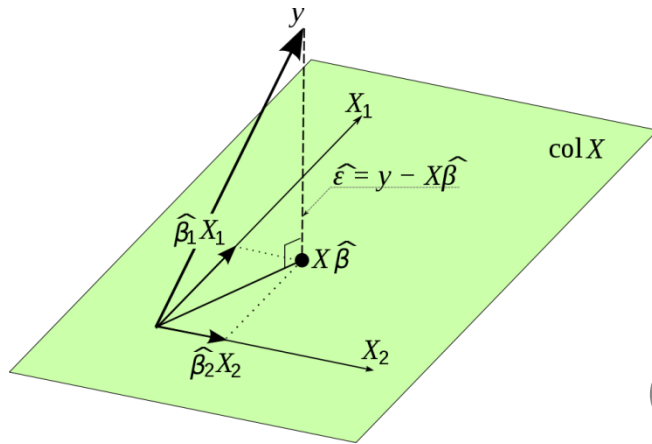
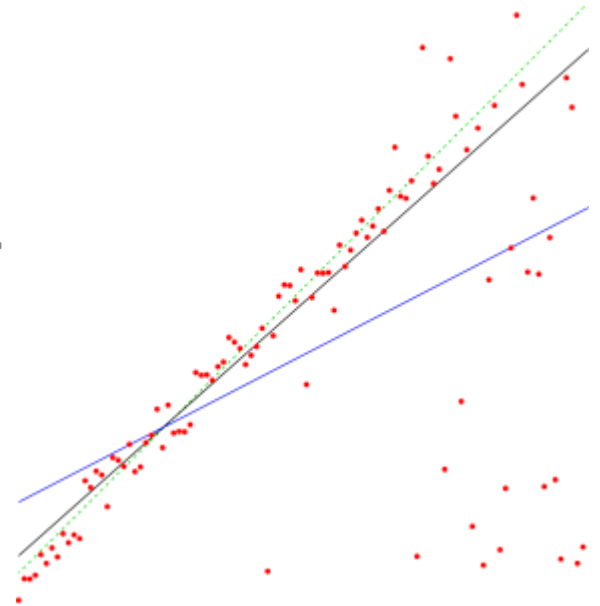


# Introduction to Panel Regression with Stata



Gustavo Mellior  
April 20<sup>th</sup> 2015



# Intended audience:

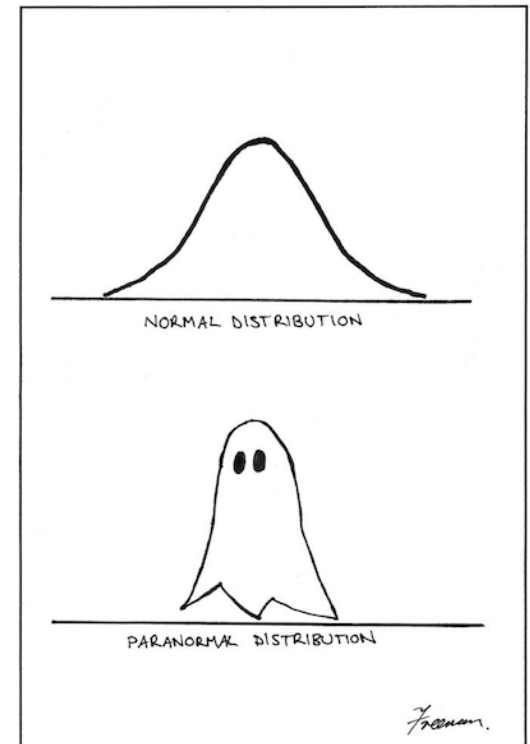
- You have used Stata before but you do not need to be an expert
- You want to know how to manage panel data
- You want to know how to implement panel regressions
- You want to have some hands-on experience on which model to use: pooled, FE or RE

# References

- *Microeconometrics Using Stata* by Cameron and Trivedi
- *Introduction to Modern Econometrics Using Stata* by Baum
- *Should I Use Fixed or Random Effects?* By Clark and Linzer
- [http://faculty.ucr.edu/~hanneman/linear\\_models/c4.html](http://faculty.ucr.edu/~hanneman/linear_models/c4.html)
- <http://dss.princeton.edu/training/Panel101.pdf#page=2>

# Overview

- 1. What is panel data
- 2. Useful pre-regression commands
- 2. Pooled regression and bias
- 3. FE vs RE
- 4. FE part 1
- 5. FE part 2
- 6. More Stata examples



# What is panel data?

- Cross-section
- Time-series
- Cross-sectional and time series data:

We are following the **same cross-section through time** (states, countries, firms, people, etc.). NLSY, SOEP, BHPS, PSID

This is not the same as an independently pooled cross-sections (CPS)

person	year	income	age	sex
1	2001	1300	27	1
1	2002	1600	28	1
1	2003	2000	29	1
2	2001	2000	38	2
2	2002	2300	39	2
2	2003	2400	40	2

Data Editor (Browse) - [psid1]

