Jayoung Kang

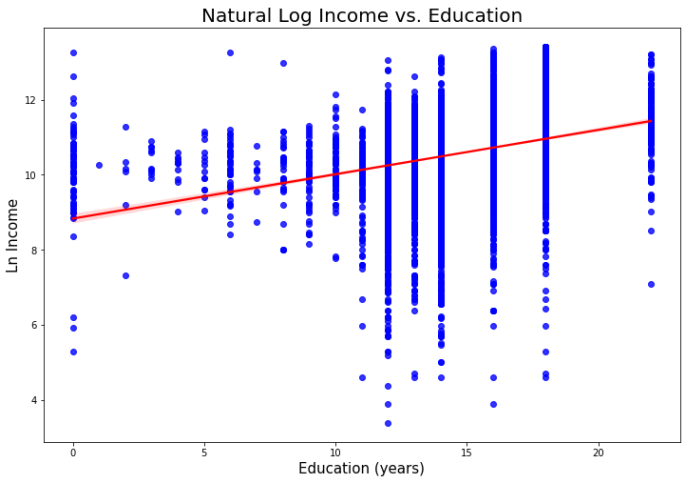
Lab Mini-Project 1

4. Data Analysis

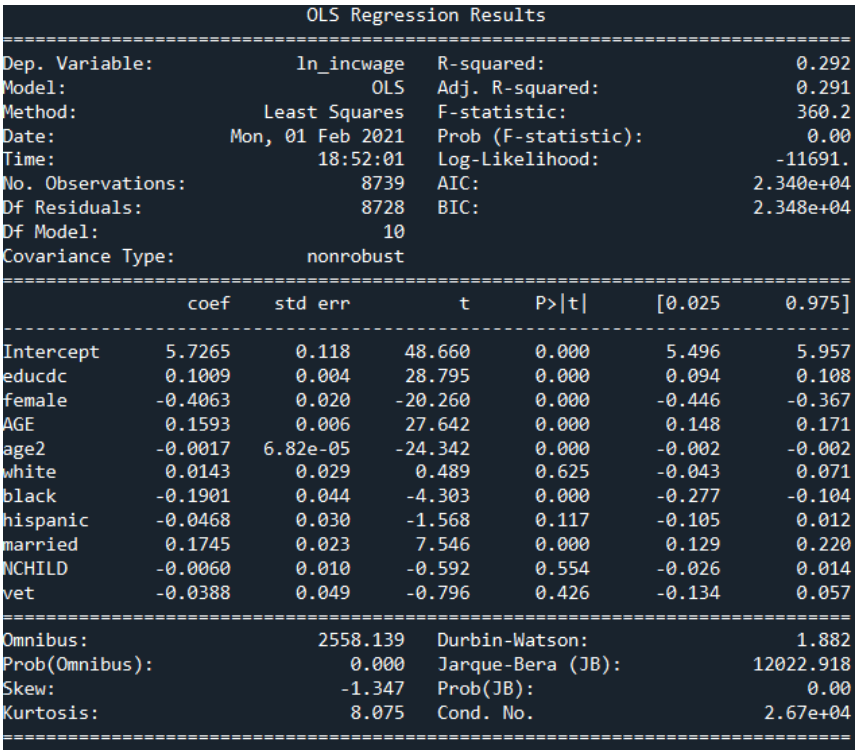
1. *Compute descriptive (summary) statistics for the following variables: year, incwage, lnincwage, educdc, female, age, age2, white, black, hispanic, married, nchild, vet, hsdip, coldip, and the interaction terms. In other words, compute sample means, standard deviations, etc.*

```
In [229]: print(desc)
             YEAR       NCHILD          AGE        INCWAGE       educdc  \
count  8739.0  8739.000000  8739.000000    8739.000000  8739.000000
mean   2019.0     0.797231    41.820918   59610.430255    14.206431
std       0.0     1.116200    13.280883   71107.095217     2.928334
min    2019.0     0.000000    18.000000      30.000000     0.000000
25%    2019.0     0.000000    31.000000   22000.000000    12.000000
50%    2019.0     0.000000    42.000000   42000.000000    14.000000
75%    2019.0     1.000000    53.000000   70000.000000    16.000000
max    2019.0     9.000000    65.000000  665000.000000    22.000000

              hsdip       coldip        white        black     hispanic  \
count  8739.000000  8739.000000  8739.000000  8739.000000  8739.000000
mean      0.549033     0.385971     0.775375     0.084106     0.140634
std       0.497618     0.486852     0.417359     0.277562     0.347663
min       0.000000     0.000000     0.000000     0.000000     0.000000
25%       0.000000     0.000000     1.000000     0.000000     0.000000
50%       1.000000     0.000000     1.000000     0.000000     0.000000
75%       1.000000     1.000000     1.000000     0.000000     0.000000
max       1.000000     1.000000     1.000000     1.000000     1.000000

            married       female          vet  hsdip_educdc  coldip_educdc  \
count  8739.000000  8739.000000  8739.000000   8739.000000    8739.000000
mean      0.557501     0.487012     0.044399      7.136514       6.543998
std       0.496711     0.499860     0.205991      6.505189       8.303040
min       0.000000     0.000000     0.000000      0.000000       0.000000
25%       0.000000     0.000000     0.000000      0.000000       0.000000
50%       1.000000     0.000000     0.000000     12.000000       0.000000
75%       1.000000     1.000000     0.000000     13.000000      16.000000
max       1.000000     1.000000     1.000000     14.000000      22.000000

              age2   ln_incwage
count  8739.000000  8739.000000
mean   1925.350841    10.510322
std    1115.265522     1.095992
min     324.000000     3.401197
25%     961.000000     9.998798
50%    1764.000000    10.645425
75%    2809.000000    11.156251
max    4225.000000    13.407542
```

