Estimating Gender Discrimination in the Workplace

In this exercise we'll use data from the 2018 US Current Population Survey (CPS) to try and estimate the effect of being a woman on workplace compensation.

Note that our focus will be *only* on differential compensation in the work place, and as a result it is important to bear in mind that our estimates are not estimates of *all* forms of gender discrimination. For example, these analyses will not account for things like gender discrimination in terms of *getting* jobs. We'll discuss this in more detail below.

Exercise 1:

Begin by downloading and importing 2018 CPS data from https://github.com/nickeubank/MIDS_Data/tree/master/Current_Population_Survey [https://github.com/nickeubank/MIDS_Data/tree/master/Current_Population_Survey]. The file is called `morg18.dta` and is a Stata dataset. Additional data on the dataset can be found by following the links in the README.txt file in the folder, but for the moment it is sufficient to know this is a national survey run in the United States.

The survey does include some survey weights we won't be using (i.e. not everyone in the sample was included with the same probability), so the numbers we estimate will not be perfect estimates of the gender wage gap in the United States, but they are pretty close.

Exercise 2:

Because our interest is only in-the-workplace wage discrimination among full-time workers, we need to start by subsetting our data for people currently employed (and "at work", not "absent") at the time of this survey using the `lfsr94` variable, who are employed full time (meaning that their usual hours per week—`uhourse`—is 35 or above).

As noted above, this analysis will miss many forms of gender discrimination. For example, in dropping anyone who isn't working, we immediately lose any women who couldn't get jobs, or who chose to lose the workforce because the wages they were offered (which were likely lower than those offered men) were lower than they were willing / could accept. And in focusing on full time employees, we miss the fact women may not be offered full time jobs at the same rate as men.

Exercise 3

Now let's estimate the basic wage gap for the United States!

Earnings per week worked can be found in the `earnwke` variable. Using the variable `sex` (1=Male, 2=Female), estimate the gender wage gap in terms of wages per hour worked!

(You may also find it helpful, for context, to estimate the average hourly pay by dividing weekly pay by `uhourse`.)

Exercise 4

Assuming 48 work weeks in a year, calculate annual earnings for men and women. Report the difference in dollars and in percentage terms.

Exercise 5

We just compared all full-time working men to all full-time working women. For this to be an accurate *causal* estimate of the effect of being a woman in the work place, what must be true of these two groups? What is one reason that this may *not* be true?

Exercise 6

One answer to the second part of Exercise 5 is that working women are likely to be younger, since a larger portion of younger women are entering the workforce as compared to older generations.

To *control* for this difference, let's now regress annual earnings on gender, age, and age-squared (the relationship between age and income is generally non-linear). What is the implied average annual wage difference between women and men? Is it different from your raw estimate?

Exercise 7

In running this regression and interpreting the coefficient on `female`, what is the implicit comparison you are making? In other words, when we run this regression and interpreting the coefficient on `female`, we're basically pretending we are comparing two groups and assuming they are counter-factuals for one another. What are these two groups?

Exercise 8

Now let's add to our regression an indicator variable for whether the respondent has at least graduated high school, and an indicator for whether the respondent at least

has a BA.

In answering this question, use the following table of codes for the variable `grade92`.

Education is coded as follows:

Less than 1st grade	31
1st - 4th grade	32
5th or 6th	33
7th or 8th	34
9th	35
10^{th}	36
11^{th}	37
12^{th} grade NO DIPLOMA	38
High school graduate, diploma or GED	39
Some college but no degree	40
Associate degree -- occupational/vocational	41
Associate degree -- academic program	42
Bachelor's degree (e.g. BA,AB,BS)	43
Master's degree (e.g. MA,MS,MEng,Med,MSW,MBA)	44
Professional school deg. (e.g. MD,DDS,DVM,LLB,JD)	45
Doctorate degree (e.g. PhD, EdD)	46

Exercise 9

In running this regression and interpreting the coefficient on `female`, what is the implicit comparison you are making? In other words, when we run this regression and interpreting the coefficient on `female`, we are once more basically pretending we are comparing two groups and assuming they are counter-factuals for one another. What are these two groups?

Exercise 10

Given how the coefficient on `female` has changed between Exercise 6 and Exercise 8, what can you infer about the educational attainment of the women in your survey data (as compared to the educational attainment of men)?

Exercise 11

What does that tell you about the *potential outcomes* of men and women before you added education as a control?

Exercise 12

Finally, let's include *fixed effects* for the type of job held by each respondent.

Fixed effects are a method used when we have a nested data structure in which respondents belong to groups, and those groups may all be subject to different pressures. In this context, for example, we can add fixed effects for the industry of each respondent—since wages often vary across industries, controlling for industry is likely to improve our estimates. Use `ind02` to control for industry.

(Note that fixed effects are very similar in principle to hierarchical models. There are some differences you will read about [../fixed_effects_v_hierarchical.html] for our next class, but they are designed to serve the same role, just with slightly different mechanics).

When we add fixed effects for groups like this, our interpretation of the other coefficients changes. Whereas in previous exercises we were trying to explain variation in men and women's wages *across all respondents*, we are now effectively comparing men and women's wages *within each employment sector*. Our coefficient on `female`, in other words, now tells us how much less (on average) we would expect a woman to be paid than a man *within the same industry*, not across all respondents.

(Note that running this regression will result in lots of coefficients popping up you don't care about. We'll introduce some more efficient methods for adding fixed effects that aren't so messy in a later class – for now, you can ignore those coefficients!)

Exercise 13

Now that we've added industry fixed effects, what groups are we implicitly treated as counter-factuals for one another now?

Exercise 14

What happened to your estimate of the gender wage gap when you added industry fixed effects? What does that tell you about the industries chosen by women as opposed to men?

When you're done, please come read this discussion [discussion_regressions_incomeineq.html].