UNIVERSITY CERTIFICATE IN ECONOMETRICS

2017 · Edition 2

Course : Microeconometrics for policy evaluation

Part 1– Panel Data Models

Vincent Vandenberghe

Outline

1. Introduction
2. The omitted variable/endogeneity problem
3. Experimental methods
4. Quasi-experimental methods
   - Short panel analysis (day 1)
     - What are panels,
     - Panels and unobserved time and individual effects models
     - **Fixed effects (FE) models**: first differencing, mean centering
     - Assessing the relevance of FE (Hausman , Mundlak tests, …) vs random effect models
     - Beyond fixed effects using panels: dynamic models
   - Policy evaluation/treatment analysis (day 2)

Main Stata commands

1. **reg**
2. **xtreg/areg [in combination with ttset/xtset]**
3. **xtdescribe, xtline….**
4. **hausman**

Useful references

Cameron, A. C. & Trivedi, P. K. (2010). *Microeconometrics using Stata*. College Station: Stata Press.

1. INTRODUCTION

- The aim of this course is to review (& implement with STATA 14) some of the most commonly used methods to infer causal relationship using non experimental data
- Key is to identify the *causal* impact of some variable X^T on y
  - y the outcome variable (wage, health, score, GDP per capita…)
  - X^T the "treatment" ie the variable (or the policy) of interest (eg. one extra year of education, employment vs. unemployment, transfers to an underdeveloped territory…)
- Practical examples (ie. base on "real" micro evidence), including some directly related to our research
- Detailed STATA code + results available
- And students are invited to exercise

Material @ your disposal

MOODLE@UCL: LECME2FC

TOPIC 2\Panels\

ECcourse1.ppt

Code…\Stata_code\

#1EC_data.do

#1EC_Ex.do

#1EC_Ex_corr.do

#1EC_Extra.do

#1EC_FE.do ______________________

(+ corrected version at the end)

Data.zip

the various data sets @ your disposal

via the web: https://perso.uclouvain.be/vincent.vandenberghe/Stata_EC/Stata_EC1.html

List of topical issues addressed

* Does education contribute to firms' productivity? And how much?

* Is there gender wage discrimination in the Belgian private economy?
An how important is it?

* Do wages impact firm-level employment?

2. The endogeneity problem

Mincer suggests human capital impacts wage *W*. It is aquired via two channels

- Schooling (*S*)
- On-the-job learning/experience (NB: $EXP=t-S$))

[Eq. 1] $lnW= \alpha+\beta.S+\gamma.EXP+\delta.\ EXP^2+\epsilon$

and is β a good approximation of the **return** of an additional year of schooling as

[Eq. 2] $\beta = \partial lnW/\partial S=(\partial W/W)/\partial S \approx (W_{S+1}-W_s)/W_s$ for $dS=1$

A crucial (unrealistic?) assumption in Mincer equation is that the term ε_t *is a* pure random shock (i.e. its mean is equal to zero)

In truth, it could contain unmeasured/unobserved differences in innate ability

Econometricans show that β estimates can be biased if two conditions hold true

*there is an omitted variable that is a significant determinant of the dependent variable (e.g. ability, motivation influences wages);

* and it is correlated with one or more of the included independent variables (e.g. schooling)

Consider a log linear (true) model *($y=logW$)* of the form

[Eq. 3] $y = X\beta + Z\delta + \mu$
where

* X is a vector containing explanatory variables (=> schooling variable S);

* Z is omitted (unobserved) data [e.g. motivation, ability…] which is potentially partially correlated with y_i (i.e. partial correlation $\delta \neq 0$) and X (=> S)

* the error terms u is an unobservable but random variable having expected value 0 (conditionally on X **and Z**);

The problem is that the OLS estimated parameters based only on the observed *X,Y* vectors of values (but omitting *Z*), is given by:

[Eq. 4] $$\hat{\beta} = (X'X)^{-1}X'Y$$

Substituting for *Y* based on the true/assumed linear model => Eq.5,

[Eq. 5]
$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(X\beta + Z\delta + U) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'U \\ &= \beta + \boxed{(X'X)^{-1}X'Z\delta} + (X'X)^{-1}X'U.\end{aligned}$$

Taking expectations, $E((X'X)^{-1}X')E(U)$ falls out => $X'U$ has zero expectation (no correlation between U and X)

Remains in addition to the true β

[Eq. 6] $\boxed{E(X'X)^{-1}X'Z)\ E(\delta)}$

the omitted variable bias

Its magnitude is function of

i) δ =>correlation between *y* and *Z*

ii) $(X'X)^{-1}X'Z$ => partial correlation between *X* (that comprises S_i) and *Z*

More specifically, if $\delta > 0$ (earnings and ability are positively correlated and $(X'X)^{-1}X'Z > 0$ (the higher the ability, the higher the chosen level of education) ; OLS would be *upward* biased.

3. EXPERIMENTAL METHODS

Experimental research design offer the most plausibly unbiased estimates

But experiments are frequently infeasible due to cost or moral objections – e.g no one proposes to randomly assign smoking to individuals to assess health risks or to randomly assign divorce status to parents so as to measure the impacts on their children

4. Quasi-experimental: Panels

4.1. What are short panels?

Panel= time series where "individuals" (persons, firms, countries…) are observed several times consecutively (y_{it}, X_{it})

Short (vs. long) panel :

not many time periods (*t: 1…..T*) but many individuals (*i=1……..N*) ; small *T* but large *N*

	vatid	year	lnay	medu
1.	200068636	2008	4.942388	.
2.	200068636	2009	5.105524	11.44
3.	200068636	2010	5.106514	11.51
4.	200068636	2011	4.932135	11.95
5.	200068636	2012	5.148784	12.15
6.	200362111	2008	4.768067	12.49
7.	200362111	2009	4.82522	12.49
8.	200362111	2010	4.921506	12.36
9.	200362111	2011	4.876973	12.66
10.	200362111	2012	4.890286	12.76
11.	200362210	2008	4.52434	.
12.	200362210	2009	.	.
13.	200362210	2010	.	.
14.	200362210	2011	.	.
15.	200362210	2012	.	.
16.	200952524	2008	5.000585	11.41
17.	200952524	2009	5.063892	11.58
18.	200952524	2010	4.983079	11.49
19.	200952524	2011	5.028475	11.53

4.2. Panels as a way to account for unobserved individual fixed effects (FE)

The idea of using panel methods to identify a causal impact of "treatment" is to use an individual *i* as its own control, by including information from multiple points in time

Suppose that the omitted variable Z_i *a)* varies only across "individuals" and *b)* for, a given "individual", is constant over the duration of the panel => it is a fixed effect (FE)

[Eq. 7] $y_{it} = X^T{}_{it}\beta + \varepsilon_{it}$

where $\varepsilon_{it} = Z_i + u_{it}$

Mean-centering [or first differencing] of all data ($y_{it} - y_{i\cdot}$, $X^T_{it} - X^T_{i\cdot}$ ....) amonts to "purging" (unobserved) fixed effects Z_i

[Eq. 8] $\varepsilon_{it} - \varepsilon_{i\cdot} = Z_i - Z_i + \mu_{it} - \mu_{i\cdot}$

where, by definition, the average of time-invariant constant Z_i is equal to that constant … and disappears

The results from the FE estimation can be interpreted as follows; treatment matters if on average, **within** "individuals", a change of the intensity of the "treatment" ($X^T_{it} - X^T_{i\cdot}$), results in a statistically significant change of outcome ($y_{it} - y_{i\cdot}$)

➔ *#1EC_FE.do/1/Case 1*

```
list vatid year lnay lnak medu if _n<60 & medu~=.
```

	vatid	year	lnay	lnak	medu
2.	200068636	2009	5.105524	7.511604	11.44
3.	200068636	2010	5.106514	7.564138	11.51
4.	200068636	2011	4.932135	7.612374	11.95
5.	200068636	2012	5.148784	7.669642	12.15
6.	200362111	2008	4.768067	6.816887	12.49
7.	200362111	2009	4.82522	6.843645	12.49
8.	200362111	2010	4.921506	6.800828	12.36
9.	200362111	2011	4.876973	6.829672	12.66
10.	200362111	2012	4.890286	6.825442	12.76
16.	200952524	2008	5.000585	6.922398	11.41
17.	200952524	2009	5.063892	6.85603	11.58
18.	200952524	2010	4.983079	6.832728	11.49
19.	200952524	2011	5.028475	6.768981	11.53
20.	200952524	2012	4.955201	6.759745	11.57
21.	201105843	2008	4.007389	6.412219	13.92
22.	201105843	2009	4.116449	6.441791	13.93
23.	201105843	2010	4.008789	6.447455	13.85
24.	201105843	2011	4.163989	7.256665	13.95
25.	201105843	2012	4.20403	7.324895	13.93
26.	201107922	2008	3.468315	3.967794	12
27.	201107922	2009	[illegible]	[illegible]	[illegible]
28.	201107922	2010	[illegible]	[illegible]	[illegible]
29.	201107922	2011	[illegible]	[illegible]	[illegible]
30.	201107922	2012	[illegible]	[illegible]	[illegible]
31.	201258172	2008	[illegible]	[illegible]	[illegible]

Log of value added per worker

Log of capital per worker

Mean number of years of education among the workforce

```
. use f_edu, clear
(Belfirst: firm-level data on productivity & educ. attainment of workforce, 2008-)

. xtset vatid year
       panel variable:  vatid (strongly balanced)
        time variable:  year, 2008 to 2012
                delta:  1 unit
```

POOLED DATA/ OLS

```
reg lnay lnak medu i.year  /*Nb the inclusing of time fixed effect as a set of dummy variables*/
```

Source	SS	df	MS
Model	40599.4236	6	6766.5706
Residual	46834.2461	227,831	.205565731
Total	87433.6697	227,837	.383755359

Number of obs	=	227,838
F(6, 227831)	=	32916.82
Prob > F	=	0.0000
R-squared	=	0.4643
Adj R-squared	=	0.4643
Root MSE	=	.45339

lnay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnak	.3354548	.0007919	423.61	0.000	.3339027	.3370069
medu	.0158046	.0003226	48.99	0.000	.0151723	.016437
year						
2009	-.0114149	.0031199	-3.66	0.000	-.0175298	-.0053001
2010	.0049495	.0031037	1.59	0.111	-.0011336	.0110326
2011	.0164147	.0030946	5.30	0.000	.0103494	.02248
2012	.0138437	.0031247	4.43	0.000	.0077193	.0199681
_cons	2.300404	.0052393	439.06	0.000	2.290135	2.310673

