_i] \neq 0$	Inconsistent	Consistent

Then,

- Under H_0, $\widehat{\beta}_{RE} - \widehat{\beta}_{FE}$ should be close to zero
- Under H_1, $\widehat{\beta}_{RE} - \widehat{\beta}_{FE}$ should be different from zero
- It can be shown that in large samples, under H_0, the test statistic

$$(\widehat{\beta}_{FE} - \widehat{\beta}_{RE})'(\widehat{Var}[\beta_{FE}] - \widehat{Var}[\beta_{RE}])^{-1}(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}) \xrightarrow{d} \chi^2_k$$

  where k is the number of time-varying regressors.
- We may reject the null hypothesis of "random effects" and stick with the less efficient, but consistent fixed effects specification

Random effects regression

```
. quietly: xtreg nat_rate repdem , fe i(muniID)

. estimates store FE

.
. quietly: xtreg nat_rate repdem , re i(muniID)

. estimates store RE

. hausman FE RE

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
      repdem |     3.0228      2.859397        .1634027        .0304517
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from xtreg
           B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       28.79
```

Hausman test

- Hausman test does not test if the fixed effect model is correct; the test assumes that the fixed effects estimator is consistent!
- Conventional Hausman test assumes homoskedastic model and does not allow for clustering
- There are Haumsman like tests that allow for clustered standard errors

```
. * hausman test with clustering
. quietly: xtreg nat_rate repdem , re i(muniID) cl(muniID)

. xtoverid

Test of overidentifying restrictions: fixed vs random effects
Cross-section time-series model: xtreg re  robust cluster(muniID)
Sargan-Hansen statistic  26.560  Chi-sq(1)    P-value = 0.0000
```

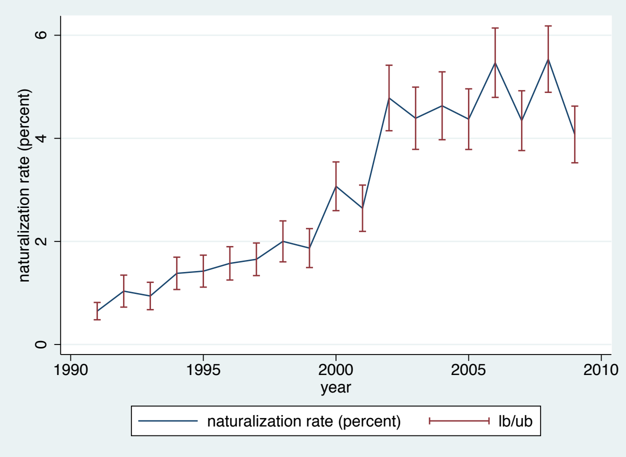
Adding Time Effects

- Reconsider our unobserved effects model:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \;\; t = 1, 2, \ldots, T$$

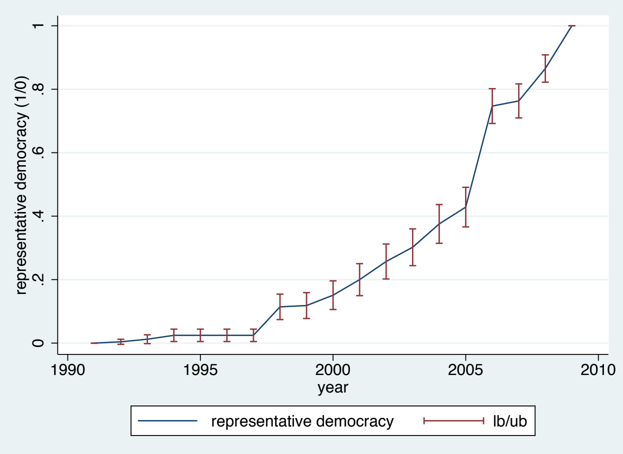
  - Fixed effects assumption: $E[\varepsilon_{it}|x_i, c_i] = 0; t = 1, 2, \ldots, T$: regressors are strictly exogenous conditional on the unobserved effect
  - Typical violation: Common shocks that affect all units in the same way and are correlated with x_{it}
    - Trends in farming technology or climate affect productivity
    - Trends in immigration inflows affect naturalization rates
  - We can allow for such common shocks by including time effects into the model