File Edit Data Tools

id[1] 48

	id	age	educ	union	laborinc	hours	tenure
1	48	36	14	0	13855	2680	.9
2	48	37	14	0	14437	2072	0
3	48	38	14	0	13011	1924	2
4	48	39	14	0	17546	2717	0
5	48	40	14	0	7062	2844	0
6	86	36	14	1	22088	2300	4
7	86	36	14	1	25987	2232	5
8	86	38	14	1	28810	2384	7
9	86	39	14	1	29197	2300	8
10	86	39	14	1	30810	2430	8
11	95	36	15	0	53213	3120	9
12	95	36	15	0	37536	2380	10
13	95	38	15	0	36245	2880	12

**Balanced panel:** each individual has an observation for each period

If you wish to replicate the following tables,  
graphs and regression outputs please use:

mus08psidextract.dta

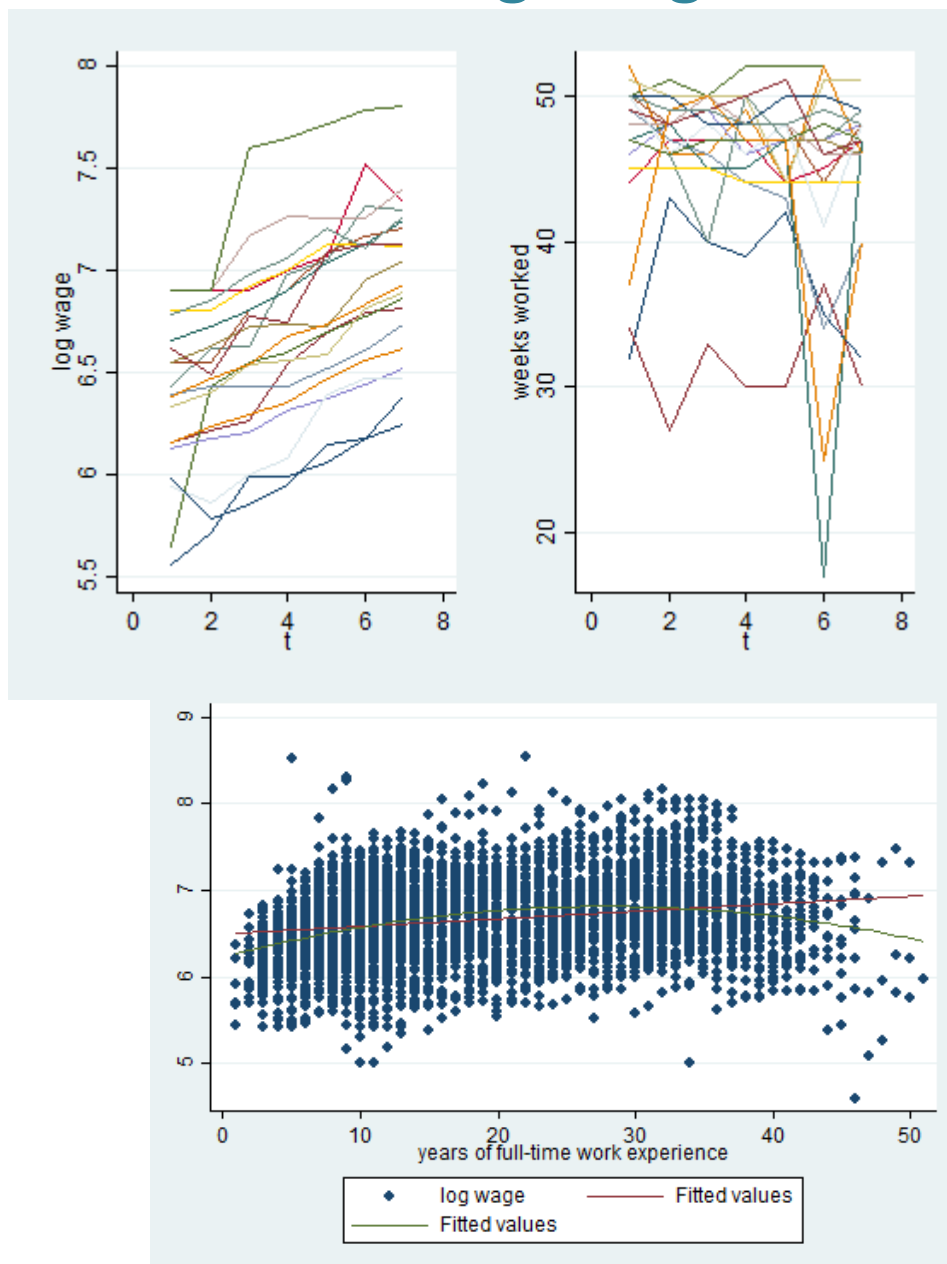
paneldocamtriv.do

These are the data set and do files, based on  
Cameron and Trivedi's *Microeconometrics Using  
Stata*, Chapter 8

