Causal Inference

Justin Grimmer

Associate Professor
Department of Political Science
University of Chicago

May 7th, 2018

Panel Setup

- Let y and $\mathbf{x} \equiv (x_1, x_2, ..., x_K)$ 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

$$E[y|x_1, x_2, ..., x_K, c]$$

- We observe a sample of $i = 1, 2, ..., N$ cross-sectional units for $t = 1, 2, ... T$ time periods (a balanced panel)
  - For each unit i, we denote the observable variables for all time periods as $\{(y_{it}, \mathbf{x}_{it}) : t = 1, 2, ..., T\}$
  - $\mathbf{x}_{it} \equiv (x_{it1}, x_{it2}, ..., x_{itK})$ is a $1 \times K$ vector

- Typically assume that cross-sectional units are i.i.d. draws from the population: $\{\mathbf{y}_i, \mathbf{x}_i, c_i\}_{i=1}^N \sim i.i.d.$ (cross-sectional independence)
  - $\mathbf{y}_i \equiv (y_{i1}, y_{i2}, ..., y_{iT})'$ and $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$
  - Consider asymptotic properties with T fixed and $N \to \infty$

Panel Setup

Single unit:

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T \times 1} \qquad \mathbf{X}_i = \begin{pmatrix} x_{i,1,1} & x_{i,1,2} & x_{i,1,j} & \cdots & x_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{i,t,1} & x_{i,t,2} & x_{i,t,j} & \cdots & x_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{i,T,1} & x_{i,T,2} & x_{i,T,j} & \cdots & x_{i,T,K} \end{pmatrix}_{T \times K}$$

Panel with all units:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_N \end{pmatrix}_{NT \times 1} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_N \end{pmatrix}_{NT \times K}$$

Unobserved Effects Model: Farm Output

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

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

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

Pooled OLS

- When we ignore the panel structure and regress y_{it} on $\mathbf{x}_{it}$ we get

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \qquad t = 1, 2, ..., T$$

  with composite error $v_{it} \equiv c_i + \varepsilon_{it}$

- Main assumption to obtain consistent estimates for $\boldsymbol{\beta}$ is:
  - $E[v_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}] = E[v_{it}|\mathbf{x}_{it}] = 0$ for $t = 1, 2, ..., T$
    - $\mathbf{x}_{it}$ are strictly exogenous: the composite error v_{it} in each time period is uncorrelated with the past, current, and future regressors
    - But: labour input $\mathbf{x}_{it}$ likely depends on soil quality c_i and so we have omitted variable bias and $\hat{\beta}$ is not consistent
  - No correlation between $\mathbf{x}_{it}$ and v_{it} implies no correlation between unobserved effect c_i and $\mathbf{x}_{it}$ for all t
    - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)

Pooled OLS

- When we ignore the panel structure and regress y_{it} on $\mathbf{x}_{it}$ we get

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \qquad t = 1, 2, ..., T$$

  with composite error $v_{it} \equiv c_i + \varepsilon_{it}$

- Main assumption to obtain consistent estimates for $\boldsymbol{\beta}$ is:
  - $E[v_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}] = E[v_{it}|\mathbf{x}_{it}] = 0$ for $t = 1, 2, ..., T$
    - $\mathbf{x}_{it}$ are strictly exogenous: the composite error v_{it} in each time period is uncorrelated with the past, current, and future regressors
    - But: labour input $\mathbf{x}_{it}$ likely depends on soil quality c_i and so we have omitted variable bias and $\hat{\boldsymbol{\beta}}$ is not consistent
  - No correlation between $\mathbf{x}_{it}$ and v_{it} implies no correlation between unobserved effect c_i and $\mathbf{x}_{it}$ for all t
    - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)

Pooled OLS

- When we ignore the panel structure and regress y_{it} on $\mathbf{x}_{it}$ we get

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \qquad t = 1, 2, ..., T$$

with composite error $v_{it} \equiv c_i + \varepsilon_{it}$

- Main assumption to obtain consistent estimates for $\boldsymbol{\beta}$ is:
  - $E[v_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}] = E[v_{it}|\mathbf{x}_{it}] = 0$ for $t = 1, 2, ..., T$
    - $\mathbf{x}_{it}$ are strictly exogenous: the composite error v_{it} in each time period is uncorrelated with the past, current, and future regressors
    - But: labour input $\mathbf{x}_{it}$ likely depends on soil quality c_i and so we have omitted variable bias and $\hat{\boldsymbol{\beta}}$ is not consistent
  - No correlation between $\mathbf{x}_{it}$ and v_{it} implies no correlation between unobserved effect c_i and $\mathbf{x}_{it}$ for all t
    - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)

Unobserved Effects Model: Program Evaluation

- Program evaluation model:

$$y_{it} = prog_{it}\,\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - 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
  - $\boldsymbol{\beta}$: 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 $prog_{it}$? $\hat{\beta}$ 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, pooled OLS is problematic
- Additional problem: $v_{it} \equiv c_i + \varepsilon_{it}$ 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:

$$y_{it} = prog_{it}\,\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - 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
  - $\boldsymbol{\beta}$: 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 $prog_{it}$? $\hat{\boldsymbol{\beta}}$ 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, pooled OLS is problematic
- Additional problem: $v_{it} \equiv c_i + \varepsilon_{it}$ are serially correlated for same i since c_i is present in each t and thus pooled OLS standard errors are invalid

Fixed Effect Regression

- Our unobserved effects model is:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- 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:

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

this amounts to including N farm dummies in regression of y_{it} on $\mathbf{x}_{it}$

Derivation: Fixed Effects Regression

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

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

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

and

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

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

Derivation: Fixed Effects Regression

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

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

where

$$\bar{\mathbf{x}}_i \equiv \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{it}, \qquad \bar{y}_i \equiv \frac{1}{T} \sum_{t=1}^{T} y_{it}.$$

Plug this result into the first FOC to obtain:

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

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

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

Fixed Effects Regression

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

Fixed effects estimator is often called the within estimator because it only uses the time variation within each cross-sectional unit.

