Exam - Applied Econometrics for Researchers, PhD

13 December 2022

Practical guidelines

The exam starts December 13, 2022, at 9.00 in K2.53.

Kindly bring your own laptop with Stata installed (or the equivalent software you plan to use for the exam; if you plan to use a different software, the hand-in requirements below applies to a code file and a log file as produced by that software).

You will have access to the internet via eduroam and you may use all books, notes, etc. during the exam, but you cannot help others or receive help from others during the exam.

Consider the assignments below. Respond to all questions contained in the assignments. More detailed answers, answers consistent with econometric reasoning, and clear accounts of actions are rewarded in the evaluation and grading. All assignments are solvable using the tools taught during the course.

For each problem, there is an indicative weight in the overall evaluation.

Files for the exam can be found in the folder “Exam” on the CANVAS page for this course.

You should use the provided .do-file AERexam.do. **Rename the file as AERfinal.do** and fill in own programming where appropriate.

At the end of the exam, you are kindly asked to hand in

- Your completed AERfinal.do file
- The .log file (AERexam.log) generated after running the complete program a final time
- A document with your final answers to each question, describing and interpreting the results as required in the assignments.

Hand in your answers at **13.00** at the latest. All hand-ins should be send to: hck.si@cbs.dk and to vr.si@cbs.dk

Good luck!

This exam is based on the study by Beaudry and Lewis (2014) (B&L). They examine the determinants of a decrease in the male-female wage gap observed in the US over the period 1980-2010.[1]

They focus on potential common explanations for the decrease in the male-female wage gap and the increase in the "returns to education" (the wage gap between high- and low-educated individuals) that largely happened over the same period – see the figure below, retrieved from their paper.

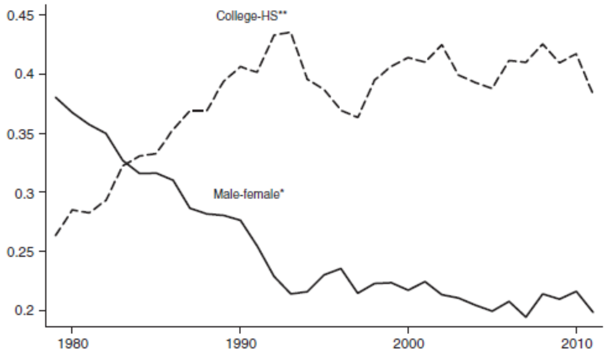


FIGURE 1. ADJUSTED MALE-FEMALE*, COLLEGE-HS** WAGE GAPS, 1979–2011

B&L motivate their analysis as follows: *"The conjecture pursued ... is that* ***the diffusion of IT [information technology]*** *increased the relative price of [cognitive or interpersonal skills] versus [physical skills], and given that women and more highly educated individuals appear ... to be relatively better endowed with this attribute, this* ***should have caused the male-female wage differential to decrease most and the return to education to increase most in cities that adopted IT most intensively."***

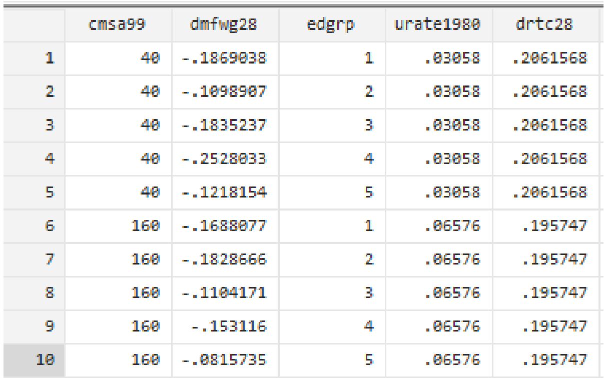
Specifically, they argue that the introduction of personal computers (PCs) into the workplace can be seen as a skill-specific shock that differentially affected the demand for high-skilled and low-skilled labour, and that women were generally better positioned to benefit from the change in demand having a comparative advantage in terms of "cognitive" and "people" skills.

To test their conjecture, B&L constructed male-female wage gaps separately for 230 cities ("metropolitan areas") in the US and related the variations in the changing male-female wage gaps across cities to variations in how intensively firms in a given city adopted PCs (personal computers) during the period 1980-2000.

[1] Paul Beaudry and Ethan Lewis (2014). "Do Male-Female Wage Differentials Reflect Differences in the Return to Skill? Cross-City Evidence from 1980–2000" *American Economic Journal: Applied Economics, 6(2): 178–194*. The replication data for the study can be downloaded from https://www.openicpsr.org/openicpsr/project/113886/version/V1/view.

