Machine Learning - Mini Project 1 Solutions

PPHA 30545 - Professor Clapp

Winter 2023

In [61]:
```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression as lm
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf
from statsmodels.stats.anova import anova_lm
```

In [62]:
```
acs_data = pd.read_csv('usa_00001.csv')
```

3. Preparing the data

3.1. Familiarizing with the data

In [63]:
```
acs_data.head()
```

Out[63]:

	YEAR	SAMPLE	SERIAL	CBSERIAL	HHWT	CLUSTER	STRATA	GQ	PERNUM	PERWT	...	RACED	HISPAN	HISPAND	EDUC	EDUCD	EMPSTAT	EMPSTATD	INCWAGE	VETSTAT	VETSTATD
0	2021	202101	1902	2021010114983	5304.0	2021000019021	160001	4	1	5304.0	...	100	0	0	7	71	1	10	10000	1	11
1	2021	202101	2994	2021000021366	25116.0	2021000029941	270201	1	2	29172.0	...	200	0	0	6	63	1	10	1000	1	11
2	2021	202101	3150	2021000032187	14664.0	2021000031501	100001	1	1	14664.0	...	100	0	0	6	63	1	10	21000	1	11
3	2021	202101	3306	2021000042884	2964.0	2021000033061	250001	1	1	3120.0	...	200	0	0	4	40	1	10	24000	1	11
4	2021	202101	3618	2021000063494	13260.0	2021000036181	130301	1	1	13104.0	...	100	0	0	11	114	1	10	85000	1	11

5 rows × 26 columns

In [64]:
```
acs_data.describe()
```

Out[64]:

	YEAR	SAMPLE	SERIAL	CBSERIAL	HHWT	CLUSTER	STRATA	GQ	PERNUM	PERWT	...	RACED	HISPAN	HISPAND	EDUC	EDUCD	EMPSTAT	EMPSTATD	INCWAGE	VETSTAT	VETSTATD
count	8556.0	8556.0	8.556000e+03	8.556000e+03	8556.000000	8.556000e+03	8.556000e+03	8556.000000	8556.000000	8556.000000	...	8556.000000	8556.000000	8556.000000	8556.000000	8556.000000	8556.0	8556.000000	8556.000000	8556.000000	8556.000000
mean	2021.0	202101.0	7.208495e+05	2.021001e+12	16262.124825	2.021007e+12	4.677905e+05	1.063114	1.694016	16624.410940	...	261.667017	0.326905	34.298621	7.886746	81.183263	1.0	10.080295	60561.317204	1.041608	11.401823
std	0.0	0.0	4.206382e+05	1.391498e+06	13530.554382	4.206382e+06	9.381907e+05	0.427287	0.953687	13964.445118	...	268.836697	0.913734	98.078562	2.352989	23.529964	0.0	0.491879	74458.147968	0.199704	1.803708
min	2021.0	202101.0	1.902000e+03	2.021000e+12	312.000000	2.021000e+12	1.000100e+04	1.000000	1.000000	156.000000	...	100.000000	0.000000	0.000000	0.000000	2.000000	1.0	10.000000	0.000000	1.000000	11.000000
25%	2021.0	202101.0	3.517320e+05	2.021000e+12	7956.000000	2.021004e+12	9.001700e+04	1.000000	1.000000	8112.000000	...	100.000000	0.000000	0.000000	6.000000	63.000000	1.0	10.000000	20000.000000	1.000000	11.000000
50%	2021.0	202101.0	7.195800e+05	2.021001e+12	12480.000000	2.021007e+12	2.200270e+05	1.000000	1.000000	12792.000000	...	100.000000	0.000000	0.000000	7.000000	71.000000	1.0	10.000000	42000.000000	1.000000	11.000000
75%	2021.0	202101.0	1.090470e+06	2.021001e+12	19968.000000	2.021011e+12	4.103360e+05	1.000000	2.000000	20280.000000	...	359.000000	0.000000	0.000000	10.000000	101.000000	1.0	10.000000	75000.000000	1.000000	11.000000
max	2021.0	202101.0	1.440846e+06	2.021010e+12	175968.000000	2.021014e+12	5.930851e+06	4.000000	9.000000	175812.000000	...	990.000000	4.000000	498.000000	11.000000	116.000000	1.0	15.000000	682000.000000	2.000000	20.000000