Return of 1 extra year of educ. = 1.58 %

First Differences

```
. reg D.(lnay lnak medu) i.year

      Source |       SS           df       MS      Number of obs   =   173,150
-------------+----------------------------------   F(5, 173144)    =   4099.85
       Model |  2138.39043         5  427.678087   Prob > F        =    0.0000
    Residual |  18061.6115   173,144  .104315549   R-squared       =    0.1059
-------------+----------------------------------   Adj R-squared   =    0.1058
       Total |  20200.0019   173,149  .116662539   Root MSE        =    .32298

------------------------------------------------------------------------------
      D.lnay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnak |
         D1. |   .3430397    .002428   141.29   0.000      .338281    .3477985
             |
        medu |
         D1. |   .0012418   .0004387     2.83   0.005     .0003819    .0021016
             |
        year |
       2010  |   .0264739   .0022573    11.73   0.000     .0220497    .0308981
       2011  |   .0237632   .0022471    10.57   0.000     .0193589    .0281676
       2012  |   .0000718   .0022639     0.03   0.975    -.0043653     .004509
             |
       _cons |   -.014772   .0016675    -8.86   0.000    -.0180402   -.0115037
------------------------------------------------------------------------------
```

Return of 1 extra year of educ. = 0.04%

Mean Centering

```
. xtreg lnay lnak medu i.year, fe

Fixed-effects (within) regression               Number of obs      =    227,838
Group variable: vatid                           Number of groups   =     52,687

R-sq:                                           Obs per group:
     within  = 0.1114                                         min =          1
     between = 0.4899                                         avg =        4.3
     overall = 0.4601                                         max =          5

                                                F(6,175145)        =    3659.47
corr(u_i, Xb)  = 0.0948                         Prob > F           =     0.0000

------------------------------------------------------------------------------
        lnay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnak |   .3118138   .0021284   146.50   0.000     .3076423    .3159854
        medu |   .0019448   .0004258     4.57   0.000     .0011103    .0027794
             |
        year |
       2009  |  -.0126081   .0018124    -6.96   0.000    -.0161603   -.0090559
       2010  |   .0024383   .0018131     1.34   0.179    -.0011153    .0059919
       2011  |   .0142525   .0018179     7.84   0.000     .0106895    .0178155
       2012  |   .0031366   .0018485     1.70   0.090    -.0004864    .0067596
             |
       _cons |   2.571336   .0113542   226.47   0.000     2.549082     2.59359
-------------+----------------------------------------------------------------
     sigma_u |  .43266913
     sigma_e |   .2584925
         rho |  .73695723   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(52686, 175145) = 9.98               Prob > F = 0.0000
```

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

Return of 1 extra year of educ. = 0.19%

4.3. Assessing the relevance of FE

Are we sure fixed effects Z_i are correlated to X_{it} (and not random)?

If they are not correlated, then pooled OLS/fGLS (known as random effet estimation (RE) [ie. Z_i are randomly distributed, but not correlated with $X^T{}_{it}$]) will be preferable to FE because they use total varation (and not just within var.)

➔ **Hausman test**

Under the null hyp. that individual effects are random, FE and RE estimators should deliver the same coef. β. The Hausman test assesses the probability that the estimated coeffients are equal https://www.youtube.com/watch?v=54o4-bN9By4

```
. xtreg lnay lnak medu i.year

Random-effects GLS regression                   Number of obs     =    227,838
Group variable: vatid                           Number of groups  =     52,687

R-sq:                                           Obs per group:
     within  = 0.1109                                         min =          1
     between = 0.4922                                         avg =        4.3
     overall = 0.4624                                         max =          5

                                                Wald chi2(6)      =   73125.42
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        lnay |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnak |   .3292132   .0012381   265.90   0.000     .3267865    .3316399
        medu |   .0064203   .0003616    17.76   0.000     .0057115     .007129
             |
        year |
        2009 |  -.0127687   .0018029    -7.08   0.000    -.0163023   -.0092351
        2010 |   .0021555   .0018017     1.20   0.232    -.0013758    .0056868
        2011 |   .0135468   .0018044     7.51   0.000     .0100102    .0170835
        2012 |   .0037978   .0018329     2.07   0.038     .0002054    .0073902
             |
       _cons |   2.427189    .007192   337.48   0.000     2.413092    2.441285
-------------+----------------------------------------------------------------
     sigma_u |  .40774138
     sigma_e |   .2584925
         rho |  .71331372   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Return of 1 extra year of educ. = 0.64%

```
qui: xi: xtreg lnay lnak medu i.year, fe

estimates store fe

qui: xi: xtreg lnay lnak medu i.year, re

estimates store re

hausman fe re
```

	Coefficients (b) fe	(B) re	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
lnak	.3118138	.3292132	-.0173994	.0017312
medu	.0019448	.0064203	-.0044754	.0002249
_Iyear_2009	-.0126081	-.0127687	.0001606	.000185
_Iyear_2010	.0024383	.0021555	.0002828	.0002026
_Iyear_2011	.0142525	.0135468	.0007057	.0002208
_Iyear_2012	.0031366	.0037978	-.0006612	.0002399

```
                       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(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =      499.63
            Prob>chi2 =      0.0000
```

We thus reject the idea that FE are irrelevant

=>The Mundlak idea

The key to the Mundlak approach is to determine if unobservable fixed effect Z_i and x_{it} are correlated.

[Eq. 9] $y_{it}=\alpha+\beta x_{it}+Z_i+\varepsilon_{it}$

His idea is that such a correlation can be represented as a linear relation between Z_i and the time-invariant part (eg. mean) of the observed regressors

[Eq. 10] $Z_i=\gamma+\vartheta x_{i.}+ v_i$ where $x_{i.}$ is the mean x_{it} ; v_i a time-invariant random term

Putting the two equations together we get

[Eq. 11] $y_{it}=\alpha^{£}+ \beta x_{it}+ \vartheta x_{i.}+ \nu_i +\varepsilon_{it}$

And if $\vartheta=0$ then Z_i and the covariates are uncorrelated=> thus the random effect model dominates the fixed effect model

