Student Name: **Luke Nguyen**
Student ID: **D5850A**

Statistical Methods and Data Analysis (EN.625.603)
Final Exam

Description: The datafile contains data for 2015 for full-time workers with a high school diploma or B.A./B.S. as their highest degree. See the pdf attachment for an overview of the data and variable descriptions. In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and higher earnings.)

a. Run a regression of average hourly earnings (AHE) on *age* (Age), *gender* ($Female$), and *education* ($Bachelor$). If *age* increases from 25 to 26, how are earnings expected to change? If *age* increases from 33 to 34, how are earnings expected to change?
Solution:
Running regression model of ahe on $age, female, bachelor$ in Python yields the following results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ahe   R-squared:                       0.190
Model:                            OLS   Adj. R-squared:                  0.189
Method:                 Least Squares   F-statistic:                     553.4
Date:                Mon, 21 Aug 2023   Prob (F-statistic):          3.46e-323
Time:                        16:15:49   Log-Likelihood:                -27036.
No. Observations:                7098   AIC:                         5.408e+04
Df Residuals:                    7094   BIC:                         5.411e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.0448      1.355      1.509      0.131      -0.611       4.700
age            0.5313      0.045     11.788      0.000       0.443       0.620
female        -4.1435      0.266    -15.583      0.000      -4.665      -3.622
bachelor       9.8456      0.262     37.519      0.000       9.331      10.360
==============================================================================
Omnibus:                     2458.198   Durbin-Watson:                   1.936
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11294.166
Skew:                           1.629   Prob(JB):                         0.00
Kurtosis:                       8.252   Cond. No.                         312.
==============================================================================
```

which is equivalent to the following linear regression equation:

$$ahe = \beta_0 + \beta_1(age) + \beta_2(female) + \beta_3(bachelor)$$
$$ahe = 2.0448 + 0.5313(age) - 4.1435(female) + 9.8456(bachelor)$$

The coefficient for age is $\beta_1 = 0.5313$. This means that for every one unit increase in age, ahe is expected to increases by **0.5313 dollars**. This applies to both the change from age 25 to 26 and the change from age 33 to 34.

b. Run a regression of the logarithm of average hourly earnings, $\ln(AHE)$, on Age, $Female$, and $Bachelor$. If age increases from 25 to 26, how are earnings expected to change? If age increases from 33 to 34, how are earnings expected to change?
Solution:
Running regression model of $\ln(ahe)$ on $age, female, bachelor$ in Python yields the following results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ln_ahe   R-squared:                       0.208
Model:                            OLS   Adj. R-squared:                  0.208
Method:                 Least Squares   F-statistic:                     622.4
Date:                Mon, 21 Aug 2023   Prob (F-statistic):               0.00
Time:                        16:49:38   Log-Likelihood:                -4821.9
No. Observations:                7098   AIC:                             9652.
Df Residuals:                    7094   BIC:                             9679.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.0274      0.059     34.220      0.000       1.911       2.143
age            0.0242      0.002     12.273      0.000       0.020       0.028
female        -0.1776      0.012    -15.274      0.000      -0.200      -0.155
bachelor       0.4615      0.011     40.212      0.000       0.439       0.484
==============================================================================
Omnibus:                      185.302   Durbin-Watson:                   1.943
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              309.107
Skew:                          -0.236   Prob(JB):                     7.55e-68
Kurtosis:                       3.906   Cond. No.                         312.
==============================================================================
```

which is equivalent to the following linear regression equation:

$$\ln(ahe) = \beta_0 + \beta_1(age) + \beta_2(female) + \beta_3(bachelor)$$
$$\ln(ahe) = 2.0274 + 0.0242(age) - 0.1776(female) + 0.4615(bachelor)$$

The coefficient for age is $\beta_1 = 0.0242$. This means that for every one unit increase in age, $\ln(ahe)$ is expected to increases by 0.0242. As the model is an exponential model for ahe, we can say that for every one unit increase in age, ahe is expected to increases by **2.42%**.

c. Run a regression of the logarithm of average hourly earnings, $ln(AHE)$, on $ln(Age)$, $Female$, and $Bachelor$. If age increases from 25 to 26, how are earnings expected to change? If age increases from 33 to 34, how are earnings expected to change?