Data:

The dataset “stackcity.dta” contains information on 230 unique metropolitan areas and five education groups - high school dropouts (DO), high school graduates (HS), those with some college education (but less than four years, SC), four-year college graduates (CO), and graduates with advanced degrees (AD). There is a total of 1,150 observations (5 rows of data, corresponding to 5 different education groups, for each of the 230 metropolitan areas). **Note that some variables are not measured separately for each education group and hence repeated within each area.** The snapshot below gives a preview of the dataset structure. For illustration, the metropolitan area “40” has five observations, one for each education group (**edgrp**); **dmfwg28** varies for each **edgrp**, but other variables like **urate1980** and **drtc28** do not, being thus repeated within each **cmsa99**.



	cmsa99	dmfwg28	edgrp	urate1980	drtc28
1	40	-.1869038	1	.03058	.2061568
2	40	-.1098907	2	.03058	.2061568
3	40	-.1835237	3	.03058	.2061568
4	40	-.2528033	4	.03058	.2061568
5	40	-.1218154	5	.03058	.2061568
6	160	-.1688077	1	.06576	.195747
7	160	-.1828666	2	.06576	.195747
8	160	-.1104171	3	.06576	.195747
9	160	-.153116	4	.06576	.195747
10	160	-.0815735	5	.06576	.195747

The table below briefly describes the key variables in the file:

Variables in the stackcity.dta dataset

Variable name	**Variable description**
cmsa99	Code identifying each metropolitan area
dmfwg28	Change in the male-female gap between 1980-2000
edgrp	Categorical variable taking values 1 to 5, each of which corresponds to different education groups (1=DO, 2=HS, 3=SC, 4=CO, 5=AD)
urate1980	Unemployment rate in the metropolitan area in 1980
drtc28	Change in wage gap between short (HS) and long-educated (CO) workers
pcemp_m_c2000	PCs per worker in 2000
HIGHpcemp_m_c2000	Dummy = 1 if the metropolitan area has pcemp_m_c2000 above the sample median, 0 otherwise
durmf1980	Percentage of employment in durable manufacturing in 1980
nondmf1980	Percentage of employment in non-durable manufacturing in 1980
durmf1980lo	durmf1980 x workers with some college education or less
nondmf1980lo	nondmf1980 x workers with some college education or less
pctbl1980	Percentage of Black employees in the metropolitan area in 1980
lnlf1980	Ln(labor force) in 1980
erate_f1980lo	Rate of low-educated (some college or less) women in the population, 1980
erate_f1980hi	Rate of highly-educated (college or more) women in the population, 1980
lncs1980	Ln(Coll/HS Equiv Hrs), 1980 (measure of education mix)
dlncs28	Change in education mix, 1980-2000
mexish1980	Share of Mexican born, 1980

Question 1 (30%)

a) Obtain a descriptive table to summarize the 1980-2000 change of the male-female wage gap across cities, **dmfwg28**, for different educational groups (**edgrp)**. Which educational group has seen the largest decline in the gender wage gap? Is there a significant difference between the gender wage gap developments for workers with a high school degree (HS) and workers with a 4-year college degree (CO)?

b) To investigate if the rate of PC adoption by firms in the early 2000s relates to the initial labor market situation in a city, start by dividing metropolitan areas into two groups: those that had a 1980 unemployment rate (**urate1980**) below the national median and those with an 1980 unemployment rate above the median. After constructing this variable, test whether the average rate of PC adoption, **pcemp_m_c2000,** was different between these two groups of cities. How many observations are you basing this test on? What do you conclude based on this test?

c) Consider the raw correlation across cities between the change in the male-female wage gap, **dmfwg28**, and the intensity of PC use in firms, **pcemp_m_c2000.** Compute this correlation separately for each education group (**edgrp**). Consider also the correlation between PC use and the change in the earnings differential between short- and long-educated workers, **drtc28.** Do your findings support the B&L conjecture described on page 2?

d) Run a linear regression that tests B&L's conjecture more formally, including all education groups. The dependent variable is the change in the male-female wage gap within a particular education group, **dmfwg28.** The explanatory variable is the measure of intensity of PC use in firm, **pcemp_m_c2000.** Include dummy variables for each education group, defined by **edgrp**. Interpret the coefficient estimates. Are there statistically significant differences between education groups?

e) Now consider the "returns to education" and run a regression for the change in the earnings differential between short- and long-educated workers, **drtc28**, on the measure of intensity of PC use in firm, **pcemp_m_c2000**. How many observations should you use to fit this regression?

f) Going back to the male-female wage gap regression estimated in d), the main analysis of B&L consists of a single regression that included all five education groups, imposing equal effects of the intensity of PC use on the male-female wage gap across groups. Estimate an extended version of the model that allows the marginal effect of PC use to differ across educational groups. Provide a graphical illustration to show the estimated marginal effects of the educational groups and comment on your graph. Implement a test of B&L's assumption that groups have equal effects. Are the marginal effects of PC use on the gender wage gap equal across education groups?

