Problem Set - 2

In [3]:
```
# Import the packages and libraries

import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf
import seaborn as sns
```

In [4]:
```
#Loading dataset using read_csv

pd.set_option('display.max_columns', None)


data = pd.read_csv("E:\\Winter'23\\ML\\PS2\\usa_00001.csv")
```

Question 3.2.a

In [5]:
```
crosswalk = pd.read_csv("E:\\Winter'23\\ML\\PS2\\PPHA_30545_MP01-Crosswalk.csv")
crosswalk = crosswalk.set_index('educd').T.to_dict('list')


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


crosswalk = pd.read_csv("E:\\Winter'23\\ML\\PS2\\PPHA_30545_MP01-Crosswalk.csv")
data = data.merge(crosswalk, left_on='EDUCD', right_on='educd')
```

Question 3.2.b

In [7]:
```
# Question 3.2.b.i
#hsdip dummy
#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
data['educdc'].unique()
data['hsdip'] = np.where((data['educdc'] >= 12)
                         & (data['educdc'] < 16),1,0)
```

Out[7]:
```
array([14., 12., 10., 18., 16.,  6.,  9., 13.,  8., 22., 11.,  0.,  5.,
        7.,  3.,  2.,  4.,  1.])
```

In [8]:
```
# Question 3.2.b.ii
#coldip dummy
data['coldip'] = np.where((data['educdc'] >= 16),1,0)
```

In [9]:
```
# Question 3.2.b.iii
# Whites are 1, Blacks are 2
data['white'] = np.where(data['RACE'] == 1, 1, 0)
```

In [10]:

```python
# Question 3.2.b.iv
data['black'] = np.where(data['RACE'] == 2, 1, 0)
```

In [11]:

```python
# Question 3.2.b.v
#hispanic dummy
data['HISPAN'].unique()
data['hispanic'] = np.where(data['HISPAN'] != 0, 1, 0)
```

In [12]:

```python
# Question 3.2.b.vi
#married dummy
data['married'] = np.where((data['MARST'] == 1)
                           | (data['MARST'] == 2),1,0)
```

In [13]:

```python
# Question 3.2.b.vii
#female dummy
data['SEX'].unique()
data['female'] = np.where(data['SEX'] == 2, 1, 0)
```

In [14]:

```python
# Question 3.2.b.viii
#veteran dummy
data['VETSTAT'].unique()
data['vet'] = np.where(data['VETSTAT'] == 2, 1, 0)
```

In [15]:

```python
# Question 3.2.c
#hsdip & educdc
data['hsdipeducdc'] = data['hsdip']*data['educdc']

#coldip & educdc
data['coldipeducdc'] = data['coldip']*data['educdc']
```

In [16]:

```python
# Question 3.2.d.i

data['agesq'] = np.power(data['AGE'], 2)
data.head()
```

Out[16]:

	YEAR	SAMPLE	SERIAL	CBSERIAL	HHWT	CLUSTER	STRATA	GQ	PERNUM	PERWT
0	2021	202101	1902	2021010114983	5304.0	2021000019021	160001	4	1	5304.C
1	2021	202101	3930	2021000087465	19188.0	2021000039301	160001	1	3	21528.C
2	2021	202101	5022	2021000164616	46644.0	2021000050221	70001	1	1	46800.C
3	2021	202101	11574	2021000611655	8892.0	2021000115741	240001	1	1	8892.C
4	2021	202101	12822	2021000695047	44460.0	2021000128221	20001	1	1	44460.C

In [17]:

```python
# Question 3.2.d.ii
data = data[data.INCWAGE != 0]
data['lnincwage'] = np.log(data['INCWAGE'])
```

Question 4 Data Analysis Questions

In [96]:

```
#Question 4.1) Compute descriptive (summary) statistics for year, incwage, lnincwage
# educdc, f emale, age, age2 , white, black, hispanic, married, nchild, vet,
# hsdip, coldip, and
#the interaction terms

data.describe()


#For loop
#column_list = ['year', 'incwage', 'lnincwage','educdc', 'female', 'age', 'age2' , '
#               , 'black', 'hispanic', 'married', 'nchild', 'vet', 'hsdipeducdc', 'c

#for i in column_list:
#   data[i].describe()
```

Out[96]:

	YEAR	SAMPLE	SERIAL	CBSERIAL	HHWT	CLUSTER	STRATA	
count	8143.0	8143.0	8.143000e+03	8.143000e+03	8143.000000	8.143000e+03	8.143000e+03	81
mean	2021.0	202101.0	7.204368e+05	2.021001e+12	16265.188751	2.021007e+12	4.729771e+05	
std	0.0	0.0	4.201259e+05	1.419170e+06	13558.811169	4.201259e+06	9.486353e+05	
min	2021.0	202101.0	1.902000e+03	2.021000e+12	312.000000	2.021000e+12	1.000100e+04	
25%	2021.0	202101.0	3.515760e+05	2.021000e+12	8112.000000	2.021004e+12	9.001800e+04	
50%	2021.0	202101.0	7.190340e+05	2.021001e+12	12480.000000	2.021007e+12	2.200420e+05	
75%	2021.0	202101.0	1.088910e+06	2.021001e+12	19812.000000	2.021011e+12	4.104375e+05	
max	2021.0	202101.0	1.440846e+06	2.021010e+12	175968.000000	2.021014e+12	5.930851e+06	

In [21]:

```
data['YEAR'].describe()
```

Out[21]:

```
count    8143.0
mean     2021.0
std         0.0
min      2021.0
25%      2021.0
50%      2021.0
75%      2021.0
max      2021.0
Name: YEAR, dtype: float64
```

In [22]:

```
data['INCWAGE'].describe()
```

Out[22]:

```
count      8143.000000
mean      63632.890826
std       75031.705812
min          30.000000
25%       24000.000000
50%       45000.000000
75%       76000.000000
max      682000.000000
Name: INCWAGE, dtype: float64
```