# Handy Stata commands before running a regression

describe  
summarize - sum  
list *varlist*  
xtset id time  
xtdescribe  
xtsum  
xttab *varlist*  
xttrans *varlist*  
xtline and overall plots  
sort year  
by year: sum varname if

Example with  
mus08psidextract.dta and  
paneldocamtriv.do





### **One-way effect model**

$$y_{it} = \alpha_i + X'_{it}\beta + \varepsilon_{it}$$

### **Two-way effect model**

$$y_{it} = \alpha_i + \gamma_t + X'_{it}\beta + \varepsilon_{it}$$

### **Pooled model**

$$y_{it} = \alpha + X'_{it}\beta + \varepsilon_{it}$$

where  $i=1,2,\dots,N$  and  $t=1,2,\dots,T$

A short panel fixes  $T$  and lets  $N$  be large. Long panels have small  $N$  and large  $T$ . We can have panels that have both big  $N$  and  $T$ . Each may require different estimation techniques. This workshop will focus on short panels. The model errors are assumed to be independent across individuals (can be relaxed).

Correction for OLS standard errors will be necessary.  $E(\varepsilon_{it}, \varepsilon_{js}) = 0, i \neq j$ .  $E(\varepsilon_{it}, \varepsilon_{is})$  may be unrestricted and  $\varepsilon_{it}$  may be heteroskedastic. Fortunately Stata can correct for this.

# Pooled OLS regression

$$y_{it} = \alpha + X'_{it}\beta + v_{it}$$

$$\text{where } v_{it} = (\alpha_i - \alpha + \varepsilon_{it})$$

$$y_{it} = \alpha + X'_{it}\beta + (\alpha_i - \alpha + \varepsilon_{it})$$

$$lwage_{it} = \alpha + \beta_1 exp_{it} + \beta_2 (exp)^2_{it} + \beta_3 ed_{it} + \beta_4 weeks_{it} + v_{it}$$

```
. reg lwage exp exp2 wks ed
```

Source	SS	df	MS	Number of obs = 4165		
Model	251.491445	4	62.8728613	F( 4, 4160) = 411.62		
Residual	635.413457	4160	.152743619	Prob > F = 0.0000		
Total	886.904902	4164	.212993492	R-squared = 0.2836		
				Adj R-squared = 0.2829		
				Root MSE = .39082		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.044675	.0023929	18.67	0.000	.0399838	.0493663
exp2	-.0007156	.0000528	-13.56	0.000	-.0008191	-.0006121
wks	.005827	.0011827	4.93	0.000	.0035084	.0081456
ed	.0760407	.0022266	34.15	0.000	.0716754	.080406
_cons	4.907961	.0673297	72.89	0.000	4.775959	5.039963

# Pooled OLS with clustered standard errors

```
. reg lwage exp exp2 wks ed, vce(cluster id)
```

Linear regression

Number of obs = 4165  
 F( 4, 594) = 72.58  
 Prob > F = 0.0000  
 R-squared = 0.2836  
 Root MSE = .39082

(Std. Err. adjusted for 595 clusters in id)

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.044675	.0054385	8.21	0.000	.0339941	.055356
exp2	-.0007156	.0001285	-5.57	0.000	-.0009679	-.0004633
wks	.005827	.0019284	3.02	0.003	.0020396	.0096144
ed	.0760407	.0052122	14.59	0.000	.0658042	.0862772
_cons	4.907961	.1399887	35.06	0.000	4.633028	5.182894

$$lwage_{it} = \alpha + \beta_1 exp_{it} + \beta_2 (exp)^2_{it} + \beta_3 ed_{it} + \beta_4 weeks_{it} + v_{it}$$

$$\frac{\partial E(lwage_{it} | regressors_{it})}{\partial exp_{it}} = 0.0447 - 2 * 0.0007 exp$$

$$\frac{\partial E(lwage_{it} | regressors_{it})}{\partial ed_{it}} = 0.076$$

# Pooled regression and bias

## Omitted variable bias

$$Y = X' B + Z' \delta + U$$

$$\hat{\beta} = (X' X)^{-1} (X' Y)$$

$$\hat{\beta} = (X' X)^{-1} X' (X' \beta + Z' \delta + U)$$

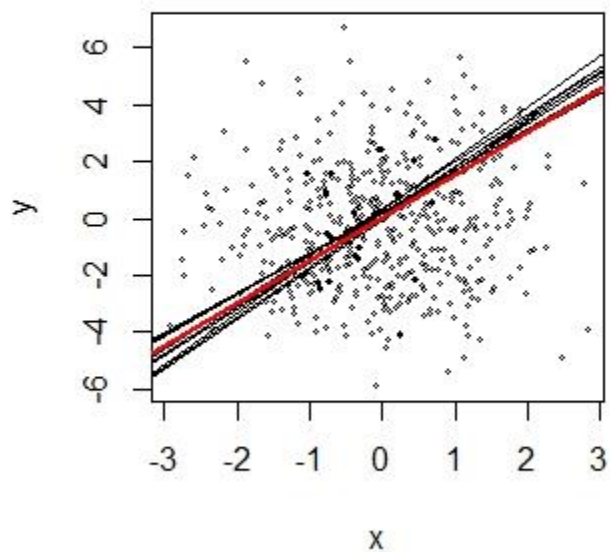
$$\hat{\beta} = \beta + (X' X)^{-1} X' Z \delta + (X' X)^{-1} X' U$$