2. *Scatter plot ln(incwage) and education. Include a linear fit line. Be sure to label all axes and include an informative title.*



3. *Estimate the model*



OLS Regression Results

Dep. Variable:	ln_incwage	R-squared:	0.292
Model:	OLS	Adj. R-squared:	0.291
Method:	Least Squares	F-statistic:	360.2
Date:	Mon, 01 Feb 2021	Prob (F-statistic):	0.00
Time:	18:52:01	Log-Likelihood:	-11691.
No. Observations:	8739	AIC:	2.340e+04
Df Residuals:	8728	BIC:	2.348e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.7265	0.118	48.660	0.000	5.496	5.957
educdc	0.1009	0.004	28.795	0.000	0.094	0.108
female	-0.4063	0.020	-20.260	0.000	-0.446	-0.367
AGE	0.1593	0.006	27.642	0.000	0.148	0.171
age2	-0.0017	6.82e-05	-24.342	0.000	-0.002	-0.002
white	0.0143	0.029	0.489	0.625	-0.043	0.071
black	-0.1901	0.044	-4.303	0.000	-0.277	-0.104
hispanic	-0.0468	0.030	-1.568	0.117	-0.105	0.012
married	0.1745	0.023	7.546	0.000	0.129	0.220
NCHILD	-0.0060	0.010	-0.592	0.554	-0.026	0.014
vet	-0.0388	0.049	-0.796	0.426	-0.134	0.057

Omnibus:	2558.139	Durbin-Watson:	1.882
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12022.918
Skew:	-1.347	Prob(JB):	0.00
Kurtosis:	8.075	Cond. No.	2.67e+04

a. What fraction of the variation in log wages does the model explain?

- The R-squared is 0.292 and the adjusted R-squared is 0.291, meaning that approximately 29% of the variation in log wages is explained by the model.

b. Test the hypothesis with alpha = 0.10

- The null hypothesis is stating that none of the x-variables help explain ln(incwage), stating that they are jointly not significant.
- To test for joint significance, we have to look at the F-statistic. Since the F-statistic is 360.2 we can reject the null hypothesis in favor of the alternative hypothesis

c. What is the return to an additional year of education? Is this statistically significant? Is it practically significant? Briefly explain.

- The additional return to an additional year of education is 10.09% based on the coefficient of educdc. This is statistically significant at the 5% level because the p-value is below 0.05. This is also practically significant because it is a relatively high increase in wage proportion even when weighed against the cost of an additional year of education. The cost associated with an additional year invested in education will be offset by the increase in wage levels as the cost can be paid off from the years in the workforce.

d. At what age does the model predict an individual will achieve the highest wage?

- The FOC by taking the derivative with respect to age shows the following equation.

$$\frac{\partial Model}{\partial Age} = 0.1593 - 2 * 0.0017 Age$$

- Wage will therefore be highest when: $0.1593 - 2 * 0.0017 Age = 0$
- The equation solves for: **Age = 0.1593/(2*0.0017) = 46.85**

e. *Does the model predict that men or women will have higher wages, all else equal? Briefly explain why we might observe this pattern in the data.*

- The model predicts that men will earn higher wages than women all else equal because the coefficient for female is -0.4063. We can observe the same result when using the
- We might observe this pattern in the data because it reflects reality. It could be that women are paid less than men because of wage discrimination or because women are systematically employed in lower paying jobs.

