Yale JACKSON SCHOOL OF GLOBAL AFFAIRS

Assignment 1: Understanding the Gender Wage Gap Due Friday, Jan 30

GLBL 5021 Applied Methods of Analysis II

Learning goals:

i) able to conduct different versions of multilinear regressions
ii) able to interpret the results and explain it to a broader audience
iii) gain practice in using statistical software to run analysis

Question of interest: How large is the gender wage gap in industry ____ after controlling for variables of interest? What factors are important in reducing the gap?

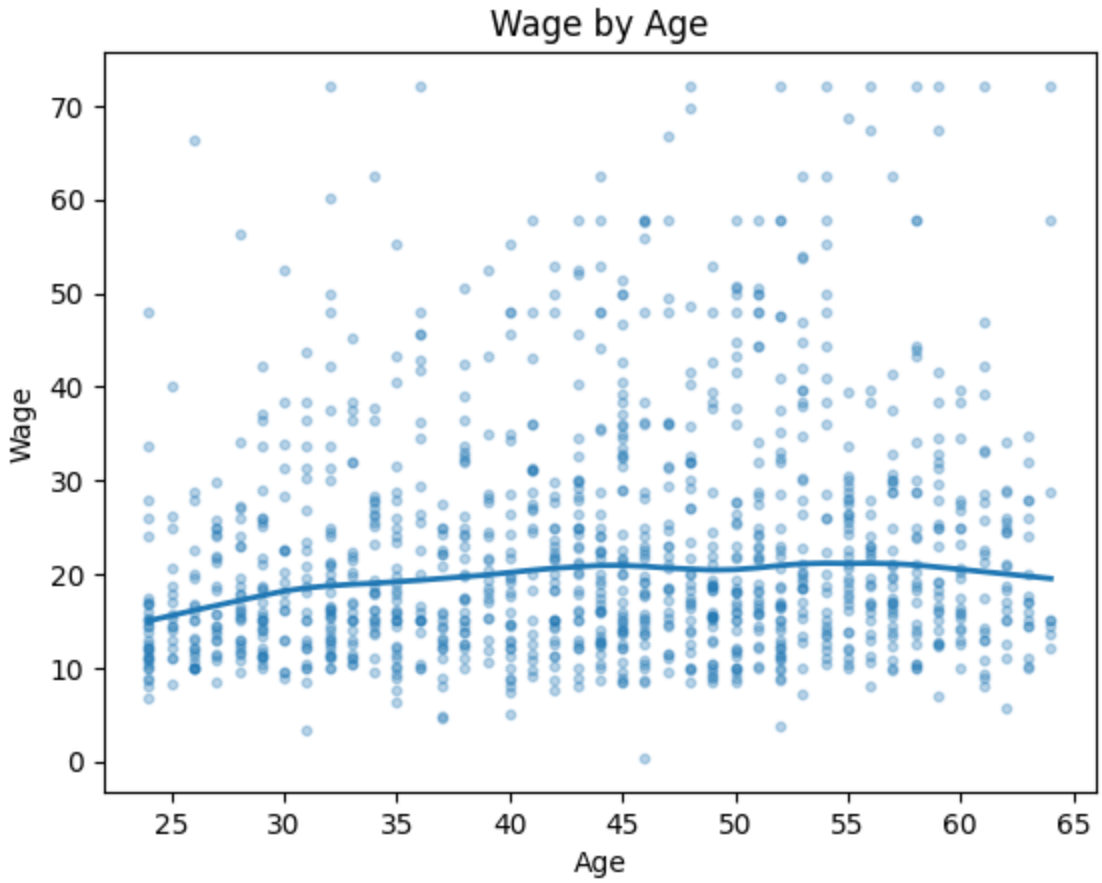
The assignment will guide you step-by-step to replicate the Case Study "*Understanding the Gender Wage Gap*" with some modifications. Download "HW1_wage_start" code from canvas in your preferred language. The code provides you with a start and some example codes to help you. Continue the code while following the instructions below. Some questions just require you to write in your code and nothing else is needed. Others may require you to screenshot graphs or write results. The question will state if a screenshot and/or written answers are required. Once you finish your code, copy and paste it at the end of your submission. Submit the pdf of your assignment on Gradescope (canvas).

