Silvia Mendolia

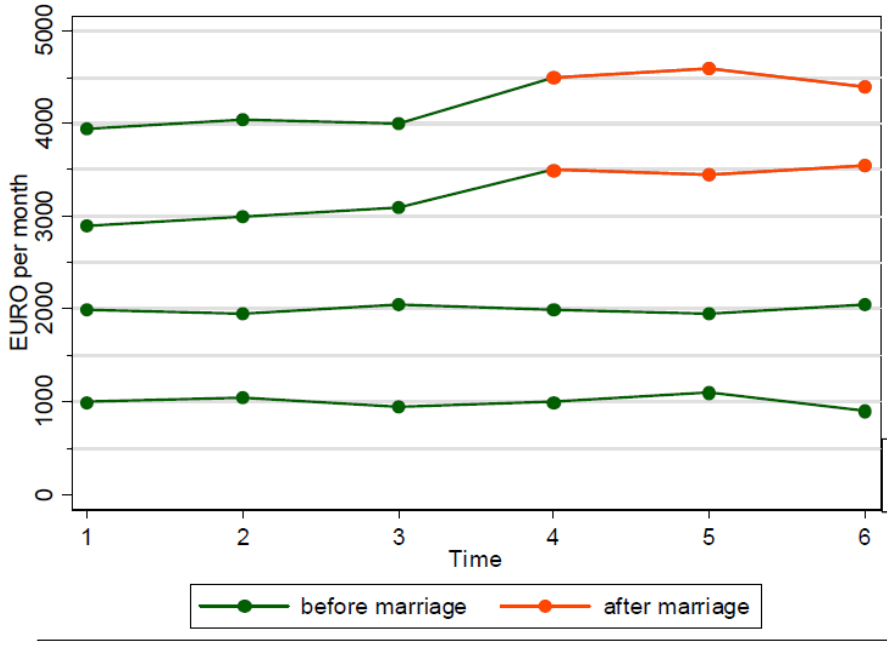
Labour market, health and well-being: economic analysis using panel data

Plan for this lecture

- Between and within estimator
- Fixed effects estimator
- Practical example of a fixed effects regression in Stata

Between estimation

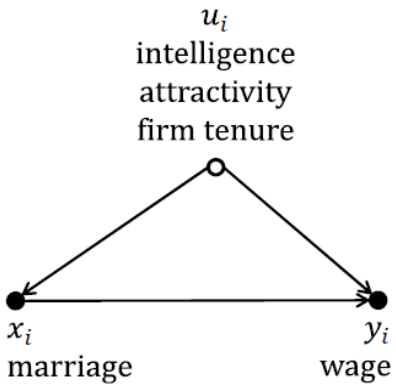
- Example: is there a marriage premium for



- Problem: men who marry and those who don’t are different in many characteristics

Between estimation

- Treatment is non random in this example
- **Men can self-select into marriage**
- Results from a cross section estimate at the last period will produce biased results!
- The estimated marital wage premium will be much higher than what it actually is!



Within estimation

$$Y_{it}=\alpha + x_{it}\beta + v_{it}$$

$$v_{it}= a_i + \varepsilon_{it}$$

- Remember the error decomposition
- a_i is person-specific time-constant error term
- ε_{it} is a time-varying error term (or idiosyncratic error term)
- Pooled OLS is unbiased only if x_{it} is independent from both error components
- $E(a_i | x_{it})=0$ – No time constant unobserved heterogeneity
- $E(\varepsilon_{it} | x_{it})=0$ No time varying unobserved heterogeneity

First Difference Estimator

- We can take the first difference in our original model and we can eliminate a_i
- $Y_{it} = x_{it}\beta + a_i + \varepsilon_{it}$
- $Y_{it-1} = x_{it-1}\beta + a_i + \varepsilon_{it-1}$
- $\mathbf{\Delta y_{it} = \beta \Delta x_{it} + \Delta \varepsilon_{it}}$
- **In this way, we have eliminated the person-specific time invariant error term**
- We can apply pooled OLS to these transformed data and we get the FD estimator