```
. reg lnay lnak medu m_lnak m_medu

      Source |       SS           df       MS      Number of obs   =   227,838
-------------+----------------------------------   F(4, 227833)    =  49526.96
       Model |  40665.9668         4  10166.4917   Prob > F        =    0.0000
    Residual |   46767.703   227,833  .205271857   R-squared       =    0.4651
-------------+----------------------------------   Adj R-squared   =    0.4651
       Total |  87433.6697   227,837  .383755359   Root MSE        =    .45307

------------------------------------------------------------------------------
        lnay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnak |   .3236533   .0036484    88.71   0.000     .3165025    .3308041
        medu |   .0019353   .0007446     2.60   0.009     .0004759    .0033948
      m_lnak |   .0107666   .0037329     2.88   0.004     .0034503    .0180829
      m_medu |   .0170558   .0008261    20.65   0.000     .0154367    .0186749
       _cons |   2.275325    .005024   452.90   0.000     2.265478    2.285172
------------------------------------------------------------------------------

. estimates store munlack

.
 test m_lnak m_medu

 ( 1)  m_lnak = 0
 ( 2)  m_medu = 0

       F(  2,227833) =  219.27
            Prob > F =    0.0000
```

We reject the idea of no correlation with fixed effect i.e; *ϑ=0*

$\Rightarrow$ **Making use of xtreg ressources**

xtreg,fe (using estimated α and β) delivers estimates of fixed effects

[Eq. 12] $Z^{£}_{i} = Y_{i.} - \beta X_{i.} - \alpha$

That can be used to assess the degree of correlation between Z_i and X_{it} and/or Y_{it}

```
. predict z_i, u          // compute empirical fixe
> ted predicted average
(73,277 missing values generated)

. list vatid year lnay z_i  if _n<10
```

	vatid	year	lnay	z_i
1.	200068636	2008	4.942388	.
2.	200068636	2009	5.105524	.1107292
3.	200068636	2010	5.106514	.1107292
4.	200068636	2011	4.932135	.1107292
5.	200068636	2012	5.148784	.1107292
6.	200362111	2008	4.768067	.1316205
7.	200362111	2009	4.82522	.1316205
8.	200362111	2010	4.921506	.1316205
9.	200362111	2011	4.876973	.1316205

```
. corr medu z_i  //compute correla
(obs=227,838)
```

	medu	z_i
medu	1.0000	
z_i	0.1192	1.0000

```
. corr lnay z_i  // compute correl
> tion)
(obs=227,838)
```

	lnay	z_i
lnay	1.0000	
z_i	0.6987	1.0000

#1EC_Ex 1.do

Case study : assessing gender wage discrimination using panel micro data

Outline

1. Introduction: stylized facts & key concepts about gender wage discrimination (GWD)

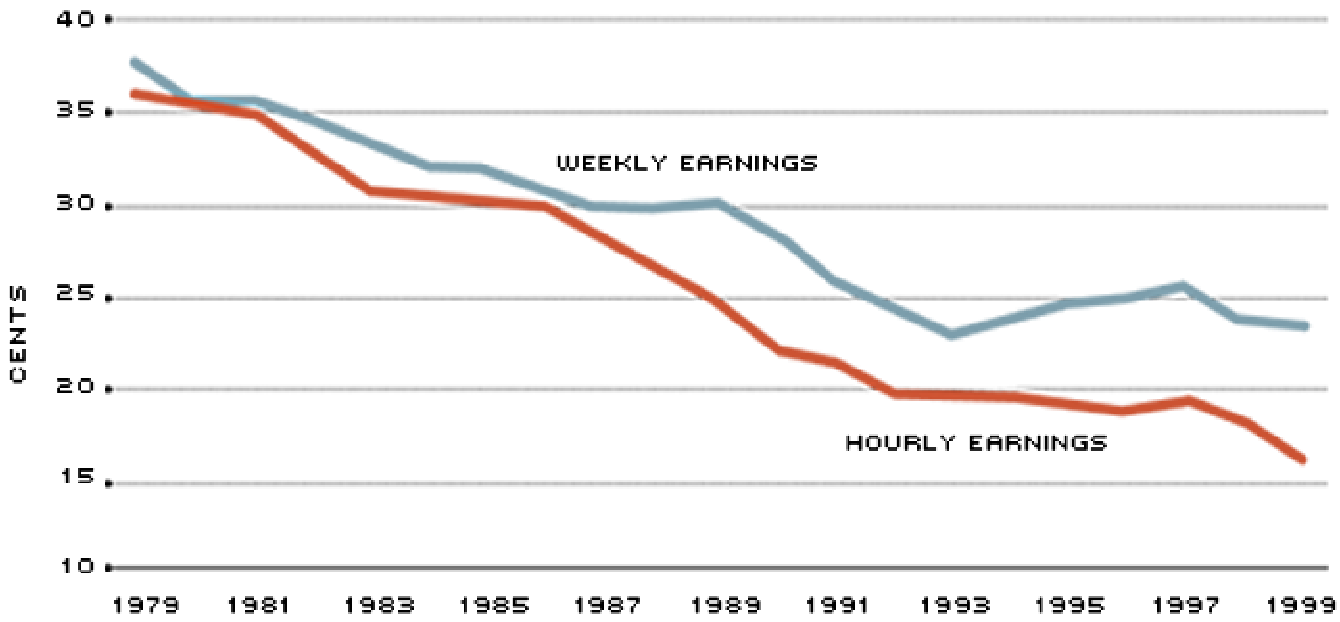
2. Estimating GWD using individual-level wage data
   - Framework
   - Implementation using Social Security individual data on gross wage

3. Estimating GWB using firm-level evidence (and fixed effects)
   - Framework
   - Implementation using Bel-first firm-level data on *i)* productivity *ii)* labour cost and *iii)* gross profits (or the inverse of unit labour cost)

1. Introduction: concepts & stylized facts

Evidence of substantial average earning differences between categories (men/women, race, country of origin…) – the Gender Wage Gap (GWG) — is a persistent social outcome in the labour markets of most developed economies

USA- The Gender Wage Gap, 1979-99



SOURCE: U.S. Department of Labor

- In 1999, the gross pay differential between women and men in the **EU-27** was, on average, 16% (European Commission, 2007) (weekly earnings)

- In **the U.S**. this figure amounted to 23.5% (weekly earnings)

- **Belgian** statistics (Institut pour l'égalité des Femmes et des Hommes, 2013) = > "*Women earn on average 10% less per hour of work then men. Many women work part-time, so that the annual gender wage gap is 23%* "

For most sociologists wage discrimination manifests itself by a lower pay for a minority group with respect to the majority group

Strictly speaking however, **for economists**, wage discrimination requires more that wage differences between groups

It implies that equal labour services provided by **equally productive workers have a sustained price/wage difference**

The standard empirical approach among economists to the measuring gender wage discrimination consists of estimating earning equations (cfr Oaxaca-Blinder in A.1). Wage discrimination is measured as the average mark-up on individual compensation (hourly, monthly wages...), associated to gender, controlling for individual productivity-related characteristics

e.g

[Eq. 12] $Ln\ W_i = \alpha + \beta DF_i + X'_i\gamma + \epsilon_i$

where

W_i = compensation

DF_i = female(1)/male(0) dummy

X'_i = vector of productivity-related characteristics (experience, education…)

And with a log linear speficification, β is a good approximation of **the conditional gender wage gap in percentage points**

[Eq. 13] $\beta = \partial lnW_i / \partial dF_i = (\partial W_i / W_i) / \partial dF_i$

$\approx (W_i^{dFi=1} - W_i^{dFi=0}) / W_i^{dFi=0}$ for $dF_i = 1$

➔ *#1EC_FE.do/1/Case 2/Part 1*

```
/*Econmetrics*/
scalar drop _all
use w_db , clear
*A. OLS + industry fixed effects (within industry identification)
qui: xi:reg lnw dum2 year        /*raw difference */
scalar gwg1=_b[dum2]
qui:reg lnww dum2 year            /*+ accounting for quadrimestrial working time differences*/
scalar gwg2=_b[dum2]
qui:reg lnww dum2 agex agex2      /*+ accounting for age (proxy of experience) with age squared*/
scalar gwg3=_b[dum2]
qui:areg  lnww dum2 agex agex2 , absorb(nace2)   /*+ accounting for industry 2 digits*/
scalar gwg4=_b[dum2]
    /*NB: areg ..., absor(nace2) equivalent to
   xi: reg lnww dum2 agex agexsq year i.nace2*/
qui:areg  lnww dum2 agex agex2 , absorb(nace)   /*+ accounting for industry 5 digits*/
scalar gwg5=_b[dum2]

*B. Display key results
scalar list gwg1 gwg2 gwg3 gwg4 gwg5

*C. Extention: GWD stable over time/years ?
scalar drop _all
sort year

local y 2002 2003 2004  2005 2006 2007 2008 2009 2010
foreach i of local y  {
use w_db if year==`i' , clear
qui: areg lnww dum2 agex agex2, absorb(nace)
qui: scalar gwg_`i'=_b[dum2]
scalar list gwg_`i'
}
```

```
*B. Display key results
scalar list gwg1 gwg2 gwg3 gwg4 gwg5
     gwg1 = -.40733842
     gwg2 = -.21057828
     gwg3 =  -.1949094
     gwg4 = -.15787496
     gwg5 = -.14407816
```

```
gwg_2002 = -.16309768
gwg_2003 = -.15211776
gwg_2004 = -.15171536
gwg_2005 = -.14717418
gwg_2006 = -.14349608
gwg_2007 = -.14008865
gwg_2008 =  -.1465544
gwg_2009 = -.13148378
gwg_2010 = -.13042656
```

3. Using firm-level data

What is missing from the above studies is an independent measure of productivity

By contrast, with firm-level data, the idea is to use firm-level **direct** measures of gender productivity and wage differentials via, the estimation of a productivity and a labour cost equations, both expanded by the specification of a labour-quality index à-la-Hellerstein & Neumark (2004)

3.1. The Hellerstein-Neumark framework

In order to estimate labour productivity, following Hellerstein *et al.*, 1999 we consider a Cobb-Douglas production function

[Eq 14] $Y_{jt} = A_{jt}\, \boldsymbol{QL_{jt}}^{\alpha} K_{jt}^{\beta}$

where Y_{jt} is output/ production in firm j at time t, K_{jt} is the stock of capital

The variable that reflects the gender heterogeneity of the workforce is *the quality of labour index* QL_{jt}

Let L_{jlt} be the number of workers of type l (men/women…) in firm j at time t, and μ_l be their marginal relative productivity* (supposedly uniform across firms). We assume that workers of various types are substitutable with different marginal products. Focusing on gender, labour quality indce can be specified as:

[Eq 15] $QL_{jt} = \sum_l \mu_l L_{jlt} = \mu_M L_{jMt} + \mu_F L_{jFt}$

[Eq 16] $Y_{jt} = A\,(QL_{jt})^{\alpha} K_{jt}^{\beta} = A\,[\mu_M L_{jMt} + \mu_F L_{jFt}]^{\alpha} K_{jt}^{\beta}$

Dropping t and j...

*$MLP_M \equiv \delta Y / \delta L_M = A\,\alpha\,[\mu_M L_M + \mu_F L_F]^{\alpha-1} \mu_M K_i^{\beta}$

*$MLP_F \equiv \delta Y / \delta L_F = A\,\alpha\,[\mu_M L_M + \mu_F L_F]^{\alpha-1} \mu_F K_i^{\beta}$

…

thus relative $MLP \equiv (\delta Y / \delta L_F)/(\delta Y / \delta L_M) = \mu_F / \mu_M$

Let us now consider labour productivity per worker in logs

[Eq. 17] $ln\ (Y_{jt}\ /L_{jt}) = lnA + \alpha\ ln\ QL_{jt} + \beta lnK_{jt} - lnL_{jt}$

And lets transform the labour quality index

[Eq. 18] $QL_{jt} = \mu_M\ L_{jt} + (\mu_F - \mu_M)\ L_{jFt}$

where male workers= ref.

Mult/div. rhs term by $\mu_M L$ and taking logarithms

[Eq. 19] $ln\ QL_{jt} = ln\ \mu_M + lnL_{jt} + ln\ (1 + (\lambda - 1)\ P_{jFt})$

where $\lambda \equiv \mu_F/\mu_M$ is the relative marginal productivity of women and $P_{jFt} \equiv L_{jFt}/L_{jt}$ the proportion/share of females in firm j.

Since $ln(1+x) \approx x$, for small values of x we can approximate Eq. 10 by:

[Eq. 20] $Ln\ QL_{jt} = ln\ \mu_M + ln\ L_{jt} + (\lambda - 1)\ P_{jFt}$

and the production function becomes:

[Eq. 21] $ln(Y_{jt}/L_{jt}) = lnA + \alpha\ [ln\mu_M + ln\ L_{jt} + (\lambda - 1)\ P_{jFt}] + ß\ lnK_{jt} - lnL_{jt}$

or, equivalently

$$\boxed{\text{[Eq. 22] } ln\ (Y_{jt}/L_{jt}) = B + (\alpha - 1) l_{jt} + \boldsymbol{\eta}\ \boldsymbol{P_{jFt}} + ß\ k_{jt}}$$

where:

- $B = lnA + \alpha ln\ \mu_M$; $\lambda = \mu_F/\mu_M$; $\boldsymbol{\eta} = \alpha(\lambda - 1)$;
- $l_{jt} = lnL_{jt}$, $k_{jt} = lnK_{jt}$

NB: Eq. 13 , being loglinear in *P*, coefficients ***η/10*=>**the percentage change of average labour productivity due to a 1/10 unit (i.e 10 percentage points) change of women' share

Similarly, for labour cost per worker

[Eq. 23] $W_{jt}/L_{jt} = \pi_M + (\pi_F - \pi_M)\, L_{jFt}/ L_{jt}$

Mult/div rhs term by π_M/L_{jt} , taking the logs and using $log(1+x) \approx x$, we get

[Eq. 24] $ln(W_{jt}/L_{jt}) = ln\pi_M + (\Phi - 1)\, P_{jFt}$

where $\Phi \equiv \pi_F/\pi_M$ is the rel. remuneration of women

[Eq. 25] $ln\,(W_{jt}/L_{jt}) = B^w + \boldsymbol{\eta^w P_{jFt}}$

where: $B^w = ln\pi_M$; $\eta^w = \Phi - 1$

Like in the productivity equation, coefficients $\boldsymbol{\eta^w}$ capture the sensitivity to changes of the gender structure (P_{jMt})

A key hypothesis test can now be formulated.

No gender wage discrimination => alignment of rel. productivity and rel. labour costs ⇔

$\eta = \eta^w$

This test that can easily implemented, if we adopt strictly equivalent econometric specifications for productivity & labour cost equations

[Eq. 26] $ln(Y_{jt}/L_{jt})=B+(\alpha-1)l_{jt}+\eta P_{jFt}+....+ ß\, k_{jt}+\varepsilon_{jt}$

[Eq. 27] $ln\,(W_{jt}/L_{jt})=B^{w}+(\alpha^{W}-1)l_{jt}+\eta^{W}P_{jFt}{}^{....+} ß^{\,w}k_{jt}+\varepsilon^{w}{}_{jt}$

And if, if we take the *difference* between we get a direct expression of the productivity-labour cost gap (ratio)= gross profits as a linear function of its workforce determinants.

[Eq. 28] $ln(Y_{jt}/L_{jt})-ln(W_{jt}/L_{jt})=B^{G}+(\alpha^{G}-1)l_{jt}+\eta^{G}P_{jFt}...+ ß^{\,G}k_{jt}+\varepsilon^{G}{}_{jt}$

where: $B^{G}=B-B^{w}$; $\alpha^{G}=\alpha-\alpha^{W}$, $\eta^{G}{}_{1}=\eta-\eta^{w}$;... $ß^{\,G}= ß - ß^{\,w}$; $\varepsilon^{G}=\varepsilon-\varepsilon^{w}$

Conclusion

if $\eta^{G}=0$ <=> no gender wage discrimination

if $\eta^{G}>0$ <=> negative gender wage discrimination (women are underpaid)

if $\eta^{G}<0$ <=> positive gender wage discrimination (women are overpaid)

3.2. HN and panel (firm-level) data: econometric identification

As to proper identification of the causal links, one of the challenges consists of dealing with the various constituents of the residual ε_{jt}

Assume that the latter has a structure that comprises two elements:

[Eq. 29] $\varepsilon_{jt} = \vartheta_j + \sigma_{jt}$

where: $COV(\vartheta_j, P_{jF,t}) \neq 0$, $CORR(\vartheta_j, Y_{jt}) \neq 0$

In other words, the OLS sample-error term potentially consists of *i)* an unobservable firm fixed effect ϑ_j; *ii)* a purely random term σ_{jt}.

Econometric identification

ϑ_{j}. represents firm-specific characteristics that are unobservable but driving labour productivity. And these might be correlated with gender mix, biasing OLS results (cfr omitted variable bias). Men for instance might be overrepresented among in sectors/firms with higher TFP embedded in used technology (eg. manufacturing vs services/commerce)

Solution

Using the panel structure of data and estimating a fixed effect model

<=> mean-centering of all data $(Y_{jt}\text{-}Y_{j\cdot};L_{jt}\text{-}L_{j\cdot}\ldots)$ => purging fixed effects and thus coping with unobserved heterogeneity terms ϑ_{j}

[Eq. 30] $\varepsilon_{jt} - \varepsilon_{j\cdot} = (\vartheta_{j} - \vartheta_{j}) + (\sigma_{jt} - \sigma_{j\cdot})$

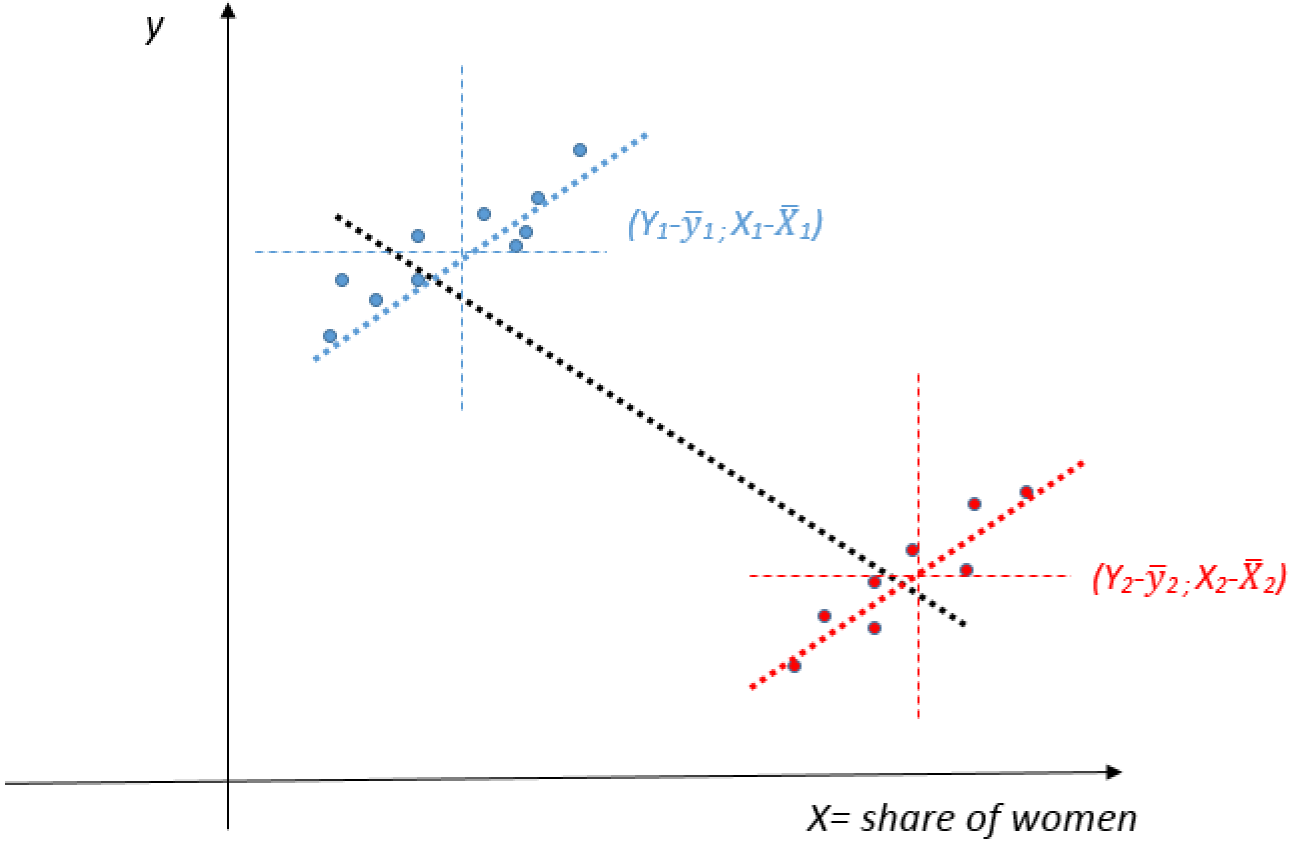


Illustration of the importance of accounting for firm fixed effects

The results from the fixed-effect estimation can be interpreted as follows:

a group (male or female) is estimated to be more (less) productive/costly/profitable if, within firms, an increase of that group's share in the overall workforce translates into productivity /labour cst/profit gains (loss).

➔ *#WS_FE.do/1/Case 2/Part 2*