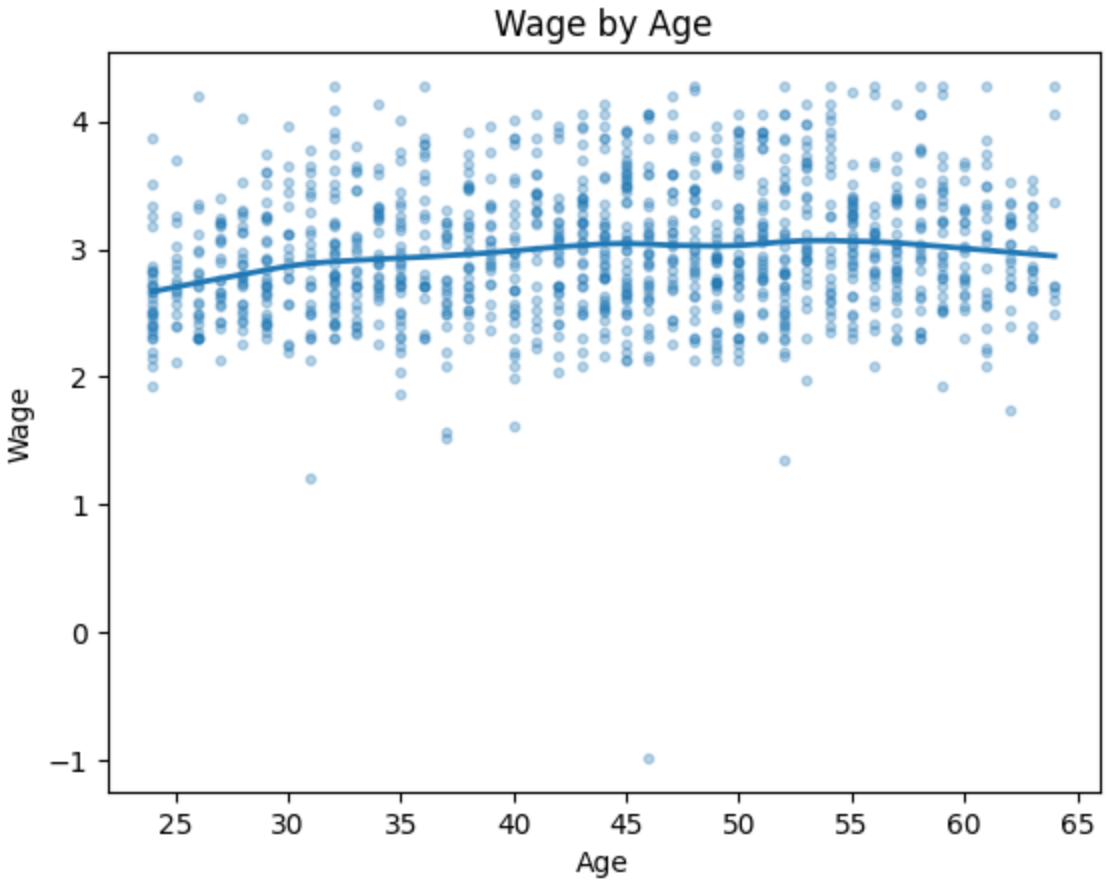
Part I: Investigating the data, sample selection, and creating key variables

Run the start of the code from "HW1_wage_start". Read more about the data and the variables available here. Notice that the code opens the data for you, conducts some cleaning, and sample selection. We will modify this a bit and create additional variables for practice.

1.1 First, at the very end of the code, the code selects an industry using `ind02`. The code is set to "Banking and related activities". Investigate the data and choose an industry that interests you. Make sure it has a large enough number of observations (aim for more than 700). Change the code to select the industry of your choice.

1.2 (Graphing Practice) Let's investigate the earnings and age patterns for this industry. Create two scatterplots with a lowess curve, one with age on the x-axis and wage in the y-axis. Then a second with log wage in the y-axis. Provide the graph below and write down one to three sentences of the pattern observed.


