

Econ 695
Final Class Project
Due: Wednesday December 17 – 7:05PM Central Time

EACH GROUP MUST SUBMIT ELECTRONICALLY via CANVAS BY THE DUE DATE/TIME.
LATE PROJECTS WILL NOT BE ACCEPTED.

In this project you will use data on men and women who are observed over several years on exactly two separate jobs. The data are from records from Country X, which is quite similar to the US.

The goal of the project is to prepare a report that will focus on trying to explain the gender wage gap using this data set. 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 new variable, which is a measure of the mean (log) wages of the other **co-workers** at a person's job. This is measured by averaging the log wages of all the other workers at the same firm during the time a person works there.

- *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 list of tables and figures you are expected to include in your report. You can add more material or expand the analysis as you see fit. Bonus points will be rewarded for answering the bonus questions or presenting your own ideas beyond the basic analysis.*

The project will be graded based on the quality of your empirical analysis and your final report. Make sure to also submit a completed **project_code.ipynb** or scripts you use to create the tables and figures.

- Correctness of your results (Tables and Figures): **30 points.**
- Clear specifications of your econometric models (e.g., regressions, equations for decomposition...): **10 points.**
- Presentation of your results (table and figure formatted nicely; try LaTeX): **10 points.**
- Quality of the **narrative** – your *interpretations* of the empirical findings (write clearly and concisely): **30 points.**
- Quality of the overall writing: **10 points** (tips: add an introduction that provides an overview of your analysis, summarize what you present at the beginning of each section in 1-2 sentences, and provide a conclusion).
- Completeness of your code for producing all tables and figures (i.e., submit project_code.ipynb): **10 points** (including **2 points** for an AI disclaimer).
- Creativity in bonus questions or additional results you include in your report: **Bonus 15 points.**

Note the report+code will account for 25% towards your grade, and another 5% is determined by peer evaluation. Everyone gets full 5% by default, but if you disagree with the default for your teammates because of their lack of participation, email me by 12/17 with your suggested rating (ranging from 1 to 5) for your teammates. I will take the average rating by your teammates.

Data Description

The dataset (projectdata.csv) contains one record per person, but each record has information on wages and the characteristics of the person's employer in multiple years. There are a total of 16,969 observations – 10,575 for men and 6,394 for women.

The timing convention for the measurement of all data is that period 0 is the first year the person is observed on his/her **second** job. All people are observed for at least 3 years on job 1 and 3 years on job 2. The sample also includes only people who move between jobs where there are 4 or more co-workers.

The following variables are available and hold information about period 0:

- age
- educ (=years of education)
- female
- exp (=potential experience based on age and education)
- y = log real hourly wage, first year on the second job ("period 0")

The following additional wage information is available for other periods:

- yp1 = log wage in year 1 (this is the second year on the second job)
- yp2 = log wage in year 2 (this is the third year on the second job)
- yl1 = log wage in year -1 (this is the wage on the last year on the first job)
- yl2 = log wage in year -2 (this is the wage on the first job, 2 years before the change)
- yl3 = log wage in year -3 (this is the wage on the first job, 3 years before the change)

In addition the following information is available about the wages of **other workers** on the first job and the second job:

- owage1 = mean log wage of other workers at the first job (held in period -1 and -2)
- owage2 = mean log wage of other workers at the second job (held in period 0, period 1 and period 2)

Here is a list of tables and figures that you should include in your report. As noted above, you can add more material or expand the analysis as you see fit.

1 Overview: Female Workers versus Male Workers

Table 1: Compare the characteristics and wages of male and female workers, focusing on period 0.

The main characteristics of interest are:

- education and age. Note that education takes on only 4 values: 6, 9, 12 and 16, corresponding to only elementary schooling; lower secondary schooling, upper secondary schooling and university). Define dummies for education levels and summarize the fraction of workers at each level.
- log wage (y)
- mean log wage for co-workers as of period 0 (i.e. owage2)

A suggested format for **Table 1** is to have 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)

In addition to comparing means you could distinguish fractions in various intervals. For example, if you find the terciles of log wages at period 0 (variable “y”) for all workers, you could compare the fractions of men and women in each tercile.