Random effects regression




```
xtgraph nat_rate
```

Random effects regression




```
xtgraph repdem
```

Fixed effects: adding time effects

- Linear time trend:

$$y_{it} = x_{it}\beta + c_i + t + \varepsilon_{it}; \;\; t = 1, 2, \ldots, T$$

    - Linear time trend common to all units
- Time fixed effects:

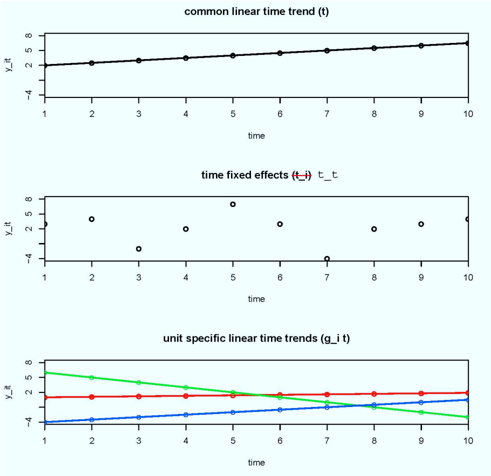
$$y_{it} = x_{it}\beta + c_i + t_t + \varepsilon_{it}; \;\; t = 1, 2, \ldots, T$$

    - Common shock in each time period
    - Generalized difference-in-difference model
- Unit specific linear time trends:

$$y_{it} = x_{it}\beta + c_i + g_i \cdot t + t_t + \varepsilon_{it}; \;\; t = 1, 2, \ldots, T$$

    - Linear time trends that vary by unit

Modeling time effects



Fixed effects: adding time effects

```
. egen time = group(year)

. list muniID muni_name year time in 20/40 ,ab(20)
```

	muniID	muni_name	year	time
20.	2	Affoltern A.A.	1991	1
21.	2	Affoltern A.A.	1992	2
22.	2	Affoltern A.A.	1993	3
23.	2	Affoltern A.A.	1994	4
24.	2	Affoltern A.A.	1995	5
25.	2	Affoltern A.A.	1996	6
26.	2	Affoltern A.A.	1997	7
27.	2	Affoltern A.A.	1998	8
28.	2	Affoltern A.A.	1999	9
29.	2	Affoltern A.A.	2000	10
30.	2	Affoltern A.A.	2001	11

Fixed effects: linear time trend

```
. xtreg nat_rate repdem time , fe cl(muniID) i(muniID)

Fixed-effects (within) regression               Number of obs      =      4655
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.1604                         Obs per group: min =        19
       between = 0.0005                                        avg =      19.0
       overall = 0.1350                                        max =        19

                                                F(2,244)           =    247.57
corr(u_i, Xb)  = -0.0079                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 245 clusters in muniID)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      repdem |   .8247928   .2590615     3.18   0.002     .3145106    1.335075
        time |   .2313692   .0171752    13.47   0.000     .1975386    .2651997
       _cons |   .3892908   .1309232     2.97   0.003     .1314069    .6471747
-------------+----------------------------------------------------------------
     sigma_u |  1.6271657
     sigma_e |   3.584409
         rho |  .17086519   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Fixed effects: year fixed effects

```
. xtreg nat_rate repdem i.time , fe cl(muniID) i(muniID)

Fixed-effects (within) regression               Number of obs      =      4655
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.1885                         Obs per group: min =        19
       between = 0.0005                                        avg =      19.0
       overall = 0.1575                                        max =        19

                                                F(19,244)          =     31.48
corr(u_i, Xb)  = -0.0168                        Prob > F           =    0.0000

                               (Std. Err. adjusted for 245 clusters in muniID)
```

nat_rate	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
repdem	1.203658	.3031499	3.97	0.000	.6065335	1.800783
time						
2	.3829173	.1723225	2.22	0.027	.0434879	.7223468
3	.2789777	.1514124	1.84	0.067	-.0192644	.5772198
4	.7034078	.167466	4.20	0.000	.3735443	1.033271

Fixed effects: unit specific time trends

```
. xtreg nat_rate repdem muniID#c.time i.time  , fe cl(muniID) i(muniID)
note: 19.time omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      4655
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.2650                         Obs per group: min =        19
       between = 0.5185                                        avg =      19.0
       overall = 0.2864                                        max =        19

                                                F(18,244)          =         .
corr(u_i, Xb)  = -0.3963                        Prob > F           =         .