In [23]:
```
data['lnincwage'].describe()
```
Out[23]:
```
count    8143.000000
mean       10.561771
std         1.133858
min         3.401197
25%        10.085809
50%        10.714418
75%        11.238489
max        13.432785
Name: lnincwage, dtype: float64
```
In [24]:
```
data['educdc'].describe()
```
Out[24]:
```
count    8143.000000
mean       14.231610
std         3.023473
min         0.000000
25%        12.000000
50%        14.000000
75%        16.000000
max        22.000000
Name: educdc, dtype: float64
```
In [25]:
```
data['female'].describe()
```
Out[25]:
```
count    8143.000000
mean        0.481027
std         0.499671
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         1.000000
Name: female, dtype: float64
```
In [26]:
```
data['AGE'].describe()
```
Out[26]:
```
count    8143.000000
mean       41.526096
std        13.178825
min        18.000000
25%        31.000000
50%        42.000000
75%        53.000000
max        65.000000
Name: AGE, dtype: float64
```
In [27]:
```
data['agesq'].describe()
```
Out[27]:
```
count    8143.000000
mean     1898.076753
std      1104.537492
min       324.000000
25%       961.000000
50%      1764.000000
75%      2809.000000
max      4225.000000
Name: agesq, dtype: float64
```

In [28]:
```
data['white'].describe()
```
Out[28]:
```
count    8143.000000
mean        0.663269
std         0.472621
min         0.000000
25%         0.000000
50%         1.000000
75%         1.000000
max         1.000000
Name: white, dtype: float64
```
In [29]:
```
data['black'].describe()
```
Out[29]:
```
count    8143.000000
mean        0.081051
std         0.272931
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: black, dtype: float64
```
In [30]:
```
data['hispanic'].describe()
```
Out[30]:
```
count    8143.000000
mean        0.162348
std         0.368792
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: hispanic, dtype: float64
```
In [31]:
```
data['married'].describe()
```
Out[31]:
```
count    8143.000000
mean        0.533833
std         0.498885
min         0.000000
25%         0.000000
50%         1.000000
75%         1.000000
max         1.000000
Name: married, dtype: float64
```
In [32]:
```
data['NCHILD'].describe()
```
Out[32]:
```
count    8143.000000
mean        0.823898
std         1.151690
min         0.000000
25%         0.000000
50%         0.000000
75%         2.000000
max         9.000000
Name: NCHILD, dtype: float64
```

In [33]:

```
data['vet'].describe()
```

Out[33]:

```
count    8143.000000
mean        0.041754
std         0.200038
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max         1.000000
Name: vet, dtype: float64
```

In [34]:

```
data['hsdip'].describe()
```

Out[34]:

```
count    8143.000000
mean        0.541815
std         0.498279
min         0.000000
25%         0.000000
50%         1.000000
75%         1.000000
max         1.000000
Name: hsdip, dtype: float64
```

In [35]:

```
data['coldip'].describe()
```

Out[35]:

```
count    8143.000000
mean        0.406607
std         0.491230
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         1.000000
Name: coldip, dtype: float64
```

In [36]:

```
data['hsdipeducdc'].describe()
```

Out[36]:

```
count    8143.000000
mean        7.009333
std         6.483140
min         0.000000
25%         0.000000
50%        12.000000
75%        13.000000
max        14.000000
Name: hsdipeducdc, dtype: float64
```

In [37]:

```
data['coldipeducdc'].describe()
```

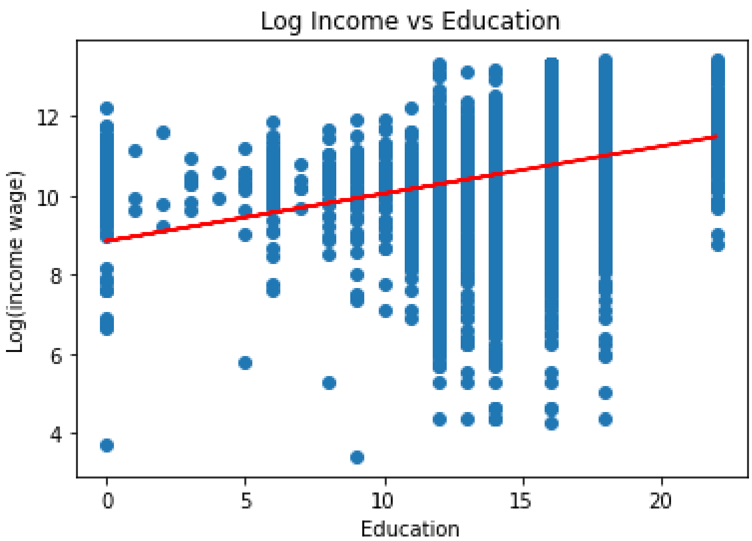
Out[37]:

```
count    8143.000000
mean        6.883949
std         8.364936
min         0.000000
25%         0.000000
50%         0.000000
75%        16.000000
max        22.000000
Name: coldipeducdc, dtype: float64
```

Question 4.2) Scatter plot ln(incwage) and education. Include a linear fit line. Be sure to label all axes and include an informative title

In [38]:

```
x = data['educdc']
y = data['lnincwage']
plt.scatter(x, y)

z = np.polyfit(x,y,1)
p = np.poly1d(z)
plt.plot(x, p(x), "red")
#plt.plot(x, z[0] * np.array(x) + z[1], color='red')
plt.xlabel('Education')
plt.ylabel('Log(income wage)')
plt.title('Log Income vs Education')
plt.show()
```



Question 4.3

In [40]:

```
model = smf.ols('lnincwage ~ educdc + female + AGE + agesq +  white + black + hispan
                + married + NCHILD + vet', data = data)
reg1 = model.fit()

print(reg1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              lnincwage   R-squared:                       0.283
Model:                            OLS   Adj. R-squared:                  0.282
Method:                 Least Squares   F-statistic:                     321.1
Date:                Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                        20:36:31   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
agesq         -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:                   1.872
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 spe
cified.
[2] The condition number is large, 2.62e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Question 4.3.a

In [45]:

```python
print("The fraction of the variation in log wages that the model explains is 0.283 o
      R-squared is a statistical measure that represents the proportion of the varia
      dependent variable that's explained by an independent variable or variables in
      It ranges from 0 to 1, where a higher value indicates a better fit of the mode
      This value is represented by the R-squared value in the summary output.")
```

```
The fraction of the variation in log wages that the model explains is 0.283 or 28.
3%. R-squared is a statistical measure that represents the proportion of the varianc
e for a dependent variable that's explained by an independent variable or variables
in a regression model. It ranges from 0 to 1, where a higher value indicates a bette
r fit of the model.This value is represented by the R-squared value in the summary o
utput.
```

In [ ]:

```python
#Question 4.3.b
```

In [46]:

```python
# Question 4.3.b

# What is the return to an additional increase in "educd"  by 1 ?

print("The return to an additional increase in 'educd'  by 1 is 0.1043 in 'log incom
      which is equivalent to a change of 10.43% in income wages.")