Fixed effects estimator

- Basic idea: For each individual, calculate the mean of x and the mean of y. Then run OLS on a transformed dataset where each y_{it} is replaced by $y_{it}-\bar{y}_i$ and each x_{it} is replaced by $x_{it}-\bar{x}_i$

- Few assumptions are required for FE to be consistent: u_i is allowed to correlate with x_i

- Disadvantage: **can't estimate the effects of any time-invariant variables**

Fixed effects estimator

- From the original model:

$Y_{it}=\alpha + x_{it}\beta + a_i + \varepsilon_{it}$

- The new model is:

$Y_{it} - \overline{Y_i}=(x_{it}\beta - xi)\overline{+} (\varepsilon_{it}-\overline{\varepsilon_i})$

- **In Stata: xtreg y x, fe**

- **Demeaning** wipes out person-specific time-constant unobserved heterogeneity! Only within variation is left.

Fixed effects estimator

- The FE estimator is unbiased under the **strict exogeneity assumption**:
- $E(\varepsilon_{it} | x_{it}) = 0$
- Time-constant heterogeneity is allowed
- **Time-varying heterogeneity is not allowed**

- Can you think of possible violations of this assumption?
  - Time-varying shocks

Within and Between variation

- A between regression uses between variation
- It is heavily affected by self-selection
- **A within-regression (FE) uses within variation only**
- The causal effect is identified by the deviation from the person-specific mean
- **Self-selection does not bias the results**

Least square Dummy variable estimator (LSDV)

- The FE estimator is equivalent to an OLS estimator including a dummy variable for each individual
- This is called LSDV
- LSDV is practical only when N is small
- Both FD and LSDV are also widely used but share the same caveats of FE (see end of

Practical Example in Stata

- Let’s open our Understanding Society teaching Dataset
- Use longitudinal_td.dta
- Let’s spend some time familiarising with the data
- Individual id? Time id? Interesting variables?

Practical Example in Stata

- Is this a balanced panel data? Why or why not?
- What is the household id and why could it be useful?
- Search for useful information using lookfor

```
. lookfor qualif

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
hiqual_dv       byte    %8.0g      hiqual_dv
                                              Highest qualification ever reported
```

Practical Example in Stata

- Tell Stata this is a panel data

```
xtset pidp wave
      panel variable:  pidp (unbalanced)
       time variable:  wave, 1 to 9
               delta:  1 unit
```

- Inspect some variables using xttab
- E.g. life satisfaction, mental health etc

Practical Example in Stata

- Let's investigate the **relationship between marital status and life satisfaction**
- Use a simple pooled OLS first (no panel data estimation)

- Which variables do you want to include?
- Why?
- reg sclfsato……

Practical Example in Stata

- Use xi: to allow creating/including binary variables

- Stata omits the first category by default
- Think of your omitted group!
- If you want to create dummy variables from a categorical variable, use tabulate with the gen option
- Tab jbstat, gen (job)

- xi: reg selfsate i.jbstat

Practical Example in Stata

- If you want to use employed people as your base category, you can create dummies and type:
- reg sclfsato_dv job1 job3 job4……..
- You can also group some categories (eg create a category for people out of the labour force)

- Try creating binary variables for employed, self-employed, unemployed and out of the labour force

Practical Example in Stata

Estimate a simple pooled OLS model:

reg sclfsato age_dv i.sex_dv married separated widow i.hiqual_dv self_employed unemployed outlforce

Source	SS	df	MS	Number of obs	=	174,062
				F(13, 174048)	=	565.84
Model	15349.2603	13	1180.71233	Prob > F	=	0.0000
Residual	363177.119	174,048	2.08664919	R-squared	=	0.0406
				Adj R-squared	=	0.0405
Total	378526.379	174,061	2.17467657	Root MSE	=	1.4445

sclfsato	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age_dv	.0088284	.0002802	31.51	0.000	.0082793	.0093775
sex_dv						
Female	.0472584	.0072211	6.54	0.000	.0331051	.0614117
married	.2939085	.0105449	27.87	0.000	.2732408	.3145763
separated	-.274586	.01473	-18.64	0.000	-.3034563	-.2457156
widow	.1480015	.0179077	8.26	0.000	.1129028	.1831002
hiqual_dv						
Other higher	-.0912292	.0117557	-7.76	0.000	-.1142702	-.0681883
A level etc	-.1326591	.0107458	-12.35	0.000	-.1537205	-.1115977
GCSE etc	-.2244654	.0105176	-21.34	0.000	-.2450796	-.2038512
Other qual	-.2744082	.0130797	-20.98	0.000	-.3000441	-.2487722
No qual	-.2962486	.0126034	-23.51	0.000	-.3209509	-.2715463
self_employed	-.0336817	.0134971	-2.50	0.013	-.0601358	-.0072276
unemployed	-.6263114	.0182772	-34.27	0.000	-.6621342	-.5904886
outlforce	-.0242729	.0088931	-2.73	0.006	-.0417031	-.0068427
_cons	4.715428	.0154199	305.80	0.000	4.685206	4.745651

Practical Example in Stata

- What are the empirical problems of estimating this model with pooled OLS?
- What is the possible role of unobserved heterogeneity in this framework?
- How can we solve (at least some) of these problems?

Practical Example in Stata

Estimate a model with FE:

xtreg sclfsato age_dv i.sex_dv i.mstat_dv i.hiqual_dv self_employed unemployed outlforce, fe

```
Fixed-effects (within) regression               Number of obs      =       174,062
Group variable: pidp                            Number of groups   =        33,896

R-sq:                                           Obs per group:
     within  = 0.0021                                         min =             1
     between = 0.0005                                         avg =           5.1
     overall = 0.0001                                         max =             9

                                                F(12,140154)       =         24.28
corr(u_i, Xb)  = -0.1317                        Prob > F           =        0.0000
```

Practical Example in Stata

Estimate a model with FE:

xtreg sclfsato age_dv i.sex_dv married separated widow i.hiqual_dv self_employed unemployed outlforce, fe

```
corr(u_i, Xb)  = -0.1317                        F(12,140154)      =     24.28
                                                Prob > F          =     0.0000
```

sclfsato	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age_dv	-.0064231	.0012287	-5.23	0.000	-.0088313	-.004015
sex_dv						
Female	0	(omitted)				
married	.1608527	.0241254	6.67	0.000	.1135673	.2081381
separated	-.0299583	.0308096	-0.97	0.331	-.0903444	.0304279
widow	-.041907	.0402849	-1.04	0.298	-.1208646	.0370506
hiqual_dv						
Other higher	.1084236	.0559798	1.94	0.053	-.0012957	.2181428
A level etc	.1734337	.0460586	3.77	0.000	.0831597	.2637076
GCSE etc	.1809429	.0600086	3.02	0.003	.0633272	.2985586
Other qual	.0805996	.083207	0.97	0.333	-.0824844	.2436837
No qual	.0876955	.088886	0.99	0.324	-.0865194	.2619103
self_employed	.0259308	.0197811	1.31	0.190	-.0128397	.0647013
unemployed	-.2111927	.0204734	-10.32	0.000	-.2513201	-.1710654
outlforce	.0205248	.013382	1.53	0.125	-.0057037	.0467532
_cons	5.330894	.0800015	66.63	0.000	5.174092	5.487695
sigma_u	1.2126787					
sigma_e	1.1344293					
rho	.53330161	(fraction of variance due to u_i)				

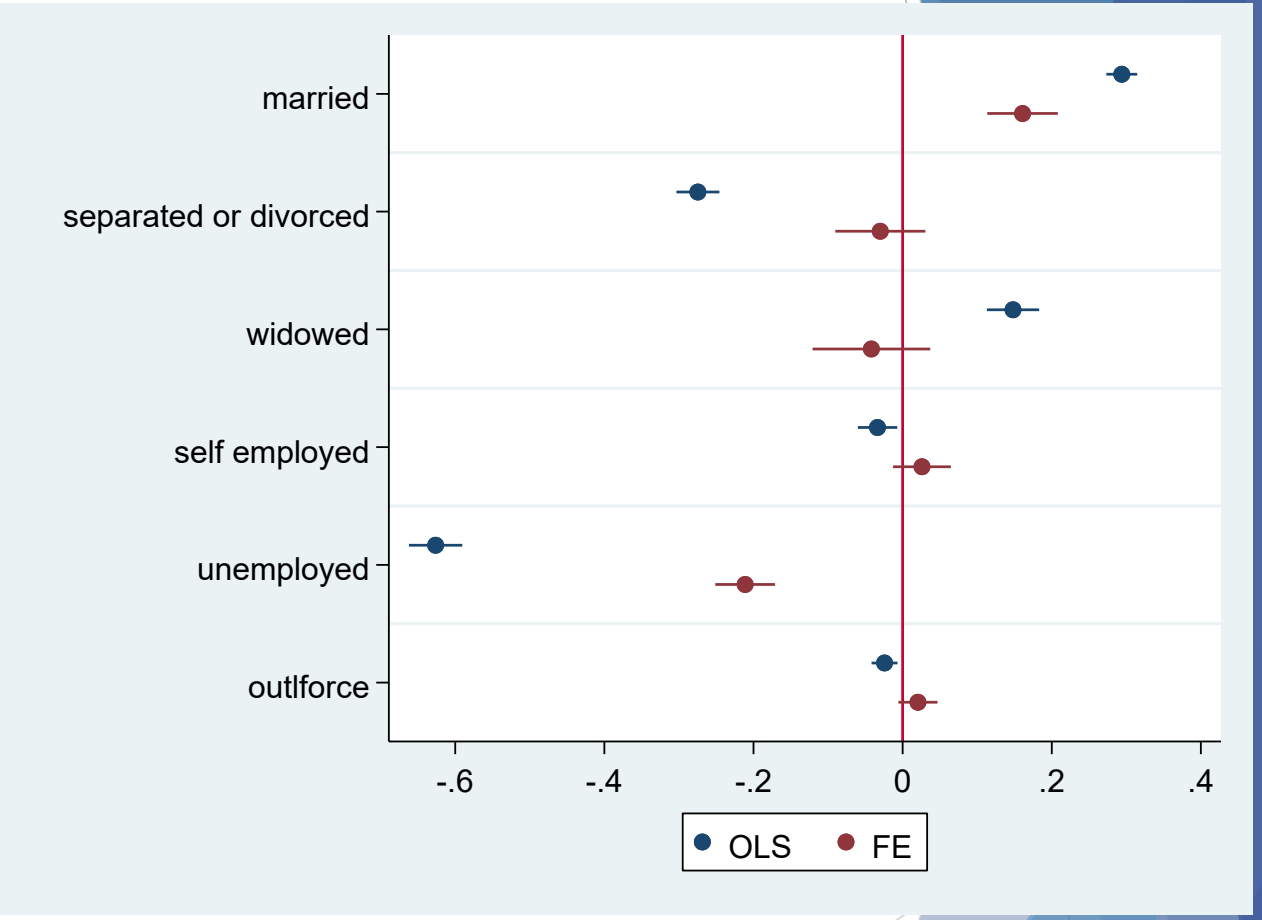
```
F test that all u_i=0: F(33895, 140154) = 4.19                     Prob > F = 0.0000
```

Practical Example in Stata

- What differences can you see between these two sets of results?
- What happens to time-constant variables (eg sex)? Why?
- What happens to the coefficients of marital status and employment? Why?
- Can you think to potential sources of bias in this analysis? (Hint: think of the strict exogeneity assumption)

Practical Example in Stata

- We can compare the OLS and FE results in a very efficient way using coefplot



- First, store results using est store

coefplot OLS FE, keep (married separated widow self_employed unemployed outlforce)

Estimation Sample

- Defining the estimation sample is very important:
  - We can include only individuals who switch from being single to being married
  - The “never-treated” individuals can be a control group
  - The “already treated” individuals may bias the results
    - If the treatment varies over time the age effect of the “already treated” may be problematic (old people in the control and young people in the treatment group

How to include age

- For simplicity, we included age as a linear variable
- This is probably not an optimal choice
- An alternative is to include age dummies (use i.age) or age groups (eg <20;20-30; 30-40 etc)

Limitations of Fixed Effects estimation

- It ignores variation across cross-sections (between variation)
- It does not allow us to estimate the coefficients of time-invariant regressors (gender, education…).
- Differenced regressors may be more susceptible to measurement error.
- It does not solve the problem of time-varying omitted variables