

Economics 142
Final Class Project, Fall 2016
Due: December 14 – 11.00 am

YOU MUST SUBMIT ELECTRONICALLY via BCOURSES BY THE DUE DATE/TIME or SUBMIT A WRITTEN COPY TO THE GSI's (ALESSANDRA/KEVIN) BY THE DUE DATE/TIME. *LATE PROJECTS WILL **NOT** BE ACCEPTED*

Note: we do not need to see your code. Just tables, figures and narrative.

In this project you will use data on the wages of men and women from Country X, which is a Western European country broadly similar to the US. The goal of the project is to prepare a report using these data that will focus on documenting and analyzing the gender wage gap. You will estimate a series of wage models that include the standard control variables (education and experience). You will also estimate models that include a measure of the **productivity** of each worker's employer, to evaluate the importance of this variable in determining pay. Finally, you will investigate the possibility that the correlation between employer productivity and wages is biased by unobserved worker characteristics, using information on how wages change as workers move between jobs at more and less productive employers.

Your final report will include tables and figures that summarize your model estimation results, as well as a written narrative in which you interpret your findings. Below is a suggested list of tables and figures for your report. You can add more material or expand the analysis as you see fit. Points will be awarded for creative insights or ideas beyond the basic analysis suggested here.

The data set **project2016.csv** contains 1 record per person for 23,144 observations – 15,058 males and 8,086 females. The basic information pertains to the “reference year”. There is also some information for the years before and after the reference year. The following variables provide information about the reference year:

- year =calendar year, ranging from 2004 to 2008
- age = age in the reference year, ranging from 22 to 59
- educ =years of education, which can be 4, 6, 9, 12, 16
- female
- y = log real hourly wage (after taxes, measured in Euros/hour) on the job in the reference year.
- va = log of output per worker at the employer in the reference year (this is measured in thousands of Euros per person)

The sample consists of people who started a new job in the reference year, so y is their log wage at the new job, and va is log output per worker at their employer in this new job. We also have information on the wage in earlier years

at their previous employer, and in the year after the reference year. The sample is selected so they held the previous job for at least 2 years before changing jobs, and at least 2 years after. Here are the variables with this extra information:

- y_{l1} = wage **one year before** the reference year (“l1” means “lag once”) – so this is the wage the person had at their old job, one year before the change.
- y_{l2} = wage **two years before** the reference year (“l2” means “lag twice”)
- y_{p1} = wage one year **after** the reference year (“p1” means “plus 1”) - so this is the wage on the new job, one year after the change
- $va_previous$ = log of output per worker at the previous employer (i.e., for the job held for the two years before the move).
- $year_previous$ = year-1
- $dy = y - y_{l1}$ = change in wages between old job (in year_previous) to new job in reference year
- $dy_pre = y_{l1} - y_{l2}$ = change in wages from 2 years before to 1 year before the reference year. Note: this change occurs while the worker is at his/her **old** job, since in both years he/she was working at the previous employer
- $dy_post = y_{p1} - y$ = change in wages from reference year to reference year +1. Note: this change occurs while the worker is at his/her **new** job, since in both years he/she was working at the same employer.

The tables at the end of this exam present means of the variables across all observations and for males and females separately.

The following is a suggested list of tables and figures for your report. As noted above, you can add more material or expand the analysis as you see fit.

Table 1 and Figure 1

In this table and figure you will compare the characteristics and wages of male and female workers, focusing on the reference period. A suggested format for the table is 4 columns:

- column 1 = characteristics for all workers
- column 2 = characteristics for female workers
- column 3 = characteristics for male workers
- column 4 = test statistic comparing females and males (eg t-test)

The main characteristics of interest are:

- education and age
- log of real hourly wage (y)
- productivity of current employer (i.e. va)

In addition to comparing means you could distinguish fractions in various intervals. For example, the fractions of each group in each education group. Or, if you find the quartiles of log wages for all workers, you could compare the fractions of men and women in each quartile.