f. *Interpret the coefficients on the white, black, and hispanic variables.*

- Black: Wage decreases on average by 19.01% compared to other races, except white and Hispanics, when the person is black, all else equal.
- White: Although the coefficient seems to indicate that by being white wage increases 1.43% compared to other races except black and Hispanics, since this coefficient is not significant at the 5% level because the p-value is above 0.05, this coefficient is not meaningful.
- Hispanic: Although the coefficient seems to indicate that by being Hispanic wage decreases 4.68% compared to other races except black and white, since this coefficient is not significant at the 5% level because the p-value is above 0.05, this coefficient is not meaningful

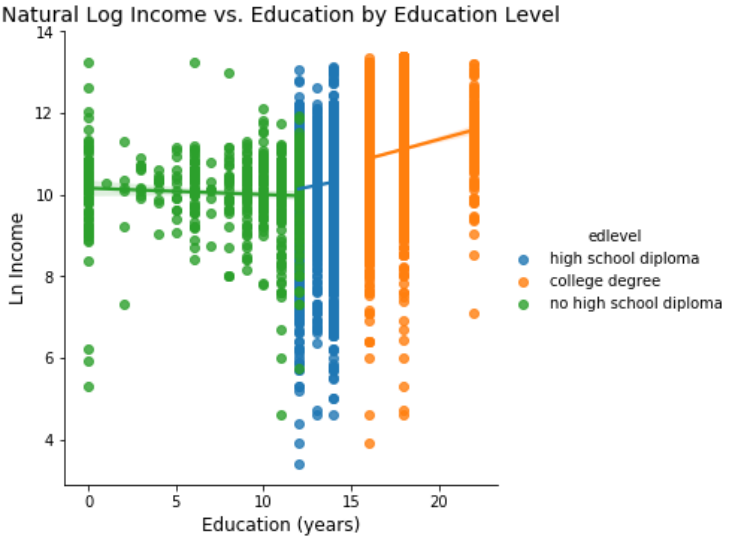
g. *Test the hypothesis that race has no effect on wages. Be sure to explicitly state the null and alternative hypotheses and show your calculations.*

$$H_0: \beta_{black} = \beta_{white} = \beta_{hispanic} = 0$$
$$H_A: \beta_j \neq 0 \; for \; some \; j, where \; j \; is \; black, white, hispanic$$

- Using the partial F-test we can see that we have enough evidence to reject the null because the full model is significantly better than the reduced model (the model where black, white and Hispanic is not included), because the F-statistic is 10.849812.

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	8728.0	7429.833705	0.0	NaN	NaN	NaN
1	8731.0	7457.636001	-3.0	-27.802295	10.849812	NaN

4. *Graph ln(incwage) and education. Include a three distinct linear fit lines specific to individuals with no high school diploma, a high school diploma, and a college degree. Be sure to label all axis and include an informative title.*



5. *Since the President is considering new education legislation, she asks you to determine whether a college degree is a strong predictor of wages. Write down a model that will allow the returns to education to vary by degree acquired (use the three categories in the previous question). Be sure to include the controls from question 3. Explain/justify why you think your model is the best possible representation of the way the world works.*
   - $\ln(incwage) = \beta_0 + \beta_1 educdc + \beta_2 female + \beta_3 age + \beta_4 age^2 + \beta_5 white + \beta_6 black + \beta_7 hispnaic + \beta_8 married + \beta_9 nchild + \beta_{10} vet + \beta_{11} hsdip + \beta_{12} coldip + \beta_{13} hsdip\_educdc + \beta_{14} coldip\_educdc + \varepsilon$
   - In order to allow the returns to education to vary by degree acquired I added the hsdip and coldip variables. I did not include a variable for no high school diploma because we need a base dummy variable. This model is better than utilizing the incwage variable as the y-variable because using incwage faces the issue of heteroskedasticity. Heteroskedasticity can be an issue because it violates the OLS assumption that residuals should have constant variance.
   - I included the interaction terms because I thought that the effect of educdc depended somewhat on the value of hsdip or coldip. This is because in the job market there are limitations to who can apply based on whether they have a certain level of education degree. So the effect of an additional year of education can be dependent on whether that additional year is associated with such a degree.