#Is this statistically significant?

print('Since the p-value of 0.00, which is significantly less than 0.05, thus educdc


#Is it practically significant? Briefly explain?
print('It maybe not be practically significant if we consider further characteristic
      in the analysis. For example for a change of 1 year of education at a lower le
      i.e from educdc 5 to 6, a 10% change in income wage maynot be practical. Howev
      practical at a higher level of educdc such as a change in educdc from 15 to 16
      i.e no college degree to college degree, a 10% increase would be practical')
```

```
The return to an additional increase in 'educd'  by 1 is 0.1043 in 'log income wage
s' which is equivalent to a change of 10.43% in income wages.
Since the p-value of 0.00, which is significantly less than 0.05, thus educdc is sta
tistiacally significant
It maybe not be practically significant if we consider further characteristics that
are involved in the analysis. For example for a change of 1 year of education at a l
ower level of educdc i.e from educdc 5 to 6, a 10% change in income wage maynot be p
ractical. However, it maybe practical at a higher level of educdc such as a change i
n educdc from 15 to 16, i.e no college degree to college degree, a 10% increase woul
d be practical
```

Question 4.3.c

In [49]:

```python
print('FOC: 0 = Beta_3*Age + Beta_4*Age2')
print('Highest Age = ', -0.1603/(2*-0.0017))

print('At age 47.147 , the model predict an individual will achieve the highest wage
```

```
FOC: 0 = Beta_3*Age + Beta_4*Age2
Highest Age =  47.14705882352941
At age 47.147 , the model predict an individual will achieve the highest wage
```

Question 4.3.d

In [51]:

```python
#Does the model predict that men or women will have higher wages, all else equal? Br

print('All else equal, the model predicts that, men will have higher wages than wome
      The female coefficient is -0.4020, which means that, on average, women are pre
      to have a 40.20% lower income than men, for a given set of independent variabl
      The reason might be due to a variety of factors, including gender discriminati
      segregation, and differences in the sex that takes more time-off, work experie
```

```
All else equal, the model predicts that, men will have higher wages than women. The
female coefficient is -0.4020, which means that, on average, women are predicted to
have a 40.20% lower income than men, for a given set of independent variables. The r
eason might be due to a variety of factors, including gender discrimination, segrega
tion, and differences in the sex that takes more time-off, work experience and educa
tion.
```

Question 4.3.e

In [52]:

```python
#(e) Interpret the coefficients on the white and black variables and their significa

print('The coefficient for the white is 0.0604 and the coefficient for the black var
      These coefficients represent the relationship between race/ethnicity and ln(in
      holding all other constant. The coefficient for white suggests that, on averag
      are predicted to have a 6.04% higher income than non-whites for a given set of
      variables. The coefficient for black suggests that, on average, blacks are pre
      have a 21.62% lower income than non-black individuals for a given set of varia
      This can be an indication of the racial wage gap that is a pertinent issue.')
```

```
The coefficient for the white is 0.0604 and the coefficient for the black variable i
s -0.2162. These coefficients represent the relationship between race/ethnicity and
ln(incwage) while holding all other constant. The coefficient for white suggests tha
t, on average, whites are predicted to have a 6.04% higher income than non-whites fo
r a given set of independent variables. The coefficient for black suggests that, on
```

average, blacks are predicted to have a 21.62% lower income than non-black individua
ls for a given set of variables. This can be an indication of the racial wage gap th
at is a pertinent issue.

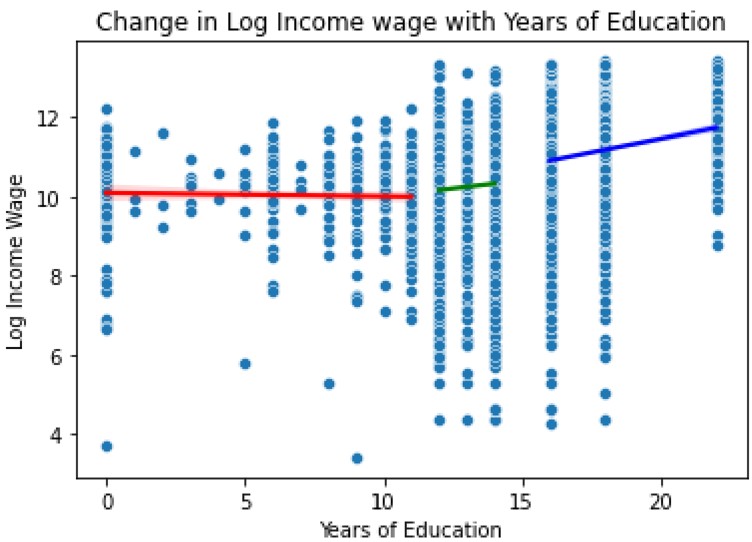
Question 4.4

In [62]:

```
#x = data['educd']
#y = data['lnincwage']
#sns.scatterplot(x=x, y=y, data=data)
#df_1 = data[data['educd'] <= 61]
#df_2 = data[(data['educd'] > 61) & (data['educd'] <= 100)]
#df_3 = data[data['educd'] > 100]
#model_1 = sns.regplot(x='educd', y='lnincwage', data=df_1, scatter=False, color='r'
#model_2 = sns.regplot(x='educd', y='lnincwage', data=df_2, scatter=False, color='g'
#model_3 = sns.regplot(x='educd', y='lnincwage', data=df_3, scatter=False, color='b'



# Using years of education variable educdc
# Assume that it takes someone 12 years to graduate high school and 16 years to grad
x = data['educdc']
y = data['lnincwage']

sns.scatterplot(x=x, y=y, data=data)

df_1 = data[data['educdc'] < 12]
df_2 = data[(data['educdc'] >= 12) & (data['educdc'] < 16)]
df_3 = data[data['educdc'] >= 16]

# fit a linear regression model for each educdc range
model_1 = sns.regplot(x='educdc', y='lnincwage', data=df_1, scatter=False, color='r'
model_2 = sns.regplot(x='educdc', y='lnincwage', data=df_2, scatter=False, color='g'
model_3 = sns.regplot(x='educdc', y='lnincwage', data=df_3, scatter=False, color='b'

plt.title('Change in Log Income wage with Years of Education')
plt.xlabel('Years of Education')
plt.ylabel('Log Income Wage')
```

Out[62]:

```
Text(0, 0.5, 'Log Income Wage')
```



Question 4.5

A tool that can be used to predict income wages for those considering a college degree is multiple linear regression, which can be used to predict the income of an individual based on education, characteristics such as age/race/sex/married etc and any other relevant variables.

Another type of model is gradiant boosting, random forest etc (not discussed in class yet). These models require large dataset of individuals to train with, and can be tested on an individual to predict income wage.

I would pick the differential intercept model due to the below reasons.

Example of differential intercept model:

$$\ln(incwage) = \beta 0 + \beta 1\,educdc + \beta 2\,female + \beta 3\,age + \beta 4\,age2 + \beta 5\,white + \beta 6\,black$$

$$+ \beta 8*hispanic + \beta 9*married + \beta 10*nchild + \beta 11*vet + \varepsilon$$

$$\ln(incwage|educdc < 12) = \alpha 0 + \beta 1\,educdc + \beta 2\,female + \beta 3\,age + \beta 4\,age2$$

$$+ \beta 5*white + \beta 6*black + \beta 8*hispanic + \beta 9*married$$

$$+ \beta 10*nchild + \beta 11*vet + \varepsilon$$

$$\ln(incwage|12 <= educdc <= 15) = \alpha 1 + \beta 1\,educdc + \beta 2\,female + \beta 3\,age + \beta 4\,age2$$

$$+ \beta 5*white + \beta 6*black + \beta 8*hispanic + \beta 9*married$$

$$+ \beta 10*nchild + \beta 11*vet + \varepsilon$$

$$\ln(incwage|educdc > 15) = \alpha 2 + \beta 1\,educdc + \beta 2\,female + \beta 3\,age + \beta 4\,age2 + \beta 5*white$$

$$+ \beta 6*black + \beta 8*hispanic + \beta 9*married + \beta 10*nchild$$

$$+ \beta 11*vet + \varepsilon$$

This model will have different intercepts for each educdc range, however a same slope for each independent variable.

Example of differential slope model:

$$\ln(incwage) = \beta 0 + \beta 1\,educdc + \beta 2\,female + \beta 3\,age + \beta 4\,age2 + \beta 5\,white + \beta 6\,black$$

$$+ \beta 8*hispanic + \beta 9*married + \beta 10*nchild + \beta 11*vet + \varepsilon$$

$$\ln(incwage|educdc < 12) = \beta 00 + \beta 01\,educdc + \beta 02\,female + \beta 03\,age + \beta 04\,age2$$

$$+ \beta 05*white + \beta 06*black + \beta 08*hispanic + \beta 09*married$$

+ β10*nchild + β11*vet + ε

ln(incwage|12 <= educdc <= 15) = β10 + β11*educdc* + *β12*female + β13*age* + *β14*age2

+ β15*white + β16*black +

β18h*ispanic + β19*married

+ β20*nchild + β21*vet + ε

ln(incwage|educdc > 15) = β20 + β21*educdc* + *β22*female + β23*age* + *β24*age2

+ β25*white + β26*black + β28*hispanic +

β29*married

+ β30*nchild + β31*vet + ε

I would pick the differential intercept model due to the aforementioned reasons. This model has different slopes for each independent variable depending on the educdc range, however the same intercept for each educdc range.

The model with greater possibility of overfitting will have a higher number of variables and interactions. However it also depends on the sample size. This is because complexity increases which leads to over dependency of unwanted noise from the data

As we saw above , differential slope model includes a greater number of variables and interactions. So, it has overfitting possibility. Due to the complexity of the model , it will also require a large data set compared to the intercept model, which may not be practical with the data set we have or adds unnecessary cost in real-world

I would pick the differential intercept model due to the aforementioned reasons.

Question 4.6

Question 4.6.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?

In [67]:

```
#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
#Reference variable or category is "no high school diploma"
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)
data['hsdip'] = data['educdc'].apply(lambda x: 1 if (x >= 12) & (x < 16) else 0)
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + white + black +
                hispanic + married + NCHILD + vet  + hsdip + col', data=data)
reg3 = model.fit()
print(reg3.summary())

print(' \n Fraction of variation observed in log wages that the model explains is 0.
      It almost the same as the variation observed in model from question 3.')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lnincwage   R-squared:                       0.299
Model:                             OLS   Adj. R-squared:                  0.298
Method:                  Least Squares   F-statistic:                     288.6
Date:                 Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                         22:07:35   Log-Likelihood:                -11132.
No. Observations:                 8143   AIC:                         2.229e+04
Df Residuals:                     8130   BIC:                         2.238e+04
Df Model:                           12
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.4359      0.136     47.344      0.000       6.169       6.702
educdc         0.0602      0.007      8.513      0.000       0.046       0.074
female        -0.4022      0.021    -18.773      0.000      -0.444      -0.360
AGE            0.1506      0.006     24.554      0.000       0.139       0.163
agesq         -0.0016   7.25e-05    -21.772      0.000      -0.002      -0.001
white          0.0782      0.030      2.621      0.009       0.020       0.137
black         -0.1709      0.047     -3.674      0.000      -0.262      -0.080
hispanic      -0.0139      0.036     -0.388      0.698      -0.084       0.056
married        0.1724      0.025      6.949      0.000       0.124       0.221
NCHILD        -0.0001      0.010     -0.011      0.991      -0.021       0.020
vet            0.1020      0.054      1.902      0.057      -0.003       0.207
hsdip         -0.0990      0.067     -1.485      0.138      -0.230       0.032
col            0.3106      0.089      3.508      0.000       0.137       0.484
==============================================================================
Omnibus:                     2762.611   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13331.164
Skew:                          -1.575   Prob(JB):                         0.00
Kurtosis:                       8.420   Cond. No.                     2.87e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The condition number is large, 2.87e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

 Fraction of variation observed in log wages that the model explains is 0.299. It al
most the same as the variation observed in model from question 3.
```

Using Differential Slope Model

In [90]:

```
#Differential slope model
#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
#Reference variable or category is "no high school diploma"
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)
data['hsdip'] = data['educdc'].apply(lambda x: 1 if (x >= 12) & (x < 16) else 0)
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + white + \
                black + hispanic + married + NCHILD + vet + hsdip*female + hsdip*AGE
                hsdip*white + hsdip*black + hsdip*hispanic + hsdip*married + \
                hsdip*NCHILD + hsdip*vet + col*female + col*AGE + col*agesq +  \
                col*white + col*black + col*hispanic + col*married + col*NCHILD + co
reg5 = model.fit()
print(reg5.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lnincwage   R-squared:                       0.303
Model:                             OLS   Adj. R-squared:                  0.300
```

```
Method:                 Least Squares   F-statistic:                     117.5
Date:                Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                        22:46:25   Log-Likelihood:                -11108.
No. Observations:                8143   AIC:                         2.228e+04
Df Residuals:                    8112   BIC:                         2.250e+04
Df Model:                          30
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept        6.183e+04  9.14e+05      0.068      0.946   -1.73e+06   1.85e+06
educdc              0.0621     0.007      8.302      0.000       0.047      0.077
female           -822.6296  1.22e+04     -0.068      0.946   -2.47e+04    2.3e+04
AGE              1.336e+06  1.97e+07      0.068      0.946   -3.74e+07      4e+07
agesq           -3.124e+07  4.62e+08     -0.068      0.946   -9.37e+08   8.74e+08
white           -4530.9404    6.7e+04     -0.068      0.946   -1.36e+05   1.27e+05
black             484.8914  7169.634      0.068      0.946   -1.36e+04   1.45e+04
hispanic        -4943.2932  7.31e+04     -0.068      0.946   -1.48e+05   1.38e+05
married          5924.7092  8.76e+04      0.068      0.946   -1.66e+05   1.78e+05
NCHILD          -9393.4906  1.39e+05     -0.068      0.946   -2.82e+05   2.63e+05
vet             -1506.5548  2.23e+04     -0.068      0.946   -4.52e+04   4.22e+04
nohsdip         -6.183e+04  9.14e+05     -0.068      0.946   -1.85e+06   1.73e+06
nohsdip:female    821.9674  1.22e+04      0.068      0.946    -2.3e+04   2.46e+04
nohsdip:AGE     -1.336e+06  1.97e+07     -0.068      0.946      -4e+07   3.74e+07
nohsdip:agesq    3.124e+07  4.62e+08      0.068      0.946   -8.74e+08   9.37e+08
nohsdip:white    4530.8915    6.7e+04      0.068      0.946   -1.27e+05   1.36e+05
nohsdip:black    -484.9791  7169.634     -0.068      0.946   -1.45e+04   1.36e+04
nohsdip:hispanic 4943.3515  7.31e+04      0.068      0.946   -1.38e+05   1.48e+05
nohsdip:married -5924.5905  8.76e+04     -0.068      0.946   -1.78e+05   1.66e+05
nohsdip:NCHILD   9393.5737  1.39e+05      0.068      0.946   -2.63e+05   2.82e+05
nohsdip:vet      1506.8722  2.23e+04      0.068      0.946   -4.22e+04   4.52e+04
hsdip           -6.183e+04  9.14e+05     -0.068      0.946   -1.85e+06   1.73e+06
hsdip:female      822.2265  1.22e+04      0.068      0.946    -2.3e+04   2.47e+04
hsdip:AGE       -1.336e+06  1.97e+07     -0.068      0.946      -4e+07   3.74e+07
hsdip:agesq      3.124e+07  4.62e+08      0.068      0.946   -8.74e+08   9.37e+08
hsdip:white      4531.1532    6.7e+04      0.068      0.946   -1.27e+05   1.36e+05
hsdip:black      -484.9402  7169.633     -0.068      0.946   -1.45e+04   1.36e+04
hsdip:hispanic   4943.3950  7.31e+04      0.068      0.946   -1.38e+05   1.48e+05
hsdip:married   -5924.5322  8.76e+04     -0.068      0.946   -1.78e+05   1.66e+05
hsdip:NCHILD     9393.4880  1.39e+05      0.068      0.946   -2.63e+05   2.82e+05
hsdip:vet        1506.6523  2.23e+04      0.068      0.946   -4.22e+04   4.52e+04
col             -6.183e+04  9.14e+05     -0.068      0.946   -1.85e+06   1.73e+06
col:female        822.2570  1.22e+04      0.068      0.946    -2.3e+04   2.47e+04
col:AGE         -1.336e+06  1.97e+07     -0.068      0.946      -4e+07   3.74e+07
col:agesq        3.124e+07  4.62e+08      0.068      0.946   -8.74e+08   9.37e+08
col:white        4530.8992    6.7e+04      0.068      0.946   -1.27e+05   1.36e+05
col:black        -485.1737  7169.633     -0.068      0.946   -1.45e+04   1.36e+04
col:hispanic     4943.1154  7.31e+04      0.068      0.946   -1.38e+05   1.48e+05
col:married     -5924.5486  8.76e+04     -0.068      0.946   -1.78e+05   1.66e+05
col:NCHILD       9393.4804  1.39e+05      0.068      0.946   -2.63e+05   2.82e+05
col:vet          1506.6480  2.23e+04      0.068      0.946   -4.22e+04   4.52e+04
==============================================================================
Omnibus:                     2762.331   Durbin-Watson:                   1.914
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13445.913
Skew:                          -1.571   Prob(JB):                         0.00
Kurtosis:                       8.455   Cond. No.                     3.39e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The smallest eigenvalue is 5.03e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Fraction of variation observed in log wages that the model explains is 0.303.It almost the same as the variation observed in model from question 3

Question 4.6.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 schooldiploma and an all else equal individual with a college diploma. Assume that it takessomeone 12 years to graduate high school and 16 years to graduate college.