Figure 1a

Plot the smoothed histograms of log hourly wages for men and women on one figure (hist function with bw choices). Develop some interesting ways to illustrate the fact that women's wages are lower than men's.

Figure 1b

Plot the smoothed histograms of va for men and women. You will notice that females tend to work at less productive firms.

Narrative: Briefly discuss the main differences between men and women, using the table and figure to make your main points.

Table 2 and Figures 2-4

In this table you will fit a series of standard wage models, and construct Oaxaca decompositions of the wage gap between men and women. Table 2 should report the results from estimating four models, S1-S4. The first two columns should report two models that fit to pooled data for men and women:

- S1: including only a constant and a female dummy
- S2: including a constant, education, a cubic in age, and a female dummy

The next 2 columns should report two models that are fit separately for men (S3) and women (S4), and that include a constant, education, and a cubic in age.

a) discuss the differences between the estimated female coefficients in model S1 and S2.

b) discuss the differences between the estimated coefficients for education and age in models S3 and S4. Then, use these two models to construct Oaxaca style decompositions of the mean log wage gap between males and females, as in Lecture 8. Comment on the contributions of the covariates (education and age) versus the coefficients on these variables.

Figure 2

An important aspect of the difference between men and women is the difference in age profiles. Graph the relationship between wages and age for men and women who have each of the 5 levels of education, and show the fit of your regression models (Show 5 panels, one for each education group). For this exercise, use the predictions from models fit separately by age.

As we showed in Lecture 22, a cubic function may not be the best way to describe the age profiles for men and women. Using the data for men and women with exactly 12 years of education, compare (for each gender) a cubic polynomial with a cubic spline function with K knots, where $K=3,4,5,\dots,10$. Use a simple 2-sample cross-validation method (by selecting a random 1/2 sample) to compare the MSE of the cubic and the spline with K knots. What is the best choice of K (i.e., the one that minimizes MSE in the hold-back sample) for females? Is it the same for males? Plot the MSE's for males and females in the

hold-back samples for the various choices of K in **Figure 3**. Finally, graph the actual wages of each gender along with the cubic fit and your “best” spline in **Figures 4a** and **4b**.

Table 3

Now we are going to extend the Oaxaca decomposition by adding information on employer productivity (va). Table 3 should report 4 models:

- Fit two alternative models using the pooled data for men and women:
 - M1: include a constant, education, a cubic in age, a female dummy
 - M2: include a constant, education, a cubic in age, a female dummy, and va
- Fit separate models (M3 and M4) for men and women that include a constant, education, and a cubic in age, and va .

Begin by comparing models M1 and M2. How much of the gender gap is “explained” by the fact that women work at less productive firms?

Next, discuss the estimated effect of va on wages. Specifically, consider a **causal** wage model where part of the “residual” is due to an unobserved characteristic of workers that makes them more or less productive, e.g.:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i \\ u_i &= \alpha_i + \epsilon_i \end{aligned} \tag{1}$$

where y_i is the wage of worker i , x_i is a vector of observed characteristics of i (including a dummy if i is female, and va , the productivity of his or her employer), u_i is the part of wages not explained by observed characteristics, which consists of two parts: α_i , which we will interpret as some factor that makes the worker more productive, and may be correlated with x_i , and ϵ_i which we will assume is pure noise and is uncorrelated with x_i .

What will happen in this model if more productive employers (with higher values of va) hire workers who have higher productivity?

HINT: focus on the coefficient on va : call this β_{va} . How do you expect the **OLS estimate** of β_{va} to differ from the causal effect of working at a more profitable employer?

Now use the models fit separately by gender to perform decompositions of the gender gap including the effect of employer productivity. You will see that working at more a productive firm has a smaller effect for women than men. This means that when you perform an Oaxaca style decomposition using models M3 and M4, there will be a component of the gender gap that arises because of the difference in “returns” to va .

Recall from Lecture 8 that the part of the gender gap that is explained by different coefficients for men and women on the va variable will vary if we “re-normalize” va . (Re-read lecture 8 and the discussion on slides 22-24 of what

happens if we re-normalize education as education-8). Re-do your decomposition, re-normalizing va by subtracting off the lowest value of va observed in the data.