Solution:
Running regression model of $ln(ahe)$ on $ln(age), female, bachelor$ in Python yields the following results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ln_ahe   R-squared:                       0.209
Model:                            OLS   Adj. R-squared:                  0.208
Method:                 Least Squares   F-statistic:                     623.4
Date:                Mon, 21 Aug 2023   Prob (F-statistic):               0.00
Time:                        17:19:02   Log-Likelihood:                -4820.8
No. Observations:                7098   AIC:                             9650.
Df Residuals:                    7094   BIC:                             9677.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3233      0.196      1.649      0.099      -0.061       0.708
ln_age         0.7154      0.058     12.368      0.000       0.602       0.829
female        -0.1775      0.012    -15.268      0.000      -0.200      -0.155
bachelor       0.4615      0.011     40.220      0.000       0.439       0.484
==============================================================================
Omnibus:                      184.684   Durbin-Watson:                   1.943
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              307.770
Skew:                          -0.236   Prob(JB):                     1.47e-67
Kurtosis:                       3.904   Cond. No.                         130.
==============================================================================
```

which is equivalent to the following linear regression equation:

$$\ln(ahe) = \beta_0 + \beta_1(\ln(age)) + \beta_2(female) + \beta_3(bachelor)$$
$$\ln(ahe) = 0.3233 + 0.7154(\ln(age)) - 0.1775(female) + 0.4615(bachelor)$$

The coefficient for $\ln(age)$ is $\beta_1 = 0.7154$.
For age increases from 25 to 26, age increases by $\frac{26-25}{25} = 4\%$, thus $\ln(ahe)$ is expected to increases by $0.7154 \times 4\% = \mathbf{2.8616\%}$.
For age increases from 33 to 34, age increases by $\frac{34-33}{33} = 3.03\%$, thus $\ln(ahe)$ is expected to increases by $0.7154 \times 3.03\% = \mathbf{2.1679\%}$.

d. Run a regression of the logarithm of average hourly earnings, $\ln(AHE)$, on Age, Age^2, $Female$, and $Bachelor$. If age increases from 25 to 26, how are earnings expected to change? If age increases from 33 to 34, how are earnings expected to change?

Solution:
Running regression model of $\ln(ahe)$ on $age, (age^2), female, bachelor$ in Python yields the following results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ln_ahe   R-squared:                       0.209
Model:                            OLS   Adj. R-squared:                  0.209
Method:                 Least Squares   F-statistic:                     468.6
Date:                Mon, 21 Aug 2023   Prob (F-statistic):               0.00
Time:                        17:51:18   Log-Likelihood:                -4819.1
No. Observations:                7098   AIC:                             9648.
Df Residuals:                    7093   BIC:                             9682.
Df Model:                           4
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const           0.4187      0.672      0.623      0.533      -0.899       1.736
age             0.1341      0.046      2.929      0.003       0.044       0.224
age_squared    -0.0019      0.001     -2.403      0.016      -0.003      -0.000
female         -0.1774      0.012    -15.256      0.000      -0.200      -0.155
bachelor        0.4616      0.011     40.236      0.000       0.439       0.484
===============================================================================
Omnibus:                      182.315   Durbin-Watson:                   1.944
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              302.731
Skew:                          -0.234   Prob(JB):                     1.83e-66
Kurtosis:                       3.897   Cond. No.                     1.07e+05
==============================================================================
```

which is equivalent to the following linear regression equation:

$$\ln(ahe) = \beta_0 + \beta_1(age) + \beta_2(age^2) + \beta_3(female) + \beta_4(bachelor)$$
$$\ln(ahe) = 0.4187 + 0.1341(age) - 0.0019(age^2) - 0.1774(female) + 0.4616(bachelor)$$