g) B&L suggest controlling for some alternative explanations because *"[the decrease in the male-female wage gap] could be due to differences in industry mix. The decline in blue-collar wages, thought to perhaps account for a quarter of the decline in the male-female wage gap in the 1980s ... will have been more of a factor in initially more blue-collar locations."* Specifically, they add controls for the 1980 shares of durable and nondurable manufacturing in a city interacted with broad education categories (workers having some college education or less *versus* workers having a college degree or more). Run a regression in which you extend your model d) with the following variables: **durmf1980, nondmf1980, durmf1980lo, nondmf1980lo**. Is this a relevant extension of the model? How is your estimate of the effect of PC use on the male-female wage gap affected by this extension?

Question 2 (25%)

Let us now focus on the introduction of PCs in the workplace – some cities did so to a much higher extent than others. The database includes a dummy variable **HIGHpcemp_m_c2000**, which distinguishes metropolitan areas that introduced more PCs per worker than the sample median from those with more modest computer adoption patterns.

a) Estimate a binary model predicting the probability that a metropolitan area introduces more PCs per worker than the sample median and use as explanatory variables the following initial conditions of the city, measured in 1980: **urate1980**, **pctbl1980**, **lnlf1980**, **erate_f1980lo**, and **erate_f1980hi**. Make sure to use only one observation per metropolitan area in your estimation. What do you conclude based on this output, i.e. which regional conditions in 1980 contributed to a more or less pronounced computer adoption at work in the early 2000s?

b) Based on the model estimated in a), how do you interpret the coefficient of **erate_f1980lo**, i.e. the relationship between the average employment rate among women with some college education or less in the population in 1980 and the probability that the focal metropolitan area adopts PCs at work at a rate above the sample median?

c) Using Propensity Score Matching and the city-level initial characteristics listed in a) as matching variables, estimate the ATET of **HIGHpcemp_m_c2000** on the 1980-2000 change in the male-female wage gap (**dmfwg28)** within high school dropouts (i.e. **edgrp** = 1). Repeat this analysis using at least three "nearest neighbors" in your analysis. What do you conclude based on these two estimates, and why do they differ slightly? How do these effects compare to the estimate you would obtain in an OLS regression for **dmfwg28** on **HIGHpcemp_m_c2000**, controlling for the remaining variables listed in a)? Why is it important to account for these initial characteristics of the city, either as control variables in a linear regression or as matching variables in PSM?

d) Are metropolitan areas with high and low computer adoption at work more comparable in other observable characteristics after the matching you have performed in c), using three "nearest neighbors"? In which initial conditions do you still spot substantial differences between the two groups of cities and how would you try to improve the balance between them, if you had the chance?

e) What do you conclude regarding the validity of the overlap assumption in this case?

Question 3 (25%)

B&L use an instrumental variables approach to address the possible endogeneity of IT adoption. They use insights from the endogenous technology adoption literature to suggest that if PCs complement cognitive skills, then areas that were initially more educated should experience faster adoption of PCs. They suggest using the skill mix in 1980 in different US cities, **lncs1980,** as an instrumental variable. This is a measure of pre-PC era education mix, computed as the log ratio of hours worked by college graduates to hours worked by people with a high school equivalent education.

NOTE: In this exercise, we limit the analysis to workers with a college degree (**edgrp** = 4) and no longer need to include the dummy variables distinguishing between different education groups.

a) The motivation for B&L's use of an IV approach was a comment from Reviewer 1. The reviewer criticized that B&L were relying on a specification that relates the *change* of the male-female wage gap *over the period from 1980 to 2000* to the *actual use* of PCs in firms *in the particular year* 2000. Is this a relevant critique in your opinion? Why (not)?

b) Follow the authors' suggestion and implement an IV regression for the change of the male-female wage gap, **dmfwg28**, on the measure of intensity of PC use in firms, **pcemp_m_c2000**, using **lncs1980** as an instrument. Include as controls the initial industry mix (**durmf1980** and **nondmf1980**) in your regressions. What is the estimate for the coefficient of **pcemp_m_c2000** with this approach?

c) Evaluate their choice of instrument. Is it a relevant and valid instrument? Provide arguments and explicit tests to support your conclusions.

d) Reviewer 2 now suggests that B&L could do even better by adding more instrumental variables to their IV regression. In particular, the reviewer suggested that, in addition to **lncs1980**, they employ variables related to i) the size of the city's initial labor force, **lnlf1980,** to capture agglomeration effects, and ii) the composition of the city's labor force related to the extent of immigration, in particular immigration from Mexico (**mexish1980**). (How) would you implement the reviewer's suggestions? Provide arguments and explicit tests (i.e. for the relevance and validity of the instruments) to support your recommendations.

e) A third reviewer raises yet another concern with the analyses conducted by B&L – the reviewer argues that there could be an increasing positive selection of women into the labor market, so that women that are more skilled enter the workforce more often over time. Because the dependent variable only takes into account men and women who actually work, the reviewer wonders about sample selection bias in this case. Elaborate a brief response to the reviewer in which you explain why one should (not) care about sample selection bias in this study, and how you would address it if you could.

