

PANEL DATA ANALYSIS USING STATA

Tim Birkenbach
(MEA)

19/10/2017

(I) Basic panel commands in Stata

- `xtset`
- `xtdescribe`
- `reshape`

(II) Panel analysis popular in Economics

- Pooled OLS
- Fixed-Effects Model & Difference-in-Difference
- Random Effects Model

mergeid	wave	health	female
AT-004855-01	1	1	0
AT-004855-01	2	1	0
AT-004855-02	1	4	1
AT-004855-02	2	2	1
AT-004855-02	3	.	1
AT-004855-02	4	4	1
AT-004855-02	5	2	1
AT-004855-02	6	1	1

```
// declare panel data structure  
xtset panelvar timevar
```

```
xtset id wave  
  panel variable:  id (unbalanced)  
  time variable:  wave, 1 to 6, but with gaps  
                 delta: 1 unit
```

```
. xtdescribe
```

```

      id:  1, 2, ..., 120568          n =      120568
     wave: 1, 2, ..., 6              T =           6
      Delta(wave) = 1 unit
      Span(wave)  = 6 periods
      (id*wave uniquely identifies each observation)

```

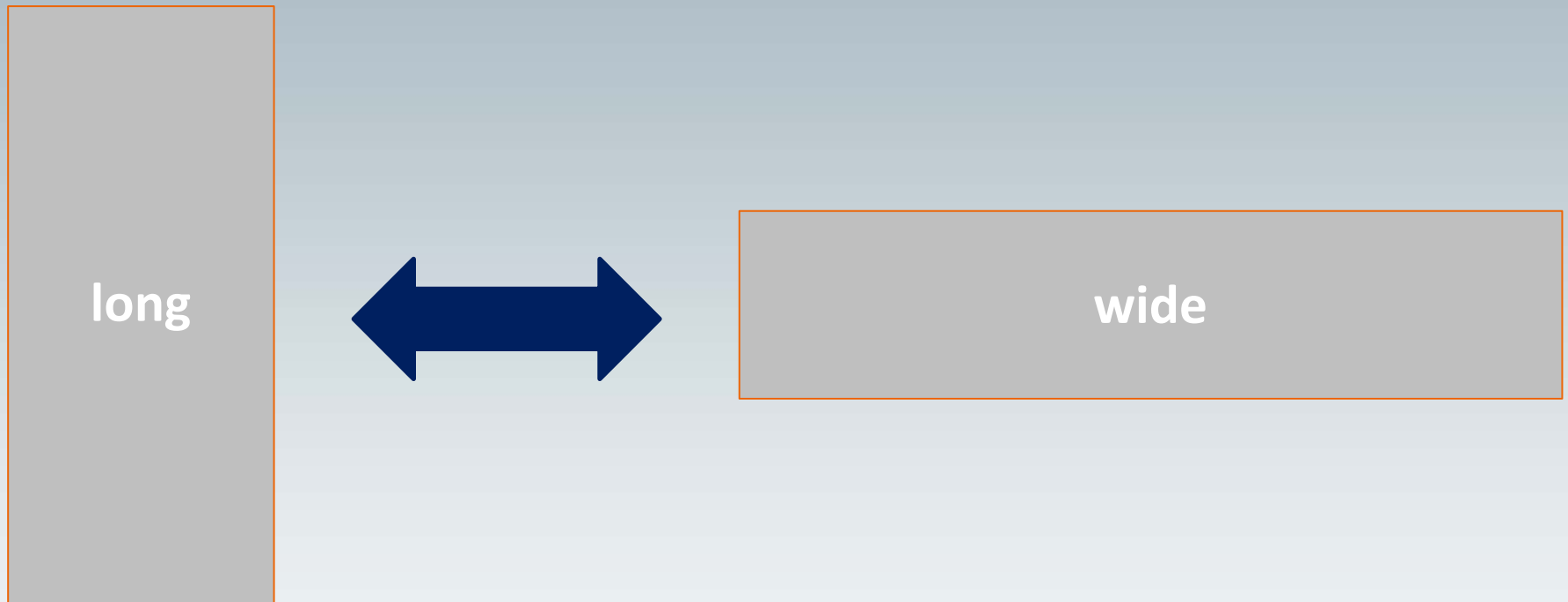
```

Distribution of T_i:  min      5%      25%      50%      75%      95%      max
                    1         1         1         2         3         6         6

```

Freq.	Percent	Cum.	Pattern
18089	15.00	15.00	...111
15254	12.65	27.6511
12242	10.15	37.811
10195	8.46	46.26	...1..
7772	6.45	52.711.
6577	5.46	58.17	1.....
6372	5.28	63.45	111111
5703	4.73	68.18	...11.
3990	3.31	71.49	.1....
34374	28.51	100.00	(other patterns)
120568	100.00		XXXXXX

reshape



reshape

```
// reshape format: long -> wide
```

```
reshape wide health, i(mergeid) j(wave)
```

mergeid	wave	health	female
AT-004855-01	1	1	0
AT-004855-01	2	1	0
AT-004855-02	1	4	1
AT-004855-02	2	2	1
AT-004855-02	3	.	1
AT-004855-02	4	4	1
AT-004855-02	5	2	1
AT-004855-02	6	1	1



mergeid	health1	health2	health3	health4	health5	health6	female
AT-004855-01	1	1	0
AT-004855-02	4	2	.	4	2	1	1

reshape

```
// reshape format: wide -> long  
reshape long health, i(mergeid) j(wave)
```

mergeid	health1	health2	health3	health4	health5	health6	female
AT-004855-01	1	1	0
AT-004855-02	4	2	.	4	2	1	1