                             (Std. Err. adjusted for 245 clusters in muniID)
```

nat_rate	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
repdem	.9865241	.322868	3.06	0.002	.3505601	1.622488
muniID#c.time						
1	.333343	.024298	13.72	0.000	.2854823	.3812036
2	.2914274	.024298	11.99	0.000	.2435667	.339288
3	.248985	.024298	10.25	0.000	.2011244	.2968457

Unit specific time trends often eliminate "results"

TABLE 5.2.3
Estimated effects of labor regulation on the performance of firms in Indian states

	(1)	(2)	(3)	(4)
Labor regulation (lagged)	−.186	−.185	−.104	.0002
	(.064)	(.051)	(.039)	(.020)
Log development expenditure per capita		.240	.184	.241
		(.128)	(.119)	(.106)
Log installed electricity capacity per capita		.089	.082	.023
		(.061)	(.054)	(.033)
Log state population		.720	0.310	−1.419
		(.96)	(1.192)	(2.326)
Congress majority			−.0009	.020
			(.01)	(.010)
Hard left majority			−.050	−.007
			(.017)	(.009)
Janata majority			.008	−.020
			(.026)	(.033)
Regional majority			.006	.026
			(.009)	(.023)
State-specific trends	No	No	No	Yes
Adjusted R^2	.93	.93	.94	.95

Notes: Adapted from Besley and Burgess (2004), table IV. The table reports regression DD estimates of the effects of labor regulation on productivity. The

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \;\; t = 1, 2, \ldots, T$$

- Model recognizes the effect of change in x may occur with a late
  - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
  - Consider **temporary increase** in x_{it} from level m to $m+1$ at t which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \;\; t = 1, 2, \ldots, T$$

- Model recognizes the effect of change in x may occur with a late
  - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
  - Consider **temporary increase** in x_{it} from level m to $m+1$ at t which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- $\beta_0 = y_t - y_{t-1}$ immediate change in y due to temporary one-unit increase in x (impact propensity)

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \;\; t = 1, 2, \ldots, T$$

- Model recognizes the effect of change in x may occur with a late
  - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
  - Consider **temporary increase** in x_{it} from level m to $m+1$ at t which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- $\beta_1 = y_{t+1} - y_t$ change in y one period after temporary one-unit increase in x

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \ \ t = 1, 2, \ldots, T$$

- Model recognizes the effect of change in x may occur with a late
  - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
  - Consider **temporary increase** in x_{it} from level m to $m+1$ at t which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- $\beta_2 = y_{t+2} - y_{t-1}$ change in y two periods after temporary one-unit increase in x

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \;\; t = 1, 2, \ldots, T$$

- Model recognizes the effect of change in x may occur with a late
  - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
  - Consider **temporary increase** in x_{it} from level m to $m+1$ at t which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- $\beta_3 = y_{t-1}$ change in y is zero three periods after temporary one-unit increase in x

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \;\; t = 1, 2, \ldots, T$$

- Interpretation of coefficients:
    - Consider **permanent increase** in x_{it} from level m to $m+1$ at t, i.e., $(x_s = m, s < t$ and $x_s = m+1, s \geq t)$
        - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
        - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
        - $y_{t+1} = (m+1)\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
        - $y_{t+2} = (m+1)\beta_0 + (m+1)\beta_1 + (m+1)\beta_2 + c_i$
        - $y_{t+3} = (m+1)\beta_0 + (m+1)\beta_1 + (m+1)\beta_2 + c_i$
- After one period y has increased by $\beta_0 + \beta_1$, after two periods y has increased by $\beta_0 + \beta_1 + \beta_2$ and there are no further increases after two periods
- Long-run increase in y : $\beta_0 + \beta_1 + \beta_2$ (long-run propensity)

Lagged effects of direct democracy

```
. xtreg nat_rate repdem L1.repdem L2.repdem  L3.repdem i.year, fe cl(muniID) i(muniID)

Fixed-effects (within) regression               Number of obs      =      3920
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.1536                         Obs per group: min =        16
       between = 0.0012                                        avg =      16.0
       overall = 0.1235                                        max =        16

                                                F(19,244)          =     21.63
corr(u_i, Xb)  = -0.0206                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 245 clusters in muniID)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      repdem |
         --. |   .6364802   .3593924     1.77   0.078    -.0714272    1.344388
         L1. |   1.201065   .4233731     2.84   0.005      .367133    2.034998
         L2. |  -.1648692   .4697434    -0.35   0.726    -1.090139    .7604003
         L3. |  -.5245206   .4109918    -1.28   0.203    -1.334065    .2850239
```

Long-run effect of direct democracy

```
. lincom repdem + L1.repdem + L2.repdem + L3.repdem

 ( 1)  repdem + L.repdem + L2.repdem + L3.repdem = 0
```

nat_rate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	1.294485	.4426322	2.92	0.004	.4226175	2.166353

Lags and Leads model

$$y_{it} = x_{i,t+1}\beta_{-1} + x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it};\ \ t = 1, 2, \ldots, T$$

- Can use estimate of β_{-1} to test for anticipation effects
  - Consider **temporary increase** in x_{it} from level m to $m+1$ at t
    - $y_{t-2} = \beta_{-1}m + m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t-1} = \beta_{-1}(m+1) + m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- Anticipation effect: $\beta_{-1} = y_{t-1} - y_{t-2}$ change in y in period $t-1$, the period before the temporary one-unit increase in x
- Placebo test: if x causes y, but y does not cause x, then β_{-1} should be close to zero

Leads and Lags

```
. xtreg nat_rate F1.repdem repdem L1.repdem L2.repdem L3.repdem  i.year, fe cl(muniID) i(muniID)

Fixed-effects (within) regression               Number of obs      =      3675
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.1621                         Obs per group: min =        15
       between = 0.0010                                        avg =      15.0
       overall = 0.1269                                        max =        15

                                                F(19,244)          =     20.34
corr(u_i, Xb)  = -0.0353                        Prob > F           =    0.0000

                               (Std. Err. adjusted for 245 clusters in muniID)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      repdem |
         F1. |   .1707685   .3212906     0.53   0.596    -.4620886    .8036255
         --. |   .6975731   .4397095     1.59   0.114    -.1685376    1.563684
         L1. |   .8723962   .4619322     1.89   0.060    -.0374873     1.78228
         L2. |    .014941   .4583628     0.03   0.974    -.8879119    .9177939
         L3. |  -.2904252   .4108244    -0.71   0.480     -1.09964    .5187895
```

The Autor Test

- Let D_{it} be a binary indicator equaling 1 if unit i **switched** from control to treatment between t and $t-1$; 0 otherwise
  - Lags: $D_{i,t-1}$: unit switched between $t-1$ and $t-2$
  - Leads: $D_{i,t+1}$: unit switches between $t+1$ and t
- Include lags and leads into the fixed effects model:

$$y_{it} = D_{i,t+2}\beta_{-2} + D_{i,t+1}\beta_{-1} + D_{it}\beta_0 + D_{i,t-1}\beta_1 + D_{i,t-2}\beta_2 + c_i + \varepsilon_{it}$$

- Interpretation of coefficients:
  - Leads β_{-1}, β_{-2}, etc. test for anticipation effects
  - Switch β_0 tests for immediate effect
  - Lags β_1, β_2, etc. test for long-run effects
    - highest lag dummy can be coded 1 for all post-switch years

Lags and Leads of Switch to Representative Democracy

```
. list muni_name year repdem switch_t sw_lag1 sw_lag2 sw_lag3  ///
>                         sw_lead1 sw_lead2 sw_lead3 in 806/817
```

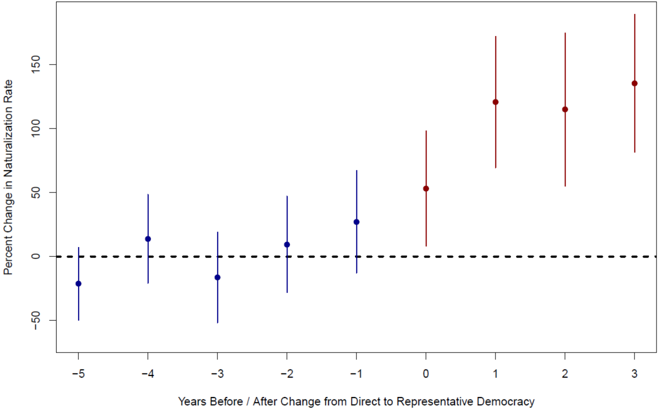
	muni_n~e	year	repdem	switch_t	sw_lag1	sw_lag2	sw_lag3	sw_lead1	sw_lead2	sw_lead3
806.	Stäfa	1998	0	0	0	0	0	0	0	0
807.	Stäfa	1999	0	0	0	0	0	0	0	0
808.	Stäfa	2000	0	0	0	0	0	0	0	0
809.	Stäfa	2001	0	0	0	0	0	0	0	0
810.	Stäfa	2002	0	0	0	0	0	0	0	1
811.	Stäfa	2003	0	0	0	0	0	0	1	0
812.	Stäfa	2004	0	0	0	0	0	1	0	0
813.	Stäfa	2005	1	1	0	0	0	0	0	0
814.	Stäfa	2006	1	0	1	0	0	0	0	0
815.	Stäfa	2007	1	0	0	1	0	0	0	0
816.	Stäfa	2008	1	0	0	0	1	0	0	0
817.	Stäfa	2009	1	0	0	0	1	0	0	0

Dynamic Effect of Switching to Representative Democracy

```
. xtreg  nat_rate sw_lag3 sw_lag2 sw_lag1 switch_t ///
>        sw_lead1 sw_lead2 sw_lead3 sw_lead4 sw_lead5 i.year, fe cluster(muniID) i(muniID)

Fixed-effects (within) regression               Number of obs      =      4655
Group variable: muniID                          Number of groups   =       245

R-sq:  within  = 0.1913                         Obs per group: min =        19
       between = 0.0011                                        avg =      19.0
       overall = 0.1601                                        max =        19

                                                F(27,244)          =     23.76
corr(u_i, Xb)  = -0.0162                        Prob > F           =    0.0000

                               (Std. Err. adjusted for 245 clusters in muniID)
```

nat_rate	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
sw_lag3	1.160345	.5080271	2.28	0.023	.1596665	2.161023
sw_lag2	1.743682	.5395212	3.23	0.001	.680969	2.806396
sw_lag1	1.881663	.4880099	3.86	0.000	.9204133	2.842913
switch_t	.7564792	.428627	1.76	0.079	-.0878019	1.60076
sw_lead1	.2138757	.3899881	0.55	0.584	-.5542971	.9820485
sw_lead2	.0843676	.3575292	0.24	0.814	-.61987	.7886051
sw_lead3	.1440446	.3194086	0.45	0.652	-.4851054	.7731945
sw_lead4	.0750194	.2990359	0.25	0.802	-.5140018	.6640405
sw_lead5	-.0942415	.2599789	-0.36	0.717	-.6063307	.4178477

Dynamic Effect of Switching to Representative Democracy



Lagged Dependent Variable

$$y_{it} = \alpha y_{i,t-1} + c_i + \varepsilon_{it}, \;\; t = 1, 2, \ldots, T$$

- y_{it} could be capital stock of firm i at time t and α the capital depreciation rate
- For simplicity, we assume that ε_{it} are uncorrelated in time (as well as across individuals)
- Note that
$$y_{i,t-1} = \alpha y_{i,t-2} + c_i + \varepsilon_{i,t-1}$$
- So we have $Cov[y_{i,t-1}, c_i] \neq 0$ and therefore we need to include fixed effects c_i into the regression
- Does this work though?

Lagged dependent variable

With $T = 3$ we have

$$
\begin{aligned}
y_{i3} &= \alpha y_{i2} + c_i + \varepsilon_{i3} \\
y_{i2} &= \alpha y_{i1} + c_i + \varepsilon_{i2}
\end{aligned}
$$

and we can take time differences to eliminate c_i (similar to fixed effects)

$$
\begin{aligned}
y_{i3} - y_{i2} &= \alpha(y_{i2} - y_{i1}) + (c_i - c_i) + (\varepsilon_{i3} - \varepsilon_{i2}) \\
\Delta y_{i3} &= \alpha \Delta y_{i2} + \Delta \varepsilon_{i3}
\end{aligned}
$$

Since ε_{i2} affects both $\Delta y_{i2} = y_{i2} - y_{i1}$ and $\Delta \varepsilon_{i3} = \varepsilon_{i3} - \varepsilon_{i2}$ we get

$$Cov[\Delta y_{i2}, \Delta \varepsilon_{i3}] \neq 0 \text{ and thus still have endogeneity}$$

Models with fixed effects and lagged y do not produce consistent estimators Might use past levels y_{i1} as an instrument for Δy_{i2}, but this

requires strong assumptions (e.g., no serial correlation in ε_{it})