Even better, the regression with the time-demeaned variables is consistent for $\boldsymbol{\beta}$ even when $Cov[\mathbf{x}_{it}, c_i] \neq 0$, because time-demeaning eliminates the unobserved effects:

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it} \\ \bar{y}_i &= \bar{\mathbf{x}}_i\boldsymbol{\beta} + c_i + \bar{\varepsilon}_i \end{aligned}$$

$$\begin{aligned} (y_{it} - \bar{y}_i) &= (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (c_i - c_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \\ \ddot{y}_{it} &= \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{\varepsilon}_{it} \end{aligned}$$

Fixed Effects Regression: Main Results

- Identification assumptions:
  1. $E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, c_i] = 0, \;\; t = 1, 2, ..., T$
     - regressors are strictly exogenous conditional on the unobserved effect
     - allows $\mathbf{x}_{it}$ to be arbitrarily related to c_i
  2. $rank(\sum_{t=1}^{T} E[\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}]) = K$
     - regressors vary over time for at least some i and are not collinear

- Fixed effects estimator:
  1. Demean and regress $\ddot{y}_{it}$ on $\ddot{\mathbf{x}}_{it}$ (need to correct degrees of freedom)
  2. Regress y_{it} on $\mathbf{x}_{it}$ and unit dummies (dummy variable regression)
  3. Regress y_{it} on $\mathbf{x}_{it}$ with canned fixed effects routine
     - Stata: `xtreg y x , fe i(PanelID) cl(PanelID)`

- Properties (under assumptions 1-2):
  - $\hat{\boldsymbol{\beta}}_{FE}$ is consistent: $\operatorname*{plim}_{N\to\infty} \hat{\boldsymbol{\beta}}_{FE,N} = \boldsymbol{\beta}$
  - $\hat{\boldsymbol{\beta}}_{FE}$ also unbiased conditional on $\mathbf{X}$

Fixed Effects Regression: Main Results

- Identification assumptions:
  1. $E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, c_i] = 0, \;\; t = 1, 2, ..., T$
     - regressors are strictly exogenous conditional on the unobserved effect
     - allows $\mathbf{x}_{it}$ to be arbitrarily related to c_i
  2. $rank(\sum_{t=1}^{T} E[\ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it}]) = K$
     - regressors vary over time for at least some i and are not collinear
- Fixed effects estimator:
  1. Demean and regress $\ddot{y}_{it}$ on $\ddot{\mathbf{x}}_{it}$ (need to correct degrees of freedom)
  2. Regress y_{it} on $\mathbf{x}_{it}$ and unit dummies (dummy variable regression)
  3. Regress y_{it} on $\mathbf{x}_{it}$ with canned fixed effects routine
     - R: `plm(y~x , model = within, data = data)`
- Properties (under assumptions 1-2):
  - $\hat{\boldsymbol{\beta}}_{FE}$ is consistent: $\underset{N\to\infty}{\text{plim}}\, \hat{\boldsymbol{\beta}}_{FE,N} = \boldsymbol{\beta}$
  - $\hat{\boldsymbol{\beta}}_{FE}$ also 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 $\boldsymbol{\beta}$, 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 $\widehat{c_i} = \bar{y}_i - \bar{\mathbf{x}}_i \widehat{\boldsymbol{\beta}}$
    - `predict c_i , u`

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.
    - R: `coeftest(mod, vcov=function(x) vcovHC(x, cluster="group", type="HC1"))`
  - Yields valid inference as long as number of clusters is reasonably large
- Typically we care about $\boldsymbol{\beta}$, 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 $N \rightarrow \infty$)
  - `plm` 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 $\widehat{c}_i = \bar{y}_i - \bar{\mathbf{x}}_i \widehat{\boldsymbol{\beta}}$
    - `fixef(mod)`

Example: Direct Democracy and Naturalizations

- Do minorities fare worse under direct democracy than under representative democracy?
- Hainmueller and Hangartner (2016, AJPS) 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 = $\#$ naturalizations$_{it}$ / eligible foreign population $_{it-1}$
  - x_{it}: 1 if municipality used representative democracy, 0 if municipality used direct democracy in year t

Naturalization Panel Data Long Format

```
> d <- read.dta("Swiss_Panel_long.dta")
> print(d[30:40,],digits=2)
   muniID      muni_name year nat_rate repdem
30      2 Affoltern A.A. 2001     3.21      0
31      2 Affoltern A.A. 2002     4.64      0
32      2 Affoltern A.A. 2003     4.84      0
33      2 Affoltern A.A. 2004     5.62      0
34      2 Affoltern A.A. 2005     4.39      0
35      2 Affoltern A.A. 2006     8.12      1
36      2 Affoltern A.A. 2007     7.07      1
37      2 Affoltern A.A. 2008     8.98      1
38      2 Affoltern A.A. 2009     6.12      1
39      3     Bonstetten 1991     0.83      0
40      3     Bonstetten 1992     0.84      0
```

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

      Source |       SS       df       MS              Number of obs =    4655
-------------+------------------------------           F(  1,  4653) =  376.17
       Model |  5958.63488     1  5958.63488           Prob > F      =  0.0000
    Residual |  73705.2336  4653  15.8403683           R-squared     =  0.0748
-------------+------------------------------           Adj R-squared =  0.0746
       Total |  79663.8685  4654  17.1172902           Root MSE      =    3.98

------------------------------------------------------------------------------
    nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      repdem |   2.503318     .12907    19.40   0.000     2.250279    2.756356
       _cons |   2.222683   .0690427    32.19   0.000     2.087326    2.358039
------------------------------------------------------------------------------
```

Pooled OLS

```
> summary(lm(nat_rate~repdem,data=d))

Call:
lm(formula = nat_rate ~ repdem, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-4.726 -2.223 -1.523  1.411 21.915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.22268    0.06904   32.19   <2e-16 ***
repdem       2.50332    0.12907   19.39   <2e-16 ***
```

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 |
```

Time-Demeaning for Fixed Effects: $y_{it} \rightarrow \ddot{y}_{it}$

```
> library(plyr)
> d <- ddply(d, .(muniID), transform,
+                        nat_rate_demean = nat_rate - mean(nat_rate),
+                        nat_rate_mean   = mean(nat_rate),
+                        repdem_demean   = repdem - mean(repdem))
>
> print(d[20:38,
+       c("muniID","muni_name","year","nat_rate","nat_rate_mean","nat_rate_demean","repdem","repdem_demean")
+       ],digits=2)
   muniID      muni_name year nat_rate nat_rate_mean nat_rate_demean repdem repdem_demean
20      2 Affoltern A.A. 1991     0.22           3.6           -3.38      0         -0.21
21      2 Affoltern A.A. 1992     0.95           3.6           -2.65      0         -0.21
22      2 Affoltern A.A. 1993     1.05           3.6           -2.55      0         -0.21
23      2 Affoltern A.A. 1994     0.83           3.6           -2.76      0         -0.21
24      2 Affoltern A.A. 1995     2.00           3.6           -1.59      0         -0.21
25      2 Affoltern A.A. 1996     1.78           3.6           -1.82      0         -0.21
26      2 Affoltern A.A. 1997     1.86           3.6           -1.73      0         -0.21
27      2 Affoltern A.A. 1998     2.05           3.6           -1.54      0         -0.21
28      2 Affoltern A.A. 1999     2.40           3.6           -1.19      0         -0.21
29      2 Affoltern A.A. 2000     2.20           3.6           -1.40      0         -0.21
30      2 Affoltern A.A. 2001     3.21           3.6           -0.39      0         -0.21
31      2 Affoltern A.A. 2002     4.64           3.6            1.04      0         -0.21
32      2 Affoltern A.A. 2003     4.84           3.6            1.25      0         -0.21
33      2 Affoltern A.A. 2004     5.62           3.6            2.03      0         -0.21
34      2 Affoltern A.A. 2005     4.39           3.6            0.79      0         -0.21
35      2 Affoltern A.A. 2006     8.12           3.6            4.52      1          0.79
36      2 Affoltern A.A. 2007     7.07           3.6            3.47      1          0.79
37      2 Affoltern A.A. 2008     8.98           3.6            5.38      1          0.79
38      2 Affoltern A.A. 2009     6.12           3.6            2.52      1          0.79
```