mergeid	wave	health	female
AT-004855-01	1	1	0
AT-004855-01	2	1	0
AT-004855-01	3	.	0
AT-004855-01	4	.	0
AT-004855-01	5	.	0
AT-004855-01	6	.	0
AT-004855-02	1	4	1
AT-004855-02	2	2	1
AT-004855-02	3	.	1
AT-004855-02	4	4	1
AT-004855-02	5	2	1
AT-004855-02	6	1	1



reshape

```
gen original = 1  
reshape wide health original, i(mergeid) j(wave)  
reshape long health original, i(mergeid) j(wave)  
keep if original == 1
```

mergeid	wave	health	female	original
AT-004855-01	1	1	0	1
AT-004855-01	2	1	0	1
AT-004855-01	3	.	0	.
AT-004855-01	4	.	0	.
AT-004855-01	5	.	0	.
AT-004855-01	6	.	0	.
AT-004855-02	1	4	1	1
AT-004855-02	2	2	1	1
AT-004855-02	3	.	1	1
AT-004855-02	4	4	1	1
AT-004855-02	5	2	1	1
AT-004855-02	6	1	1	1





(I) Basic panel commands in Stata

- `xtset`
- `xtdescribe`
- `reshape`

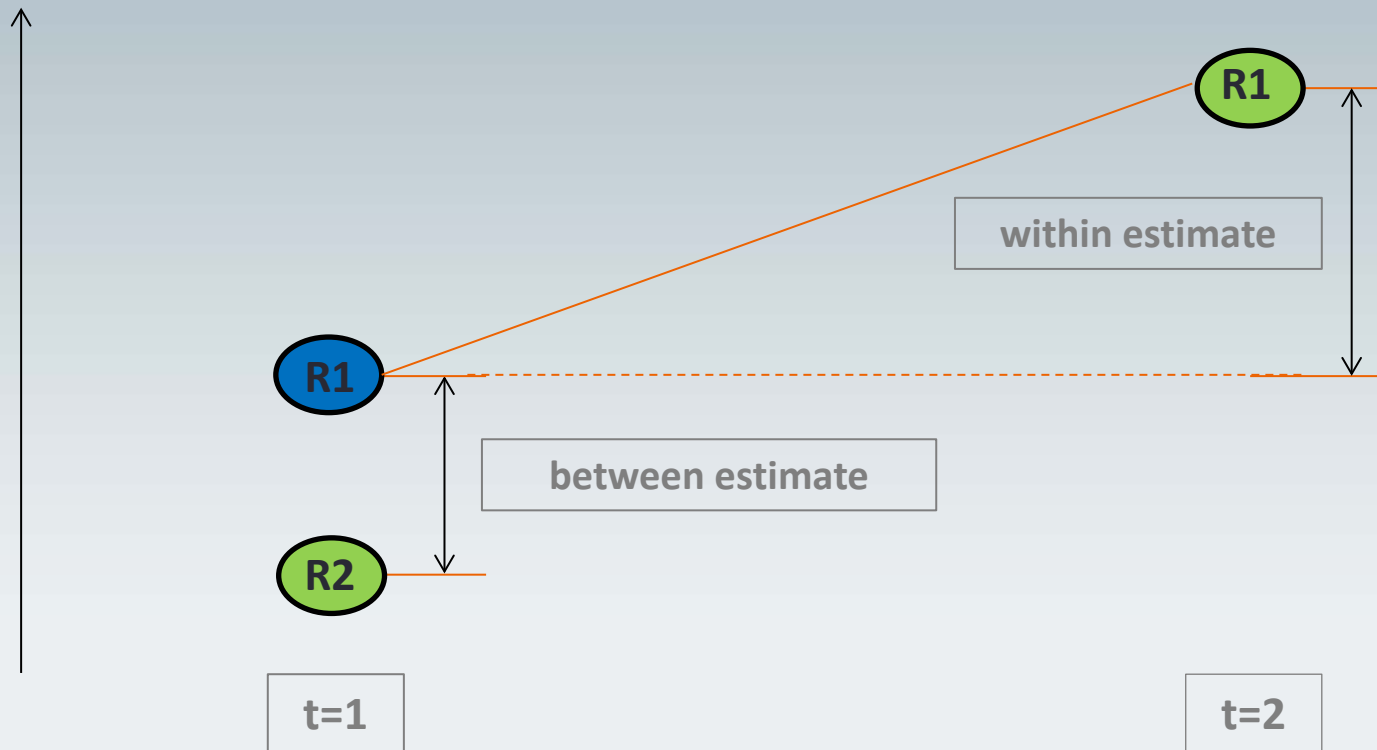
(II) Panel analysis popular in Economics

- Pooled OLS
- Fixed-Effects Model & Difference-in-Difference
- Random Effects Model

between & within estimator

-  X (working)
-  X (retired)

Y (health)



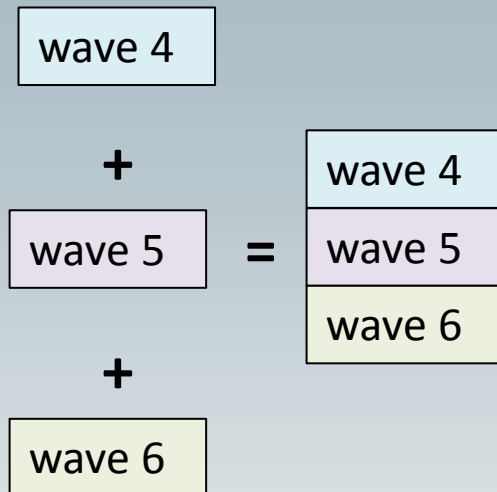
(I) Basic panel commands in Stata

- `xtset`
- `xtdescribe`
- `reshape`

(II) Panel analysis popular in Economics

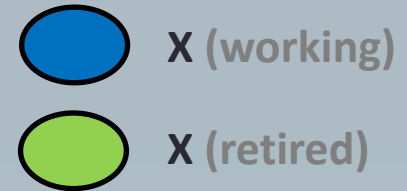
- Pooled OLS
- Fixed-Effects Model & Difference-in-Difference
- Random Effects Model

Pooled Ordinary Least Squares (POLS)

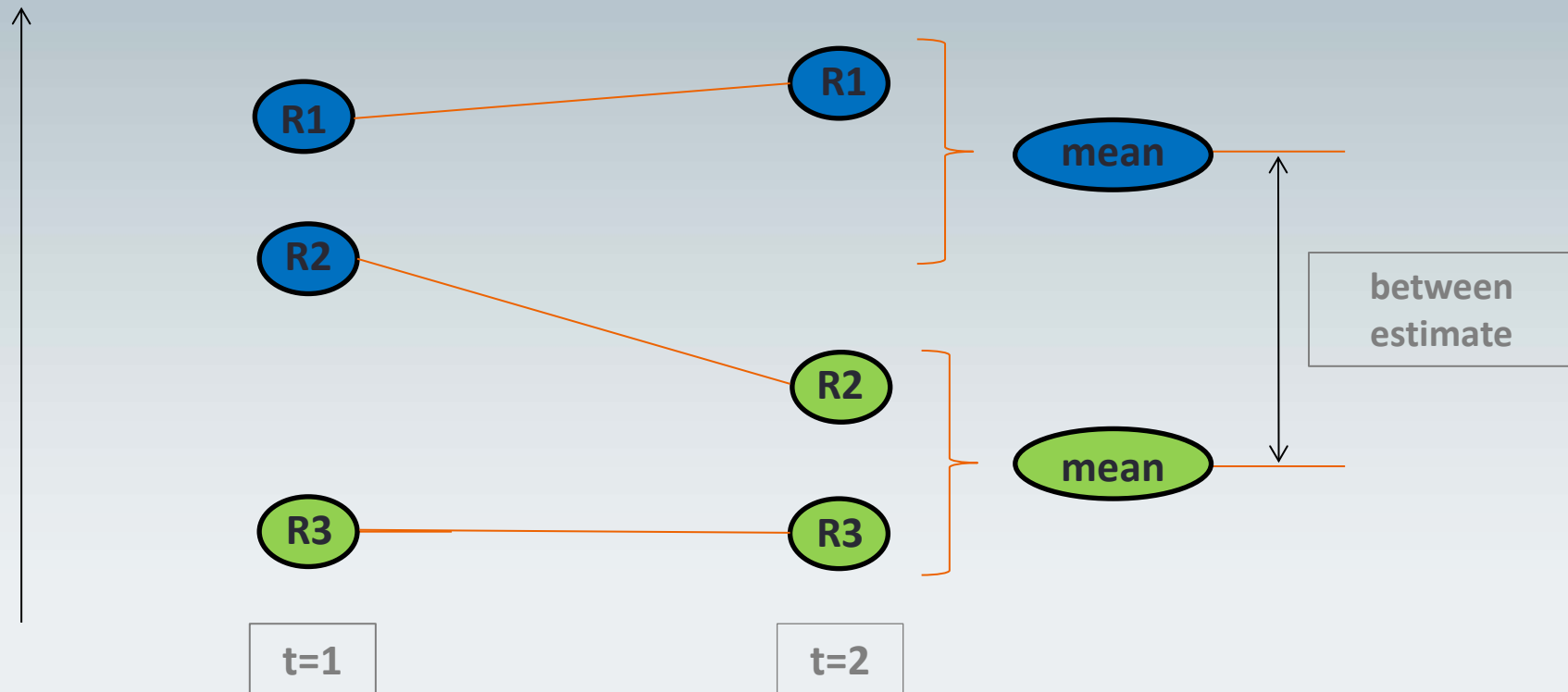


mergeid	wave	health	retired
AT-077700-01	4	3	1
AT-077700-01	5	3	1
AT-077700-01	6	4	1
AT-077718-01	4	4	0
AT-077718-01	5	3	0
AT-077788-01	4	3	0
AT-077788-01	5	4	1
AT-077788-01	6	4	1

Pooled Ordinary Least Squares (POLS)



Y (health)



Pooled Ordinary Least Squares (POLS)