Wage by Age
Wage
Age
0
10
20
30
40
50
60
70
25
30
35
40
45
50
55
60
65



We can see that wage slowly increases for older workers. However, after a certain age is reached, wage might start falling again.

1.3 Let's create additional variables to add to your regression.

- Create age^2 and age^3
- The explanation for grade92 values is below. Create five categories for the education variable: Less than high school (12^{th} grade no diploma and below), high school (+diploma or GED), some college or associate, bachelor's degree, advanced higher education (MA and higher).

```
Less than 1st grade                                31
1st - 4th     grade                                32
5th or 6th                                         33
7th or 8th                                         34
9th                                                35
10th                                               36
11th                                               37
12th 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
```

1.4 Come up with a simple code to investigate the share of individuals in each education category. Provide the output of that code below. Explain what you found in one to three sentences.

```
educ
hs_grad                  0.413882
some_college_associate   0.295630
college_grad             0.154242
post_grad                0.076264
less_than_hs             0.059983
Name: proportion, dtype: float64
```

Most people (41.4%) have high school as the highest level of education, 29.6% of people have some college/associate degree, 15.4% are college grads, 7.6% have post graduate degrees, and 6% of people don't even have high school education in the dataset.

Part II: Run several regression models and provide interpretation

Let's start with simpler regressions and practice interpreting them.

2.1 Run four regression models and show the side-by-side table results: (1) ln wage on female (2) ln wage on female and age (3) ln wage on female, age, and age-squared, and (4) ln wage on female, age, age-squared, and age cubed.

a. Interpret the female and age coefficient in Model 2.
b. Explain which regression you would prefer.

	ln wage			
	(1)	(2)	(3)	(4)
female	-0.213***	-0.225***	-0.231***	-0.231***
	(0.032)	(0.032)	(0.032)	(0.032)
age		0.008***	0.050***	0.073
		(0.001)	(0.011)	(0.065)
age2			-0.000***	-0.001
			(0.000)	(0.002)
age3				0.000
				(0.000)
Constant	3.048***	2.707***	1.859***	1.548
	(0.018)	(0.056)	(0.215)	(0.862)
Observations	1167	1167	1167	1167
S.E. type	hetero	hetero	hetero	hetero
R^2	0.033	0.059	0.071	0.071
Adj. R^2	0.032	0.058	0.068	0.068

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

a)

female: On average female automotive workers have 22.5% lower wages than male automotive workers, holding age constant.

age: The average predicted increase in wages as age increases by one is 0.8% among people, controlling for their gender (in this case, this would be male automotive workers).

b)

I would prefer the second regression if I was to analyse a possible gender gap, as it would allow me to capture a possible confounding variable (age). Adding age_squared and age_cubed does not improve the predictive power. Plus, the results in Model 4 are not significant.

2.2 Run a regression of lwage on female and the education category you created. Write a short paragraph explaining how wages vary by education level.

	ln wage
	(1)
female	-0.191***
	(0.028)
educ[T.hs_grad]	-0.441***
	(0.043)
educ[T.less_than_hs]	-0.784***
	(0.055)
educ[T.post_grad]	0.363***
	(0.064)
educ[T.some_college_associate]	-0.291***
	(0.044)
Constant	3.330***
	(0.038)
Observations	1167
S.E. type	hetero
R^2	0.285
Adj. R^2	0.282

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

On average, automotive workers have 78.4% lower wages when not being high school graduates compared to college grads, controlling for their gender (in this case, it is male automotive workers). Similarly, it is 44.1% less for high school grads, 29.1% less for people with some college/associate degree, and 36.3% more for people with post-graduate degrees compared to college grads, controlling for their gender.

2.3 Run three regressions side-by-side. (1) ln wage on female, (2) ln wage on female and age variables of choice, (3) ln wage on female, age variable(s) of choice, and education category. Explain what happens to the coefficient on female across the models and what can we infer about the gender gap from these results?

	ln wage		
	(1)	(2)	(3)
female	-0.213*** (0.032)	-0.225*** (0.032)	-0.202*** (0.028)
age		0.008*** (0.001)	0.007*** (0.001)
educ[T.hs_grad]			-0.454*** (0.042)
educ[T.less_than_hs]			-0.791*** (0.055)
educ[T.post_grad]			0.333*** (0.065)
educ[T.some_college_associate]			-0.300*** (0.043)
Constant	3.048*** (0.018)	2.707*** (0.056)	3.016*** (0.059)
Observations	1167	1167	1167
S.E. type	hetero	hetero	hetero
R^2	0.033	0.059	0.309
Adj. R^2	0.032	0.058	0.306

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

The gender wage gap (the absolute value of the female coefficient) increases when adding age as a variable in the multilinear regression, indicating that age acted as a suppressor of the magnitude of the wage gap.