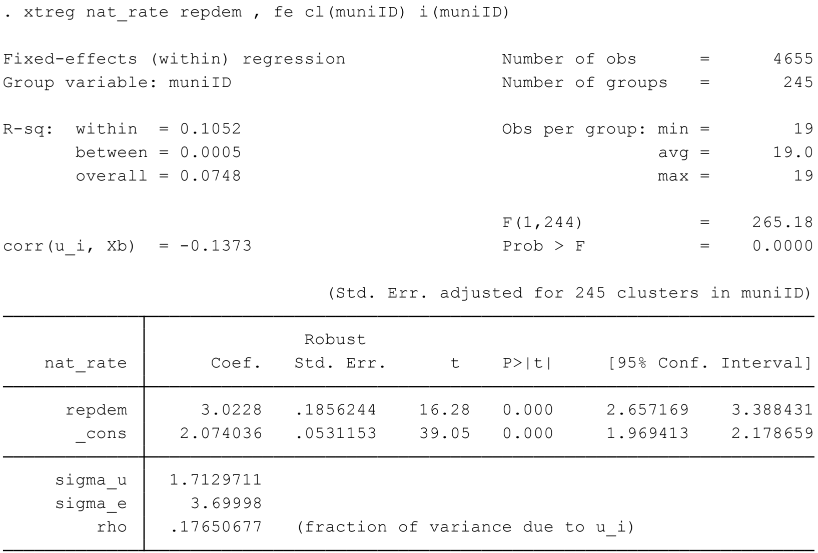
Fixed Effects Regression with Demeaned Data

```
. 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 Demeaned Data

```
> summary(lm(nat_rate_demean~repdem_demean,data=d))

Call:
lm(formula = nat_rate_demean ~ repdem_demean, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4712 -2.0883 -0.5978  1.0841 21.3076 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.266e-16  5.279e-02    0.00        1    
repdem_demean 3.023e+00  1.293e-01   23.39   <2e-16 ***
```

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 Canned Routine

```
> library(plm)
> library(lmtest)
> d <- plm.data(d, indexes = c("muniID", "year"))
> mod_fe <- plm(nat_rate~repdem,data=d,model="within")
> coeftest(mod_fe,
vcov=function(x) vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

       Estimate Std. Error t value  Pr(>|t|)
repdem  3.02280    0.18525  16.318 < 2.2e-16 ***
---
```

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
```

Fixed Effects Regression with Dummies

```
> mod_du <- plm(nat_rate~repdem+as.factor(muniID),data=d,model="pooling")
> coeftest(mod_du, vcov=function(x) vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                         Estimate  Std. Error    t value  Pr(>|t|)
(Intercept)            1.5922e+00  4.0068e-02  3.9737e+01 < 2.2e-16 ***
repdem                 3.0228e+00  1.9032e-01  1.5883e+01 < 2.2e-16 ***
as.factor(muniID)2     1.3674e+00  1.4249e-08  9.5960e+07 < 2.2e-16 ***
as.factor(muniID)3     1.2923e+00  1.4283e-08  9.0472e+07 < 2.2e-16 ***
as.factor(muniID)9     1.2847e+00  1.3404e-08  9.5837e+07 < 2.2e-16 ***
as.factor(muniID)10    1.2718e+00  1.4182e-08  8.9675e+07 < 2.2e-16 ***
as.factor(muniID)13    3.2655e-01  1.2597e-08  2.5922e+07 < 2.2e-16 ***
as.factor(muniID)25    5.6413e-02  3.0051e-02  1.8772e+00 0.0605523 .
as.factor(muniID)26    3.1257e+00  1.0017e-02  3.1204e+02 < 2.2e-16 ***
as.factor(muniID)29    3.1797e+00  3.0051e-02  1.0581e+02 < 2.2e-16 ***
as.factor(muniID)33    3.2293e+00          NA          NA        NA
as.factor(muniID)34    1.7467e+00  3.0051e-02  5.8123e+01 < 2.2e-16 ***
.
```

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

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- 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?
  - $\hat{\beta}_{FE}$ inconsistent unless strict exogeneity conditional on c_i holds
    - $E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, c_i] = 0, \ t = 1, 2, ..., T$
    - 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
  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!

Problems that (even) fixed effects do not solve

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- 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?
  - $\hat{\boldsymbol{\beta}}_{FE}$ inconsistent unless strict exogeneity conditional on c_i holds
    - $E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, c_i] = 0,\ \ t = 1, 2, ..., T$
    - 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
  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:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Cannot use the fixed effects regression to estimate effects of time-constant regressors in $\mathbf{x}_{it}$ (e.g. soil quality, farm location, etc.)
    - Since fixed effect estimator allows c_i to be correlated with $\mathbf{x}_{it}$, we cannot distinguish the effects of time-invariant regressors from the time-invariant unobserved effect c_i
    - If a regressor does not change much in time, the standard errors of the coefficients in the fixed effects regression will be large (because there is little variation in the demeaned regressor $\ddot{\mathbf{x}}_{it} \equiv \mathbf{x}_{it} - \bar{\mathbf{x}}_i i$)
  - Need orthogonality assumption: $Cov[\mathbf{x}_{it}, c_i] = 0, \quad t = 1, ..., T$
    - Strong assumption: Unobserved effects c_i are uncorrelated with each explanatory variable in $\mathbf{x}_{it}$ in each time period.
    - For example, if we include soil quality in $\mathbf{x}_{it}$, we have to assume it is uncorrelated with all other time-invariant inputs.

Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

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

Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

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

Assumption 4 implies $\Omega = E[\boldsymbol{v}_i\boldsymbol{v}_i'|\mathbf{x}_i] =$

$$\begin{pmatrix} \sigma_c^2 + \sigma_\varepsilon^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_\varepsilon^2 & \dots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2 + \sigma_\varepsilon^2 \end{pmatrix}_{T \times T}$$

Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

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

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

Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

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

- Random effects estimator $\hat{\boldsymbol{\beta}}_{RE}$ exploits this serial correlation in a generalized least squares (GLS) framework
    - $\hat{\boldsymbol{\beta}}_{RE}$ is consistent under assumptions 1-3: $\operatorname*{plim}_{N\rightarrow\infty} \hat{\boldsymbol{\beta}}_{RE,N} = \boldsymbol{\beta}$
    - $\hat{\boldsymbol{\beta}}_{RE}$ is asymptotically efficient given assumption 4 (in the class of estimators consistent under $E[\mathbf{v}_i|\mathbf{x}_i] = \mathbf{0}$)

Random Effects Estimator

- Consider the transformation parameter:

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

  - $\sigma_\varepsilon^2 = Var[\varepsilon_{it}]$: variance of idiosyncratic error
  - $\sigma_c^2 = Var[c_i]$: variance of unobserved effect

- $\hat{\boldsymbol{\beta}}_{RE}$ is equivalent to pooled OLS on quasi-demeaned data:

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

  - As $\lambda \rightarrow 1$, $\hat{\boldsymbol{\beta}}_{RE} \rightarrow \hat{\boldsymbol{\beta}}_{FE}$
  - As $\lambda \rightarrow 0$, $\hat{\boldsymbol{\beta}}_{RE} \rightarrow \hat{\boldsymbol{\beta}}_{Pooled\ OLS}$
    - $\lambda \rightarrow 1$ as $T \rightarrow \infty$ or if variance of c_i is large relative to variance of ε_{it}

- λ can be estimated from data $\widehat{\lambda} = 1 - (\widehat{\sigma}_\varepsilon^2 / (\widehat{\sigma}_\varepsilon^2 + T\widehat{\sigma}_c^2))^{1/2}$

- 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)
------------------------------------------------------------------------------
```

Random Effects Regression

```
> mod_re <- plm(nat_rate˜repdem,data=d,model="random")
> coeftest(mod_re, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 2.120793   0.097108  21.840 < 2.2e-16 ***
repdem      2.859397   0.189008  15.129 < 2.2e-16 ***
---
```

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 estimator and pooled OLS are consistent under assumptions 1-2, and allow for time-invariant regressors
  - Given homoscedasticity 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

Given the homoskedastic model (RE assumptions 1-4):

	$\widehat{\beta}_{RE}$	$\widehat{\beta}_{FE}$
$H_0 : \text{Cov}[\mathbf{x}_{it}, c_i] = 0$	consistent and efficient	consistent
$H_1 : \text{Cov}[\mathbf{x}_{it}, c_i] \neq 0$	inconsistent	consistent

Then,

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

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

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

Hausman Test

```
. 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

```
> phtest(mod_fe, mod_re)

  Hausman Test

data:  nat_rate ~ repdem
chisq = 28.7935, df = 1, p-value = 8.052e-08
alternative hypothesis: one model is inconsistent
```

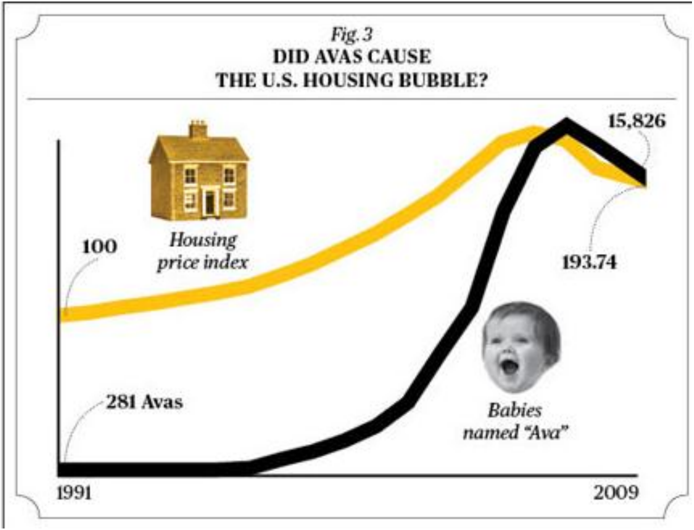
- Hausman test does not test if the fixed effects model is correct, the test assumes that the fixed effects estimator is consistent!
- Conventional Hausman test assumes homoscedastic model and does not allow for clustering

Hausman Test

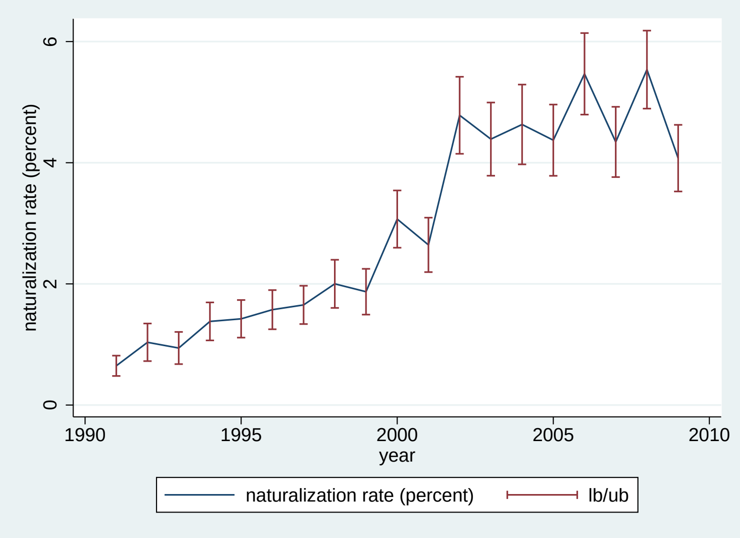
- Hausman test does not test if the fixed effects model is correct, the test assumes that the fixed effects estimator is consistent!
- Conventional Hausman test assumes homoscedastic model and does not allow for clustering
- There are Hausman like tests that allow for clustered standard errors

```
. * 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
```

Common Shocks or Causation?

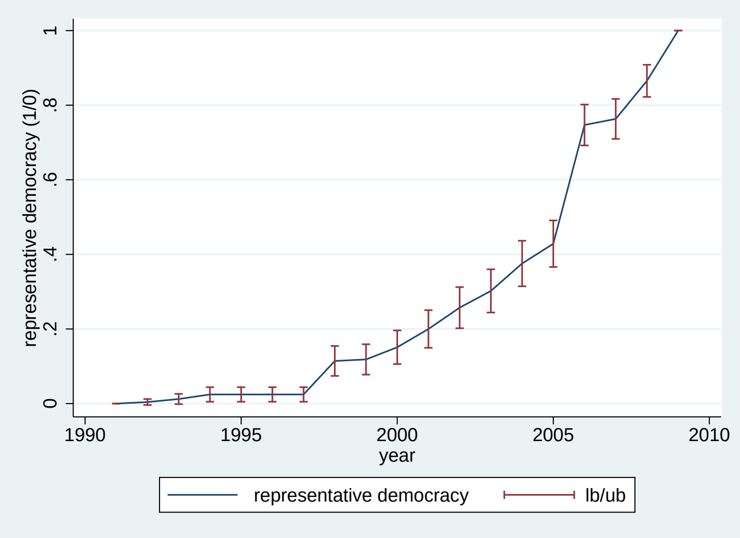


Naturalization Rate Over Time



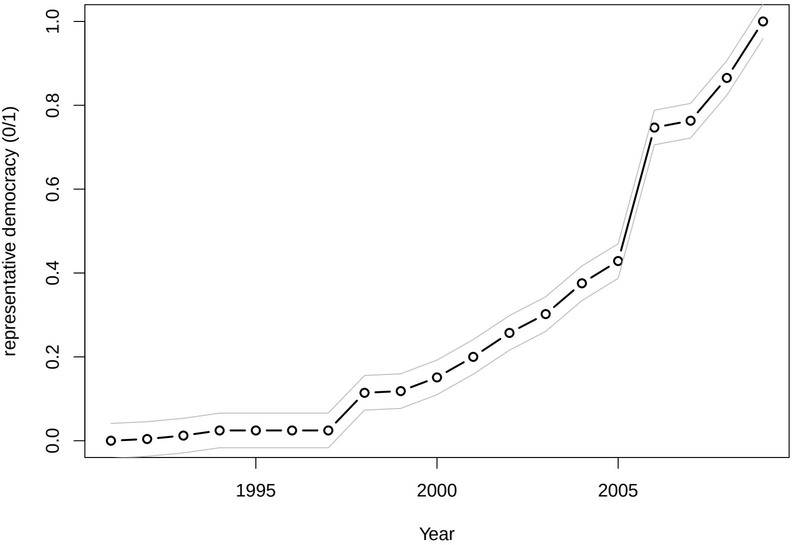
`xtgraph nat_rate`

Representative Democracy Over Time



```
xtgraph repdem
```

Representative Democracy Over Time



Adding Time Effects

- Reconsider our unobserved effects model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Fixed effects assumption: $E[\varepsilon_{it}|\mathbf{x}_i, c_i] = 0,\ t = 1, 2, ..., T$: 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 $\mathbf{x}_{it}$.
    - 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

Fixed Effects Regression: Adding Time Effects

- Linear time trend:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Linear time trend common to all units

- Time fixed effects:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + t_t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

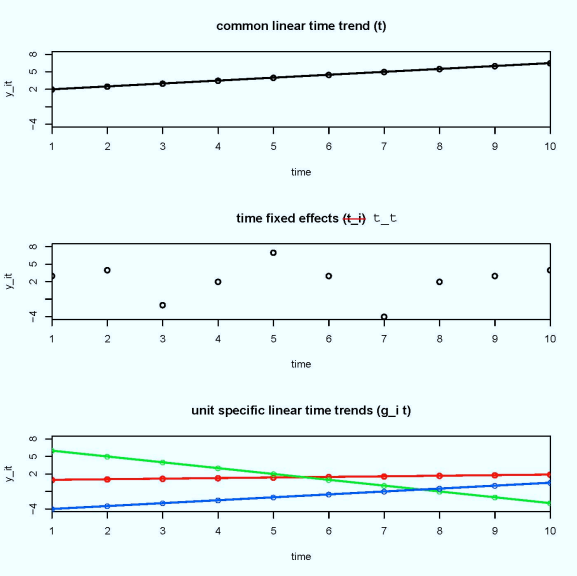
  - Common shock in each time period
  - Generalized difference-in-differences model

- Unit specific linear time trends:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + g_i \cdot t + t_t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Linear time trends that vary by unit

Modeling Time Effects



Modeling Time Effects

```
> d$time <- as.numeric(d$year)
> d[1:21,c("muniID","muni_name","year","time")]
   muniID      muni_name year time
1       1    Aeugst A.A. 1991    1
2       1    Aeugst A.A. 1992    2
3       1    Aeugst A.A. 1993    3
4       1    Aeugst A.A. 1994    4
5       1    Aeugst A.A. 1995    5
6       1    Aeugst A.A. 1996    6
7       1    Aeugst A.A. 1997    7
8       1    Aeugst A.A. 1998    8
9       1    Aeugst A.A. 1999    9
10      1    Aeugst A.A. 2000   10
11      1    Aeugst A.A. 2001   11
12      1    Aeugst A.A. 2002   12
13      1    Aeugst A.A. 2003   13
14      1    Aeugst A.A. 2004   14
15      1    Aeugst A.A. 2005   15
16      1    Aeugst A.A. 2006   16
17      1    Aeugst A.A. 2007   17
18      1    Aeugst A.A. 2008   18
19      1    Aeugst A.A. 2009   19
20      2 Affoltern A.A. 1991    1
21      2 Affoltern A.A. 1992    2
```

Fixed Effects Regression: 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 Regression: 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 Regression: Linear Time Trend

```
> mod_fe <- plm(nat_rate~repdem+time,data=d,model="within")
> coeftest(mod_fe, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

       Estimate Std. Error t value  Pr(>|t|)
repdem  0.82479    0.25853  3.1903  0.001431 **
time    0.23137    0.01714 13.4987 < 2.2e-16 ***
```

Fixed Effects Regression: 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)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   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 Regression: Year Fixed Effects

```
> mod_fe <- plm(nat_rate˜repdem+year,data=d,model="within")
> coeftest(mod_fe, vcov=function(x) vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

          Estimate Std. Error t value  Pr(>|t|)
repdem     1.20366    0.30253  3.9786 7.043e-05 ***
year1992  0.38292    0.17197  2.2266   0.02602 *
year1993  0.27898    0.15110  1.8463   0.06492 .
year1994  0.70341    0.16712  4.2089 2.617e-05 ***
year1995  0.74591    0.17827  4.1841 2.919e-05 ***
year1996  0.89693    0.18345  4.8892 1.049e-06 ***
year1997  0.97570    0.18661  5.2285 1.788e-07 ***
year1998  1.21550    0.22506  5.4007 6.988e-08 ***
year1999  1.08051    0.21430  5.0419 4.794e-07 ***
year2000  2.23993    0.23968  9.3457 < 2.2e-16 ***
year2001  1.75531    0.24790  7.0807 1.662e-12 ***
year2002  3.82573    0.32672 11.7096 < 2.2e-16 ***
year2003  3.37837    0.32664 10.3428 < 2.2e-16 ***
year2004  3.53176    0.34285 10.3012 < 2.2e-16 ***
year2005  3.20837    0.31097 10.3171 < 2.2e-16 ***
year2006  3.92057    0.39023 10.0468 < 2.2e-16 ***
year2007  2.77646    0.36884  7.5276 6.237e-14 ***
year2008  3.84780    0.40135  9.5872 < 2.2e-16 ***
year2009  2.22388    0.41997  5.2953 1.246e-07 ***
---
```