$$\hat{\beta} = \beta + (X' X)^{-1} X' Z \delta$$

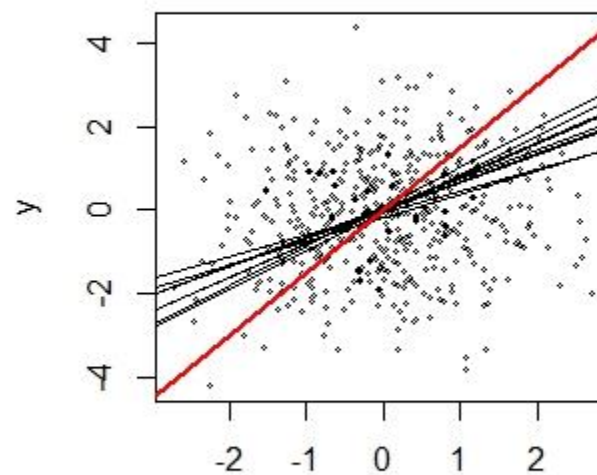
$$\hat{\beta} = \beta + \text{corr}(X, Z) \frac{\sigma_Z}{\sigma_X} \delta$$

$$\text{where } (X' X)^{-1} (X' Z) = \frac{\text{COV}(X, Z)}{\sigma_X^2} = \text{corr}(X, Z) \frac{\sigma_Z}{\sigma_X}$$

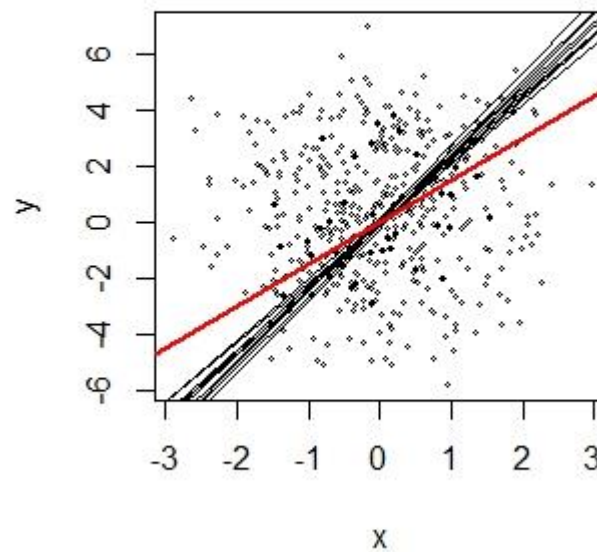
True beta=1.5 in red corr=0



True beta=1.5 in red corr=-0.9



True beta=1.5 in red corr=0.9



$$\hat{\beta} = \beta + \text{corr}(X, Z) \frac{\sigma_Z}{\sigma_X} \delta$$

# If the error term is correlated with the explanatory variable...

- We will get a biased coefficient
- Pooled regression will not be the way to go

## Examples of when this can happen:

Wages, education, gender, ability and race

GDP growth and life-expectancy

- Use **Fixed Effects** (more on this in a bit)
- If  $\text{corr}(X, Z) = 0$ ... pooled or **Random Effects**

*“Fixed-effects models have the added complication that regressors may be correlated with the individual-level effects so that consistent estimation of regression parameters requires eliminating or controlling for the fixed effects.”*

Cameron and Trivedi, *Microeconometrics Using Stata*

*“...the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements **that are correlated with the regressors** in the model, **not whether these effects are stochastic or not**”* Greene, 2008, p.183

# Random Effects

The RE model assumes that the individual effects are distributed independently of the regressors. The benefit is that we can calculate the individual effects. The cost is a restrictive assumption that if unrealistic, will bias our results.

Sometimes RE is called partial pooling or random intercept model



```
. xtreg lwage exp exp2 wks ed, re vce(cluster id)
```

```
Random-effects GLS regression           Number of obs   =       4165
Group variable: id                     Number of groups  =       595

R-sq:  within  = 0.6340                 Obs per group: min =        7
       between  = 0.1716                                     avg  =       7.0
       overall  = 0.1830                                     max  =        7

                                           Wald chi2(4)      =    1598.50
corr(u_i, X)   = 0 (assumed)           Prob > chi2       =     0.0000
```

(Std. Err. adjusted for 595 clusters in id)

lwage	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
exp	.0888609	.0039992	22.22	0.000	.0810227	.0966992
exp2	-.0007726	.0000896	-8.62	0.000	-.0009481	-.000597
wks	.0009658	.0009259	1.04	0.297	-.000849	.0027806
ed	.1117099	.0083954	13.31	0.000	.0952552	.1281647
_cons	3.829366	.1333931	28.71	0.000	3.567921	4.090812
sigma_u	.31951859					
sigma_e	.15220316					
rho	.81505521	(fraction of variance due to u_i)				