Question 4 (20%)

In their paper, B&L find empirical support for their argument that the decline in the male-female wage gap and the increase in the returns to education have a common cause: the spread of IT in the workplace. There are, however, theories to suggest that, for example, changes in labor supply linked to immigration flows to a region play a similar role. Specifically, we will assess whether developments of the male-female wage gap can be linked to different immigration patterns in US cities after 2000.

For this analysis, we will use a subset of the previous data, reorganized as a two-period "long" panel dataset. We will limit the analysis to workers with a college degree (**edgrp** = 4). The cross-sectional units are the metropolitan areas, and we distinguish between two sub-periods: the period from 1980 until 2000 (the "before" period, **per** = 0), and the period from 2001 until 2010 (the "after" period, **per** = 1).

The variable **dmfwga** measures the average yearly change in the male-female wage gap over each sub period. The variable **drtca** is similarly the average yearly change in the returns to education.

The migration literature has shown that immigrant flows are persistent, as immigrants to a country tend to settle in cities and regions where immigrants are already well-represented in the population. Based on this insight, we will assume that post-2000 immigration patterns are largely predetermined. We distinguish between high- and low-immigration cities by defining the dummy variable **highimm**, which equals 1 for cities that had a proportion of immigrants in 1980 larger than 1.5 per cent, 0 otherwise.

a) The following 2x2 table summarizes the average change of the male-female wage gap, **dmfwga,** before/after 2000 in high/low immigration cities, respectively. Based on the values provided below, calculate and interpret the difference-in-differences estimate of the effect of a high immigrant proportion on the average change in male-female wage gap.

Average change in the male-female wage gap (**dmfwga**)	**per** = 0	**per** = 1
highimm = 0	-0.0027016	-0.001361
highimm = 1	-0.0050217	-0.000507

b) The output below is a pooled OLS regression of **dmfwga** on **per, highimm,** and their interaction. Based on this ouput:

i. How do you interpret the constant term in this model?

ii. Was there a significant difference between low- and high-immigration cities in how much the gender wage gap changed annually during the period 1980-2000?

iii. Was the change in the male-female wage gap any different in low-immigration cities after 2000 (compared to 1980-2000)?

iv. Was there a difference in terms of the evolution of the gender wage gaps in high- and low-immigration cities over time?

Specify which coefficients you look at when answering questions ii-iv.

```
reg dmfwga i.per##i.highimm, cluster(cmsa99)

Linear regression                               Number of obs     =        460
                                                F(3, 229)         =      13.06
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0826
                                                Root MSE          =     .00681

                              (Std. err. adjusted for 230 clusters in cmsa99)
------------------------------------------------------------------------------
             |               Robust
      dmfwga | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       1.per |   .0013406    .001549     0.87   0.388    -.0017116    .0043927
  1.highimm |  -.0023201    .000659    -3.52   0.001    -.0036187   -.0010216
             |
 per#highimm |
        1 1  |   .0031745   .0017359     1.83   0.069    -.0002459    .0065949
             |
       _cons |  -.0027016   .0005761    -4.69   0.000    -.0038366   -.0015665
------------------------------------------------------------------------------
```

c) We now repeat the analysis in b) for the returns to education variable, **drtca**. Provide an interpretation of all the coefficients reported in this output. Discuss if these overall findings support the claim that post-2000 immigration patterns could be significantly linked to changes in the male-female wage gap and the returns to education.

```
reg drtca i.per##i.highimm, cluster(cmsa99)

Linear regression                               Number of obs     =        460
                                                F(3, 229)         =      28.39
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1429
                                                Root MSE          =     .00422

                             (Std. err. adjusted for 230 clusters in cmsa99)
------------------------------------------------------------------------------
             |               Robust
       drtca | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       1.per |  -.0021269   .0009953    -2.14   0.034    -.0040882   -.0001657
   1.highimm |   .0023403   .0004532     5.16   0.000     .0014472    .0032333
             |
 per#highimm |
        1 1  |  -.0012209    .001102    -1.11   0.269    -.0033924    .0009505
             |
       _cons |   .0066729   .0003966    16.82   0.000     .0058914    .0074544
------------------------------------------------------------------------------
```