The gender wage gap decreases when adding the education category, indicating that some of the gap can be explained by differing education levels between men and women, with men potentially having higher levels of education (which can drive their wage growth).

Holding education constant and age constant, however, the wage gap is still 20.2%.

2.4 Now let's practice interaction terms. Run three regressions: (1) Just for women, ln wage on age; (2) the same but for men; (3) For all observation, ln wage on female and age. Provide the side-by-side output below. Provide an interpretation of the key coefficients and explain how the three regressions are connected.

	ln wage		
	(1)	(2)	(3)
age	0.002 (0.002)	0.010*** (0.001)	0.010*** (0.001)
female			0.120 (0.131)
age:female			-0.008** (0.003)
Constant	2.744*** (0.114)	2.624*** (0.065)	2.624*** (0.065)
Observations	319	848	1167
S.E. type	hetero	hetero	hetero
R^2	0.002	0.042	0.064
Adj. R^2	-0.001	0.041	0.062

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

1)

As age increases, the average wage of women increases by 0.2%. However, the coefficient is not significant.

2)

As age increases, the average wage of men increases by 1%

3)

As age increases, the average wage of men increases by 1% (because we're holding gender constant).

The female coefficient tells us that on average, females have 12% higher wages than male, holding age constant. However, the results are not significant.

As age increases, the average change in wage will be 0.8 percentage points less for women than for men.

The models are connected thus: The age coefficient in model 3 is the same as the age coefficient in model 2, because in model 3, we're holding gender constant (meaning, we're calculating the coefficient for male workers). The age:female interaction term (-0.008) is the same as taking the age

coefficient for the female sample (0.002) and subtracting the age coefficient for the male sample (0.01).

Part III: Playing around with a more complex model

3.1 Create three regressions. The first one you already did above, basic ln wage and female. The second one adds age and education categories. For the last one, add more variables of your choice. See variables that have been created at the start of the code. You can even add other variables from the data itself. Show the results side-by-side below. The next question will ask you to analyze the results. Tip: see also if you can add dummy variables for type of occupation (`occ2012`).

	ln wage		
	(1)	(2)	(3)
female	-0.213***	-0.202***	-0.189***
	(0.032)	(0.028)	(0.028)
age		0.007***	0.007***
		(0.001)	(0.001)
educ[T.hs_grad]		-0.454***	-0.446***
		(0.042)	(0.043)
educ[T.less_than_hs]		-0.791***	-0.794***
		(0.055)	(0.055)
educ[T.post_grad]		0.333***	0.316***
		(0.065)	(0.066)
educ[T.some_college_associate]		-0.300***	-0.294***
		(0.043)	(0.043)
married			0.075**
			(0.028)
ownchild			0.019
			(0.012)
Constant	3.048***	3.016***	2.956***
	(0.018)	(0.059)	(0.063)
Observations	1167	1167	1167
S.E. type	hetero	hetero	hetero
R^2	0.033	0.309	0.317
Adj. R^2	0.032	0.306	0.312

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

3.2 From the output above, write two short paragraphs. First, discuss different variables that are important in determining the gender wage gap in your choice industry. Interpret variables that you think are interesting. Second, discuss the magnitude of the gender wage gap before and after these variables are controlled for.

As can be seen from the output above, gender, age, education levels and possibly marriage status are important in determining the wage gap between men and women in the automotive sector. Marriage is especially interesting as a variable, as it still plays a role even when controlling for age. On average, married people earn more than unmarried people, holding other variables (including age) constant. Having kids, on the other hand, does not play a significant role in determining a wage gap.

Before controlling for all these variables, the gender wage gap was 21.3%. After controlling, it is 18.9%.