# xttest0 Pooled OLS or RE?

```
. xttest0
```

Breusch and Pagan Lagrangian multiplier test for random effects

```
lwage[id,t] = Xb + u[id] + e[id,t]
```

Estimated results:

	Var	sd = sqrt(Var)
lwage	.2129935	.4615122
e	.0231658	.1522032
u	.1020921	.3195186

Test: Var(u) = 0

```
chibar2(01) = 5192.13  
Prob > chibar2 = 0.0000
```



Reject null, RE is better than pooled OLS. There are individual effects

# Fixed Effects

- Need healthy dose of within variation, otherwise imprecise estimates. Think of demeaning with little within variation
- Do not have to worry about time invariant observable or unobservable variables that are correlated with regressors
- But cannot estimate the impact of those coefficients (or do inference on them).

```
. xtreg lwage exp exp2 wks ed, fe vce(cluster id)
note: ed omitted because of collinearity
```

```
Fixed-effects (within) regression           Number of obs   =       4165
Group variable: id                         Number of groups =       595

R-sq:  within = 0.6566                     Obs per group:  min =        7
        between = 0.0276                               avg  =       7.0
        overall = 0.0476                               max  =        7

                                           F(3,594)         =    1059.72
corr(u_i, Xb)  = -0.9107                     Prob > F          =     0.0000
```

(Std. Err. adjusted for 595 clusters in id)

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
exp	.1137879	.0040289	28.24	0.000	.1058753	.1217004
exp2	-.0004244	.0000822	-5.16	0.000	-.0005858	-.0002629
wks	.0008359	.0008697	0.96	0.337	-.0008721	.0025439
ed	0	(omitted)				
_cons	4.596396	.0600887	76.49	0.000	4.478384	4.714408
sigma_u	1.0362039					
sigma_e	.15220316					
rho	.97888036	(fraction of variance due to u_i)				

Variable	Pooled_~b	FE	FE_rob	RE	RE_rob
exp	0.0447	0.1138	0.1138	0.0889	0.0889
	0.0054	0.0025	0.0040	0.0028	0.0040
exp2	-0.0007	-0.0004	-0.0004	-0.0008	-0.0008
	0.0001	0.0001	0.0001	0.0001	0.0001
wks	0.0058	0.0008	0.0008	0.0010	0.0010
	0.0019	0.0006	0.0009	0.0007	0.0009
ed	0.0760	(omitted)	(omitted)	0.1117	0.1117
	0.0052			0.0061	0.0084
_cons	4.9080	4.5964	4.5964	3.8294	3.8294
	0.1400	0.0389	0.0601	0.0936	0.1334
N	4165	4165	4165	4165	4165
r2	0.2836	0.6566	0.6566		
r2_0					
r2_b		0.0276	0.0276	0.1716	0.1716
r2_w		0.6566	0.6566	0.6340	0.6340
sigma_u		1.0362	1.0362	0.3195	0.3195
sigma_e		0.1522	0.1522	0.1522	0.1522
rho		0.9789	0.9789	0.8151	0.8151

# How and when to choose

There is a variance-bias trade-off... *“The fixed effects model will produce unbiased estimates of  $\beta$ , but those estimates can be subject to high sample-to-sample variability. The random effects model will, except in rare circumstances, introduce bias in estimates of  $\beta$ , but can greatly constrain the variance of those estimates—leading to estimates that are closer, on average, to the true value in any particular sample.”* Clark and Linzer

## Hausman Test

# Hausman Test

```
. hausman FE RE, sigmamore
```

	—— Coefficients ——		(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
	(b) FE	(B) RE		
exp	.1137879	.0888609	.0249269	.0012778
exp2	-.0004244	-.0007726	.0003482	.0000285
wks	.0008359	.0009658	-.0001299	.0001108

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)  
 = 1513.02  
 Prob>chi2 = 0.0000

Null hypothesis is that RE is correct and that the difference between the FE and RE estimates are not significant. We can see that the Hausman test rejects the null hypothesis

# More on Fixed-Effects

- Show that LSDV=FE=Demeaned regression=Differenced regression (for short panels)
- Other useful tricks such as creating multiple dummies, demeaning by id, etc.