8 rows × 26 columns

3.2. For our analysis, we'll need to use the codebook we saved to clean and create a few variables:

a) Education

In [65]:
```
# Create a continuous education variable
crosswalk = pd.read_csv('PPHA_30545_MP01-Crosswalk')
crosswalk = crosswalk.set_index('educd').T

acs_data['EDUCDC'] = acs_data['EDUCD']
acs_data = acs_data.replace({'EDUCDC': crosswalk})
```

b) Dummy variables

In [66]:
```
# i. High school diploma
acs_data['hsdip'] = ((acs_data['EDUCDC'] >= 12) & (acs_data['EDUCDC'] < 16)).astype(int)
# ii. College degree
acs_data['coldip'] = (acs_data['EDUCDC'] >= 16).astype(int)
# iii. white
acs_data['White'] = np.where(acs_data['RACE'] == 1, 1, 0)
# iv. black
acs_data['Black'] = np.where(acs_data['RACE'] == 2, 1, 0)
# v. hispanic
acs_data['hispanic'] = ((acs_data['HISPAN'] != 0) & (acs_data['HISPAN'] != 9)).astype(int)
# vi. married
acs_data['married'] = ((acs_data['MARST'] == 1) | (acs_data['MARST'] == 2)).astype(int)
# vii. female
acs_data['female'] = (acs_data['SEX'] == 2).astype(int)
# viii. veteran
acs_data['VET'] = np.where(acs_data['VETSTAT'] == 2, 1, 0)
```

c) Interaction terms

In [67]:
```
for var in ['hsdip', 'coldip']:
    acs_data[var + '_inter_educdc'] = acs_data[var]*acs_data['EDUCDC']
```

d) Create the following

In [68]:
```
# Drop observations with zero income wage
incwage_zero_index = acs_data[acs_data['INCWAGE'] == 0].index
acs_data.drop(incwage_zero_index, inplace=True)
```

In [69]:
```
# i. Age squared
acs_data['AGE_SQ'] = np.power(acs_data['AGE'], 2)
# ii. log of income
acs_data['INCWAGE_log'] = np.log(acs_data['INCWAGE'])
```

4. Data Analysis

1. Compute descriptive statistics

In [70]:
```
acs_data[["YEAR", "INCWAGE", "INCWAGE_log", 'EDUCDC', 'female', 'AGE', 'AGE_SQ',
          'White', 'Black', 'hispanic', 'married', 'NCHILD', 'VET', 'hsdip', 'coldip']].describe()
```

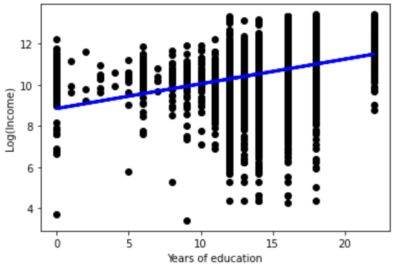
Out[70]:

	YEAR	INCWAGE	INCWAGE_log	EDUCDC	female	AGE	AGE_SQ	White	Black	hispanic	married	NCHILD	VET	hsdip	coldip
count	8143.0	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000	8143.000000
mean	2021.0	63632.890826	10.561771	14.231610	0.481027	41.526096	1898.076753	0.663269	0.081051	0.162348	0.533833	0.823898	0.041754	0.541815	0.406607
std	0.0	75031.705812	1.133858	3.023473	0.499671	13.178825	1104.537492	0.472621	0.272931	0.368792	0.498885	1.151690	0.200038	0.498279	0.491230
min	2021.0	30.000000	3.401197	0.000000	0.000000	18.000000	324.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2021.0	24000.000000	10.085809	12.000000	0.000000	31.000000	961.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2021.0	45000.000000	10.714418	14.000000	0.000000	42.000000	1764.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	0.000000
75%	2021.0	76000.000000	11.238489	16.000000	1.000000	53.000000	2809.000000	1.000000	0.000000	0.000000	1.000000	2.000000	0.000000	1.000000	1.000000
max	2021.0	682000.000000	13.432785	22.000000	1.000000	65.000000	4225.000000	1.000000	1.000000	1.000000	1.000000	9.000000	1.000000	1.000000	1.000000

2. Scatter plot

In [71]:
```
lm_simple = lm().fit(acs_data[['EDUCDC']], acs_data[['INCWAGE_log']])
simple_y_pred = lm_simple.predict(acs_data[['EDUCDC']])
plt.scatter(acs_data[['EDUCDC']], acs_data[['INCWAGE_log']], color="black")
plt.plot(acs_data[['EDUCDC']], simple_y_pred, color="blue", linewidth=3)
plt.xlabel('Years of education')
plt.ylabel('Log(Income)')

plt.show()
```



3. Estimate the model

In [72]:
```
result = smf.ols('INCWAGE_log ~ EDUCDC + female + AGE + AGE_SQ + White + Black + hispanic + married + NCHILD + VET', data = acs_data)
print(result.fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.283
Model:                            OLS   Adj. R-squared:                  0.282
Method:                 Least Squares   F-statistic:                     321.1
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:46   Log-Likelihood:                -11222.
No. Observations:                8143   AIC:                         2.247e+04
Df Residuals:                    8132   BIC:                         2.254e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.6989      0.126     45.295      0.000       5.452       5.946
EDUCDC         0.1043      0.004     28.120      0.000       0.097       0.112
female        -0.4020      0.022    -18.563      0.000      -0.444      -0.360
AGE            0.1603      0.006     26.028      0.000       0.148       0.172
AGE_SQ        -0.0017   7.28e-05    -23.211      0.000      -0.002      -0.002
White          0.0604      0.030      2.007      0.045       0.001       0.119
Black         -0.2162      0.047     -4.610      0.000      -0.308      -0.124
hispanic      -0.0073      0.036     -0.202      0.840      -0.078       0.064
married        0.1894      0.025      7.562      0.000       0.140       0.239
NCHILD        -0.0022      0.011     -0.206      0.837      -0.023       0.019
VET            0.0687      0.054      1.267      0.205      -0.038       0.175
==============================================================================
Omnibus:                     2586.782   Durbin-Watson:                   1.864
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11798.652
Skew:                          -1.483   Prob(JB):                         0.00
Kurtosis:                       8.096   Cond. No.                     2.62e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.62e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

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

Answer: *The value of the R squared: 0.264*

(b) Test the hypothesis that [...]:

Answer: *This hypothesis is being tested in the default summary under the F-statistic and Prob(F-Statistic). In this case, the p-value is zero. Therefore, we can reject the null at the 90, 95 and 99% of confidence.*

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

Answer *The coefficient of years of education is 0.0903. Since the dependent variable is in logs, an additional year of education is associated with an increase of about 9.45% (=* $e^{0.0903} - 1$*) in income.*

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

Answer: *Let's take the derivative of Age*

```
d(ols)/d(AGE) = 0.1571 + 2 * - 0.0016 * AGE
d(ols)/d(AGE) = 0.1571 - 0.0032 * AGE
```

Since we know that our function is concave, our max will be located whenever the derivative is equal to zero. In other words:

```
0 = 0.1571 - 0.0032 * AGE
0.0032 * AGE = 0.1571
AGE = 0.1571/0.0032 = 49.09
```

Another way is the brute-force way:

In [73]:
```
highest_income = 0
for current_age in range(100):
    # Current income for age = current_age
    current_income = 0.1571 - 0.0032*current_age
    if highest_income > current_income:
        print("Age with highest income:", current_age - 1)
        break
```

```
Age with highest income: 49
```

Doesn't work with decimals but it's not so bad

(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 female coefficient is negative. This suggests women earn about 30% less than men

All else in the model equal, women earn 70.5% of what men earn since 100($e^{-0.3496}$*-1) is roughly -29.5%. There are many factors left out of the model such as occupational choice, preference over leisure, and willingness to negotiate compensation. However, the model's result of women earning less than men with all other attributes of the model being equal is consistent with studies that control for much more and still find women earning less than men albeit to a lesser degree.*

(f) Interpret the coefficients on the white and black, and its significance

First, it's important to establish the baseline group for comparison. The baseline group consists of people who either did not check any of the boxes or do not identify as white, black, or Hispanic. So compared to this baseline group, a person from a particular demographic J earns 100($e^{\beta J}$*-1) of what a baseline group member would earn with all else in the model equal. So all else in the model equal, a white person earns 103.33% (*$e^{0.0328}$*-1=3.33%) of what a baseline group member earns. For a black person, it's 82.73% 100(e* $^{-0.1896}$ *-1=-17.27%).*

Only the estimate for black is statistically significant, associated p-value is 0.00

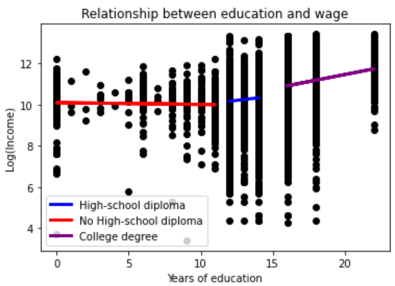
4. Graph

In [74]:
```
def get_reg_line(data, line_color, label):
    '''
    Get scatter plot with fitted OLS regression line
    Input: data- data frame
    line_color(str)- string for line color
    label(str)- line label name
    Output: Plot
    '''
    y = data[['INCWAGE_log']].values
    X = data[['EDUCDC' ]].values
    y_pred = lm().fit(X, y).predict(X)

    return plt.plot(X, y_pred, color= line_color, linewidth=3, label = label)



plt.scatter(acs_data[['EDUCDC']], acs_data[["INCWAGE_log"]], color="black")
get_reg_line(acs_data[acs_data['hsdip'] == 1], "blue", "High-school diploma")
get_reg_line(acs_data[acs_data['EDUCDC'] < 12], "red", "No High-school diploma")
get_reg_line(acs_data[acs_data['coldip'] == 1], "purple", "College degree")
plt.xlabel('Years of education')
plt.title('Relationship between education and wage')
plt.legend()
plt.ylabel('Log(Income)')
plt.show()
```



5.

Answer: *There are many ways to modify the model. One such way is to allow (i) different intercepts for the three groups (no degree, high school degree, college degree) and (ii) different slopes for the three groups. That is,*

ln(incwage) = β0+$\gamma_1$hsdip+$\gamma_2$coldip+$\gamma_3$hsdip·educdc+$\gamma_4$coldip·educdc+...

where hsdip and coldip are indicator functions for whether an individual is a high school graduate or a college graduate. In the ellipsis are controls from the original model as well as the error term

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

In [75]:
```python
result = smf.ols('INCWAGE_log ~ hsdip + coldip + EDUCDC + hsdip_inter_educdc + coldip_inter_educdc + female + AGE + AGE_SQ + White + Black + hispanic + married + NCHILD + VET', data = acs_data)
print(result.fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.302
Model:                            OLS   Adj. R-squared:                  0.301
Method:                 Least Squares   F-statistic:                     250.9
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:46   Log-Likelihood:                -11115.
No. Observations:                8143   AIC:                         2.226e+04
Df Residuals:                    8128   BIC:                         2.236e+04
Df Model:                          14
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               6.7842      0.149     45.658      0.000       6.493       7.075
hsdip                  -0.6967      0.216     -3.233      0.001      -1.119      -0.274
coldip                 -0.7336      0.218     -3.359      0.001      -1.162      -0.306
EDUCDC                  0.0128      0.011      1.180      0.238      -0.008       0.034
hsdip_inter_educdc      0.0693      0.019      3.701      0.000       0.033       0.106
coldip_inter_educdc     0.0907      0.016      5.646      0.000       0.059       0.122
female                 -0.4047      0.021    -18.881      0.000      -0.447      -0.363
AGE                     0.1488      0.006     24.274      0.000       0.137       0.161
AGE_SQ                 -0.0016   7.25e-05    -21.524      0.000      -0.002      -0.001
White                   0.0899      0.030      3.013      0.003       0.031       0.148
Black                  -0.1604      0.046     -3.453      0.001      -0.252      -0.069
hispanic               -0.0091      0.036     -0.256      0.798      -0.079       0.061
married                 0.1680      0.025      6.780      0.000       0.119       0.217
NCHILD                 -0.0026      0.010     -0.249      0.803      -0.023       0.018
VET                     0.0996      0.054      1.858      0.063      -0.005       0.205
==============================================================================
Omnibus:                     2804.569   Durbin-Watson:                   1.880
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13758.157
Skew:                          -1.595   Prob(JB):                         0.00
Kurtosis:                       8.511   Cond. No.                     5.06e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.06e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(a) What fraction of the variation in log wages does the model explain? How does this compare to the model you estimated in question 3?

The variation in the log of wages explained by the model is $R^2 = 0.286$*, this is greater than the* $R^2 = 0.264$ *of the model estimated in question 3.*

(b) 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.

The predicted wages for an with individual with that characteristics and a HS degree are approximately $10,911.87, while the predicted wages for an individual with a college degree are approximately$*18,924.54*

In [76]:
```
# High School degree
dict_hs = {'female': [1], 'AGE' : [22] ,'AGE_SQ' : [484], 'hsdip': [1], 'coldip':[0],
 'Black': [0], 'hispanic':[0], 'NCHILD': [0], 'married': [0], 'VET':[0], 'EDUCDC':[12],
 'hsdip_inter_educdc': [12], 'coldip_inter_educdc': [0], 'White': [0]}
df_pred = pd.DataFrame(data=dict_hs)
prediction1 = result.fit().get_prediction(df_pred)
prediction1 = prediction1.summary_frame(alpha=0.05)
prediction1_wage = np.exp(prediction1['mean'].values[0])
print(f'Wage with HS degree: {prediction1_wage}')

# College degree
dict_cd = {'female': [1], 'AGE' : [22] ,'AGE_SQ' : [484], 'hsdip': [0], 'coldip':[1],
 'Black': [0], 'hispanic':[0], 'NCHILD': [0], 'married': [0], 'VET':[0], 'EDUCDC':[16],
 'hsdip_inter_educdc': [0], 'coldip_inter_educdc': [16], 'White': [0]}
df_pred = pd.DataFrame(data=dict_cd)
prediction2 = result.fit().get_prediction(df_pred)
prediction2 = prediction2.summary_frame(alpha=0.05)
prediction2_wage = np.exp(prediction2['mean'].values[0])
print(f'Wage with college degree: {prediction2_wage}')
```

```
Wage with HS degree: 9772.042974449947
Wage with college degree: 18411.02083136736
```

(c) The President is concerned that citizens will be harmed (and voters unhappy) if the predictions from your model turn out to be wrong. She wants to know how confident you are in your predictions. Briefly explain.

Open-ended question. Full credit if explanation is based on the strengths or weaknesses of the model or the estimation results.

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 (without any other stipulations)? Try fitting some different models and report the results of the model that best predicts log wages that you can come up with. Use adjusted R2 as your measure of the model that produces the best prediction.

In [77]:
```
# Examining all columns
acs_data.columns
```

Out[77]:
```
Index(['YEAR', 'SAMPLE', 'SERIAL', 'CBSERIAL', 'HHWT', 'CLUSTER', 'STRATA',
       'GQ', 'PERNUM', 'PERWT', 'NCHILD', 'NCHLT5', 'SEX', 'AGE', 'MARST',
       'RACE', 'RACED', 'HISPAN', 'HISPAND', 'EDUC', 'EDUCD', 'EMPSTAT',
       'EMPSTATD', 'INCWAGE', 'VETSTAT', 'VETSTATD', 'EDUCDC', 'hsdip',
       'coldip', 'White', 'Black', 'hispanic', 'married', 'female', 'VET',
       'hsdip_inter_educdc', 'coldip_inter_educdc', 'AGE_SQ', 'INCWAGE_log'],
      dtype='object')
```

In [78]:
```
# Model 1
model1 = smf.ols('INCWAGE_log ~ hsdip + coldip', data=acs_data).fit()
print("Adjusted R2 for model: ", model1.rsquared_adj)
print(model1.summary())
```

```
Adjusted R2 for model:  0.12647115702715728
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.127
Model:                            OLS   Adj. R-squared:                  0.126
Method:                 Least Squares   F-statistic:                     590.4
Date:                Fri, 17 Feb 2023   Prob (F-statistic):          3.67e-240
Time:                        14:11:46   Log-Likelihood:                -12025.
No. Observations:                8143   AIC:                         2.406e+04
Df Residuals:                    8140   BIC:                         2.408e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.0427      0.052    194.213      0.000       9.941      10.144
hsdip          0.2049      0.054      3.786      0.000       0.099       0.311
coldip         1.0036      0.055     18.284      0.000       0.896       1.111
==============================================================================
Omnibus:                     2362.944   Durbin-Watson:                   1.858
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8706.325
Skew:                          -1.420   Prob(JB):                         0.00
Kurtosis:                       7.194   Cond. No.                         9.37
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [79]:
```python
# Model 2
model2 = smf.ols('INCWAGE_log ~ hsdip + coldip + female + White + Black', data=acs_data).fit()
print("Adjusted R2 for model: ", model2.rsquared_adj)
print(model2.summary())
```

```
Adjusted R2 for model:  0.16554784386963584
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.166
Model:                            OLS   Adj. R-squared:                  0.166
Method:                 Least Squares   F-statistic:                     324.1
Date:                Fri, 17 Feb 2023   Prob (F-statistic):          1.81e-317
Time:                        14:11:46   Log-Likelihood:                -11838.
No. Observations:                8143   AIC:                         2.369e+04
Df Residuals:                    8137   BIC:                         2.373e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.1527      0.053    193.196      0.000      10.050      10.256
hsdip          0.2074      0.054      3.867      0.000       0.102       0.312
coldip         1.0094      0.055     18.488      0.000       0.902       1.116
female        -0.4041      0.023    -17.520      0.000      -0.449      -0.359
White          0.1409      0.027      5.194      0.000       0.088       0.194
Black         -0.1579      0.047     -3.394      0.001      -0.249      -0.067
==============================================================================
Omnibus:                     2515.159   Durbin-Watson:                   1.853
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10041.436
Skew:                          -1.487   Prob(JB):                         0.00
Kurtosis:                       7.555   Cond. No.                         11.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [80]:
```python
# Model 3
model3 = smf.ols('INCWAGE_log ~ hsdip + coldip + female + White + Black + EDUCDC', data=acs_data).fit()
print("Adjusted R2 for model: ", model3.rsquared_adj)
print(model3.summary())
```

```
Adjusted R2 for model:  0.1731113322827299
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.174
Model:                            OLS   Adj. R-squared:                  0.173
Method:                 Least Squares   F-statistic:                     285.1
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:46   Log-Likelihood:                -11800.
No. Observations:                8143   AIC:                         2.361e+04
Df Residuals:                    8136   BIC:                         2.366e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      9.7228      0.072    135.004      0.000       9.582       9.864
hsdip         -0.2130      0.072     -2.956      0.003      -0.354      -0.072
coldip         0.3243      0.096      3.385      0.001       0.137       0.512
female        -0.4108      0.023    -17.883      0.000      -0.456      -0.366
White          0.1319      0.027      4.880      0.000       0.079       0.185
Black         -0.1635      0.046     -3.529      0.000      -0.254      -0.073
EDUCDC         0.0665      0.008      8.685      0.000       0.051       0.081
==============================================================================
Omnibus:                     2559.933   Durbin-Watson:                   1.853
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10387.529
Skew:                          -1.509   Prob(JB):                         0.00
Kurtosis:                       7.637   Cond. No.                         152.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [81]:
```python
# Model 4
model4 = smf.ols('INCWAGE_log ~ hsdip + coldip + female + White + Black + AGE_SQ ', data=acs_data).fit()
print("Adjusted R2 for model: ", model4.rsquared_adj)
print(model4.summary())
```

```
Adjusted R2 for model:  0.2135158001352515
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.214
Model:                            OLS   Adj. R-squared:                  0.214
Method:                 Least Squares   F-statistic:                     369.4
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:46   Log-Likelihood:                -11596.
No. Observations:                8143   AIC:                         2.321e+04
Df Residuals:                    8136   BIC:                         2.326e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      9.7418      0.054    179.592      0.000       9.635       9.848
hsdip          0.2497      0.052      4.793      0.000       0.148       0.352
coldip         1.0359      0.053     19.539      0.000       0.932       1.140
female        -0.4142      0.022    -18.495      0.000      -0.458      -0.370
White          0.0743      0.027      2.805      0.005       0.022       0.126
Black         -0.2064      0.045     -4.565      0.000      -0.295      -0.118
AGE_SQ         0.0002   1.02e-05     22.300      0.000       0.000       0.000
==============================================================================
Omnibus:                     2704.126   Durbin-Watson:                   1.850
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12140.361
Skew:                          -1.564   Prob(JB):                         0.00
Kurtosis:                       8.099   Cond. No.                     1.72e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [82]:
```python
# Model 5
model5 = smf.ols('INCWAGE_log ~ hsdip + coldip + female + White + Black + AGE_SQ + married ', data=acs_data).fit()
print("Adjusted R2 for model: ", model5.rsquared_adj)
print(model5.summary())
```

```
Adjusted R2 for model:  0.23245534613223184
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.233
Model:                            OLS   Adj. R-squared:                  0.232
Method:                 Least Squares   F-statistic:                     353.3
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:46   Log-Likelihood:                -11496.
No. Observations:                8143   AIC:                         2.301e+04
Df Residuals:                    8135   BIC:                         2.306e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      9.6635      0.054    179.386      0.000       9.558       9.769
hsdip          0.2543      0.051      4.941      0.000       0.153       0.355
coldip         0.9966      0.052     19.000      0.000       0.894       1.099
female        -0.3933      0.022    -17.737      0.000      -0.437      -0.350
White          0.0567      0.026      2.165      0.030       0.005       0.108
Black         -0.1431      0.045     -3.187      0.001      -0.231      -0.055
AGE_SQ         0.0002   1.06e-05     16.739      0.000       0.000       0.000
married        0.3394      0.024     14.204      0.000       0.293       0.386
==============================================================================
Omnibus:                     2715.608   Durbin-Watson:                   1.858
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12360.181
Skew:                          -1.566   Prob(JB):                         0.00
Kurtosis:                       8.159   Cond. No.                     1.72e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [83]:
```
# Model 6 with interaction terms
acs_data['white_married'] = acs_data['White'] * acs_data['married']
acs_data['black_married'] = acs_data['Black'] * acs_data['married']
acs_data['hispanic_married'] = acs_data['hispanic'] * acs_data['married']

model6 = smf.ols('INCWAGE_log ~ hsdip + coldip + female + White + Black + white_married + black_married + hispanic_married', data=acs_data).fit()
print("Adjusted R2 for model: ", model6.rsquared_adj)
print(model6.summary())
```

```
Adjusted R2 for model:  0.19646808604798882
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.197
Model:                            OLS   Adj. R-squared:                  0.196
Method:                 Least Squares   F-statistic:                     249.8
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:47   Log-Likelihood:                -11682.
No. Observations:                8143   AIC:                         2.338e+04
Df Residuals:                    8134   BIC:                         2.345e+04
Df Model:                           8
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept           10.0984      0.054    186.177      0.000       9.992      10.205
hsdip                0.2383      0.053      4.498      0.000       0.134       0.342
coldip               0.9970      0.054     18.423      0.000       0.891       1.103
female              -0.3845      0.023    -16.965      0.000      -0.429      -0.340
White               -0.1040      0.033     -3.168      0.002      -0.168      -0.040
Black               -0.2455      0.054     -4.567      0.000      -0.351      -0.140
white_married        0.4741      0.028     16.771      0.000       0.419       0.530
black_married        0.3707      0.086      4.299      0.000       0.202       0.540
hispanic_married     0.1470      0.047      3.097      0.002       0.054       0.240
==============================================================================
Omnibus:                     2533.911   Durbin-Watson:                   1.866
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10516.468
Skew:                          -1.484   Prob(JB):                         0.00
Kurtosis:                       7.710   Cond. No.                         12.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [84]:
```
# Model 7 with interaction terms
model7 = smf.ols('INCWAGE_log ~ hsdip + coldip + female + White + Black + white_married + black_married + hispanic_married + AGE', data=acs_data).fit()
print("Adjusted R2 for model: ", model7.rsquared_adj)
print(model7.summary())
```

```
Adjusted R2 for model:  0.24176574026011655
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.243
Model:                            OLS   Adj. R-squared:                  0.242
Method:                 Least Squares   F-statistic:                     289.5
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:47   Log-Likelihood:                -11446.
No. Observations:                8143   AIC:                         2.291e+04
Df Residuals:                    8133   BIC:                         2.298e+04
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            9.3428      0.063    148.678      0.000       9.220       9.466
hsdip                0.2704      0.051      5.253      0.000       0.170       0.371
coldip               1.0142      0.053     19.290      0.000       0.911       1.117
female              -0.4021      0.022    -18.253      0.000      -0.445      -0.359
White               -0.0840      0.032     -2.632      0.009      -0.147      -0.021
Black               -0.2551      0.052     -4.885      0.000      -0.357      -0.153
white_married        0.3013      0.029     10.551      0.000       0.245       0.357
black_married        0.1976      0.084      2.349      0.019       0.033       0.362
hispanic_married     0.0641      0.046      1.386      0.166      -0.027       0.155
AGE                  0.0194      0.001     22.067      0.000       0.018       0.021
==============================================================================
Omnibus:                     2750.393   Durbin-Watson:                   1.862
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12808.391
Skew:                          -1.580   Prob(JB):                         0.00
Kurtosis:                       8.269   Cond. No.                         363.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [85]:
```
# Model 8 with interaction terms
model8 = smf.ols('INCWAGE_log ~ hsdip + coldip + female + White + Black + white_married + AGE_SQ', data=acs_data).fit()
print("Adjusted R2 for model: ", model8.rsquared_adj)
print(model8.summary())
```

```
Adjusted R2 for model:  0.2271882053650799
                            OLS Regression Results
==============================================================================
Dep. Variable:            INCWAGE_log   R-squared:                       0.228
Model:                            OLS   Adj. R-squared:                  0.227
Method:                 Least Squares   F-statistic:                     342.9
Date:                Fri, 17 Feb 2023   Prob (F-statistic):               0.00
Time:                        14:11:47   Log-Likelihood:                -11524.
No. Observations:                8143   AIC:                         2.306e+04
Df Residuals:                    8135   BIC:                         2.312e+04
Df Model:                           7
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept         9.7983      0.054    181.534      0.000       9.693       9.904
hsdip             0.2506      0.052      4.854      0.000       0.149       0.352
coldip            1.0089      0.053     19.179      0.000       0.906       1.112
female           -0.4014      0.022    -18.062      0.000      -0.445      -0.358
White            -0.1151      0.031     -3.759      0.000      -0.175      -0.055
Black            -0.2043      0.045     -4.557      0.000      -0.292      -0.116
white_married     0.3442      0.029     12.039      0.000       0.288       0.400
AGE_SQ            0.0002   1.04e-05     18.755      0.000       0.000       0.000
==============================================================================
Omnibus:                     2718.814   Durbin-Watson:                   1.861
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12379.305
Skew:                          -1.568   Prob(JB):                         0.00
Kurtosis:                       8.162   Cond. No.                     1.72e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Inferences:

1. In our trails, the maximum Adjusted R2 = '0.022621371990750205' is acheived for Model 5 regressing 'INCWAGE_log' with 'hsdip + coldip + female + White + Black + AGE_SQ + married'.
2. Education level, Gender, Race, Age and Marital status are key determinants for wage levels. Removing these terms in Model 9 resulted in decrease of Adjusted R2 to '0.002517230323518249'.
3. We can experiement with more models.

In [ ]: