Machine Learning - Lab Mini-Project 1
Sief Salameh
04/11/2023

Data Analysis Answers

Question 1)

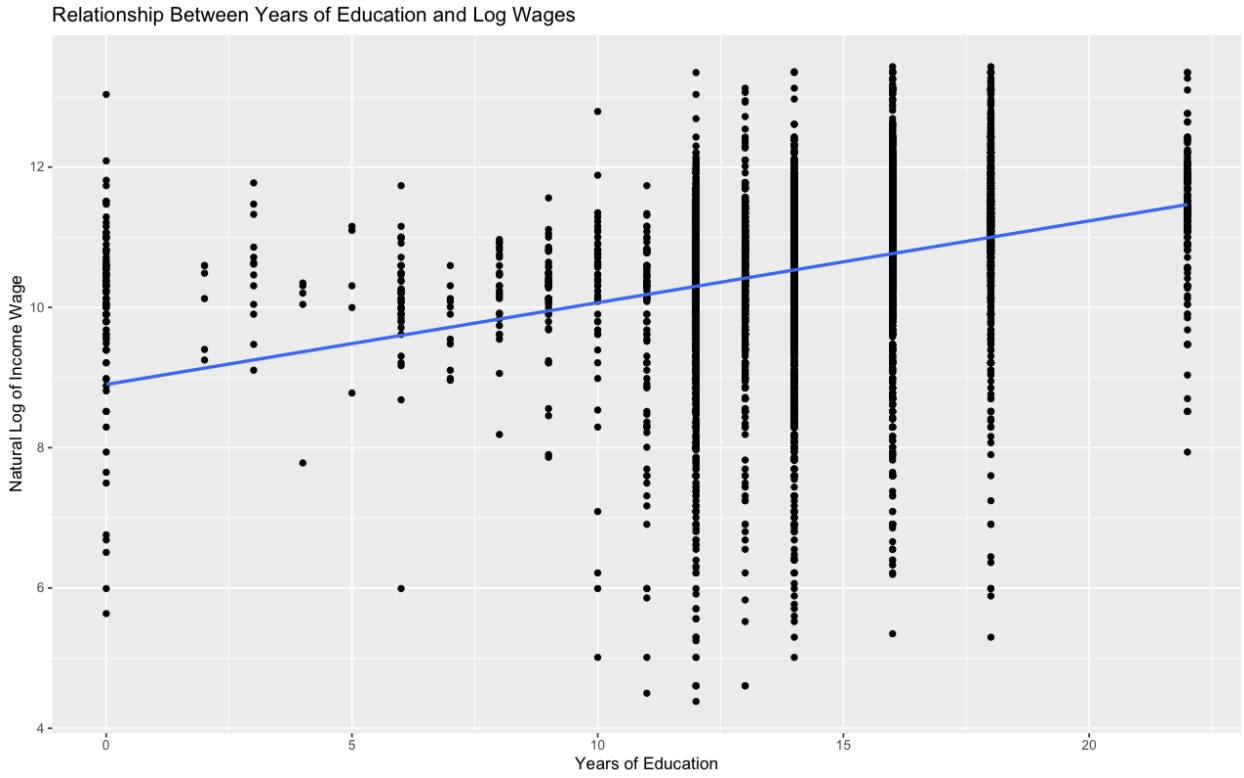
- **Mean, Median, Quantiles**

```
> summary(edu_selected)
      YEAR         INCWAGE        log_incwage         educdc          female      
 Min.   :2021   Min.   :     0   Min.   :  -Inf   Min.   : 0.0   Min.   :0.0000  
 1st Qu.:2021   1st Qu.: 20800   1st Qu.: 9.943   1st Qu.:12.0   1st Qu.:0.0000  
 Median :2021   Median : 43200   Median :10.674   Median :14.0   Median :0.0000  
 Mean   :2021   Mean   : 59702   Mean   :  -Inf   Mean   :14.3   Mean   :0.4845  
 3rd Qu.:2021   3rd Qu.: 75000   3rd Qu.:11.225   3rd Qu.:16.0   3rd Qu.:1.0000  
 Max.   :2021   Max.   :682000   Max.   :13.433   Max.   :22.0   Max.   :1.0000  
      AGE            age_sq          white            black           hispanic      
 Min.   :18.00   Min.   : 324   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.:31.00   1st Qu.: 961   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
 Median :42.00   Median :1764   Median :1.0000   Median :0.00000   Median :0.0000  
 Mean   :41.91   Mean   :1934   Mean   :0.6793   Mean   :0.07757   Mean   :0.1464  
 3rd Qu.:53.00   3rd Qu.:2809   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
 Max.   :65.00   Max.   :4225   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000  
    married           NCHILD            vet              hsdip            coldip      
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.0000   Median :0.0000   Median :0.00000   Median :1.0000   Median :0.0000  
 Mean   :0.5517   Mean   :0.7877   Mean   :0.04472   Mean   :0.5282   Mean   :0.4106  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :9.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
     no_hs         interaction_hs   interaction_clg  interaction_no_hs
 Min.   :0.00000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000  
 1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000  
 Median :0.00000   Median :12.000   Median : 0.000   Median : 0.0000  
 Mean   :0.06126   Mean   : 6.863   Mean   : 6.965   Mean   : 0.4675  
 3rd Qu.:0.00000   3rd Qu.:13.000   3rd Qu.:16.000   3rd Qu.: 0.0000  
 Max.   :1.00000   Max.   :14.000   Max.   :22.000   Max.   :12.0000  
```