$$y_{it} = \beta_0 + \delta_t + \beta_1 x_{it} + \varepsilon_{it}$$

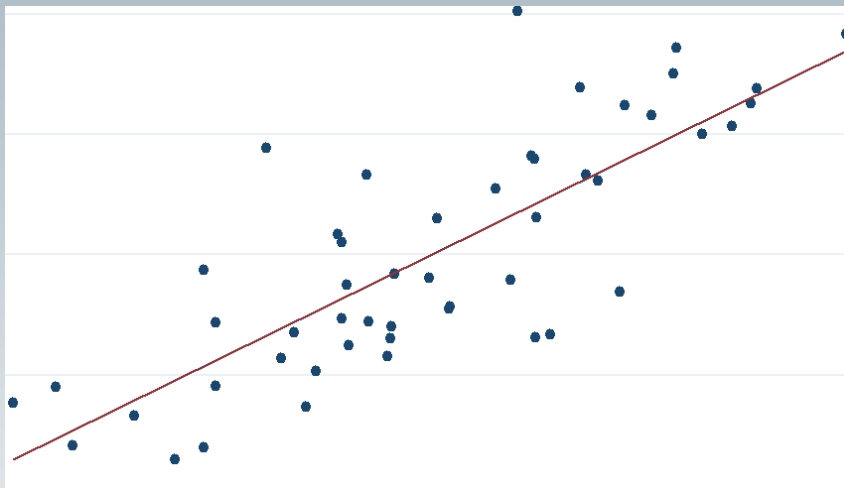
↑
multiple lines
per respondent i

↑
time effect

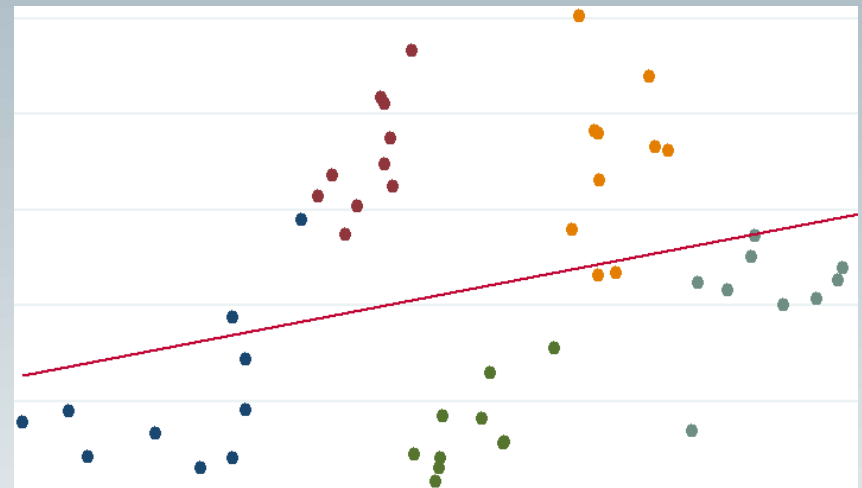
↑
 ε dependent ?

Pooled Ordinary Least Squares (POLS)

ε independent



ε dependent



- cluster by respondent
- serial correlation

Pooled Ordinary Least Squares (POLS)



```
// POLS
reg health retired

// + cluster robust inference
reg health retired          , cluster(id)

// + linear time trend
reg health retired wave     , cluster(id)

// + period effect
reg health retired i.wave   , cluster(id)
```

Pooled Ordinary Least Squares (POLS)

```
. reg health retired wave
```

Source	SS	df	MS	Number of obs	=	256,705
Model	9443.95479	2	4721.97739	F(2, 256702)	=	4130.01
Residual	293495.65	256,702	1.14333215	Prob > F	=	0.0000
Total	302939.605	256,704	1.18011252	R-squared	=	0.0312
				Adj R-squared	=	0.0312
				Root MSE	=	1.0693

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
retired	-.3684334	.0042533	-86.62	0.000	-.3767698 -.3600971
wave	-.0255196	.0012406	-20.57	0.000	-.0279512 -.0230881
_cons	3.161913	.0058617	539.42	0.000	3.150424 3.173402


```
. reg health retired wave, cluster(id)
```

Linear regression

Number of obs	=	256,705
F(2, 119003)	=	2644.20
Prob > F	=	0.0000
R-squared	=	0.0312
Root MSE	=	1.0693

(Std. Err. adjusted for 119,004 clusters in id)

health	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
retired	-.3684334	.0055348	-66.57	0.000	-.3792816 -.3575852
wave	-.0255196	.0012858	-19.85	0.000	-.0280398 -.0229995
_cons	3.161913	.0065764	480.80	0.000	3.149024 3.174803

(I) Basic panel commands in Stata

- `xtset`
- `xtdescribe`
- `reshape`

(II) Panel analysis popular in Economics

- Pooled OLS
- Fixed-Effects Model & Difference-in-Difference
- Random Effects Model

Fixed Effects Estimation (FE)

if $Cov(x_{it}, \varepsilon_{it}) \neq 0$ for any t , POLS is biased

$$y_{it} = \beta_1 x_{it} + \varepsilon_{it}$$

$$y_{it} = \beta_1 x_{it} + \underbrace{u_i + e_{it}}_{\text{error decomposition}}$$

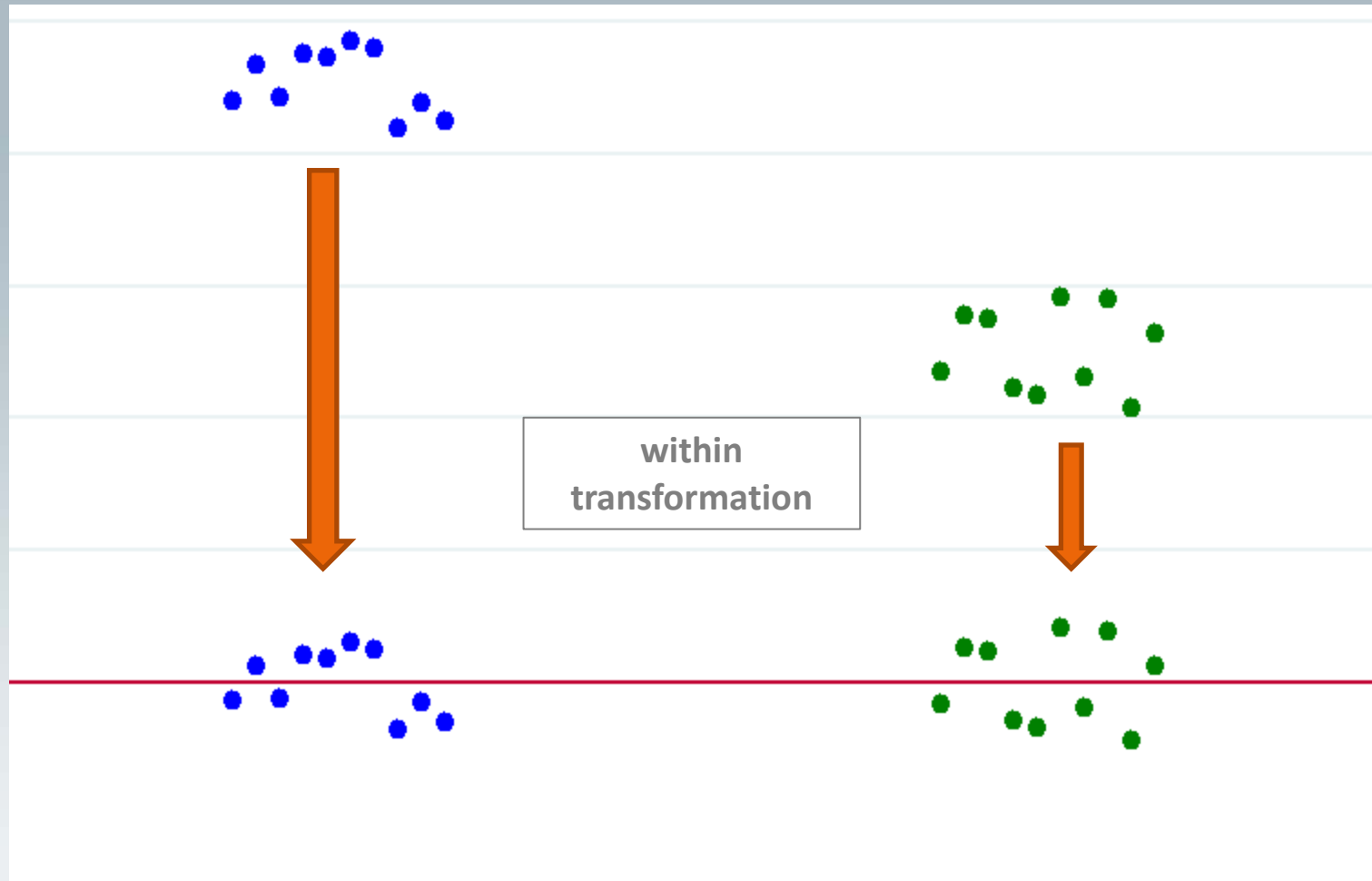
error decomposition

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (u_i - \bar{u}_i) + (e_{it} - \bar{e}_i) \leftarrow \text{within transformation}$$



$$\dot{y}_{it} = \beta_1 \dot{x}_{it} + \dot{e}_{it}$$

Fixed Effects Estimation (FE)

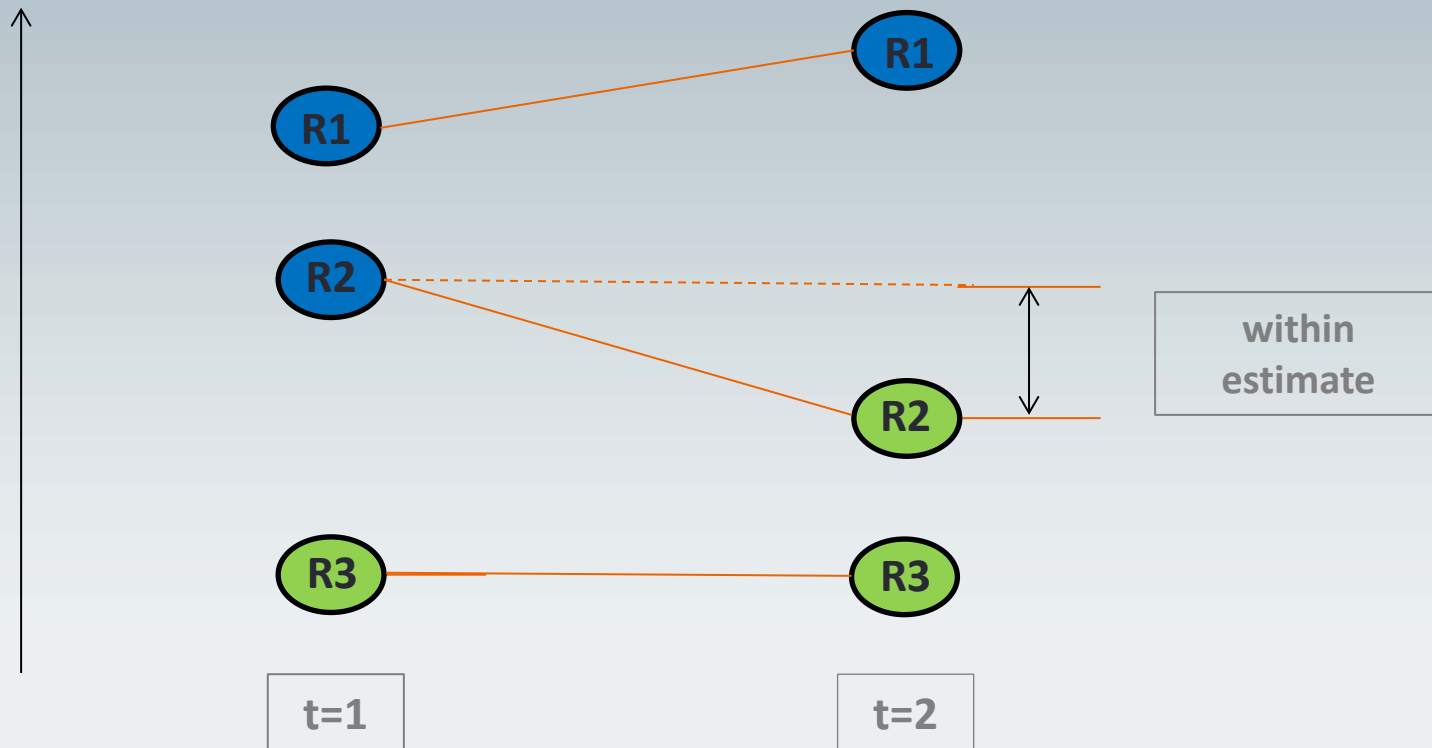
0



Fixed Effects Estimation (FE)

-  X (working)
-  X (retired)

Y (health)



Fixed Effects Estimation (FE)

```
// declare panel data structure  
xtset id wave
```

```
// FE  
xtreg health retired , fe
```

```
// + cluster robust inference  
xtreg health retired , fe cluster(id)
```

Fixed Effects Estimation (FE)

```
. xtreg health retired , fe cluster(id)

Fixed-effects (within) regression      Number of obs   =   189,835
Group variable: id                    Number of groups =    99,657

R-sq:                                  Obs per group:
    within = 0.0000                    min =           1
    between = 0.0399                   avg =           1.9
    overall = 0.0275                   max =           3



                                         F(1,99656)     =    0.28
                                         Prob > F        =    0.5987

corr(u_i, Xb) = 0.1794
```

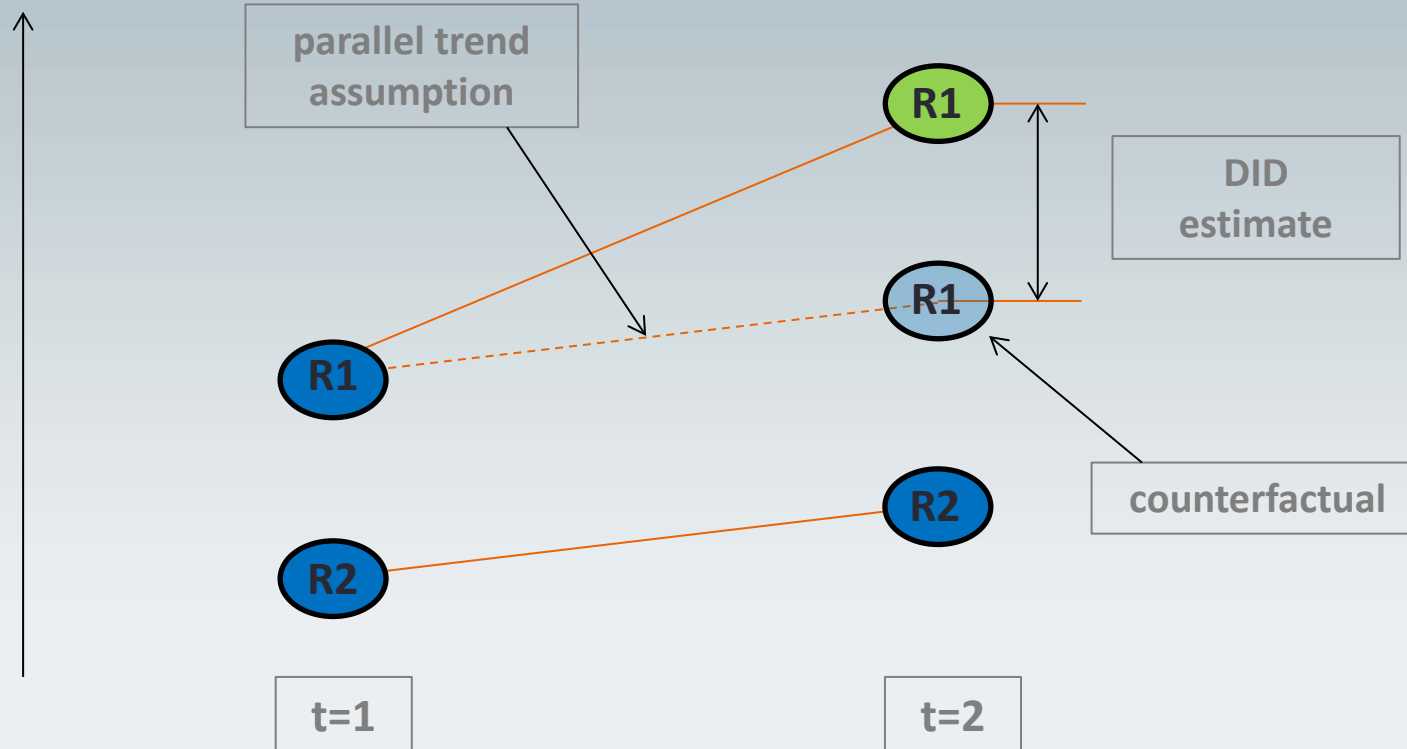
(Std. Err. adjusted for 99,657 clusters in id)

health	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
retired	-.0048006	.0091222	-0.53	0.599	-.0226799	.0130788
_cons	2.817331	.0052048	541.29	0.000	2.807129	2.827532
sigma_u	1.0221537					
sigma_e	.64192857					
rho	.71715368	(fraction of variance due to u_i)				