# LSDV=FE=Demeanded=Differenced

In order to replicate the following tables, graphs and regression outputs please use:

psid1.dta

panel.do

- The equivalence between LSDV and FE does not carry over for nonlinear models
- Avoid the dummy variable trap!

areg lnhourlywage tenure tenuresquared age agesquared black union educ i.id, absorb(id)

Linear regression, absorbing indicators	Number of obs	=	3945
	F( 5, 3151)	=	45.05
	Prob > F	=	0.0000
	R-squared	=	0.8554
	Adj R-squared	=	0.8190
	Root MSE	=	0.2345

lnhourlywage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tenure	.0185458	.0032555	5.70	0.000	.0121626	.0249289
tenuresquared	-.0005425	.0001424	-3.81	0.000	-.0008216	-.0002633
age	.0678427	.0127567	5.32	0.000	.0428304	.092855
agesquared	-.0005006	.0001715	-2.92	0.004	-.0008368	-.0001644
black	0	(omitted)				
union	.0528083	.0236289	2.23	0.025	.0064786	.099138
educ	0	(omitted)				
id						
86	0	(omitted)				
95	0	(omitted)				
162	0	(omitted)				

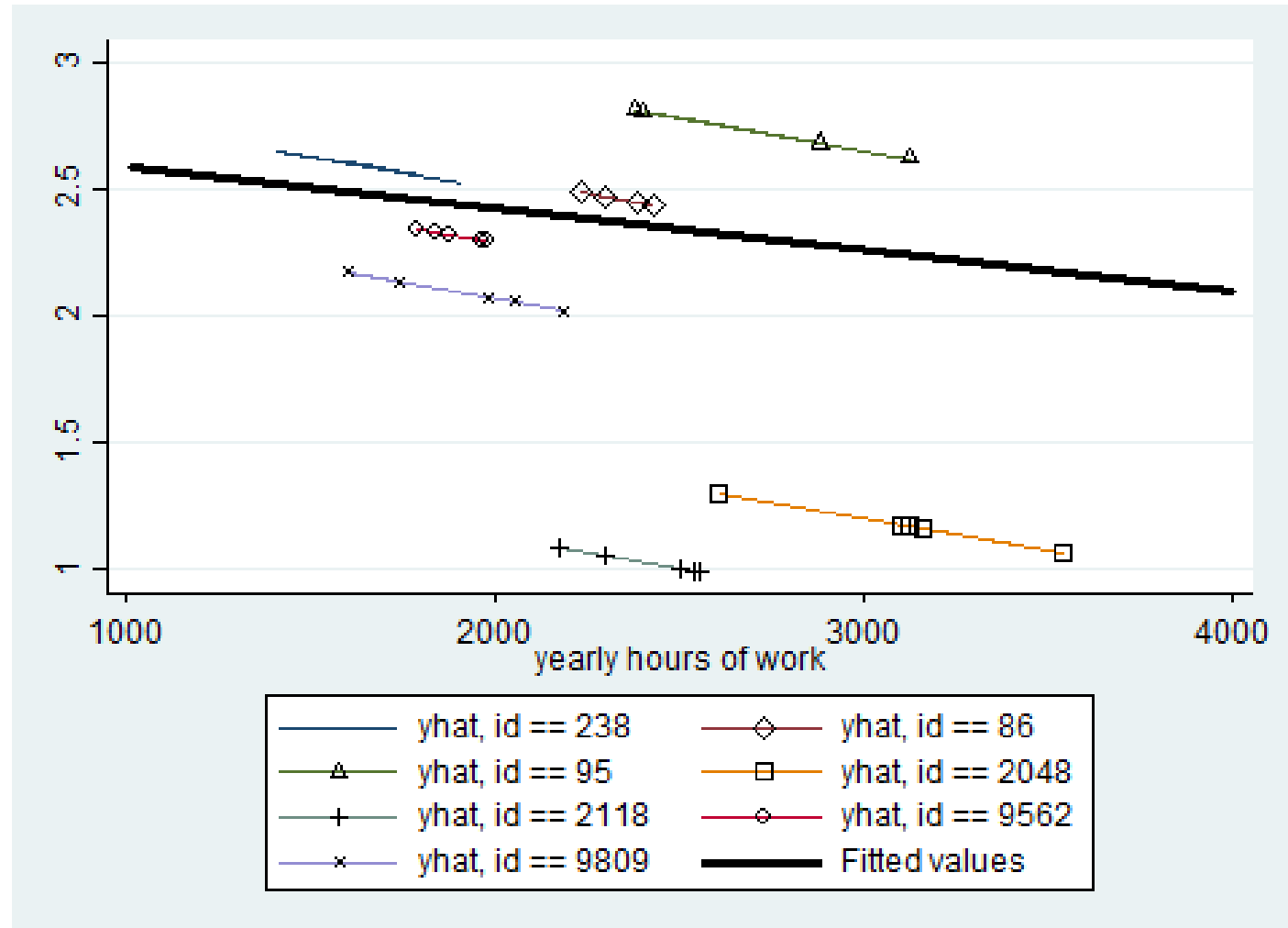
```
. reg lnhourlywage tenure tenuresquared age agesquared black union educ i.id
note: 9749.id omitted because of collinearity
note: 9830.id omitted because of collinearity
```

Source	SS	df	MS	Number of obs = 3945		
Model	1025.2034	793	1.2928164	F(793, 3151) = 23.51		
Residual	173.269153	3151	.054988624	Prob > F = 0.0000		
Total	1198.47255	3944	.303872352	R-squared = 0.8554		
				Adj R-squared = 0.8190		
				Root MSE = .2345		

lnhourlywage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tenure	.0185458	.0032555	5.70	0.000	.0121626	.0249289
tenuresquared	-.0005425	.0001424	-3.81	0.000	-.0008216	-.0002633
age	.0678427	.0127567	5.32	0.000	.0428304	.092855
agesquared	-.0005006	.0001715	-2.92	0.004	-.0008368	-.0001644
black	-1.845181	.1797857	-10.26	0.000	-2.19769	-1.492672
union	.0528083	.0236289	2.23	0.025	.0064786	.099138
educ	-.0157004	.037219	-0.42	0.673	-.0886763	.0572755
id						
86	-1.166965	.1638603	-7.12	0.000	-1.488249	-.8456811
95	-.8830107	.1578124	-5.60	0.000	-1.192436	-.5735852
162	-1.203459	.2483432	-4.85	0.000	-1.69039	-.7165286
181	-.8886226	.1959452	-4.54	0.000	-1.272816	-.5044296
183	-1.280398	.156619	-8.18	0.000	-1.587483	-.9733121

# LSDV regression of logwages on hours worked for a few industries



```
. xtreg lnhourlywage tenure tenuresquared age agesquared black union educ, fe
note: black omitted because of collinearity
note: educ omitted because of collinearity
```

```
Fixed-effects (within) regression      Number of obs      =      3945
Group variable: id                    Number of groups   =      789
```

```
R-sq:  within = 0.0667                Obs per group: min =      5
      between = 0.1310                      avg =      5.0
      overall = 0.1202                      max =      5
```

```
corr(u_i, Xb) = -0.1534                F(5, 3151)         =      45.05
                                          Prob > F           =      0.0000
```

lnhourlywage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
tenure	.0185458	.0032555	5.70	0.000	.0121626	.0249289
tenuresquared	-.0005425	.0001424	-3.81	0.000	-.0008216	-.0002633
age	.0678427	.0127567	5.32	0.000	.0428304	.092855
agesquared	-.0005006	.0001715	-2.92	0.004	-.0008368	-.0001644
black	0	(omitted)				
union	.0528083	.0236289	2.23	0.025	.0064786	.099138
educ	0	(omitted)				
_cons	.501758	.2352478	2.13	0.033	.0405035	.9630124
sigma_u	.47856639					
sigma_e	.23449653					
rho	.80638791	(fraction of variance due to u_i)				

```
F test that all u_i=0:      F(788, 3151) =      19.86      Prob > F = 0.0000
```

## If we had a small amount of dummies...

- `tabulate id, generate(dummyid)`

This would create the dummy variables for each individual. Then we can add them to the `reg y x` command.

However, remember that the dummy estimates have no out-of-sample meaning. They are not consistent. That is the advantage of RE, it can calculate and do inference on the individual effects

# Demeaned regression

by id: egen lnhourlywage\_mean=mean(lnhourlywage)  
 gen dm\_lnhourlywage =lnhourlywage- lnhourlywage\_mean

```
. reg dm_lnhourlywage dm_tenure dm_tenuresquared dm_age dm_agesquared dm_union dm_black
note: dm_black omitted because of collinearity
```

Source	SS	df	MS	Number of obs = 3945		
Model	12.3873476	5	2.47746953	F( 5, 3939) = 56.32		
Residual	173.269153	3939	.043988107	Prob > F = 0.0000		
Total	185.656501	3944	.047073149	R-squared = 0.0667		
				Adj R-squared = 0.0655		
				Root MSE = .20973		

dm_lnhourlywage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dm_tenure	.0185458	.0029117	6.37	0.000	.0128371	.0242544
dm_tenuresquared	-.0005425	.0001273	-4.26	0.000	-.0007921	-.0002928
dm_age	.0678427	.0114096	5.95	0.000	.0454735	.0902119
dm_agesquared	-.0005006	.0001534	-3.26	0.001	-.0008013	-.0001999
dm_union	.0528083	.0211337	2.50	0.013	.0113743	.0942423
dm_black	0	(omitted)				
_cons	2.01e-09	.0033392	0.00	1.000	-.0065467	.0065468

# Demeaned regression

- Instead of calculating the means and demeaning we can use `xtdata`.
- This will permanently change your variables so it is best to use `preserve – regress – restore` (see `do file`)



# Differenced regression

- We could get the same estimate if we difference  $y_{i,t} - y_{i,t-1}$  and so on for the regressors. The fixed effects drop. However, this will only work for short panels. For  $T=2$  FE and the 1<sup>st</sup> difference estimator will be identical.

Sort id year

regress D.(lnhourlywage agesquared age educ union tenure tenuresquared union black)

# Other useful panel commands

- `xtreg indepvar regressors i.year, fe`  
`testparm i.year`
- `xttest0` Pooled or RE?
- `xttest2` Contemporaneous correlation?
- `xttest3` Heteroskedasticity in FE? (corrected by robust and/or cluster)

- Remember lecture will be up on the website
- The big picture. We looked at pooled OLS and:

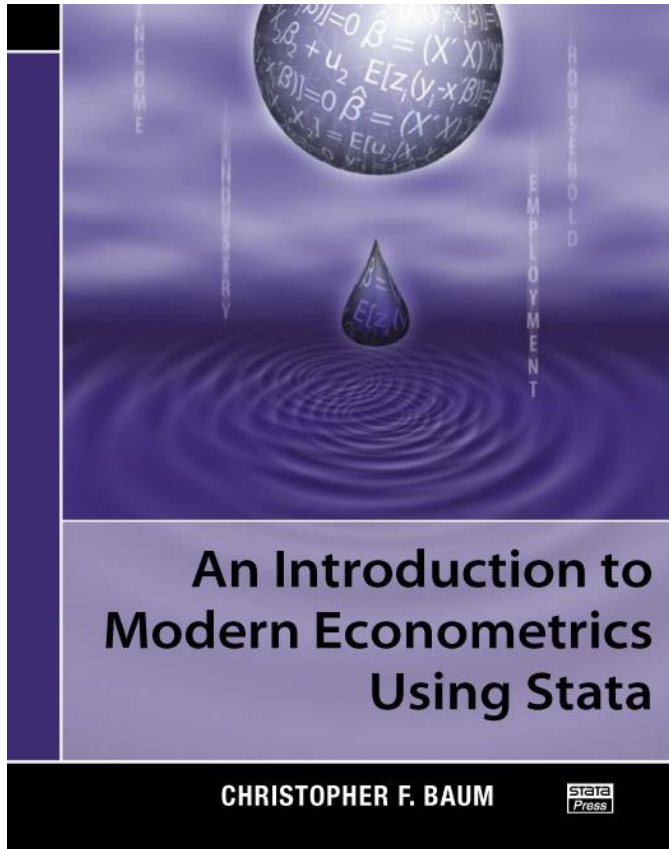
Command	Syntax
	<b><i>Entity fixed effects</i></b>
xtreg	xtreg y x1 x2 x3 x4 x5 x6 x7, fe
areg	areg y x1 x2 x3 x4 x5 x6 x7, absorb(country)
regress	xi: regress y x1 x2 x3 x4 x5 x6 x7 i.country,
	<b><i>Entity and time fixed effects</i></b>
xi: xtreg	xi: xtreg y x1 x2 x3 x4 x5 x6 x7 i.year, fe
xi: areg	xi: areg y x1 x2 x3 x4 x5 x6 x7 i.year, absorb(country)
xi: regress	xi: regress y x1 x2 x3 x4 x5 x6 x7 i.country i.year
	<b><i>Random effects</i></b>
xtreg	xtreg y x1 x2 x3 x4 x5 x6 x7, re robust

Google

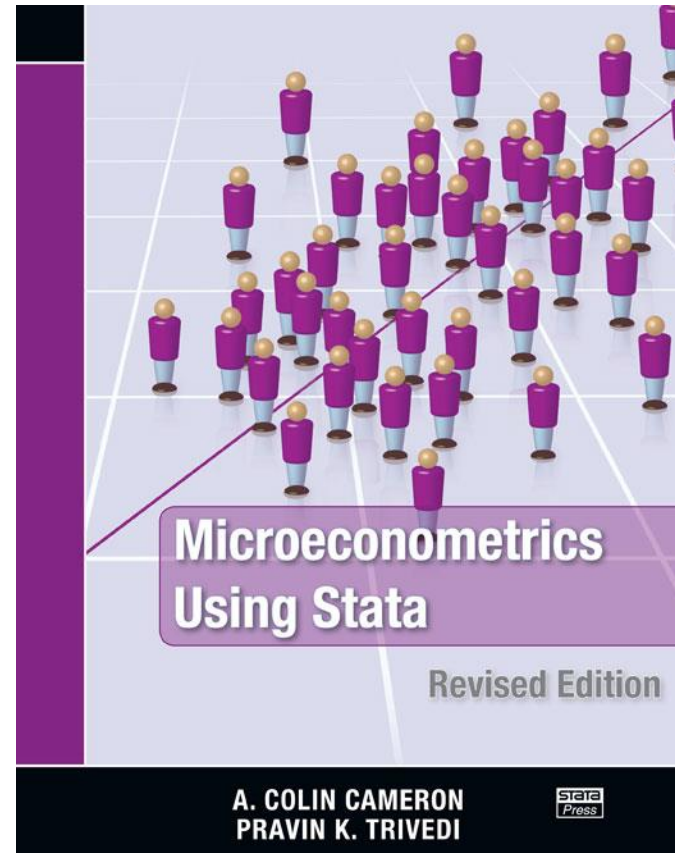
200805

Google Search

I'm Feeling Lucky



# Useful resources



help *command*

Links at the beginning

# More Stata examples and questions

- IPUMS data set and do file
- Any other questions?