In [91]:

```
#Our equation from the differential intercept is as below:

#ln(incwage) = β0 + β1*educdc + β2*female + β3*age + β4*age2 + β5*white + β6*black
#               + β7*hispanic + β8*married + β9*nchild + β10*vet
#               + β11*nohsdip + β12*hsdip + β13*col + e

#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
#Reference variable or category is "no high school diploma"
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + \
                white + black + hispanic + married + NCHILD + vet  + \
                hsdip + col', data=data)
reg6 = model.fit()
print(reg6.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              lnincwage   R-squared:                       0.299
Model:                            OLS   Adj. R-squared:                  0.298
Method:                 Least Squares   F-statistic:                     288.6
Date:                Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                        22:47:02   Log-Likelihood:                -11132.
No. Observations:                8143   AIC:                         2.229e+04
Df Residuals:                    8130   BIC:                         2.238e+04
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.4359      0.136     47.344      0.000       6.169       6.702
educdc         0.0602      0.007      8.513      0.000       0.046       0.074
female        -0.4022      0.021    -18.773      0.000      -0.444      -0.360
AGE            0.1506      0.006     24.554      0.000       0.139       0.163
agesq         -0.0016   7.25e-05    -21.772      0.000      -0.002      -0.001
white          0.0782      0.030      2.621      0.009       0.020       0.137
black         -0.1709      0.047     -3.674      0.000      -0.262      -0.080
hispanic      -0.0139      0.036     -0.388      0.698      -0.084       0.056
married        0.1724      0.025      6.949      0.000       0.124       0.221
NCHILD        -0.0001      0.010     -0.011      0.991      -0.021       0.020
vet            0.1020      0.054      1.902      0.057      -0.003       0.207
hsdip         -0.0990      0.067     -1.485      0.138      -0.230       0.032
col            0.3106      0.089      3.508      0.000       0.137       0.484
==============================================================================
Omnibus:                     2762.611   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13331.164
Skew:                          -1.575   Prob(JB):                         0.00
Kurtosis:                       8.420   Cond. No.                     2.87e+04
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The condition number is large, 2.87e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [92]:

```
# for individual above conditions, only high school diploma
dat_1 = data.loc[
      (data['educdc'] >= 12)
     & (data['educdc'] <16)
     & (data['AGE'] == 22)
     & (data['female'] == 1)
     & (data['white'] == 0)
     & (data['black'] == 0)
     & (data['hispanic'] == 0)
     & (data['married'] == 0)
     & (data['NCHILD'] == 0)
     & (data['vet'] == 0)
#     & (data['nohsdip'] == 0)
#     & (data['hsdip'] == 1)
#     & (data['col'] == 0)
]

predictions = reg6.get_prediction(dat_1)
predictions.summary_frame(alpha=0.05)[:1]

np.exp(9.327396)
# 11241.819639722586
```

Out[92]:

```
11241.819639722586
```

In [93]:

```
# For individual with characteristics as above individual but having a college degre
dat_2 = data.loc[
    (data['educdc'] >= 16)
    & (data['AGE'] == 22)
     & (data['female'] == 1)
     & (data['white'] == 0)
     & (data['black'] == 0)
     & (data['hispanic'] == 0)
     & (data['married'] == 0)
     & (data['NCHILD'] == 0)
     & (data['vet'] == 0)
 #    & (data['nohsdip'] == 0)
 #    & (data['hsdip'] == 1)
 #    & (data['col'] == 1)
]
predictions2 = reg6.get_prediction(dat_2)
predictions2.summary_frame(alpha=0.05)[:1]

np.exp(9.977857)
# 21544.094050721822
```

Out[93]:

```
21544.094050721822
```

In [75]:

```
21544.094050721822 - 11241.819639722586
```

Out[75]:

```
10302.274410999236
```

Based on the model, keeping everything else constant and as defined, the income wage for an individual characterised as above with a high school diploma, with relative to the individuals *without a high school diploma(our reference point) is 11241.819639722586, in the units defined by the dataset Based on the model, keeping everything else constant and as defined, the income wage for an individual characterised as above with a college degree, with relative to the individuals without a high school diploma(our reference point) is 21544.094050721822, in the units defined by the dataset*

Thus the diference in income wage between the individuals characterised above, one with only a high school diploma and one with a college degree is 10302.274410999236, per the units defined in the dataset

Question 4.6.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 confidentyou are in your predictions. Briefly explain.

The p-value observed 0.00, is less than the significance levels of 0.05, 0.01 and 0.001. This means that at with a 99.9% confidence, we can reject the null hypothesis that the model is not significant and conclude that the model is significant.

Now we observe the p-values of the coefficients of the variables. The p-values are less than 0.05 for most independent variables we can reject the null hypothesis and conclude that they are statistically significant.

We have tested and predicted at an alpha of 0.05 and have resulted in significant model and significant for all variables included in the model. The same applies to the confidence level of 99% and 99.9%. Thus we can help the President understand that we are atleast 95% confident and reasonably 99% confident that our model predicts the income wages correctly.

Question 4.7

There are many ways that this model could be improved. How would you do things differ ently 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.

I think polynomial regression models with different variations in variables would provide a better model to predict the log wages

Example 1: Including age, age2, and age3 $\ln(incwage) = \beta_0 + \beta_1 age + \beta_2 age2 + \beta_3 age3 + \beta_4 educdc + \beta_5 female + \beta_6 white + \beta_7 black + \beta_8 hispanic + \beta_9 married + \beta_{10} nchild + \beta_{11} vet + \beta_{12} nohsdip + \beta_{13} hsdip + \beta_{14} col + e$

In [85]:

```
data['agecube'] = np.power(data['AGE'], 3)
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)
data['hsdip'] = data['educdc'].apply(lambda x: 1 if (x >= 12) & (x < 16) else 0)
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + agecube + \
                white + black + hispanic + married + NCHILD + vet  + hsdip + col', d
reg7 = model.fit()
print(reg7.summary())

print('\n R2 value here is 0.304')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              lnincwage   R-squared:                       0.304
Model:                            OLS   Adj. R-squared:                  0.303
Method:                 Least Squares   F-statistic:                     273.2
Date:                Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                        22:31:54   Log-Likelihood:                -11101.
No. Observations:                8143   AIC:                         2.223e+04
Df Residuals:                    8129   BIC:                         2.233e+04
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.0036      0.338     11.848      0.000       3.341       4.666
educdc         0.0616      0.007      8.742      0.000       0.048       0.075
female        -0.3985      0.021    -18.669      0.000      -0.440      -0.357
AGE            0.3551      0.027     13.283      0.000       0.303       0.408
agesq         -0.0068      0.001    -10.158      0.000      -0.008      -0.006
agecube     4.223e-05   5.38e-06      7.857      0.000    3.17e-05    5.28e-05
white          0.0759      0.030      2.555      0.011       0.018       0.134
black         -0.1721      0.046     -3.713      0.000      -0.263      -0.081
hispanic      -0.0177      0.036     -0.496      0.620      -0.088       0.052
married        0.1684      0.025      6.812      0.000       0.120       0.217
NCHILD         0.0005      0.010      0.046      0.964      -0.020       0.021
vet            0.0964      0.053      1.803      0.071      -0.008       0.201
hsdip         -0.1276      0.067     -1.918      0.055      -0.258       0.003
col            0.2568      0.088      2.903      0.004       0.083       0.430
==============================================================================
Omnibus:                     2812.298   Durbin-Watson:                   1.920
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14065.924
Skew:                          -1.593   Prob(JB):                         0.00
Kurtosis:                       8.595   Cond. No.                     3.87e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The condition number is large, 3.87e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

 R2 value here is 0.304
```

In [83]:

```
#Example 2: Including age, age2, and age3 and interaction between age and educdc
#ln(incwage) = β0 + β1*age + β2*age2 + β3*age3 + β4*educdc +
#              β5*female + β6*white + β7*black + β8*hispanic + β9*married +
#              β10*nchild + β11*vet + β12*nohsdip + β13*hsdip + β14*col
#              β15*age*educdc + e


data['agecube'] = np.power(data['AGE'], 3)
data['ageeducdc'] = data['AGE']*data['educdc']
#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
#Reference variable or category is "no high school diploma"
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + agecube + \
        ageeducdc + white + black + hispanic + married + NCHILD + vet  + hsdip + col
reg8 = model.fit()
print(reg8.summary())

print('\n R2 value is 0.305')
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              lnincwage   R-squared:                       0.305
Model:                            OLS   Adj. R-squared:                  0.304
Method:                 Least Squares   F-statistic:                     255.0
Date:                Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                        22:29:31   Log-Likelihood:                -11095.
No. Observations:                8143   AIC:                         2.222e+04
Df Residuals:                    8128   BIC:                         2.232e+04
Df Model:                          14                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.3364      0.272     12.256      0.000       2.803       3.870
educdc         0.0126      0.015      0.827      0.408      -0.017       0.042
female        -0.3941      0.021    -18.443      0.000      -0.436      -0.352
AGE            0.3607      0.027     13.480      0.000       0.308       0.413
agesq         -0.0073      0.001    -10.674      0.000      -0.009      -0.006
agecube     4.568e-05   5.45e-06      8.374      0.000     3.5e-05    5.64e-05
ageeducdc      0.0011      0.000      3.625      0.000       0.001       0.002
white          0.0749      0.030      2.524      0.012       0.017       0.133
black         -0.1780      0.046     -3.841      0.000      -0.269      -0.087
hispanic      -0.0201      0.036     -0.565      0.572      -0.090       0.050
married        0.1679      0.025      6.797      0.000       0.119       0.216
NCHILD        -0.0027      0.010     -0.253      0.800      -0.023       0.018
vet            0.0942      0.053      1.764      0.078      -0.010       0.199
nohsdip        1.0636      0.093     11.389      0.000       0.881       1.247
hsdip          0.9422      0.093     10.101      0.000       0.759       1.125
col            1.3307      0.109     12.253      0.000       1.118       1.544
==============================================================================
Omnibus:                     2824.670   Durbin-Watson:                   1.919
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14175.109
Skew:                          -1.600   Prob(JB):                         0.00
Kurtosis:                       8.616   Cond. No.                     2.83e+18
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The smallest eigenvalue is 1.46e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

 R2 value is 0.305
```

In both the above regressions we observe that hispanic variable is close to being statistically insignificant. So it may not be actually helping the model. Lets remove it from the base model and test the model.

In [86]:

```python
#Example 3
#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
#Reference variable or category is "no high school diploma"
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)
data['hsdip'] = data['educdc'].apply(lambda x: 1 if (x >= 12) & (x < 16) else 0)
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + white + black \
+ married + NCHILD + vet + hsdip + col', data=data)
reg9 = model.fit()
print(reg9.summary())

print(' \n R2 is same as base model i.e 0.299. This means that hispanic variable add
no meaningful value in the regression')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lnincwage   R-squared:                       0.299
Model:                             OLS   Adj. R-squared:                  0.298
Method:                  Least Squares   F-statistic:                     314.9
Date:                 Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                         22:32:11   Log-Likelihood:                -11132.
No. Observations:                 8143   AIC:                         2.229e+04
Df Residuals:                     8131   BIC:                         2.237e+04
Df Model:                           11
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.4233      0.132     48.658      0.000       6.165       6.682
educdc         0.0604      0.007      8.561      0.000       0.047       0.074
female        -0.4021      0.021    -18.772      0.000      -0.444      -0.360
AGE            0.1507      0.006     24.568      0.000       0.139       0.163
agesq         -0.0016   7.25e-05    -21.779      0.000      -0.002      -0.001
white          0.0844      0.025      3.354      0.001       0.035       0.134
black         -0.1640      0.043     -3.817      0.000      -0.248      -0.080
married        0.1727      0.025      6.966      0.000       0.124       0.221
NCHILD        -0.0003      0.010     -0.026      0.979      -0.021       0.020
vet            0.1022      0.054      1.904      0.057      -0.003       0.207
hsdip         -0.0975      0.067     -1.464      0.143      -0.228       0.033
col            0.3124      0.088      3.533      0.000       0.139       0.486
==============================================================================
Omnibus:                     2759.914   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13305.325
Skew:                          -1.573   Prob(JB):                         0.00
Kurtosis:                       8.414   Cond. No.                     2.78e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The condition number is large, 2.78e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

