The Gender Wage Gap Myth Myth

Hariharan Jayashankar

This notebook simulates data and does a couple of regressions to see the effect of colliders and confounders on regression estimates.

This is to show that the popular idea that the the gender wage gap disappears once you control for occupation isn't true.

Loading packages

```
library(tidyverse) # easy data stuff if needed
library(stargazer) # nice output
```

Generating Data

```
# number of individuals
n = 1000

# half the population is female
female <- runif(n, min=0, max=1) > 0.5

# ability is independant of gender
ability <- rnorm(n)

# discrimination (we could use the female var, but just to make things clear)
disc <- female == TRUE

# Parameters
ability_occ <- 2
disc_occ <- -2

disc_wage <- -1
occ_wage <- 1
ability_wage <- 2

# Occupation is ranked monotonically according to specificed function
occ <- 1 + ability_occ * ability +
    disc_occ * disc +
    rnorm(n)

# wage
wage <- 1 +
    disc_wage*disc +
    occ_wage*occ +
    ability_wage*ability +
    rnorm(n)

# Making it all a table
df <- tibble(female = female,
             ability = ability,
             disc = disc,
```

```
              occ = occ,
              wage = wage)
```

There are two ways for female to affect wages - through occupation and directly.

Lets regress wages on gender (female)

```
mod_base <-          lm(wage ~ female, data=df)
mod_badcontrol <-    lm(wage ~ female + occ, data = df)
mod_god <-           lm(wage ~ female + occ + ability, data = df)

stargazer(mod_base, mod_badcontrol, mod_god, type = 'html')
```

Dependent variable:

wage

(1)

(2)

(3)

female

-2.979***

0.441***

-1.125***

(0.258)

(0.095)

(0.088)

occ

1.789***

0.982***

(0.020)

(0.031)

ability

2.021***

(0.069)

Constant

1.786***

0.291***

1.102***

(0.177)

(0.062)

(0.053)

Observations

1,000

1,000

1,000

R2

0.118

0.901

0.947

Adjusted R2

0.117

0.901

0.947

Residual Std. Error

4.074 (df = 998)

1.366 (df = 997)

1.001 (df = 996)

F Statistic

133.286*** (df = 1; 998)

4,531.701*** (df = 2; 997)

5,915.518*** (df = 3; 996)

Note:

p<0.1; ***p<0.05;*** p<0.01

Interpretting

The base coefficient which can be interpretted as the unconditional estimate of the effect of being female on wages is -2.9789699. It's negative because of 2 effects:

1. discrimination -> wages
2. discrimination -> occupation -> wages

Both of these effects work in the same direction. Being a female gives you worse jobs and being a female gives you a worse wage given the same job.

The specification with the bad control gives us the coefficient of 0.4408067. This is showing a positive coefficient. Effectively what we have done by including the control of occupation is open up the channel of disc -> occ <- ability -> wages.

Another classical way of thinking of it is that the β coefficient just doesn't have the same interpretation anymore.

What we want is $\mathbb{E}(w_{fi} - w_{mi})$ (w_{di}) refers to wages of an individual i if he/she was assigned to treatment d, where d can be m for male of f for female). In a typical randomized setting where we randomize gender (although we actually cant do that in reality), a regression of wages on gender would give us

$$\mathbb{E}(w|gender_i = female) - \mathbb{E}(w|gender_i = male) = \mathbb{E}(w_{fi}|gender_i = female) - \mathbb{E}(w_{mi}|gender_i = male) = \mathbb{E}(w_{fi} - w_{mi})$$

Last step comes from independance of gender assignment in our imaginary world to wages (outcomes)

But in the bad control case, this is how it goes.

$$\mathbb{E}(w|gender_i = female, occ_{di} = o) - \mathbb{E}(w|gender_i = male, occ_{di} = o) = \mathbb{E}(w_{fi}|gender_i = female, occ_{fi} = o) - \mathbb{E}(w_{mi}|gender_i =$$

We can't go beyond the last step because occupation and wages aren't actually independantly determined, in our case due to ability. (remember, gender -> occ <- ability). More over

Regression intuition wise, the ability variable from the true DGP has gone into the error term, but ability is correlated with gender for a given occupation, and ability is also related to wages directly.

All in all for trying to control the channel gender -> occ -> wages, we introduced gender -> occ <- ability, which might arguable be a worse estimate than what the confounding channel we eliminated.

But if we could observe ability levels, our lives are easy. Our estimate is -1.1253023 which is very close to the direct effect of gender on wages in our constructed case. This is simply because that is what the DGP actually looks like!