Difference-in-Difference (DID)

-  X (working)
-  X (retired)

Y (health)



Difference-in-Difference (DID)

period effect



$$y_{it} = \beta_1 x_{it} + \delta_t + \varepsilon_{it}$$
$$y_{it} = \beta_1 x_{it} + \delta_t + \underbrace{u_i + e_{it}}_{\varepsilon_{it}}$$

error decomposition

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (\delta_t - \bar{\delta}_t) + (u_i - \bar{u}_i) + (e_{it} - \bar{e}_i)$$

within
transformstion

$$\dot{y}_{it} = \beta_1 \dot{x}_{it} \dot{\delta}_t + \ddot{\delta}_t + \dot{e}_{it}$$

Difference-in-Difference (DID)

```
// declare panel data structure  
xtset id wave
```

```
// DID  
xtreg health retired i.wave, fe
```

```
// + cluster robust inference  
xtreg health retired i.wave, fe cluster(id)
```

Difference-in-Difference (DID)

```
. xtreg health retired 1.wave fe cluster(id)

Fixed-effects (within) regression      Number of obs   =   189,835
Group variable: id                    Number of groups =    99,657

R-sq:                                  Obs per group:
    within = 0.0074                    min =           1
    between = 0.0356                   avg =           1.9
    overall = 0.0082                   max =           3

corr(u_i, Xb) = -0.1516                F(3,99656)      =   213.28
                                        Prob > F        =    0.0000

                                        (Std. Err. adjusted for 99,657 clusters in id)
```

health	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
retired	.045631	.0093257	4.89	0.000	.0273528	.0639092
wave						
5	-.0604422	.004334	-13.95	0.000	-.0689367	-.0519476
6	-.1191141	.004721	-25.23	0.000	-.1283672	-.109861
_cons	2.851521	.0055432	514.42	0.000	2.840657	2.862386
sigma_u	1.031665					
sigma_e	.63954285					
rho	.7223907	(fraction of variance due to u_i)				

Within models (pros & cons)

// pro: within models can overcome problems that arises from unobserved heterogeneity
bias & attrition

// contra: within models only focus on a small fraction of the variance in the data
they usually have larger standard errors (lower efficiency)

// contra: within models cannot incorporate time-constant variables directly
e.g. gender, country, ethnical background

(I) Basic panel commands in Stata

- `xtset`
- `xtdescribe`
- `reshape`

(II) Panel analysis popular in Economics

- Pooled OLS
- Fixed-Effects Model & Difference-in-Difference
- Random Effects Model

Random Effects Estimation (RE)