```
/*Econometrics*/
 scalar drop _all
use f_db , clear

*A. OLS
    /*productivity*/
areg lnyha lnk lnh sfem sbcol magey p25agey p75agey spt year, absorb(nace) /*per hour*/
scalar prod_ols=_b[sfem]

    /*labour cost*/
areg lnwha lnk lnh sfem sbcol magey p25agey p75agey spt year, absorb(nace) /*per hour*/
scalar lcost_ols=_b[sfem]

    /*gross profit*/
areg lnp lnk lnh sfem sbcol magey p25agey p75agey spt year, absorb(nace)
scalar gprof_ols=_b[sfem]

scalar list prod_ols lcost_ols gprof_ols


*B. Firm fixed effects  (i.e. within-firm identification)
/*NB areg y x, absorb (vatid)  is equivalent to ...
    tsset vatid year
    xtreg y, fe */

    /*productivity per hour*/
areg lnyha lnk lnh sfem  sbcol magey p25agey p75agey spt year, absorb(vatid) /*per hour */
scalar prod_fe=_b[sfem]

    /*labour cost per hour*/
areg lnwha lnk lnh sfem sbcol magey p25agey p75agey spt year, absorb(vatid)  /*per hour*/
scalar lcost_fe=_b[sfem]

    /*gross profit per hour*/
areg lnp lnk lnh sfem sbcol magey p25agey p75agey spt year, absorb(vatid)
scalar gprof_fe=_b[sfem]


*C. Display key results
scalar list prod_ols lcost_ols gprof_ols
scalar list prod_fe lcost_fe gprof_fe
```

lnp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnk	.0209938	.0009308	22.55	0.000	.0191694	.0228182
lnh	-.040875	.0017359	-23.55	0.000	-.0442774	-.0374726
sfem	.0398353	.0099136	4.02	0.000	.0204047	.0592659
sbcol	-.0047309	.0067019	-0.71	0.480	-.0178666	.0084048
magey	-.0024131	.0018394	-1.31	0.190	-.0060183	.0011921
p25agey	-.0040573	.0009315	-4.36	0.000	-.0058831	-.0022315
p75agey	-.0036306	.0009226	-3.94	0.000	-.005439	-.0018223
spt	.0000726	.0001159	0.63	0.531	-.0001545	.0002997
year	-.0003826	.0005324	-0.72	0.472	-.0014261	.000661
_cons	1.840036	1.060883	1.73	0.083	-.2392889	3.919361
nace	F(607, 75928) =		30.315	0.000	(608 categories)	

lnp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnk	.0080074	.0018251	4.39	0.000	.0044302	.0115846
lnh	-.0701906	.0027601	-25.43	0.000	-.0756005	-.0647807
sfem	.0758839	.0176354	4.30	0.000	.0413185	.1104493
sbcol	.0248527	.0119847	2.07	0.038	.0013627	.0483426
magey	-.0083125	.0015299	-5.43	0.000	-.0113111	-.0053139
p25agey	-.0004581	.0007769	-0.59	0.555	-.0019809	.0010646
p75agey	.0008625	.0007732	1.12	0.265	-.0006529	.002378
spt	.0011662	.0001556	7.50	0.000	.0008613	.0014712
year	-.000083	.0004462	-0.19	0.853	-.0009576	.0007917
_cons	1.527152	.8694241	1.76	0.079	-.176919	3.231222
vatid	F(9309, 67226) =		19.343	0.000	(9310 categories)	

```
. *C. Display key results
. scalar list prod_ols lcost_ols gprof_ols
  prod_ols = -.14698709
 lcost_ols = -.18682239
 gprof_ols =    .0398353

. scalar list prod_fe lcost_fe gprof_fe
   prod_fe = -.04751146
  lcost_fe = -.12339532
  gprof_fe =  .07588387
```

#1EC_Ex.do/Ex 2 & 3

References

Blinder, Alan S. 1973. Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources* 8 (4): 436–455.

Hellerstein, J.K. & D. Neumark, 2004. "Production Function and Wage Equation Estimation with Heterogeneous Labor: Evidence from a New Matched Employer-Employee Data Set," NBER Working Papers 10325, National Bureau of Economic Research, Inc.

Oaxaca, Ronald L. 1973. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review* 14 (3): 693–709.

Appendix- Oaxaca-Blinder in a Nutshell

Step 1 – estimate separately two Mincer-like equations

$Ln\ W^m{}_i = \alpha^m + X^m{}_i\, \beta^m + \varepsilon^m{}_i$

$Ln\ W^f{}_i = \alpha^f + X^f{}_i\, \beta^f + \varepsilon^f{}_i$

where X is a vector of variables proxying productivity (eg. Educational attainment, experience, exp^2…)

Step 2 – use estimates and initial data to compute

$Ln\ W^m{}_i - Ln\ W^f{}_i = \hat{\alpha}^m - \hat{\alpha}^f + X^f{}_i\, (\hat{\beta}^m - \hat{\beta}^f) + (X^m{}_i - X^f{}_i)\, \hat{\beta}^f$

with

- Explained difference $(X^m{}_i - X^f{}_i)\, \hat{\beta}^f$
- Unexplained men-womend wage gap (i.e. discrimination)

  $\hat{\alpha}^m - \hat{\alpha}^f + X^f{}_i\, (\hat{\beta}^m - \hat{\beta}^f)$