Unit Specific Time Trends Often Eliminate "Results"

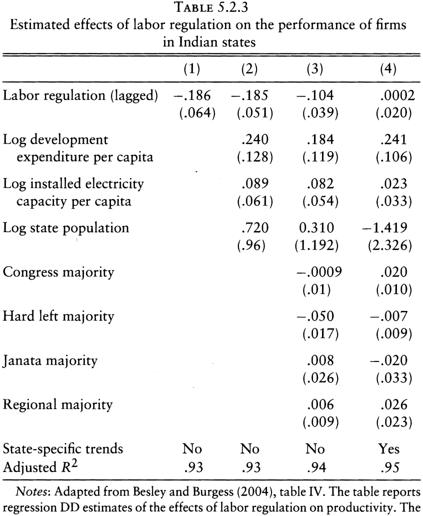


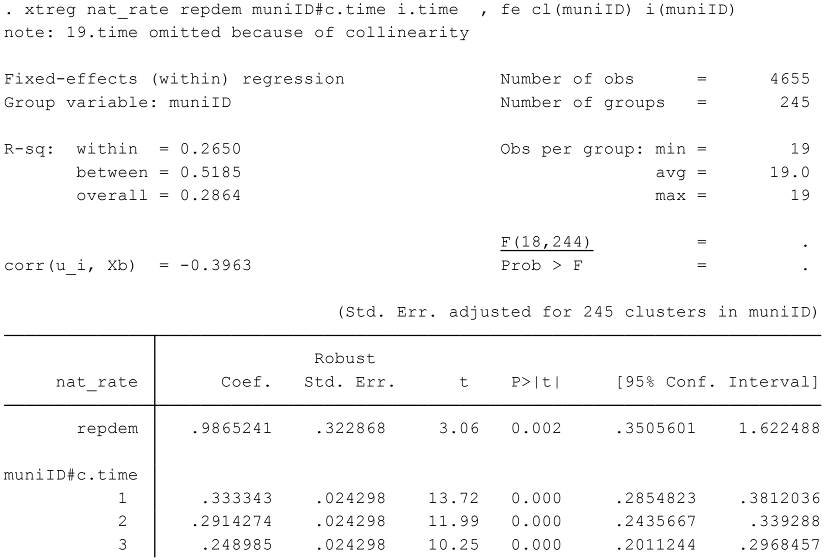
TABLE 5.2.3
Estimated effects of labor regulation on the performance of firms in Indian states

	(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

"labor regulation increased in states where output was declining anyway"

Fixed Effects Regression: 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)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   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
```

Fixed Effects Regression: Unit Specific Time Trends

```
> mod_fe <- plm(nat_rate˜
repdem+muniID*time+year,data=d,model="within")
> coeftest(mod_fe, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                  Estimate  Std. Error      t value  Pr(>|t|)
repdem           0.9865241    .322868    3.0634545  2.043e-05 ***
muniID2:time    -4.1916e-02  2.0515e-09 -2.0432e+07 < 2.2e-16 ***
muniID3:time    -8.4358e-02  2.1145e-09 -3.9896e+07 < 2.2e-16 ***
.
```

Unit Specific Quadratic Time Trends

```
. xtreg nat_rate repdem muniID#c.time muniID#c.time2 i.time,  fe cl(muniID) i(muniID)
note: 3954.muniID#c.time omitted because of collinearity
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  = 0.3138                         Obs per group: min =        19
       between = 0.4203                                        avg =      19.0
       overall = 0.2664                                        max =        19

                                                F(17,244)          =         .
corr(u_i, Xb)  = -0.6410                        Prob > F           =         .

                                 (Std. Err. adjusted for 245 clusters in muniID)
--------------------------------------------------------------------------------
               |               Robust
      nat_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
        repdem |   1.222728   .3804359     3.21   0.001     .4733704    1.972085
               |
 muniID#c.time |
             1 |  -.6710838   .1160209    -5.78   0.000     -.899614   -.4425535
             2 |   .7006871   .1160209     6.04   0.000     .4721568    .9292174
             3 |   .6175381   .1160209     5.32   0.000     .3890078    .8460683
```

Unit Specific Quadratic Time Trends

```
> d$time2 <- d$time^2
> mod_fe <- plm(nat_rate~repdem+
muniID*time+muniID*time^2+year,data=d,model="within")
> coeftest(mod_fe, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                  Estimate  Std. Error     t value  Pr(>|t|)
repdem           1.22272779   .3804359   3.212323  1.023e-05 ***
muniID2:time     1.37177084  1.034e-09    .0344+07 < 3.4e-16 ***
muniID2:time2   -0.07068432  2.2034e-09 -1.234e+07 < 5.6e-16 ***
.
```

Distributed Lag Model

$$y_{it} = x_{it}\boldsymbol{\beta}_0 + x_{it-1}\boldsymbol{\beta}_1 + x_{it-2}\boldsymbol{\beta}_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in x may occur with a lag
  - 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
    - $y_{t-1} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_t \quad = (m+1)\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+1} = m\boldsymbol{\beta}_0 + (m+1)\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+2} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + (m+1)\boldsymbol{\beta}_2 + c_i$
    - $y_{t+3} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$

Distributed Lag Model

$$y_{it} = x_{it}\boldsymbol{\beta}_0 + x_{it-1}\boldsymbol{\beta}_1 + x_{it-2}\boldsymbol{\beta}_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in x may occur with a lag
  - 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
    - $y_{t-1} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_t \quad = (m+1)\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+1} = m\boldsymbol{\beta}_0 + (m+1)\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+2} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + (m+1)\boldsymbol{\beta}_2 + c_i$
    - $y_{t+3} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$

- $\boldsymbol{\beta}_0 = y_t - y_{t-1}$ immediate change in y due to temporary one-unit increase in x (impact propensity)

Distributed Lag Model

$$y_{it} = x_{it}\boldsymbol{\beta}_0 + x_{it-1}\boldsymbol{\beta}_1 + x_{it-2}\boldsymbol{\beta}_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in x may occur with a lag
    - 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
        - $y_{t-1} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
        - $y_t \quad = (m + 1)\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
        - $y_{t+1} = m\boldsymbol{\beta}_0 + (m + 1)\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
        - $y_{t+2} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + (m + 1)\boldsymbol{\beta}_2 + c_i$
        - $y_{t+3} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$

- $\boldsymbol{\beta}_1 = y_{t+1} - y_{t-1}$ change in y one period after temporary one-unit increase in x

Distributed Lag Model

$$y_{it} = x_{it}\boldsymbol{\beta}_0 + x_{it-1}\boldsymbol{\beta}_1 + x_{it-2}\boldsymbol{\beta}_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in x may occur with a lag
  - 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
    - $y_{t-1} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_t \quad = (m+1)\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+1} = m\boldsymbol{\beta}_0 + (m+1)\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+2} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + (m+1)\boldsymbol{\beta}_2 + c_i$
    - $y_{t+3} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$

- $\boldsymbol{\beta}_2 = y_{t+2} - y_{t-1}$ change in y two periods after temporary one-unit increase in x

Distributed Lag Model

$$y_{it} = x_{it}\boldsymbol{\beta}_0 + x_{it-1}\boldsymbol{\beta}_1 + x_{it-2}\boldsymbol{\beta}_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in x may occur with a lag
  - 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
    - $y_{t-1} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_t \quad = (m+1)\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+1} = m\boldsymbol{\beta}_0 + (m+1)\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+2} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + (m+1)\boldsymbol{\beta}_2 + c_i$
    - $y_{t+3} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$

- $y_{t+3} = y_{t-1}$ change in y is zero three periods after temporary one-unit increase in x

Distributed Lag Model

$$y_{it} = x_{it}\boldsymbol{\beta}_0 + x_{it-1}\boldsymbol{\beta}_1 + x_{it-2}\boldsymbol{\beta}_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- 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 $x_s = m+1, s \geq t)$
    - $y_{t-1} = m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_t \quad = (m+1)\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+1} = (m+1)\boldsymbol{\beta}_0 + (m+1)\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t+2} = (m+1)\boldsymbol{\beta}_0 + (m+1)\boldsymbol{\beta}_1 + (m+1)\boldsymbol{\beta}_2 + c_i$
    - $y_{t+3} = (m+1)\boldsymbol{\beta}_0 + (m+1)\boldsymbol{\beta}_1 + (m+1)\boldsymbol{\beta}_2 + c_i$
- After one period y has increased by $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1$, after two periods, y has increased by $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$, and there are no further increases after two periods
- Long-run increase in y: $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$ (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
```