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

$$\varepsilon_{it} = u_i + e_{it}$$

error decomposition

$$\text{Cov}(x_{it}, u_i) = 0, \quad t = 1, 2, \dots, T$$

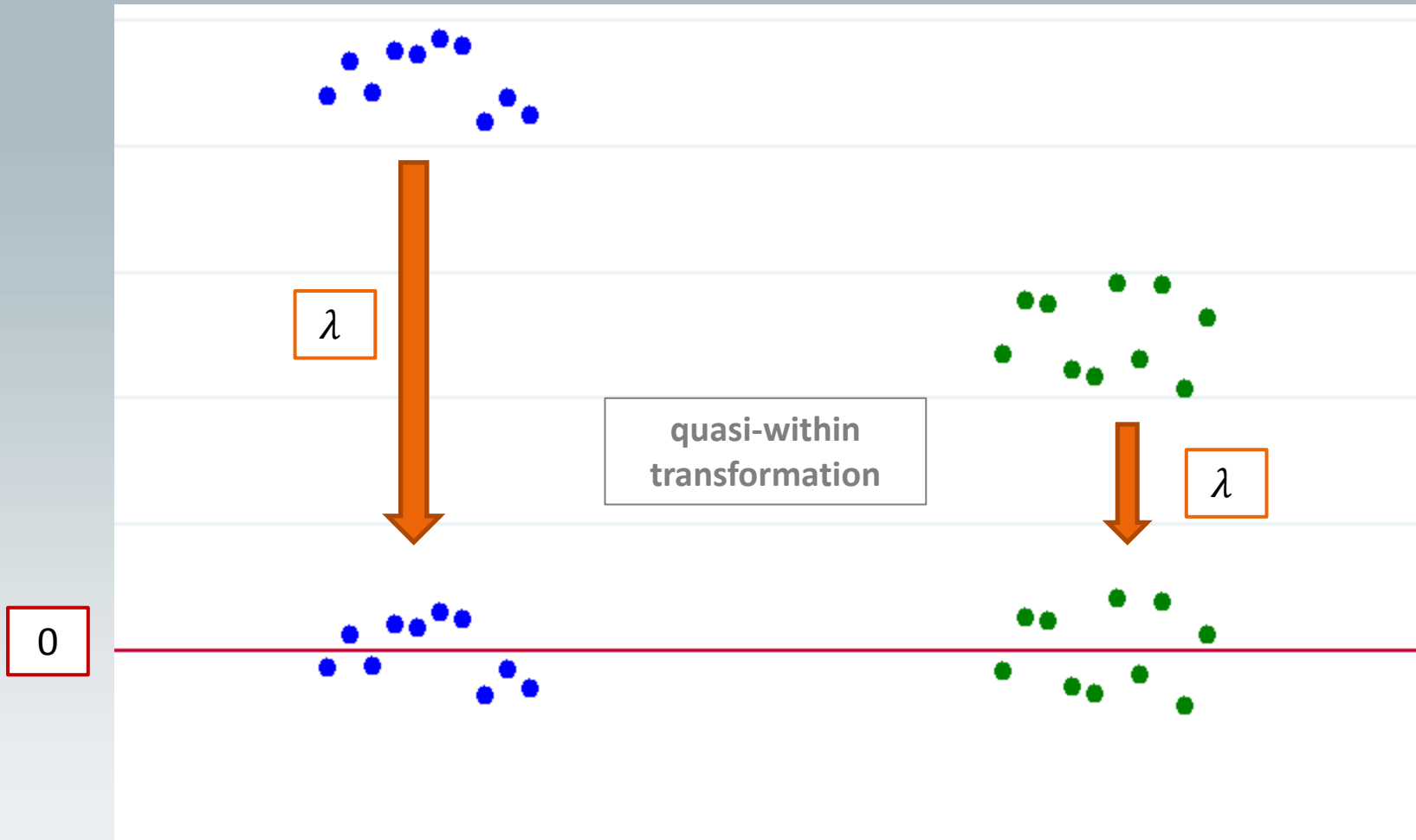
random effects assumption

$$\lambda = 1 - \sqrt{\frac{\sigma_e^2}{\sigma_e^2 + T\sigma_u^2}}$$

$$y_{it} - \lambda \bar{y}_i = \beta_0(1 - \lambda) + \beta_1(x_{it} - \lambda \bar{x}_i) + (\varepsilon_{it} - \lambda \bar{\varepsilon}_i)$$

quasi within
transformation
=
FGLS

Random Effects Estimation (RE)



Random Effects Estimation (RE)

$$y_{it} - \lambda \bar{y}_i = \beta_0(1 - \lambda) + \beta_1(x_{it} - \lambda \bar{x}_i) + (\varepsilon_{it} - \lambda \bar{\varepsilon}_i)$$

If $\lambda = 0$, then RE = POLS

If $\lambda = 1$, then RE = FE

β_{RE} lies in between β_{POLS} & β_{FE}

Random Effects Estimation (RE)

```
// declare panel data structure
xtset id wave

// RE
xtreg health retired          , re

// + time-constant explanatory variable
xtreg health retired female    , re

// + cluster robust inference & period effect
xtreg health retired female i.wave, re cluster(id)
```

Random Effects Estimation (RE)

```
. xtreg health retired female i.wave , re cluster(id)
```

```
Random-effects GLS regression      Number of obs   =   189,835
Group variable: id                 Number of groups =    99,657
```

```
R-sq:                               Obs per group:
  within = 0.0002                    min =          1
  between = 0.0420                   avg =         1.9
  overall = 0.0300                   max =          3
```

```
corr(u_i, X) = 0 (assumed)          Wald chi2(4)    =   2780.34
                                         Prob > chi2     =    0.0000
```

(Std. Err. adjusted for 99,657 clusters in id)

health	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
retired	-.2771256	.0055414	-50.01	0.000	-.2879865	-.2662647
female	-.092976	.0063474	-14.65	0.000	-.1054166	-.0805353
wave						
5	.0161024	.0041014	3.93	0.000	.0080638	.0241409
6	-.0228265	.0043211	-5.28	0.000	-.0312958	-.0143573
_cons	3.019172	.0063416	476.09	0.000	3.006743	3.031602
sigma_u	.85097906					
sigma_e	.63954285					
rho	.63905566				(fraction of variance due to u_i)	

Hausman test

$$H_0: \beta_{RE} - \beta_{FE} = 0$$

→ random effects model

$$H_A: \beta_{RE} - \beta_{FE} \neq 0$$

→ fixed effects model

Hausman test

```
// fixed effects model  
xtreg health retired i.wave , fe  
estimates store fixed
```

```
// random effects model  
xtreg health retired i.wave , re  
estimates store random
```

```
// hausman test  
hausman fixed random
```

Hausman test

```
. hausman fixed random
```

	Coefficients			
	(b) fixed	(B) random	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
retired	.045631	-.2691823	.3148133	.007254
wave				
5	-.0604422	.0159274	-.0763695	.0014623
6	-.1191141	-.0234729	-.0956412	.001847

```
      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(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)  
              = 3529.22  
Prob>chi2 = 0.0000
```

→ fixed effects model

THANK YOU !

birkenbach@mea.mpisoc.mpg.de

info@share-project.org



Appendix

time series operators

```
// generate lagged variables  
gen lag1_health = l.health  
gen lag2_health = l2.health
```

mergeid	wave	health	lag1_health	lag2_health
AT-004855-01	1	1	.	.
AT-004855-01	2	1	1	.
AT-004855-02	1	4	.	.
AT-004855-02	2	2	4	.
AT-004855-02	3	2	2	4
AT-004855-02	4	4	2	2
AT-004855-02	5	2	4	2
AT-004855-02	6	1	2	4

time series operators

```
// generate differenced variables  
gen diff1_health = d.health  
gen diff2_health = d2.health
```

mergeid	wave	health	diff1_health	diff2_health
AT-004855-01	1	1	.	.
AT-004855-01	2	1	0	.
AT-004855-02	1	4	.	.
AT-004855-02	2	2	-2	.
AT-004855-02	3	2	0	2
AT-004855-02	4	4	2	2
AT-004855-02	5	2	-2	-4
AT-004855-02	6	1	-1	1

Logit models

```
// Pooled logit
logit goodhealth retired

// declare panel data structure
xtset id wave

// FE logit
xtlogit goodhealth retired , fe

// RE logit
xtlogit goodhealth retired , re
```

methodology
&
interpretation
quite complex !