The coefficient for age is $\beta_1 = 0.1341$. and the coefficient for age^2 is $\beta_2 = -0.0019$.
For age increases from 25 to 26, age increases by

$$0.1341 - 0.0019 \times (26^2 - 25^2) = 3.72\%$$

thus, ahe is expected to increases by

3.72%

For age increases from 33 to 34, age increases by

$$0.1341 - 0.0019 \times (34^2 - 33^2) = 0.68\%$$

thus, ahe is expected to increases by

0.68%

e. Do you prefer the regression in (c) to the regression in (b)? Explain.

Solution:
The model in (c) is a better model than the model in (b).
Although the $R-squared, Adj.R-squared, AIC, BIC, p-value$ of the model (b) and (c) are very close.
The model (b) is a linear model for the change in the percentage of *ahe* with respect to the change in *age*, and it is fixed at **2.42%** for every one unit increase in *age*.
However, in the case of comparing the change in *ahe* with respect to the change in *age*, the model should account for the diminishing return of *ahe* with respect to the increase in *age*, as the workers's average hour earnings should plateau at some point, for example, between 25 and 26, the increase in *ahe* is **2.8616%**, but between 33 and 34, the increase in *ahe* is only **2.1679%**, which will be more accurately modeled by the model in (c).

f. Do you prefer the regression in (d) to the regression in (b)? Explain.
Solution:
The model in (d) is a better model than the model in (b).
We can use the same argument as in (e) to explain why the model in (d) is better than the model in (b) as the model in (d) accounts for the diminishing return of *ahe* with respect to the increase in *age*.

g. Do you prefer the regression in (d) to the regression in (c)? Explain.
Solution:
The model in (d) is a better model than the model in (c).
Although, both models in (c) and (d) account for the diminishing return of *ahe* with respect to the increase in *age*, model (d) includes the quadratic term of *age* and its coefficient is negative, thus it represents that after passing a certain age, the *ahe* will decrease with respect to the increase in *age*. Which I think is more realistic in the real world. Additionally, the increase in *ahe* with respect to the increase in *age* is more aggressive in early career which also seems to be more accurate.
This assumption is more of my personal and relative opinion than an absolute proof.

h. Run a regression of $\ln(AHE)$, on Age, Age^2 , $Female$, $Bachelor$, and the interaction term $Female \times Bachelor$. What does the coefficient on the interaction term measure? Alexis is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of $\ln(AHE)$? Jane is a 30-year-old female with a high school degree. What does the regression predict for her value of $\ln(AHE)$? What is the predicted difference between Alexis's and Jane's earnings? Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of $\ln(AHE)$? Jim is a 30-year-old male with a high school degree. What does the regression predict for his value of $\ln(AHE)$? What is the predicted difference between Bob's and Jim's earnings?