Lagged Effects of Direct Democracy

```
> mod_lag    <- plm(nat_rate˜repdem+
  lag(repdem,1)+lag(repdem,2)+lag(repdem,3)+
+                     year,data=d,model="within")
> coeftest(mod_lag, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))
t test of coefficients:
                Estimate Std. Error t value  Pr(>|t|)
repdem          0.636480   0.358658  1.7746  0.076044 .
lag(repdem, 1)  1.201065   0.422508  2.8427  0.004498 **
lag(repdem, 2) -0.164869   0.468783 -0.3517  0.725087
lag(repdem, 3) -0.524521   0.410152 -1.2788  0.201033
year1995        0.031281   0.204172  0.1532  0.878244
year1996        0.188603   0.210878  0.8944  0.371183

# long run effect
> sum(coef(mod_lag)[1:4])
[1] 1.148156
```

Long-run Effect of Direct Democracy

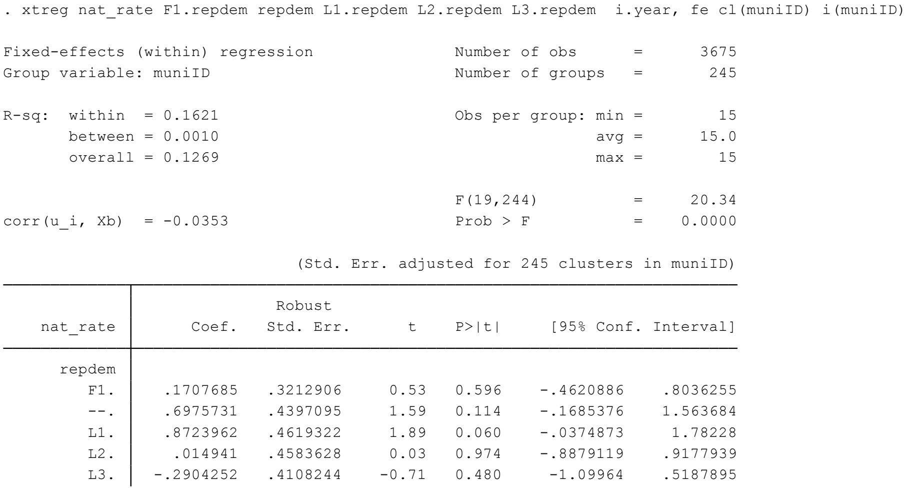
```
. lincom repdem + L1.repdem + L2.repdem + L3.repdem

 ( 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

$$y_{it} = x_{it+1}\boldsymbol{\beta}_{-1} + x_{it}\boldsymbol{\beta}_0 + x_{it-1}\boldsymbol{\beta}_1 + x_{it-2}\boldsymbol{\beta}_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Can use estimate of $\boldsymbol{\beta}_{-1}$ to test for anticipation effects
  - Consider temporary increase in x_{it} from level m to $m+1$ at t
    - $y_{t-2} = \boldsymbol{\beta}_{-1}m + m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
    - $y_{t-1} = \boldsymbol{\beta}_{-1}(m+1) + m\boldsymbol{\beta}_0 + m\boldsymbol{\beta}_1 + m\boldsymbol{\beta}_2 + c_i$
- Anticipation effect: $\boldsymbol{\beta}_{-1} = y_{t-1} - y_{t-2}$ 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 $\boldsymbol{\beta}_{-1}$ 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
```

Leads and Lags

```
> d <- ddply(
+   d, .(muniID), transform,
+   lead_repdem = c( repdem[-1],NA )
+ )
> d <- plm.data(d, indexes = c("muniID", "year"))
> mod_lagleads    <- plm(nat_rate~lead_repdem+
repdem+lag(repdem,1)+lag(repdem,2)+lag(repdem,3)+
+                      year,data=d,model="within")
> coeftest(mod_lagleads, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                Estimate Std. Error t value  Pr(>|t|)
lead_repdem     0.170768   0.320634  0.5326 0.5943479
repdem          0.697573   0.438811  1.5897 0.1119975
lag(repdem, 1)  0.872396   0.460988  1.8924 0.0585159 .
lag(repdem, 2)  0.014941   0.457426  0.0327 0.9739451
lag(repdem, 3) -0.290425   0.409985 -0.7084 0.4787574
year1995        0.032882   0.204374  0.1609 0.8721896
```

The Autor Test

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

$$y_{it} = D_{it+2}\boldsymbol{\beta}_{-2} + D_{it+1}\boldsymbol{\beta}_{-1} + D_{it}\boldsymbol{\beta}_{0} + D_{it-1}\boldsymbol{\beta}_{1} + D_{it-2}\boldsymbol{\beta}_{2} + c_i + \varepsilon_{it}$$

- Interpretation of coefficients:
    - Leads $\boldsymbol{\beta}_{-1}$, $\boldsymbol{\beta}_{-2}$, etc. test for anticipation effects
    - Switch $\boldsymbol{\beta}_0$ tests for immediate effect
    - Lags $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, 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 |
      +-------------------------------------------------------------------------------------------------+
```

Lags and Leads of Switch to Representative Democracy