6. *Estimate the model you proposed in the previous question and report your results.*

OLS Regression Results

Dep. Variable:	ln_incwage	R-squared:	0.312
Model:	OLS	Adj. R-squared:	0.310
Method:	Least Squares	F-statistic:	281.9
Date:	Mon, 01 Feb 2021	Prob (F-statistic):	0.00
Time:	22:29:03	Log-Likelihood:	-11570.
No. Observations:	8739	AIC:	2.317e+04
Df Residuals:	8724	BIC:	2.328e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.8712	0.142	48.538	0.000	6.594	7.149
educdc	0.0050	0.009	0.543	0.587	-0.013	0.023
female	-0.4112	0.020	-20.759	0.000	-0.450	-0.372
AGE	0.1481	0.006	25.831	0.000	0.137	0.159
age2	-0.0015	6.78e-05	-22.601	0.000	-0.002	-0.001
white	0.0386	0.029	1.335	0.182	-0.018	0.095
black	-0.1538	0.044	-3.520	0.000	-0.239	-0.068
hispanic	-0.0611	0.030	-2.067	0.039	-0.119	-0.003
married	0.1546	0.023	6.764	0.000	0.110	0.199
NCHILD	-0.0041	0.010	-0.407	0.684	-0.024	0.016
vet	-0.0180	0.048	-0.374	0.709	-0.112	0.076
coldip	-0.5375	0.204	-2.636	0.008	-0.937	-0.138
hsdip	-0.8444	0.202	-4.174	0.000	-1.241	-0.448
hsdip_educdc	0.0814	0.017	4.837	0.000	0.048	0.114
coldip_educdc	0.0801	0.014	5.614	0.000	0.052	0.108

   a. *Predict the wages of an 22 year old, female individual (who is neither white, black, nor Hispanic, is not married, has no children, and is not a veteran) with a high school diploma and an all else equal individual with a college diploma. Assume that it takes someone 12 years to graduate high school and 16 years to graduate college.*
   - Using the model above we can see that the log wage levels are as follows:
     - With high school diploma: 9.17

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	9.168697	0.038831	9.092579	9.244816	7.382996	10.954399

- With college diploma: 9.80

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	9.799916	0.039083	9.723304	9.876529	8.014194	11.585639

- In order to predict the actual wage in dollars I would have to change the y-variable to incwage instead of ln(incwage).
  - With high school diploma: - 1605.32 dollars → This number however is unrealistic

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	-1605.315544	2679.687566	-6858.135435	3647.504348	-124833.421437	121622.79035

  - With college diploma: 37868.63 dollars

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	37868.633791	2697.076119	32581.728233	43155.539349	-85360.929771	161098.197353

b. The President wants to know, given your results, do individuals with college degrees have higher predicted wages than those without? By how much? Briefly explain

- Those with college degrees have higher predicted wages than those without.
- When looking at the log wages difference we can see that those with college diplomas earn 63% more than those with high school degrees, all else equal
- Based on the calculations with the y-variable as incwage we can see that those with college degrees earn on average about 39473.95 dollars more than those with high school degrees all else equal.

c. The President asked you to look into this question because she is considering legislation that will expand access to college education (for instance, by increasing student loan subsidies). She will only support the legislation if there are cost offsets (if college education increases wages and therefore, future income tax revenues that help reduce the net cost of the subsidy). Given that criteria, how would you advise the President?

- I would advise the president to consider expanding access to college education because there is a strong correlation between the two and the difference in wage levels between those with college degrees and those with only high school diplomas is large even when considering the additional cost. Assuming that those in the workforce work on average between 10 years, these 10 years of higher income could provide tax revenues and reduce subsidies that would offset the cost of implementing the policy to expand access to college education.
- However, I would also advise the president to consider the possibility that the reason why we observe this pattern where wage increases when there is a college degree is not due to a real increase in ability but rather due to the stigma that deters high school graduates with no college degree to from being employed to higher paying jobs. In this case it could be more efficient to advocate a policy that creates more higher paying jobs for high school graduates and to eradicate stigma against them.
- Also, I would advise the president to target the policy to expand access to college education to those who show aptitude and willingness because it is possible that those who actually have a college are different in essence from those who do not have a college degree in unobservables. Those who finished the college degree could be more diligent and have higher innate aptitude that leads them to higher wages. It would be more effective to aim the policy to those who will therefore finish their college degree when offered the opportunity.

7. *There are many ways that this model could be improved. How would you do things differently if you were asked to predict the returns to education given the data available on IPUMS?*
    - We could include additional data about geography, vocational training and school type from the IPUMS to improve the model. By controlling for geography we can see whether there is a difference in the relationship between wage and education given geography and this could also help observe building policies tailored to the specific regions. Vocational training could be controlled for to help better understand if vocational training is a good alternative to regular education. Controlling for school type could also be helpful because it could help better understand if the impact on wage differs depending on the type of school where the education is received.
    - We could also include data not available in the IPUMS to measure the level of innate ability through factors such as aptitude tests to control for some of the unobservables.
    - Also if we are interested in predicting the returns to education we could change the y-variable to some other data such as family income or marital status to see the impact of education on other factors as well.