Solution:
Running regression model of $\ln{(ahe)}$ on $age, (age^2), female, bachelor, female \times bachelor$ in Python yields the following results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ln_ahe   R-squared:                       0.209
Model:                            OLS   Adj. R-squared:                  0.209
Method:                 Least Squares   F-statistic:                     375.1
Date:                Mon, 21 Aug 2023   Prob (F-statistic):               0.00
Time:                        19:03:45   Log-Likelihood:                -4818.5
No. Observations:                7098   AIC:                             9649.
Df Residuals:                    7092   BIC:                             9690.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.4119      0.672      0.613      0.540      -0.906       1.729
age                   0.1348      0.046      2.944      0.003       0.045       0.225
age_squared          -0.0019      0.001     -2.416      0.016      -0.003      -0.000
female               -0.1903      0.017    -10.955      0.000      -0.224      -0.156
bachelor              0.4521      0.015     30.379      0.000       0.423       0.481
female_x_bachelor     0.0235      0.023      1.004      0.315      -0.022       0.069
==============================================================================
Omnibus:                      181.391   Durbin-Watson:                   1.944
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              300.574
Skew:                          -0.234   Prob(JB):                     5.39e-66
Kurtosis:                       3.893   Cond. No.                     1.07e+05
==============================================================================
```

which is equivalent to the following linear regression equation:

$$\begin{aligned}
\ln{(ahe)} = \quad & \beta_0 + \beta_1(age) + \beta_2(age^2) \\
& + \beta_3(female) + \beta_4(bachelor) + \beta_5(female \times bachelor) \\
\ln{(ahe)} = \quad & 0.4119 + 0.1348(age) - 0.0019(age^2) \\
& - 0.1903(female) + 0.4521(bachelor) + 0.0235(female \times bachelor)
\end{aligned}$$

For Alexis, a 30-year-old female with a bachelor's degree, her $\ln{(ahe)}$ is as follows:

$$\begin{aligned}
\ln{(ahe)} &= 0.4119 + 0.1348(30) - 0.0019(30^2) - 0.1903(1) + 0.4521(1) + 0.0235(1 \times 1) \\
&= 0.4119 + 4.044 - 1.71 - 0.1903 + 0.4521 + 0.0235 \\
&= \mathbf{3.0312}
\end{aligned}$$

For Jane, a 30-year-old female with a high school degree, her $\ln{(ahe)}$ is as follows:

$$\begin{aligned}
\ln{(ahe)} &= 0.4119 + 0.1348(30) - 0.0019(30^2) - 0.1903(1) + 0.4521(0) + 0.0235(1 \times 0) \\
&= 0.4119 + 4.044 - 1.71 - 0.1903 + 0 + 0 \\
&= \mathbf{2.5556}
\end{aligned}$$

The predicted difference between Alexis's and Jane's earnings is as follows:

$$\begin{aligned}
e^{\ln{(ahe)}_{Alexis}} - e^{\ln{(ahe)}_{Jane}} &= e^{3.0312} - e^{2.5556} \\
&= 20.7221 - 12.8790 \\
&= \mathbf{7.8431}
\end{aligned}$$

For Bob, a 30-year-old male with a bachelor's degree, his $\ln(ahe)$ is as follows:

$$\begin{aligned}\ln(ahe) &= 0.4119 + 0.1348(30) - 0.0019(30^2) - 0.1903(0) + 0.4521(1) + 0.0235(0 \times 1)\\ &= 0.4119 + 4.044 - 1.71 - 0 + 0.4521 + 0\\ &= \mathbf{3.198}\end{aligned}$$

For Jim, a 30-year-old male with a high school degree, his $ln(ahe)$ is as follows:

$$\begin{aligned}\ln(ahe) &= 0.4119 + 0.1348(30) - 0.0019(30^2) - 0.1903(0) + 0.4521(0) + 0.0235(0 \times 0)\\ &= 0.4119 + 4.044 - 1.71 - 0 + 0 + 0\\ &= \mathbf{2.7459}\end{aligned}$$

The predicted difference between Bob's and Jim's earnings is as follows:

$$\begin{aligned}e^{\ln(ahe)_{Alexis}} - e^{\ln(ahe)_{Jane}} &= e^{3.198} - e^{2.7459}\\ &= 24.4835 - 15.5786\\ &= \mathbf{8.9049}\end{aligned}$$

i. Is the effect of age on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.
Solution:
We run a regression of $\ln(ahe)$ on $age, age^2, bachelor, age \times female$ in Python and obtain the following results, we skipped $female$ because it is not statistically significant:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ln_ahe   R-squared:                       0.210
Model:                            OLS   Adj. R-squared:                  0.209
Method:                 Least Squares   F-statistic:                     470.0
Date:                Mon, 21 Aug 2023   Prob (F-statistic):               0.00
Time:                        19:58:48   Log-Likelihood:                -4816.7
No. Observations:                7098   AIC:                             9643.
Df Residuals:                    7093   BIC:                             9678.
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          0.3081      0.672      0.459      0.647      -1.009       1.625
age            0.1391      0.046      3.039      0.002       0.049       0.229
age_squared   -0.0019      0.001     -2.457      0.014      -0.003      -0.000
bachelor       0.4612      0.011     40.234      0.000       0.439       0.484
age_x_female  -0.0060      0.000    -15.413      0.000      -0.007      -0.005
==============================================================================
Omnibus:                      183.172   Durbin-Watson:                   1.944
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              305.436
Skew:                          -0.234   Prob(JB):                     4.74e-67
Kurtosis:                       3.902   Cond. No.                     1.07e+05
==============================================================================
```

which is equivalent to the following linear regression equation:

$$\ln(ahe) = \beta_0 + \beta_1(age) + \beta_2(age^2) + \beta_3(bachelor) + \beta_4(age \times female)$$

$$\ln(ahe) = 0.3081 + 0.1391(age) - 0.0019(age^2) + 0.4612(bachelor) - 0.006(age \times female)$$

$\beta_4 = -0.006$ represents the difference in the effect of age on earning for female compared to male, the negative sign indicates that the effect is less for female than for male.

For example, all else equal, a 30-year-old male is expected to earn $e^{0.006\times30} = 119.7217\%$ of a 30-year-old female.

j. Is the effect of age on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.
Solution:
We run a regression of $\ln(ahe)$ on $age, age^2, female, age \times bachelor$ in Python and obtain the following results, we skipped $bachelor$ because it will make $age \times bachelor$ statistically insiginficant.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ln_ahe   R-squared:                       0.208
Model:                            OLS   Adj. R-squared:                  0.208
Method:                 Least Squares   F-statistic:                     466.9
Date:                Mon, 21 Aug 2023   Prob (F-statistic):               0.00
Time:                        20:46:11   Log-Likelihood:                -4821.7
No. Observations:                7098   AIC:                             9653.
Df Residuals:                    7093   BIC:                             9688.
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.6514      0.672      0.969      0.333      -0.667       1.969
age                0.1266      0.046      2.765      0.006       0.037       0.216
age_squared       -0.0019      0.001     -2.417      0.016      -0.003      -0.000
female            -0.1753      0.012    -15.083      0.000      -0.198      -0.153
age_x_bachelor     0.0155      0.000     40.155      0.000       0.015       0.016
==============================================================================
Omnibus:                      181.414   Durbin-Watson:                   1.941
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              301.818
Skew:                          -0.232   Prob(JB):                     2.89e-66
Kurtosis:                       3.897   Cond. No.                     1.07e+05
==============================================================================
```

which is equivalent to the following linear regression equation:

$$\ln(ahe) = \beta_0 + \beta_1(age) + \beta_2(age^2) + \beta_3(female) + \beta_4(age \times bachelor)$$
$$\ln(ahe) = 0.6514 + 0.1266(age) - 0.0019(age^2) - 0.1753(female) + 0.0155(age \times bachelor)$$

$\beta_4 = 0.0155$ represents the difference in the effect of age on earning for bachelor's degree holders, the positive sign indicates that the effect is more for bachelor's degree holders.
For example, all else equal, a 30-year-old bachelor's degree holder is expected to earn $e^{0.0155\times30} = 159.2014\%$ of a high school degree holder.

k. After running all these regressions, summarize the effect of age on earnings for young workers.
Solution:
All regressions have consistently shown that the effect of age on ahe is positive. However, by running the regression of $\ln ahe$ on age, age^2, the model suggests that the positive effect of age on ahe is decreasing and eventually plateaus. $bachelor$ showed a very strong positive effect on ahe and $female$ showed a negative effect on ahe.

Extra Credit:

Although the assignment refers to a single source of data, there are 7 different regression models from (a), (b), (c), (d), (h), (i), and (j).

Besides, having extremely hand-on experiece with applying models and reading results, we also learnt that understanding the nature of the data is very important. For example, if we did not take into consideration the diminishing effect of age on earnings, we would have concluded that the effect of age on earnings is positive and increasing linearly forever, which is pretty absurd.

Also, linking from project 1, 2 to this final exam, I noticed that the way I used Python code for modeling and analyzing data has improved, I am very comfortable using Python stats, numpy to do regression after this series of problems.