```
> d[970:989,c(1:3,5,12:ncol(d))]
    muniID year muni_name repdem switcht lag1 lag2 lag3 lead1 lead2 lead3 lead4 lead5
970    220 1991 Hagenbuch      0       0    0    0    0     0     0     0     0     0
971    220 1992 Hagenbuch      0       0    0    0    0     0     0     0     0     0
972    220 1993 Hagenbuch      0       0    0    0    0     0     0     0     0     0
973    220 1994 Hagenbuch      0       0    0    0    0     0     0     0     0     0
974    220 1995 Hagenbuch      0       0    0    0    0     0     0     0     0     0
975    220 1996 Hagenbuch      0       0    0    0    0     0     0     0     0     0
976    220 1997 Hagenbuch      0       0    0    0    0     0     0     0     0     0
977    220 1998 Hagenbuch      0       0    0    0    0     0     0     0     0     1
978    220 1999 Hagenbuch      0       0    0    0    0     0     0     0     1     0
979    220 2000 Hagenbuch      0       0    0    0    0     0     0     1     0     0
980    220 2001 Hagenbuch      0       0    0    0    0     0     1     0     0     0
981    220 2002 Hagenbuch      0       0    0    0    0     1     0     0     0     0
982    220 2003 Hagenbuch      1       1    0    0    0     0     0     0     0     0
983    220 2004 Hagenbuch      1       0    1    0    0     0     0     0     0     0
984    220 2005 Hagenbuch      1       0    0    1    0     0     0     0     0     0
985    220 2006 Hagenbuch      1       0    0    0    1     0     0     0     0     0
986    220 2007 Hagenbuch      1       0    0    0    1     0     0     0     0     0
987    220 2008 Hagenbuch      1       0    0    0    1     0     0     0     0     0
988    220 2009 Hagenbuch      1       0    0    0    1     0     0     0     0     0
989    224 1991   Pfungen      0       0    0    0    0     0     0     0     0     0
```

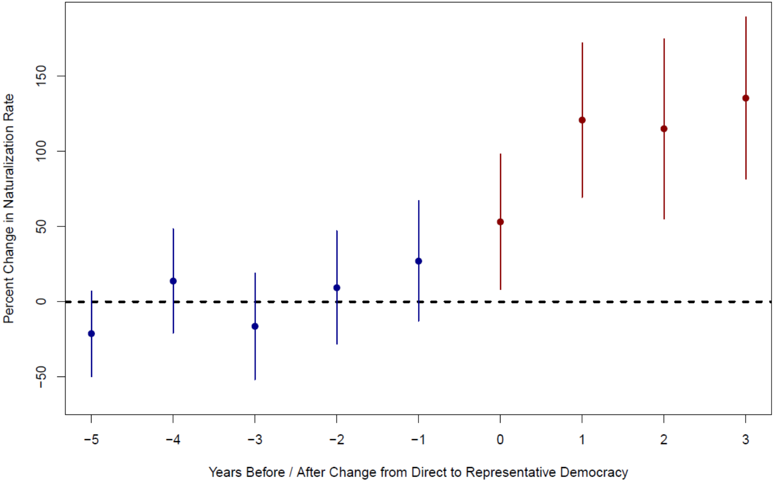
Dynamic Effect of Switching to Representative Democracy

```
. 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)
------------------------------------------------------------------------------
             |               Robust
    nat_rate |      Coef.   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
```

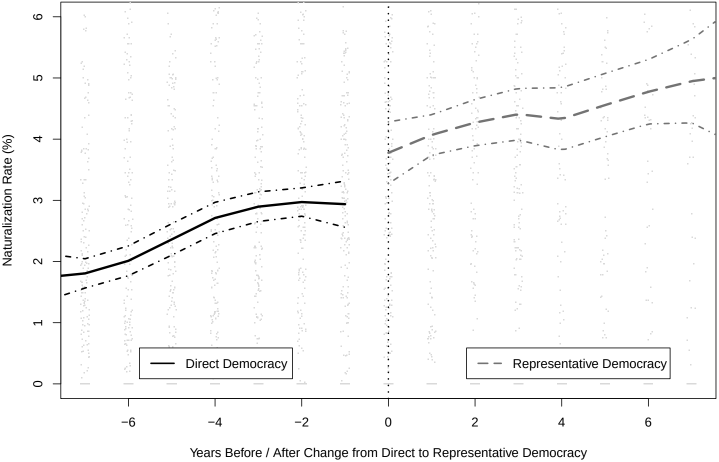
Dynamic Effect of Switching to Representative Democracy

```
> mod_all    <- plm(nat_rate˜lag3+lag2+lag1+switcht+
lead1+lead2+lead3+lead4+lead5+
+                          year,data=d,model="within")
> coeftest(mod_all, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

          Estimate Std. Error t value  Pr(>|t|)
lag3      1.160345   0.506989  2.2887 0.0221442 *
lag2      1.743682   0.538419  3.2385 0.0012105 **
lag1      1.881663   0.487013  3.8637 0.0001133 ***
switcht   0.756479   0.427751  1.7685 0.0770463 .
lead1     0.213876   0.389191  0.5495 0.5826635
lead2     0.084368   0.356799  0.2365 0.8130891
lead3     0.144045   0.318756  0.4519 0.6513661
lead4     0.075019   0.298425  0.2514 0.8015287
lead5    -0.094241   0.259448 -0.3632 0.7164439
year1992  0.385289   0.172172  2.2378 0.0252829 *
```

Dynamic Effect of Switching to Representative Democracy



Switching Plot



Lagged Dependent Variable

$$y_{it} = \alpha\, y_{it-1} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- y_{it} could be capital stock of firm i at time t, and α the capital depreciation rate
- Models with unit fixed effects and lagged y do not produce consistent estimators!
  - after taking first differences to eliminate c_i, the differenced residual $\Delta\varepsilon_{it}$ is correlated with the lagged dependent variable Δy_{it-1} by construction
- We might use past levels y_{it-2} as an instrument for Δy_{it-1}, but this requires strong assumptions (e.g. no serial correlation in ε_{it})

Heterogeneous Treatment Effects

- So far we have assumed that the treatment effect is constant across units
- Can allow for heterogeneous treatment effects by including interaction of treatment with other regressors

$$y_{it} = treat_{it}\alpha_0 + (treat_{it} \cdot x_{it})\alpha_1 + x_{it}\boldsymbol{\beta} + c_i + t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Often the treatment is interacted with a time-invariant regressor:

$$y_{it} = treat_{it}\alpha_0 + (treat_{it} \cdot x_i)\alpha_1 + c_i + t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Note: The lower order term on the time-invariant x_i is collinear with the fixed effects and drops out

Heterogeneous Effect of Direct Democracy