Figure 1 (Distribution of Wages): Plot the histogram and kernel density of log hourly wages (variable **yp1**) for each gender. Overlay the graphs in one figure, or develop some interesting ways to illustrate the fact that women’s wages are lower than men’s.

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

2 Gender Wage Gaps

2.1 Standard Models and Oaxaca Decomposition

- In **Table 2**, you will fit a series of standard wage models of log wage in period 0 (“y”) on observable characteristics, and construct an Oaxaca decomposition of the wage gap between men and women.
 - Fit a set of models using the pooled data for men and women
 - * including only a constant and a female dummy
 - * including a constant, education (categorical), a cubic polynomial in experience, and a female dummy
 - Fit separate models for men and women that include a constant, education (categorical), and a cubic polynomial in experience. Use the models to construct a decomposition of gender gap. How much can the differences in education and experience (composition) explain the gender wage gap?
- In report, type the regressions you run for Table 2. Write the equation for Oaxaca decomposition, and point out which part represents the between vs. within components.
- **Narrative:** interpret your regression estimates in Table 2, and your decomposition result.

2.2 Gender Difference in Experience Profiles

An important aspect of the difference between men and women is the difference in experience profiles.

- **Figure 2:** Plot the relationship between log wages (variable “y”) and experience (ranging from 5 years to 30 years) for men and women who have education=16. Fit the experience profile by a cubic polynomial of experience, and add the fitted line to your figure. How much can the gender wage gap among college graduates can be explained by differences in experience?
- **BONUS (Figure 2 continued):** Consider plotting “Figure 2” for the other three education groups. Within each education group, how much of the gender gap can be explained by differences in experience?
- **Narrative:** Discuss the results shown in Figure 2.

3 Gender Wage Gaps Conditional on Coworker Wages

In this part you are going to add a new control variable, which is the mean log wage of a person’s co-workers (the variable `owage2`).

- **Table 3** will report 5 models:
 - Fit a set of models using the **pooled** data for men and women
 - * including only a constant, a female dummy, and `owage2`
 - * including a constant, education, a cubic in experience, a female dummy, and `owage2`
 - * including a constant, education, a cubic in experience, a female dummy, `owage2`, and the interaction of `owage2` with the female dummy.
 - Fit **separate** models for men and women that include a constant, education, and a cubic in experience and **owage2**. Use these models to perform a new decomposition that accounts for the effect of higher-wage coworkers (that is, how much can the gender wage gap be explained by observables including the coworker wages?).

It will turn out that controlling for `owage2` makes quite a difference to the gender gap. You will see that the wage effect of working with highly paid co-workers is quite large, and in the models that allow different returns by gender, the effect is smaller for women than men.

- **Narrative: This is the most important part of your narrative**, since the idea of using co-worker wages as a control is relatively new. It will turn out that controlling for **owage2** makes quite a difference to the gender gap. You will see that the wage effect of working with highly paid co-workers is quite large, and in the models that allow different returns by gender, the effect is smaller for women than men.

- In your narrative you will discuss alternative interpretations of the effect of working with highly paid co-workers. In the final section of the paper (below) we will look at job changers to try and sort out these alternative interpretations.
- Before starting to write your narrative you will want to think carefully about possible reasons why people who work with higher-paid co-workers earn more. You will want to think about the following (economic) models:
 - * Model/Hypothesis 1: getting a job with high-paid co-workers is largely a matter of good luck or connections, and men have better connections, or search harder to find higher paid coworker jobs.
 - * Model/Hypothesis 2: getting a job with highly paid co-workers is only possible for workers who have high levels of cognitive skills or ambition.
- Based on your estimates in Table 3, how much of the gender gap can be explained by differences in co-worker wages? (Hint: you may use the pooled regression or separate regressions for your decompositions). Think about the implications of the 2 (economic) models above for your interpretations of the decompositions.
 - * Hint: is coworker-wage (owage2) simply a proxy of unobserved individual ability?

4 Event Study - Wage Changes around Moves

In this part you use the fact that we have job changers in the data to conduct some event studies, and do an analysis of wage changes as people move between jobs with higher and lower paid co-workers.

Begin by finding the terciles of **owage1**. Classify all the first jobs (held in periods -3, -2, and -1) into 3 groups based on the tercile of **owage1**. Classify all the second jobs (held in periods 0, 1 and 2) into 3 groups based on the same terciles. (You will notice that on average the second jobs have slightly higher co-worker pay, which might reflect the fact that people are trying to find better jobs as they get older).

- **Figure 3:** Conduct 9 separate event studies, plotting mean wages in period -3, -2, -1, 0, 1, 2 for people who start in each tercile of **owage1** and go to each tercile of **owage2**. Show the 9 trajectories in one figure and label them clearly (or 3 figures, one for each tercile of origin **owage1**).
 - Hint: The original data is at person level. You can create a person ID, and then create a long data at (person, l) level, where $l = -3, -2, -1, 0, 1, 2$ years relative to the move to the 2nd job.
- **Narrative:** Think carefully about the alternative models (Model 1 and Model 2) of why co-worker wages matter. Then discuss the event study graphs. Do these graphs provide more support for Model 1 or Model 2? Or come up with an alternative mechanism through which coworker wages influence a worker's own wage.

- For **Table 4**, you will model the change in wages from -1 to 0 ($y - y_{11}$) as a function of the change in the mean log wage of co-workers ($owage2 - owage1$).
 1. Fit a set of models for the change in wages using the pooled data for men and women
 - (a) including only a constant, a female dummy, and ($owage2 - owage1$)
 - (b) including a constant, quadratic polynomial in experience as of period -1, a female dummy, and ($owage2 - owage1$)
 - (c) including a constant, quadratic polynomial in experience as of period -1, a female dummy, and ($owage2 - owage1$), and the interaction of ($owage2 - owage1$) with the female dummy.
 2. Fit separate models for men and women that include a constant, quadratic polynomial in experience as of period -1, and ($owage2 - owage1$). *Note: experience in period -1 is just experience in period 0 minus 1.
- **Narrative:** The main issue in this part of the narrative is the comparison between the effect of coworker average wages in OLS models (Table 3) and first-differenced models that control for unobserved characteristics of people (Table 4).
 - One way to summarize the two sets of results is to ask: what fraction of the OLS effect of co-worker wages do we see in the first-differenced models? If, for example, the OLS model for males gives a coefficients on coworker wages of 0.66, but the differenced model gives a coefficient of 0.33, then you might conclude that one half of the OLS effect is a causal effect and the other half reflects differences in the unobserved skills of people who tend to work at high-coworker wage jobs.
- **Bonus (Table 4b):** Consider more flexible first-differenced models that control for a constant, a female dummy, interactions between ($owage2 - owage1$) and female, and interactions between ($owage2 - owage1$) and dummies for experience (as of period -1). Estimate 3 models: OLS, Ridge, and Lasso.
 - Plot the OLS coefficients on the X-axis, and on the Y-axis the Ridge and Lasso coefficients (reproducing a figure we used to illustrate the shrinkage towards 0 in Lecture 12). In the narrative, discuss if the role of coworker wages matter differently by experience, based on your estimates.
 - Estimate the Lasso model separately by gender. In the narrative, discuss if coworker wages matter more for less experienced workers, and whether it's different for females vs. males.

5 Conclusion

In the conclusion, review and summarize your findings. Discuss the progress that you have made on understanding the gender pay gap.