- **Standard Deviation**

```
> print(sd_df)
           YEAR          INCWAGE       log_incwage            educdc 
   0.000000e+00    7.127281e+04             NaN    3.009025e+00 
         female             AGE           age_sq            white 
   4.997878e-01    1.330033e+01    1.125624e+03    4.667611e-01 
          black         hispanic         married           NCHILD 
   2.675128e-01    3.534879e-01    4.973507e-01    1.110542e+00 
            vet            hsdip           coldip           no_hs 
   2.066964e-01    4.992339e-01    4.919642e-01    2.398192e-01 
  interaction_hs  interaction_clg  interaction_no_hs 
   6.521817e+00    8.398432e+00    2.127704e+00 
```

Question 2)


Relationship Between Years of Education and Log Wages
Natural Log of Income Wage
Years of Education
4
6
8
10
12
0
5
10
15
20

Question 3)

```
Call:
lm(formula = log_incwage ~ educdc + female + AGE + age_sq + white +
    black + hispanic + married + NCHILD + vet, data = edu_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1376 -0.3677  0.1340  0.5501  3.3190

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.671e+00  1.236e-01  45.877   <2e-16 ***
educdc       1.043e-01  3.614e-03  28.855   <2e-16 ***
female      -3.784e-01  2.096e-02 -18.057   <2e-16 ***
AGE          1.591e-01  6.019e-03  26.436   <2e-16 ***
age_sq      -1.667e-03  7.079e-05 -23.553   <2e-16 ***
white        6.283e-02  2.919e-02   2.153   0.0314 *
black       -1.051e-01  4.550e-02  -2.310   0.0209 *
hispanic     2.894e-02  3.605e-02   0.803   0.4221
married      2.183e-01  2.400e-02   9.096   <2e-16 ***
NCHILD      -1.144e-02  1.075e-02  -1.064   0.2874
vet          3.238e-02  5.052e-02   0.641   0.5216
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9379 on 8319 degrees of freedom
Multiple R-squared:  0.2839,    Adjusted R-squared:  0.2831
F-statistic: 329.9 on 10 and 8319 DF,  p-value: < 2.2e-16
```

a) Based on the Adjusted R-squared, the model explains 28.31% of the variation in log wages, taking into account the number of independent variables included in the model. I used the Adjusted R-squared value because it penalizes the model for every additional independent variable that is added but does not improve the model's predictive power. The lower Adjusted R-squared value indicates that this model uses a higher number of unnecessary independent variables and produces a less effective/worse fit to the data when comparing our estimated model vs. the true phenomena.

b) The null hypothesis states that all 11 regression coefficients educdc + female + AGE + age_sq + white + black + hispanic + married + NCHILD + vet are equal to zero. This means that none of the independent variables have a significant effect on the dependent variable log wage.