```
# All related code for the assignment is below:
import numpy as np
import statsmodels.api as sm
import pandas as pd


class CurrentPopulationSurveyDataFrame:
    df = None

    def __init__(self):
        file_path = 'CPS2015-1.xlsx'
        file_sheet_name = 'Data'
        self.df = pd.read_excel(file_path, sheet_name=file_sheet_name)


def part_a():
    df = CurrentPopulationSurveyDataFrame().df
    Y = df['ahe']
    X = df[['age', 'female', 'bachelor']]
    X = sm.add_constant(X)
    model = sm.OLS(Y, X).fit()
    print()
    print(model.summary())


def part_b():
    df = CurrentPopulationSurveyDataFrame().df
    df['ln_ahe'] = np.log(df['ahe'])
    Y = df['ln_ahe']
    X = df[['age', 'female', 'bachelor']]
    X = sm.add_constant(X)
    model = sm.OLS(Y, X).fit()
    print()
    print(model.summary())


def part_c():
    df = CurrentPopulationSurveyDataFrame().df
    df['ln_ahe'] = np.log(df['ahe'])
    df['ln_age'] = np.log(df['age'])
    Y = df['ln_ahe']
    X = df[['ln_age', 'female', 'bachelor']]
    X = sm.add_constant(X)
    model = sm.OLS(Y, X).fit()
    print()
    print(model.summary())

```

```python

def part_d():
    df = CurrentPopulationSurveyDataFrame().df
    df['ln_ahe'] = np.log(df['ahe'])
    df['age_squared'] = np.square(df['age'])
    Y = df['ln_ahe']
    X = df[['age', 'age_squared', 'female', 'bachelor']]
    X = sm.add_constant(X)
    model = sm.OLS(Y, X).fit()
    print()
    print(model.summary())


def part_e():
    df = CurrentPopulationSurveyDataFrame().df
    df['ln_ahe'] = np.log(df['ahe'])
    df['age_squared'] = np.square(df['age'])
    df['female_x_bachelor'] = df['female'] * df['bachelor']
    Y = df['ln_ahe']
    X = df[['age', 'age_squared', 'female', 'bachelor', 'female_x_bachelor']]
    X = sm.add_constant(X)
    model = sm.OLS(Y, X).fit()
    print()
    print(model.summary())


def part_i():
    df = CurrentPopulationSurveyDataFrame().df
    df['ln_ahe'] = np.log(df['ahe'])
    df['age_squared'] = np.square(df['age'])
    df['age_x_female'] = df['age'] * df['female']
    Y = df['ln_ahe']
    X = df[['age', 'age_squared', 'bachelor', 'age_x_female']]
    X = sm.add_constant(X)
    model = sm.OLS(Y, X).fit()
    print()
    print(model.summary())


def part_j():
    df = CurrentPopulationSurveyDataFrame().df
    df['ln_ahe'] = np.log(df['ahe'])
    df['age_squared'] = np.square(df['age'])
    df['age_x_bachelor'] = df['age'] * df['bachelor']
    Y = df['ln_ahe']
    X = df[['age', 'age_squared', 'female', 'age_x_bachelor']]
    X = sm.add_constant(X)
    model = sm.OLS(Y, X).fit()
    print()
    print(model.summary())


if __name__ == '__main__':
    part_j()
```