 R2 is same as base model i.e 0.299. This means that hispanic variable adds no meani
ngful value in the regression
```

Same as above, now lets remove the insignificant variable NCHILD and see how it affects the model

In [88]:

```
#Example 4
#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
#Reference variable or category is "no high school diploma"
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)data['hsdip'] =
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + white + black
                + married  + vet  + hsdip + col', data=data)
reg10 = model.fit()
print(reg10.summary())

print(' \n R2 is same as base model i.e 0.299. This means that NCHILD variable adds
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               lnincwage   R-squared:                       0.299
Model:                             OLS   Adj. R-squared:                  0.298
Method:                  Least Squares   F-statistic:                     346.4
Date:                 Wed, 25 Jan 2023   Prob (F-statistic):               0.00
```

```
Time:                        22:33:35   Log-Likelihood:                -11132.
No. Observations:                8143   AIC:                         2.229e+04
Df Residuals:                    8132   BIC:                         2.236e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.8718      0.106     45.982      0.000       4.664       5.079
educdc         0.0604      0.007      8.563      0.000       0.047       0.074
female        -0.4021      0.021    -18.777      0.000      -0.444      -0.360
AGE            0.1506      0.006     26.095      0.000       0.139       0.162
agesq         -0.0016    6.8e-05    -23.219      0.000      -0.002      -0.001
white          0.0844      0.025      3.359      0.001       0.035       0.134
black         -0.1640      0.043     -3.817      0.000      -0.248      -0.080
married        0.1725      0.024      7.284      0.000       0.126       0.219
vet            0.1022      0.054      1.904      0.057      -0.003       0.207
nohsdip        1.5523      0.048     32.109      0.000       1.457       1.647
hsdip          1.4548      0.041     35.376      0.000       1.374       1.535
col            1.8647      0.065     28.709      0.000       1.737       1.992
==============================================================================
Omnibus:                     2759.847   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            13304.558
Skew:                          -1.573   Prob(JB):                         0.00
Kurtosis:                       8.414   Cond. No.                     5.19e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The smallest eigenvalue is 1.46e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

 R2 is same as base model i.e 0.299. This means that NCHILD variable adds no meaning
ful value in the regression
```

In [89]:

```python
#Example 5: A model with age, age2, age3, interaction between age and educdc after r
data['agecube'] = np.power(data['AGE'], 3)
data['ageeducdc'] = data['AGE']*data['educdc']
#Assuming that it takes someone 12 years to graduate high school and 16 years to gra
#Reference variable or category is "no high school diploma"
#data['nohsdip'] = data['educdc'].apply(lambda x: 1 if x < 12 else 0)
data['hsdip'] = data['educdc'].apply(lambda x: 1 if (x >= 12) & (x < 16) else 0)
data['col'] = data['educdc'].apply(lambda x: 1 if x >= 16 else 0)
model = smf.ols(formula='lnincwage ~ educdc + female + AGE + agesq + agecube + \
                ageeducdc + white + black + married  + vet  + hsdip + col', data=dat
reg11 = model.fit()
print(reg11.summary())

print(' \n The R-square is now 0.305 and Adjusted R2 is 0.304. Same as the base mode
with age2 and age3 and interaction between age and educdc. Which means that  its tru
that hispanic and NCHILD donot add any meaningful value to the determination of inco
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              lnincwage   R-squared:                       0.305
Model:                            OLS   Adj. R-squared:                  0.304
Method:                 Least Squares   F-statistic:                     297.5
Date:                Wed, 25 Jan 2023   Prob (F-statistic):               0.00
Time:                        22:34:32   Log-Likelihood:                -11095.
No. Observations:                8143   AIC:                         2.222e+04
Df Residuals:                    8130   BIC:                         2.231e+04
Df Model:                          12
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.3288      0.271     12.281      0.000       2.797       3.860
educdc         0.0134      0.015      0.880      0.379      -0.016       0.043
female        -0.3942      0.021    -18.454      0.000      -0.436      -0.352
AGE            0.3600      0.027     13.494      0.000       0.308       0.412
agesq         -0.0073      0.001    -10.666      0.000      -0.009      -0.006
agecube     4.562e-05   5.45e-06      8.365      0.000    3.49e-05    5.63e-05
ageeducdc      0.0011      0.000      3.605      0.000       0.000       0.002
white          0.0843      0.025      3.369      0.001       0.035       0.133
black         -0.1678      0.043     -3.921      0.000      -0.252      -0.084
married        0.1664      0.024      7.052      0.000       0.120       0.213
vet            0.0945      0.053      1.770      0.077      -0.010       0.199
nohsdip        1.0589      0.093     11.407      0.000       0.877       1.241
hsdip          0.9403      0.093     10.117      0.000       0.758       1.123
col            1.3295      0.108     12.278      0.000       1.117       1.542
==============================================================================
Omnibus:                     2819.878   Durbin-Watson:                   1.919
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14125.523
Skew:                          -1.597   Prob(JB):                         0.00
Kurtosis:                       8.606   Cond. No.                     2.83e+18
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The smallest eigenvalue is 1.46e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

 The R-square is now 0.305 and Adjusted R2 is 0.304. Same as the base model with age
2 and age3 and interaction between age and educdc. Which means that  its true that h
ispanic and NCHILD donot add any meaningful value to the determination of income wag
es.
```

As we can see above, by further analysis from the regression scores, we have determined that a few variables do not actually contribute to the model. Also we have seen the R squared value to increase resulting a better model when we have included age^3 and also when we have included an interaction variable between age and education variable. This shows that tweaking the model in appropriate ways and including and excluding variations of variable results a better model. Further analysis can be done using variations such as root of age, removing the borderline significant variable black. Also futrther regression models such as Lasso or Rigid can result a better model.

In [ ]: