Ricardo Saucedo, **UCID: 12245077**
PPHA 30545
Dr. Christopher Clapp
2/1/2021

ML Lab Mini-Project 1

4 Data Analysis

---

1. Compute descriptive (summary) statistics for the following variables: *year, incwage, lnincwage, educdc, female, age, age^2, white, black, hispanic, married, nchild, vet, hsdip, coldip*, and the interaction terms. In other words, compute sample means, standard deviations, etc.

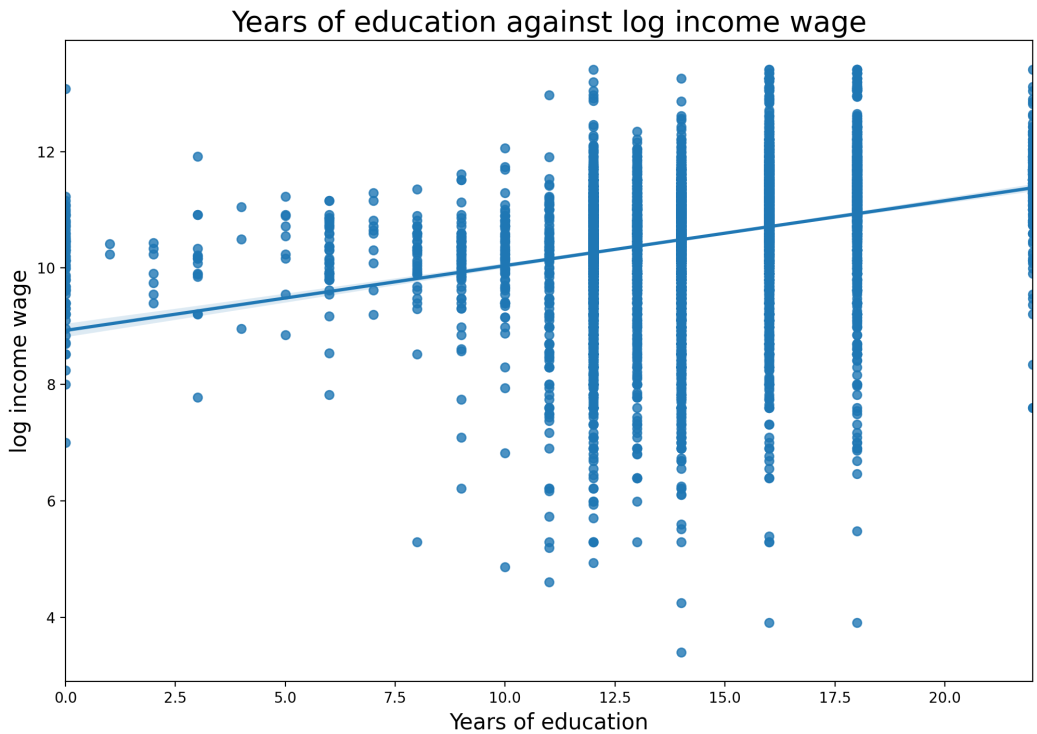
2.

```
Year summary... in case you're curious
count    8665.0
mean     2019.0
std         0.0
min      2019.0
25%      2019.0
50%      2019.0
75%      2019.0
max      2019.0
Name: YEAR, dtype: float64

Income wage summary
count      8665.000000
mean      58128.161570
std       66595.450579
min          30.000000
25%       23000.000000
50%       41000.000000
75%       70000.000000
max      665000.000000
Name: INCWAGE, dtype: float64

Log income wage summary
count    8665.000000
mean       10.510626
std         1.070592
min         3.401197
25%        10.043249
50%        10.621327
75%        11.156251
max        13.407542
Name: LNincwage, dtype: float64

Education year count summarized
count    8665.000000
mean       14.218927
std         2.940894
min         0.000000
25%        12.000000
50%        14.000000
75%        16.000000
max        22.000000
Name: EDUCDC, dtype: float64

Female summary
count    8665.000000
mean        0.487248
std         0.499866
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         1.000000
Name: female, dtype: float64

Age summary
count    8665.000000
mean       41.754183
std        13.168988
min        18.000000
25%        31.000000
50%        42.000000
75%        53.000000
max        65.000000
Name: AGE, dtype: float64
```

```
Age square summary
count    8665.000000
mean     1916.814080
std      1105.097898
min       324.000000
25%       961.000000
50%      1764.000000
75%      2809.000000
max      4225.000000
Name: ageSq, dtype: float64

White summary
count    8665.000000
mean        0.767340
std         0.422552
min         0.000000
25%         1.000000
50%         1.000000
75%         1.000000
max         1.000000
Name: white, dtype: float64

Black summary
count    8665.000000
mean        0.092672
std         0.289989
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: black, dtype: float64

Hispanic summary
count    8665.000000
mean        0.150952
std         0.358023
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: hispan, dtype: float64

Married sum
count    8665.000000
mean        0.016157
std         0.126086
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: married, dtype: float64

Number of children
count    8665.000000
mean        0.801039
std         1.098086
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         9.000000
Name: NCHILD, dtype: float64
```

```
Veteran status summary
count    8665.000000
mean        0.047086
std         0.211835
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: vet, dtype: float64

Highschool diploma summary
count    8665.000000
mean        0.244893
std         0.430049
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: hsdip, dtype: float64

College diploma summary
count    8665.000000
mean        0.242816
std         0.428809
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: coldip, dtype: float64

Interaction term summary
count    8665.000000
mean        2.938719
std         5.160584
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max        12.000000
Name: educXhs, dtype: float64

count    8665.000000
mean        3.885055
std         6.860952
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max        16.000000
Name: educXcol, dtype: float64
```

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 following model:

$$ln(incwage) = \beta_0 + \beta_1 educdc + \beta_2 female + \beta_3 age + \beta_4 age2 + \beta_5 white + \beta_6 black + \beta_8 hispanic + \beta_9 married + \beta_{10} nchild + \beta_{11} vet + \varepsilon,$$

and report your results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               LNincwage   R-squared:                       0.296
Model:                             OLS   Adj. R-squared:                  0.295
Method:                  Least Squares   F-statistic:                     363.5
Date:                 Mon, 01 Feb 2021   Prob (F-statistic):               0.00
Time:                         17:44:11   Log-Likelihood:                -11366.
No. Observations:                 8665   AIC:                         2.275e+04
Df Residuals:                     8654   BIC:                         2.283e+04
Df Model:                           10
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.4635      0.115     47.314      0.000       5.237       5.690
EDUCDC         0.0997      0.003     29.151      0.000       0.093       0.106
female        -0.4007      0.020    -20.321      0.000      -0.439      -0.362
AGE            0.1753      0.006     31.078      0.000       0.164       0.186
ageSq         -0.0018   6.71e-05    -27.247      0.000      -0.002      -0.002
white          0.0424      0.029      1.480      0.139      -0.014       0.099
black         -0.1775      0.042     -4.247      0.000      -0.259      -0.096
hispan        -0.0775      0.029     -2.713      0.007      -0.134      -0.022
married       -0.0983      0.077     -1.276      0.202      -0.249       0.053
NCHILD         0.0133      0.010      1.379      0.168      -0.006       0.032
vet           -0.0450      0.046     -0.969      0.333      -0.136       0.046
==============================================================================
Omnibus:                     2698.134   Durbin-Watson:                   1.863
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13300.836
Skew:                          -1.424   Prob(JB):                         0.00
Kurtosis:                       8.359   Cond. No.                     2.67e+04
==============================================================================
```

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

a. The R-Squared is .296, which means that the fraction of the variation in log wages is explained by approximately 30 percent of the model.

(b) Test the hypothesis that:

$$H_0: \beta_1 = \beta_2 = ... = \beta_{11} = 0$$
$$H_A: \beta_j \neq 0 \text{ for some } j \text{ with } \alpha = 0.10.$$

With an alpha of 0.1, we are able to reject the null at a 10 percent significance level. The F-Statistic is 363.5, as well as our p-value. Thus, we are certain that we are able to reject the null hypothesis. Furthermore, this means that our X variables all have predictive values; the predictors we should be careful about are **white**, **married**, **NCHILD**, and **vet**.

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

a. Because this is the log income wage, an additional year of education yields a 9.97 percent increase in one's wages. This is statistically significant, with certainty in our p-values, confidence intervals, and t-statistic. Given that this is an econometric measure, it is also practically significant when one's income is at a certain threshold. The magnitude at which the amount of years of education one has affecting their log wages will have a practical and realistic significance whenever the amount one gains in income is substantial to a 9.97 percent increase.

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

   a. When we take the derivative with respects to age in our regression model and maximize, we find that the model predicts that an individual will achieve the highest wage at 46.36 years old.

4D

$$\ln(incwage) = \beta_0 + \beta_1 educdc + \beta_2 female + \beta_3 age + \beta_4 age^2 + \beta_5 white + \beta_6 black + \beta_8 hispanic + \beta_9 married + \beta_{10} nchild + \beta_{11} vet + \varepsilon$$

$$\frac{d(\ln(incwage))}{d\, age} = \beta_3 + 2\beta_4 age$$

$$0 = .1669 + 2(-0.0018)\, age$$

$$0 = .1669 - 0.0036\, age$$

$$\frac{-.1669}{-0.0036} = \frac{-0.0036\, age}{-0.0036}$$

$$46.36 = age$$

(coefficient values may differ from work shown above due to time screenshot was taken, but same practice applies).

(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.

   a. Seeing as how the coefficient for female is negative, the model predicts that those who identify as female will reportedly have lower wages, even if we hold all else equal. Since the model does not account for tax or tax credits, the data would otherwise support the predicted values the model produced.

(f) Interpret the coefficients on the white, black, and Hispanic variables.

   a. The coefficient on the white variable is statistically insignificant, even at the alpha = .1 level. Black and hispanic variables, however, are statistically significant, both with negative effects. Therefore this would translate into receiving negative wages solely from the fact that one identifies as black or hispanic.

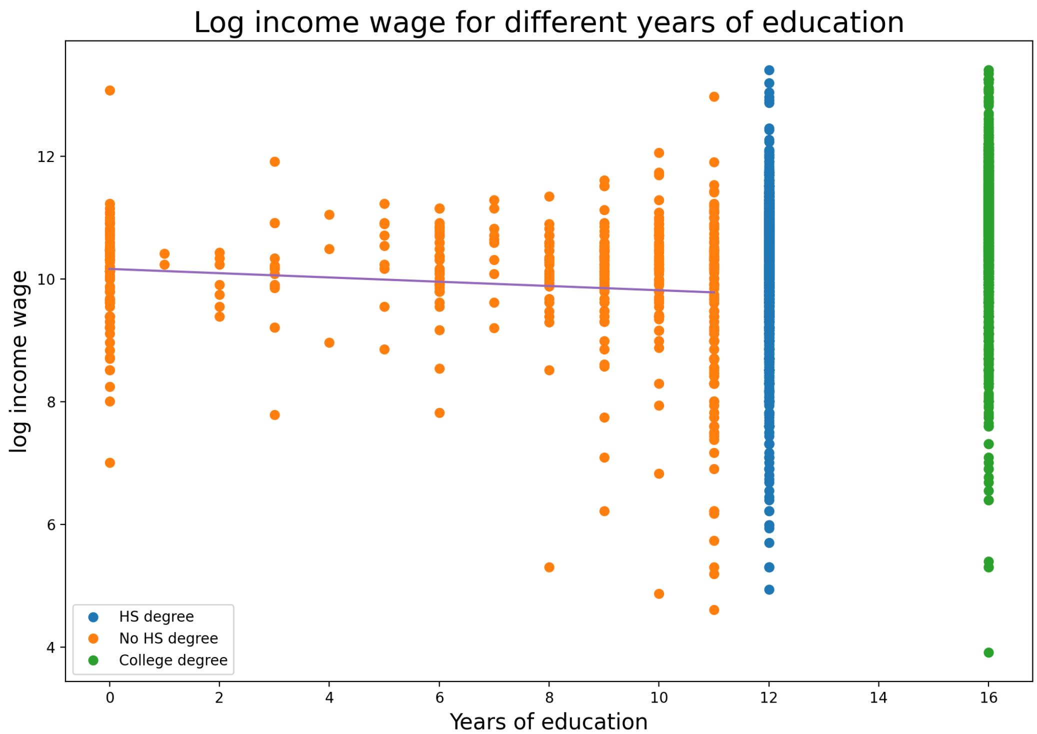
(g) Test the hypothesis that race has no effect on wages. Be sure to explicitly state the null and alternative hypotheses and show your calculations.

a. Considering the rhetoric that some would argue Hispanic not being a race, but more so an ethnicity, I will proceed with the assumption that race is equal to white and black for this question.

The Null Hypothesis would be that race has no effect on wages, namely that Beta_5 + Beta_6 = 0; thus the indicator variable would have a 0 percent change in one's wages. Based off of my model's predictions, I will take a one sided approach.
The Alternative Hypothesis would state that race does have a negative non-zero effect on one's wage, namely that Beta_5 + Beta_6 =/= 0. In my calculations, I conducted a linear regression of the influence race had on the response. Based on my model's predictions, we are able to statistically reject the null hypothesis, for we see that race still does have an effect on one's wages.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               LNincwage   R-squared:                       0.009
Model:                             OLS   Adj. R-squared:                  0.009
Method:                  Least Squares   F-statistic:                     41.15
Date:                 Mon, 01 Feb 2021   Prob (F-statistic):           1.63e-18
Time:                         18:07:05   Log-Likelihood:                -12845.
No. Observations:                 8665   AIC:                         2.570e+04
Df Residuals:                     8662   BIC:                         2.572e+04
Df Model:                            2
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.4095      0.031    340.203      0.000      10.349      10.469
white          0.1534      0.033      4.611      0.000       0.088       0.219
black         -0.1787      0.048     -3.686      0.000      -0.274      -0.084
==============================================================================
Omnibus:                      1766.509   Durbin-Watson:                   1.773
Prob(Omnibus):                   0.000   Jarque-Bera (JB):             5010.290
Skew:                           -1.075   Prob(JB):                         0.00
Kurtosis:                        6.043   Cond. No.                         6.66
==============================================================================
```

4. Graph ln(incwage) and education. Include 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.

```
Inserting our interaction terms to our model
                            OLS Regression Results
==============================================================================
Dep. Variable:              LNincwage   R-squared:                       0.300
Model:                            OLS   Adj. R-squared:                  0.299
Method:                 Least Squares   F-statistic:                     309.1
Date:                Mon, 01 Feb 2021   Prob (F-statistic):               0.00
Time:                        18:44:41   Log-Likelihood:                -11340.
No. Observations:                8665   AIC:                         2.271e+04
Df Residuals:                    8652   BIC:                         2.280e+04
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.6499      0.120     47.138      0.000       5.415       5.885
EDUCDC         0.0880      0.004     22.896      0.000       0.080       0.096
female        -0.4034      0.020    -20.487      0.000      -0.442      -0.365
AGE            0.1731      0.006     30.735      0.000       0.162       0.184
ageSq         -0.0018   6.71e-05    -26.850      0.000      -0.002      -0.002
white          0.0427      0.029      1.493      0.135      -0.013       0.099
black         -0.1663      0.042     -3.989      0.000      -0.248      -0.085
hispan        -0.0749      0.029     -2.625      0.009      -0.131      -0.019
married       -0.0836      0.077     -1.087      0.277      -0.234       0.067
NCHILD         0.0130      0.010      1.349      0.177      -0.006       0.032
vet           -0.0400      0.046     -0.863      0.388      -0.131       0.051
educXhs       -0.0064      0.002     -3.038      0.002      -0.011      -0.002
educXcol       0.0090      0.002      5.848      0.000       0.006       0.012
==============================================================================
Omnibus:                     2741.314   Durbin-Watson:                   1.869
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13671.480
Skew:                          -1.446   Prob(JB):                         0.00
Kurtosis:                       8.432   Cond. No.                     2.77e+04
==============================================================================
```

Taking our formula from 3 and adding on the two interaction terms, we still can see that white, married, NCHILD, and vet variables are insignificant. However, we see that there is a statistical significance at the $\propto = .1$ level with the addition of the two interaction terms. I believe this new model explains, despite the statistical insignificance of some variables, a realistic view of the way the world works; each of these variables contribute to one’s income wage. When we incorporate the interaction terms, we see that having a college degree is positive, which the President could interpret as being a strong predictor of wages. Also, R^2 is higher (slightly), which means our predictors explain the variation in log income wages based on the model.

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

a. Predict the wages of a 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

6a

$$\begin{aligned}\ln(incwage) &= \beta_0+\beta_1 educdc+\beta_2 female+\beta_3 age+\beta_4 age^2\\ &+ \beta_5 \cancel{white}+\beta_6 \cancel{black}+\beta_8 \cancel{hispanic}\\ &+ \beta_9 \cancel{married}+\beta_{10} \cancel{nchild}+\beta_{11} \cancel{vet}+\varepsilon,\\ &+ \beta_{13}\, educdc \times hs + \beta_{14}\, educdc \times col\\ &+ \varepsilon\end{aligned}$$

$$= 5.6499 + (0.088)(12) + (-0.4034)(1) + (.1731)(22) + (-0.0018)(22^2) + (-0.6064)(12) + (0.0090)(0)$$

$\ln(incwage) \approx 10.08370$

$$\begin{aligned}\ln(incwage) &= \beta_0+\beta_1 educdc+\beta_2 female+\beta_3 age+\beta_4 age^2\\ &+ \beta_5 \cancel{white}+\beta_6 \cancel{black}+\beta_8 \cancel{hispanic}\\ &+ \beta_9 \cancel{married}+\beta_{10} \cancel{nchild}+\beta_{11} \cancel{vet} \cancel{+\varepsilon},\\ &+ \beta_{13}\, educdc \times hs + \beta_{14}\, educdc \times col\\ &+ \varepsilon\end{aligned}$$

$$= 5.6499 + (0.088)(12) + (-0.4034)(1) + (.1731)(22) + (-0.0018)(22^2) + (-0.6064)(0) + (0.0090)(16)$$

$\ln(incwage) \approx 9.997$

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.

By my models predictions, my estimate results show that individuals with college degrees have nearly identically, yet still slightly higher percent predicted wages than those who have only received a high school diploma, by nearly a few percentage points (5-10%).

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. If I were the president, I would study or investigate the longevity of one's involvement in the workforce for those pursuing college education. If there remains a positive slope coefficient, then it might be worthwhile to pursue increasing student loan subsidies.

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?

    Depending on what parameters I would use to predict the returns to education, I would ensure some geographical context is also provided. Using data like proximal distance to city center, communal health indexes, and other more contextual specifiers that are identifiable like mental disabilities or other handicaps.