The alternative hypothesis states that at least one regression coefficient is not equal to zero. This means that at least one independent variable has a significant effect on the dependent variable log wage.

To test this hypothesis, we perform an F-test. The F-test computes the ratio of the explained variance to the unexplained variance (error term) and follows an F-distribution with degrees of freedom equal to the number of independent variables (10), and the sample size minus the number of independent variables minus 1 (8330-10-1 = 8319).

Therefore, the F-test in this linear regression model is equal to 329.9 with degrees of freedom of 10 and 8319. The associated p-value is less than 2.2e-16, which is very close to zero and much smaller than the chosen significance level of $\alpha = 0.10$.

Hence, we reject the null hypothesis that all 11 regression coefficients are equal to zero and conclude that the model is statistically significant where at least one independent variable has a significant effect on the dependent variable (log wage). The F-test and its associated p-value tell us there is evidence that at least one independent variable, including educdc, female, AGE, age_sq, white, black, married, NCHILD, and vet, is correlated with the values in the dependent variable (log wage).

c) The coefficient estimate for education is .1043, which means that the return for each additional year of education correlates with a 10.43 percentage point increase in log wages - ceteris paribus. The p-value for the coefficient estimate is less than 2e-16, which is highly statistically significant at the lowest significance level/ alpha of .001. Therefore, we can reject the null hypothesis that the true coefficient is zero - which means education has a significant effect on the values for log wages. The coefficient estimate is also practically significant because 12 years of education (a high school diploma) will increase an individual's log wages by 125.16 percentage points compared to someone with 0 years of education. Likewise, 20 years of education (a doctoral or phd education) will increase an individual’s log wages by 208.6 percentage points compared to someone with 0 years of education. Realistically, a high school graduate who makes $40,000 a year or 10.60 log wages a year, can predict their income to increase to approximately $61,000 a year or 11.01 log wages after completing a 4-year bachelor's degree. This is because the 4 additional years of education increased their income by (.1043 x 4 = .4172 log wages). Hence, we take the sum of 10.60 + .4172 = 11.01 log wages. Therefore, we can see that the additional 4 years of education increased the individual’s income by more than $20,000, thus the coefficient estimate for education is both statistically and practically significant.

d) To find the predicted age where an individual will achieve the highest wage, we have to look at the AGE and age_sq variables. The coefficient estimate for AGE tells us that there is a highly statistically significant positive effect between AGE and log wages. In other words, every additional unit for AGE increases the log wages by 0.1591 or 15.91 percentage points (ceteris paribus). However, the age squared variable tells that there is a negative correlation between age squared and log wages. It is also highly statistically significant and shows that for every additional unit increase in age squared, the log wages decrease by 0.001667 or 0.1667 percentage points (ceteris paribus). This information tells us that increasing age tends to have a larger positive effect on log wages compared to a negative effect. However, at some point in the model, the increase in age will begin to have a larger negative effect on log wages compared to a positive effect.

To compute the specific age that will achieve the highest wage, we can take the derivative of the model's function with respect to AGE. To do this, we differentiate the AGE term first and then treat all other variables as constants:

1) log_incwage = 5.671 + 0.1043 * educdc - 0.3784 * female + 0.1591 * AGE - 0.001667 * age_sq + 0.06283 * white - 0.1051 * black + 0.02894 * hispanic + 0.2183 * married - 0.01144 * NCHILD + 0.03238 * vet

2) derivative(log_incwage) / derivative(AGE) = 0 + 0 - 0 + 0.1591 - 0.001667 * 2 * AGE + 0 + 0 + 0 + 0 + 0 + 0

3) derivative(log_incwage) / derivative(AGE) = 0.1591 - 0.003334 * AGE

4) 0.1591 - 0.003334 * AGE = 0

5) 0.1591 = 0.003334 * AGE

6) AGE = 0.1591 / 0.003334

7) **AGE = 47.72**

- **Therefore, based on these computations, the model predicts that an individual will achieve the highest wage at the age of 48 years.**

CITATION: https://towardsdatascience.com/linear-regression-derivation-d362ea3884c2

CITATION:
https://towardsai.net/p/machine-learning/linear-regression-complete-derivation-with-mathematics-explained

e) The model predicts that men will have higher wages (all else equal). This is because the coefficient estimate for females is -0.3784 or -37.84 percentage points. This means that if an individual is a female, then their log wages are predicted to decrease by -0.3784 (ceteris paribus). This estimate is highly statistically significant at the lowest significance /alpha level of .001. Therefore, we can reject the null hypothesis that the true coefficient is zero - which means gender has a significant effect on the values for log wages. This model indicates that females who are in the labor market are being discriminated against through lower wages or job positions that pay lower salaries. Another explanation might suggest that females have to stay home longer and sacrifice wages in order to take care of their children or experience non-paid maternal leave. Thus, males will have predicted higher wages than females.

f) The estimated coefficient for the variable white is equal to 0.0628. This means that if an individual is white, then their predicted log income will increase by 6.28 percentage points (ceteris paribus). The p-value is equal to 0.0314, which means the coefficient estimate is statistically significant at the .05 significance /alpha level. At the given significance level of .05, we can reject the null hypothesis that the true coefficient is zero - which means being white has some positive effects on the values for log wages.

The estimated coefficient for the variable black is equal to -0.1051. This means that if an individual is black, then their predicted log income will decrease by 10.51 percentage points (ceteris paribus). The p-value is equal to 0.0209, which means the coefficient estimate is statistically significant at the .05 significance /alpha level. At the given significance level of .05, we can reject the null hypothesis that the true coefficient is zero - which means being black has some negative effects on the values for log wages.

The estimated coefficient for the variable hispanic is equal to 0.0289. This means that if an individual is hispanic, then their predicted log income will increase by 2.89 percentage points (ceteris paribus). The p-value is equal to 0.4221, which means the coefficient estimate is not statistically significant in any of the significance /alpha levels.

g)

Null hypothesis (H0): Race has no effect on wages.

Alternative hypothesis (HA): Race has an effect on wages.

Based on the initial p-values for the white and black race variables in the linear regression, the coefficient estimates are statistically significant at the .05 significance /alpha level. Therefore, we reject the null hypothesis that race has no effect on log wages.

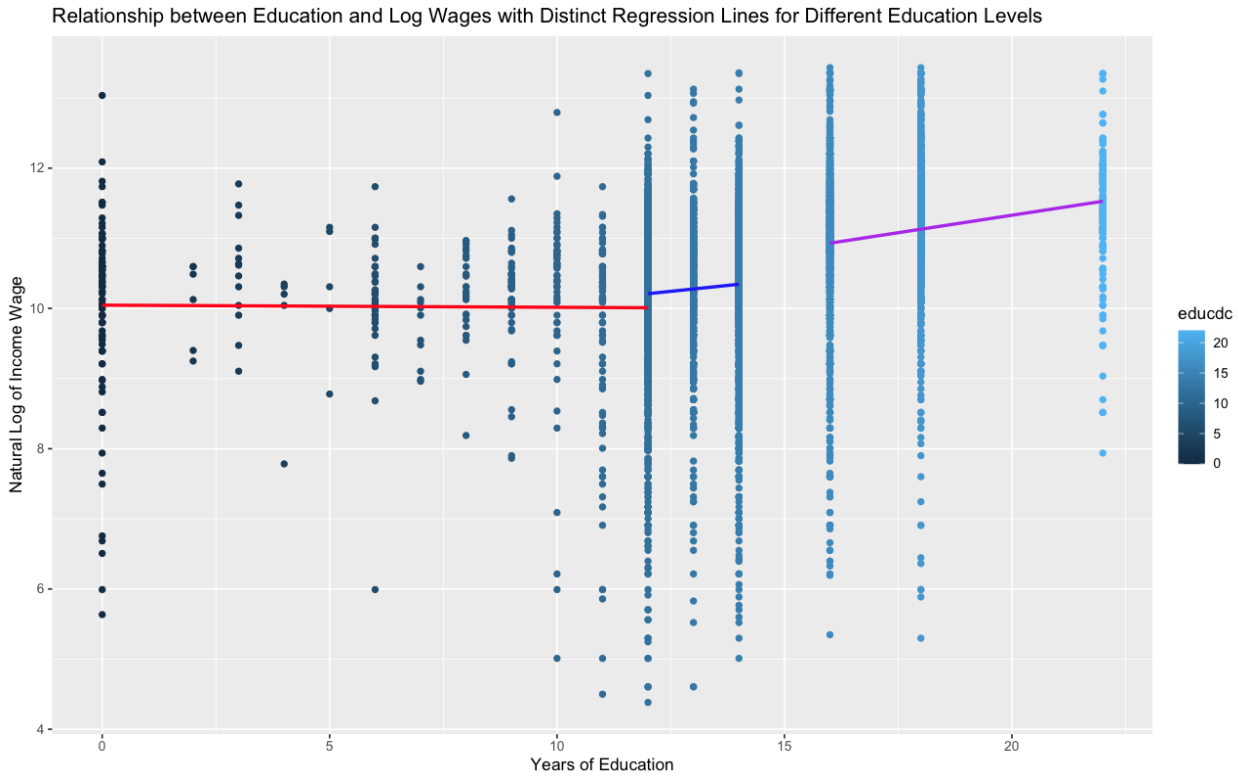
Furthermore, I conducted an Anova test between the full model (with race variables included) and the reduced model (without the race variables included), and found that the full model had a lower residual sum of squares (RSS) estimate compared to the reduced model. This means that the full model fits the data better than the reduced model.

The Anova test also showed that the F-test for the difference in the RSS estimates between the two models was 10.499 with an associated p-value equal to 2.793e-05, which is closer to zero and much smaller than the smallest significance level of 0.001. Therefore, we can firmly reject the null hypothesis that race has no effect on wages, and state that race does correlate with the values in log wages.

```
Analysis of Variance Table

Model 1: log_incwage ~ educdc + female + AGE + hispanic + age_sq + married +
    NCHILD + vet
Model 2: log_incwage ~ educdc + female + AGE + age_sq + white + black +
    hispanic + married + NCHILD + vet
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   8321 7336.7
2   8319 7318.2  2    18.472 10.499 2.793e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question 4)


Relationship between Education and Log Wages with Distinct Regression Lines for Different Education Levels
Natural Log of Income Wage
Years of Education
educdc
20
15
10
5
0
12
10
8
6
4
0
5
10
15
20

Question 5)

```
Call:
lm(formula = log_incwage ~ hsdip + coldip + female + AGE + age_sq +
    white + black + hispanic + married + NCHILD + vet, data = edu_filtered)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0240 -0.3543  0.1370  0.5522  3.2832

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.7718629  0.1243897  54.441  < 2e-16 ***
hsdip        0.3225930  0.0455769   7.078 1.58e-12 ***
coldip       0.9519757  0.0468576  20.316  < 2e-16 ***
female      -0.3794543  0.0207870 -18.254  < 2e-16 ***
AGE          0.1497433  0.0060124  24.906  < 2e-16 ***
age_sq      -0.0015584  0.0000707 -22.042  < 2e-16 ***
white        0.0844691  0.0290147   2.911  0.00361 **
black       -0.0750013  0.0452117  -1.659  0.09718 .
hispanic     0.0170313  0.0358437   0.475  0.63469
married      0.2058158  0.0238381   8.634  < 2e-16 ***
NCHILD      -0.0080881  0.0106756  -0.758  0.44869
vet          0.0566829  0.0501922   1.129  0.25880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9305 on 8318 degrees of freedom
Multiple R-squared:  0.2953,    Adjusted R-squared:  0.2944
F-statistic: 316.9 on 11 and 8318 DF,  p-value: < 2.2e-16
```

- This model allows us to estimate the log wages based on the returns of three education levels - no high school diploma, a high school diploma, and a college degree. This linear regression model also accounts for the same controls used in the previous model. To differentiate between the education levels, I used the binary variables hsdip and coldip which assign a 1 to individuals who receive either their high school diploma or bachelors degree. Or a 0 if they did not receive either of those educational attainments.

- This model is an accurate representation of the way the world works because realistically someone who did not complete a high school diploma will not be eligible to earn a bachelor's degree. For the high school diploma but no college degree variable, I included individuals who obtained an associate’s degree or post-high school training because they officially did not earn a bachelor's degree. Likewise, for individuals without a high school diploma, I included individuals who had 10, 11, and 12 years of education but did not

finish or pass high school - thus could not earn a high school diploma. These characteristics ensure the model is accurate and reflective of the true world.

Question 6)

Part 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

1) log(wage) = 6.7718629 + 0.32259301 - 0.37945431 + 0.149743322 - 0.0015584(22^2)
2) log(wage) = 7.4482893
3) wage = exp(log(wage))
4) wage = exp(7.4482893)
5) **Wage = $1819.416**

Therefore, the predicted wage for an individual who is 22 years old, female, not white, black, or Hispanic, not married, has no children, and is not a veteran, but has a high school diploma, is $1,819.42.

Predict the wages of someone with a college diploma while holding all other variables constant

1) log(wage) = 6.7719 + 0.9519757 = 7.7239
2) wage = exp(log(wage))
3) exp(7.7239) = $2,338.55
4) **Wage = $2,338.55**

Therefore, the predicted wage for an individual with a college diploma while holding all other variables constant is $2,338.55.

Part b)

Based on the results of the linear regression model, individuals with college degrees do have higher predicted log wages compared to those without a college degree, holding all other variables constant. The coefficient estimate for "coldip" (college diploma) in the model is 0.952. This means that individuals with a college degree will experience a 95.2 percentage point increase in their log wages controlling for other factors in the model. In contrast, the coefficient estimate for a high school diploma is .323. This means that individuals with a high school diploma will experience a 32.3 percentage point increase in their log wages controlling for other factors in the model. When we take the difference between the two coefficient estimates, we find that the college degree provides a .629 or 62.9 percentage point advantage for log wages compared to a high school diploma. Both coefficient estimates were highly statistically significant at the .001 significance/ alpha level.

Part c)

I would advise the President to increase college subsidies, grants, and scholarships for the following groups - women, Black and African Americans, as well as individuals with children. These groups all experienced negative log wages if they identified as a woman, Black, or if they had children to take care of. In addition to assisting these groups, I would also strongly urge the President to create more college pipelines for women. Out of all the marginalized groups listed, identifying as a woman had the strongest negative effect on log wages. Therefore, programs that financially support women to attend college and universities should be the President's priority.

Question 7)

I did not find information or details specifying the type of wages earned. So I would differentiate between a weekly, monthly, or yearly salary to standardize the comparison between different individuals and groups. Also, the information differentiating the different education levels was too vague. It was not clear if individuals who had a particular number of educational years obtained the respective degree or diploma associated with that level. For example, did individuals who have 12 years of education earn a high school diploma, or do individuals who have 16 years of education have a Masters degree? Additional clarity on how those benchmarks were set would be helpful.