Table 4

In this part of your report you will use the fact that we have data on job changers to conduct an analysis of wage changes as people move between jobs with higher and lower paid co-workers.

a) Fit a series of models in which the dependent variable is the **change** in wages (dy) from the old job held in the period before the reference period (call this period -1) and the new job held in the reference period (call this period 0).

Model C1: include age, age-squared, and $dva = va - va_{previous}$, the change in va between the employer in period -1 and the employer in period 0

- Model C2: include age, age-squared, a female dummy, and dva .

- Model C3 and C4: fit models **separately by gender** that include age, age-squared, and dva .

You should discover that the estimated coefficient on dva is positive, but smaller than the effect you estimated in models M2, M3, and M4 in Table 3. Discuss how you interpret this difference.

HINT: Suppose that the causal model (1) is now written as:

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + u_{it} \\ u_{it} &= \alpha_i + \epsilon_{it} \end{aligned} \tag{2}$$

where now we have information for different individuals in different periods $t \in \{-1, 0\}$. Consider the first-differenced version of (2).

In both models M3 and M4 in Table 3, and models C3 and C4 in Table 4, you should find that the effect of va (or dva) is **smaller** for female workers than male workers. In other words, female workers seem to benefit less than male workers from working at more productive firms.

Can you think of any possible explanations for this gap?

Table 5, Figures 5 and 6

In Table 4 we only looked at the change in wage from the year before the job change to the year after. Recall that when we are trying to establish causality, an “event study” design can be useful. In these figures we are going to conduct event studies of wages as workers move between different groups of employers.

a) Begin by finding the quartiles of $va_{previous}$, the value of log productivity for the employers in period -1 (the year before the reference period). Classify all

workers by whether their initial employer is in quartile 1, 2, 3 or 4, and compare the distributions of males and females across the quartiles in Table 5, columns 1 and 2. (Each row will be a quartile and the entries will be the fractions of the genders in each quartile).

b) Next, find the quartiles of va , the value of log productivity at the set of employers where sample members worked in period 0 (the reference period). Classify all workers by whether their reference period employer is in quartile 1, 2, 3 or 4, and compare the distributions of males and females across the quartiles in Table 5 (columns 3 and 4). Compare the **changes** in how males and females are distributed between period -1 and period 0 and report the results in Table 5 (columns 5 and 6).

c) Now, for female and males separately, estimate mean wages for workers who start in a given origin quartile and end up in a given destination quartile in each of the 4 periods from two years before the job change to one year after (i.e., get the means of $yl2$, $yl1$, y , and $yp1$). Note that $yl2$ and $yl1$ are both wages on the old job, whereas y and $yp1$ are wages on the new job. What should the wage profiles look like if employer productivity has a positive causal effect on wages?

HINT: Consider a worker who moves from a firm in quartile 1 to one in quartile 4. Compare that to a worker who starts in quartile 4 and moves to one in quartile 1.

In panel A of figure 5, plot the means of $yl2$, $yl1$, y , and $yp1$ in “event time” (i.e., periods -2, -1, 0, 1) for male workers who started in a quartile 1 firm and ended up in a quartile 1, origin quartile 2, origin quartile 3, and origin quartile 4 firm. In panels B, C, and D do the same for male workers who started in quartile 2, 3 and 4 firms, respectively.

Repeat the same 4 panels for figure 6, using data for females.

Carefully discuss the evidence from the event studies. Do you interpret the data as supporting the idea that moving to a higher productivity firm increases wages while moving to a lower productivity firm lowers wages? Can you make comparisons between the panels in Figure 5 and 6 to say anything about how moves between employers benefit females versus males?

Suggestion: you might want to think about doing some analysis using *dy-pre* and *dy-post*. Sometimes researchers estimate models that should NOT find an effect of a variable of interest. These are often called “Placebo tests”. Can you devise some Placebo tests?

Table 6

In this final section, we consider how to use the estimated causal effects of working at a more productive employer from the wage change models C3 and C4 to modify the decompositions based on models M3 and M4.

If you have a regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \dots + u_i$ and you **know** the true value of β_1 then you can get the correct estimates of the

other coefficients by estimating the model:

$$y_i - \beta_1 x_{1i} = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} \dots + u_i$$

Using this idea, re-estimate the models M3 and M4 for males and females, imposing the estimated “causal effects” of va obtained from the differenced models C3 and C4. Then use these two models and their coefficients to redo the Oaxaca style decompositions. How much of the gender gap is due to differences in the mean value of va for males versus females? How much to the different “returns” to working at a higher productivity firm for males versus females?

Conclusion

In the conclusion, review and summarize your findings. Discuss the progress that you have made on understanding the gender gap by having access to information on employer productivity.

Overall Sample

Variable	Obs	Mean	Std. Dev.	Min	Max
year	23,144	2006.237	1.43472	2004	2008
y	23,144	1.567162	.4980061	.5978258	4.340468
age	23,144	36.40473	8.370122	22	59
educ	23,144	8.723989	3.76091	4	16
female	23,144	.3493778	.4767838	0	1
yl1	23,144	1.522958	.4739945	.5864503	4.044948
yl2	23,144	1.516496	.4693498	.5842376	4.337263
yp1	23,144	1.592553	.5004793	.5888437	4.22359
dy	23,144	.044204	.2817848	-.997246	.997428
dy_pre	23,144	.0064623	.1449147	-.9782514	.999605
dy_post	23,144	.0253905	.1496668	-.985704	.99844
va	23,144	3.07154	.5777233	1.703631	4.499358
va_previous	23,144	3.001042	.5814499	1.700257	4.494887
dva	23,144	.0704976	.6383156	-2.600621	2.49009
year_previous	23,144	2005.237	1.43472	2003	2007

Males

Variable	Obs	Mean	Std. Dev.	Min	Max
year	15,058	2006.225	1.434394	2004	2008
y	15,058	1.636162	.4995021	.6053866	4.340468
age	15,058	36.45285	8.414287	22	59
educ	15,058	8.494422	3.68914	4	16
female	15,058	0	0	0	0
yl1	15,058	1.58599	.4780682	.593665	4.044948
yl2	15,058	1.580847	.4732682	.5928036	4.337263
yp1	15,058	1.662585	.5020193	.6078192	4.22359
dy	15,058	.0501722	.3002182	-.997246	.997428
dy_pre	15,058	.0051433	.1519495	-.9782514	.971534
dy_post	15,058	.0264232	.1565296	-.985704	.99844
va	15,058	3.139626	.5483631	1.704323	4.496827
va_previous	15,058	3.048284	.5668475	1.700362	4.494887
dva	15,058	.0913414	.6318657	-2.387871	2.49009
year_previous	15,058	2005.225	1.434394	2003	2007

Females

Variable	Obs	Mean	Std. Dev.	Min	Max
year	8,086	2006.258	1.435172	2004	2008
y	8,086	1.438668	.4689153	.5978258	3.949744
age	8,086	36.31511	8.287021	22	59
educ	8,086	9.151496	3.855	4	16
female	8,086	1	0	1	1
yl1	8,086	1.405578	.4430532	.5864503	3.745887
yl2	8,086	1.396659	.4374503	.5842376	3.740367
yp1	8,086	1.462135	.4706271	.5888437	3.919143
dy	8,086	.0330897	.2433988	-.992908	.982678
dy_pre	8,086	.0089188	.1307835	-.8793665	.999605
dy_post	8,086	.0234675	.1359552	-.959384	.990049
va	8,086	2.944748	.608824	1.703631	4.499358
va_previous	8,086	2.913067	.5978763	1.700257	4.494887
dva	8,086	.0316815	.6484121	-2.600621	2.470714
year_previous	8,086	2005.258	1.435172	2003	2007