Problem E8.2 from Stock and Watson

October 8, 2020

Econometrics II, Bachelor degree in Economics

Universitat Autònoma de Barcelona

Student, Roylan Martinez Vargas

Professor, Michael Creel

NIU: 1539069

E8.2 On the text website **www.pearsonglobaleditions.com/Stock_Watson** you will find a data file **CPS12**, which contains data for full-time, full-year workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS12_Description**, also available on the website. (These are the same data as in **CPS92_12**, used in Empirical Exercise 3.1, but are limited to the year 2012.) 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?

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?

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?

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?

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

f. Do you prefer the regression in (d) to the regression in (b)? Explain.

g. Do you prefer the regression in (d) to the regression in (c)? Explain.

h. Plot the regression relation between *Age* and $\ln(AHE)$ from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

i. 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?

j. 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.

k. 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.

l. After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.

[1]:
```
# TOOLS
import numpy as np
import pandas as pd
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
```

[2]:
```
# Before anything else I will create age ^ 2 column.
dataExcel = pd.read_excel("cps12.xlsx")
dataExcel['age2'] = dataExcel['age'] ** 2
dataExcel
```

[2]:

	year	ahe	bachelor	female	age	age2
0	2012	19.230770	0	0	30	900
1	2012	17.548077	0	0	29	841
2	2012	8.547009	0	0	27	729
3	2012	16.826923	0	1	25	625
4	2012	16.346153	1	1	27	729
…	…	…	…	…	…	…
7435	2012	14.423077	0	0	25	625
7436	2012	7.692307	0	0	32	1024
7437	2012	11.538462	0	0	30	900
7438	2012	9.134615	0	1	25	625
7439	2012	9.615385	1	0	33	1089

```
[7440 rows x 6 columns]
```

E8.2 - a

```
[3]: sampleA = ols("ahe ~ age + female + bachelor", data = dataExcel).fit()
     print(sampleA.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ahe   R-squared:                       0.180
Model:                            OLS   Adj. R-squared:                  0.180
Method:                 Least Squares   F-statistic:                     544.5
Date:                Thu, 08 Oct 2020   Prob (F-statistic):          6.51e-320
Time:                        17:43:54   Log-Likelihood:                -27443.
No. Observations:                7440   AIC:                         5.489e+04
Df Residuals:                    7436   BIC:                         5.492e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.8662      1.188      1.571      0.116      -0.462       4.194
age            0.5103      0.040     12.912      0.000       0.433       0.588
female        -3.8103      0.230    -16.596      0.000      -4.260      -3.360
bachelor       8.3186      0.227     36.584      0.000       7.873       8.764
==============================================================================
Omnibus:                     1975.582   Durbin-Watson:                   1.935
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6089.399
Skew:                           1.360   Prob(JB):                         0.00
Kurtosis:                       6.499   Cond. No.                         316.
==============================================================================

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

The OLS model would be:
ahe = 1.8662 + 0.5103 * **age** - 3.8103 * **female** + 8.3186 * **bachelor**
If age icreases from 25 to 26 (one year) earnings are expected to increase by 0.5103
If age icreases from 33 to 34 (one year) earnings are expected to increase by 0.5103

E8.2 - b

```
[4]: sampleB = ols("np.log(ahe) ~ age + female + bachelor", data = dataExcel).fit()
     print(sampleB.summary())
```

```
                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:            np.log(ahe)   R-squared:                       0.196
Model:                            OLS   Adj. R-squared:                  0.196
Method:                 Least Squares   F-statistic:                     605.7
Date:                Thu, 08 Oct 2020   Prob (F-statistic):               0.00
Time:                        17:43:54   Log-Likelihood:                -5066.6
No. Observations:                7440   AIC:                         1.014e+04
Df Residuals:                    7436   BIC:                         1.017e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.9414      0.059     33.083      0.000       1.826       2.056
age            0.0255      0.002     13.067      0.000       0.022       0.029
female        -0.1923      0.011    -16.953      0.000      -0.215      -0.170
bachelor       0.4378      0.011     38.964      0.000       0.416       0.460
==============================================================================
Omnibus:                      316.825   Durbin-Watson:                   1.936
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              508.141
Skew:                          -0.375   Prob(JB):                    4.56e-111
Kurtosis:                       4.037   Cond. No.                         316.
==============================================================================

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

The OLS model would be:
ln(ahe) = 1.9414 + 0.0255 * **age** - 0.1923 * **female** + 0.4378 * **bachelor**
If age icreases from 25 to 26 (one year) earnings are expected to increase 2.55%
If age icreases from 33 to 34 (one year) earnings are expected to increase 2.55%

E8.2 - c

[5]:
```
sampleC = ols("np.log(ahe) ~ np.log(age) + female + bachelor", data =
↪dataExcel).fit()
print(sampleC.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            np.log(ahe)   R-squared:                       0.197
Model:                            OLS   Adj. R-squared:                  0.196
Method:                 Least Squares   F-statistic:                     606.4
Date:                Thu, 08 Oct 2020   Prob (F-statistic):               0.00
Time:                        17:43:54   Log-Likelihood:                -5065.8
No. Observations:                7440   AIC:                         1.014e+04
Df Residuals:                    7436   BIC:                         1.017e+04
Df Model:                           3
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1495      0.194      0.769      0.442      -0.231       0.531
np.log(age)    0.7529      0.057     13.132      0.000       0.641       0.865
female        -0.1924      0.011    -16.957      0.000      -0.215      -0.170
bachelor       0.4377      0.011     38.957      0.000       0.416       0.460
==============================================================================
Omnibus:                      316.790   Durbin-Watson:                   1.936
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              508.147
Skew:                          -0.375   Prob(JB):                    4.54e-111
Kurtosis:                       4.037   Cond. No.                         131.
==============================================================================

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

The OLS model would be:
ln(ahe) = 0.1495 + 0.7529 * **ln(age)** - 0.1924 * **female** + 0.4377 * **bachelor**
If age icreases from 25 to 26 (4%) earnings are expected to increase 2.9529274933105695%
If age icreases from 33 to 34 (3.03%) earnings are expected to increase 2.2476295955395076%

E8.2 - d

[6]:
```python
sampleD = ols("np.log(ahe) ~ age + age2 + female + bachelor", data = dataExcel).
 ↪fit()
print(sampleD.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            np.log(ahe)   R-squared:                       0.197
Model:                            OLS   Adj. R-squared:                  0.196
Method:                 Least Squares   F-statistic:                     455.2
Date:                Thu, 08 Oct 2020   Prob (F-statistic):               0.00
Time:                        17:43:54   Log-Likelihood:                -5065.1
No. Observations:                7440   AIC:                         1.014e+04
Df Residuals:                    7435   BIC:                         1.017e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.7919      0.670      1.182      0.237      -0.521       2.105
age            0.1040      0.046      2.280      0.023       0.015       0.193
age2          -0.0013      0.001     -1.722      0.085      -0.003       0.000
female        -0.1924      0.011    -16.961      0.000      -0.215      -0.170
```

```
bachelor         0.4374      0.011     38.928      0.000       0.415       0.459
==============================================================================
Omnibus:                      316.471   Durbin-Watson:                   1.935
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              507.649
Skew:                          -0.375   Prob(JB):                    5.83e-111
Kurtosis:                       4.037   Cond. No.                     1.09e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.09e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The OLS model would be:
ln(ahe) = 0.7919 + 0.1040 * **age** - 0.0013 * **age^2** - 0.1924 * **female** + 0.4374 * **bachelor**
If age icreases from 25 to 26 (one year) earnings are expected to increase by 0.10140% (0.1040 - 2 * 0.0013)
If age icreases from 33 to 34 (one year) earnings are expected to increase by 0.10140% (0.1040 - 2 * 0.0013)

E8.2 - e

No, I prefer the regression model of b since, in the regression model of c, *age* is being measured using percentages and it is more appropiate to use years rather than percentages to measure changes in the *age* variable. Besides, measuring the percentage change in *ahc* given an increase in years in unit terms is crystal clear for interpretation purposes.

E8.2 - f

I prefer the model of d, the quadratic model, since it captures some possible changes in the slope of *age*. In other words, it reflects, some possible peak in *ahe* given some *age* and it also records therefore a decreasing change in *ahe* as the *age* increases after the peak since it is a concave parabola.

E8.2 - g

I prefer again the regression of d, since again the model of c takes the log of *age* and it forces us to interpret changes in percentages rather than years and it can lead to confusing interpretations since measuring the age in percentage changes makes changes in *age* everytime smaller and therefore it gives everytime a smaller change in *ahe*. Besides the d model includes a quadratic term that can improve the interpretation as I explained in the **E8.2 - f**.

E8.2 - h

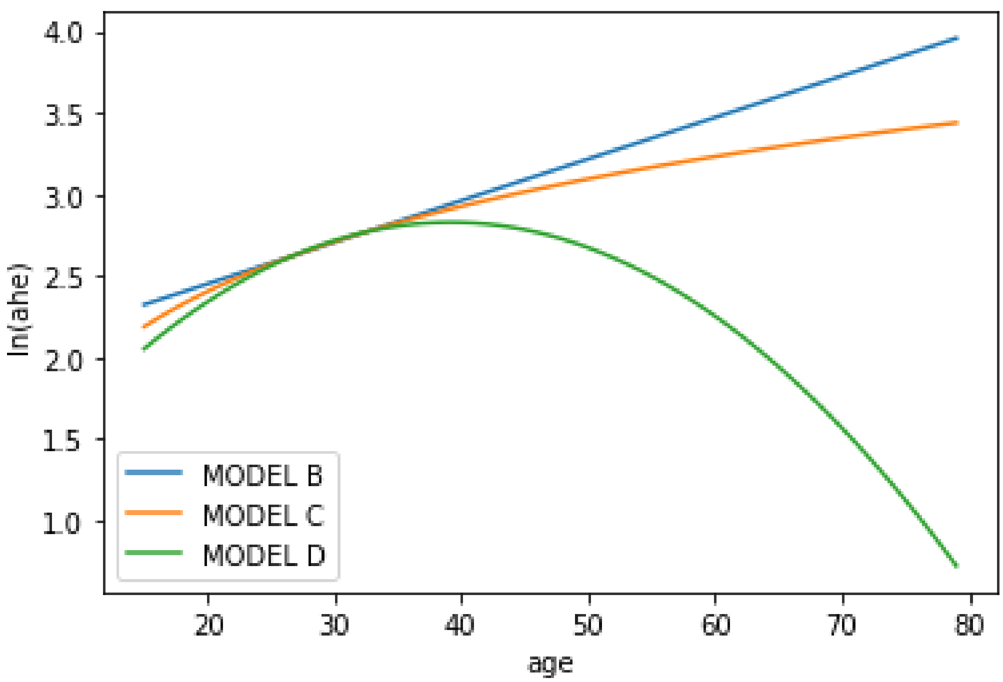
[13]:
```
sampleB1 = ols("np.log(ahe) ~ age + female + bachelor", data = dataExcel).fit()
plt.plot([x for x in range(15, 80)], [(1.941423 + 0.025518 * x) for x in␣
 ↪range(15, 80)], label = 'MODEL B')
sampleC1 = ols("np.log(ahe) ~ np.log(age) + female + bachelor", data =␣
 ↪dataExcel).fit()
```

```python
plt.plot([x for x in range(15, 80)], [(0.149532 + 0.752941 * np.log(x)) for x in range(15, 80)], label = 'MODEL C')
sampleD1 = ols("np.log(ahe) ~ age + age2 + female + bachelor", data = dataExcel).fit()
plt.plot([x for x in range(15, 80)], [(0.791882 + 0.104045 * x - 0.001328 * (x ** 2)) for x in range(15, 80)], label = 'MODEL D')
plt.legend(loc='lower left')
plt.xlabel('age')
plt.ylabel('ln(ahe)')
print(sampleB1.params, "\n\n", sampleC1.params, "\n\n", sampleD1.params)
```

```
Intercept    1.941423
age          0.025518
female      -0.192338
bachelor     0.437783
dtype: float64

 Intercept      0.149532
np.log(age)    0.752941
female        -0.192356
bachelor       0.437664
dtype: float64

 Intercept    0.791882
age          0.104045
age2        -0.001328
female      -0.192398
bachelor     0.437412
dtype: float64
```



The similarities between being a man and a woman is that they both have a decreasing AHE in the log - quadratic model and log - log model (MODEL C and MODEL D) and that they also have an additional AHE if they got a bachelor degree.
The differences is that women have a deprived AHE in comparisson to men, in other words, the models predict a coefficient regressor of -0.19 in all the three regression cases if the observation is a woman.

E8.2 - h . i

[8]:
```
sampleE = ols("np.log(ahe) ~ age + age2 + female + bachelor + female *␣
↪bachelor", data = dataExcel).fit()
print(sampleE.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            np.log(ahe)   R-squared:                       0.198
Model:                            OLS   Adj. R-squared:                  0.198
Method:                 Least Squares   F-statistic:                     367.9
Date:                Thu, 08 Oct 2020   Prob (F-statistic):               0.00
Time:                        17:43:55   Log-Likelihood:                -5057.4
No. Observations:                7440   AIC:                         1.013e+04
Df Residuals:                    7434   BIC:                         1.017e+04
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
```

```
===
                        coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---
Intercept             0.8037      0.669      1.201      0.230      -0.508
2.116
age                   0.1043      0.046      2.288      0.022       0.015
0.194
age2                 -0.0013      0.001     -1.728      0.084      -0.003
0.000
female               -0.2424      0.017    -14.249      0.000      -0.276
-0.209
bachelor              0.4004      0.015     27.370      0.000       0.372
0.429
female:bachelor       0.0899      0.023      3.940      0.000       0.045
0.135
==============================================================================
Omnibus:                      319.786   Durbin-Watson:                   1.933
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              511.678
Skew:                          -0.379   Prob(JB):                    7.77e-112
Kurtosis:                       4.038   Cond. No.                     1.09e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.09e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The interaction term shows that even though both women and men have an additional wage if they hold a bachelor degree and that being a men gives an additional AHE, being a woman with a bachelor degree gives an additional AHE in comparisson with men. **Alexis:** 0.8037 + 0.1043 * 30 - 0.0013 * 900 - 0.2424 * 1 + 0.4004 + 0.0899 = **3.0106 = ln(AHE)**
Alexis AHE = 20.299576017658516
Jane: 0.8037 + 0.1043 * 30 - 0.0013 * 900 - 0.2424 * 1 = **2.5203 ln(AHE)**
Jane AHE = 12.4323258019194
The difference between Bob and Jim earning is 7.801217983769728**

Bob: 0.8037 + 0.1043 * 30 - 0.0013 * 900 + 0.4004 = **3.1631 = ln(AHE) Bob AHE = 23.643778150284366**
Jim: 0.8037 + 0.1043 * 30 - 0.0013 * 900 = **2.7627 = ln(AHE)**
Jim AHE = 15.842560166514637
The difference between Bob and Jim earning is 7.801217983769728

E8.2 - j

[9]:
```
sampleJ = ols("ahe ~ age + female + (female * age)", data = dataExcel).fit()
print(sampleJ.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ahe   R-squared:                       0.033
Model:                            OLS   Adj. R-squared:                  0.033
Method:                 Least Squares   F-statistic:                     85.34
Date:                Thu, 08 Oct 2020   Prob (F-statistic):           2.76e-54
Time:                        17:43:55   Log-Likelihood:                -28056.
No. Observations:                7440   AIC:                         5.612e+04
Df Residuals:                    7436   BIC:                         5.615e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.3052      1.683      1.963      0.050       0.005       6.605
age            0.5929      0.056     10.503      0.000       0.482       0.704
female         3.6086      2.587      1.395      0.163      -1.462       8.679
female:age    -0.2082      0.087     -2.397      0.017      -0.379      -0.038
==============================================================================
Omnibus:                     1912.305   Durbin-Watson:                   1.845
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5016.515
Skew:                           1.384   Prob(JB):                         0.00
Kurtosis:                       5.920   Cond. No.                         775.
==============================================================================

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

The model I would use is:
ahe = 3.3052 + 0.5929 * **age** + 3.6086 * **female** - 0.2082 * **(female * age)**
This what basically states is that even though being a female gives apareantly an additional *ahe*, the increase of the *ahe* regarding a change in the *age* is higher for men than for women as is shown in the *female* age* variable.

E8.2 - k

[10]:
```
sampleK = ols("ahe ~ age + bachelor + (bachelor * age)", data = dataExcel).fit()
print(sampleK.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    ahe   R-squared:                       0.151
Model:                            OLS   Adj. R-squared:                  0.151
```

```
Method:                 Least Squares   F-statistic:                     440.6
Date:                Thu, 08 Oct 2020   Prob (F-statistic):          1.66e-263
Time:                        17:43:55   Log-Likelihood:                -27573.
No. Observations:                7440   AIC:                         5.515e+04
Df Residuals:                    7436   BIC:                         5.518e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        4.3139      1.735      2.486      0.013       0.912       7.715
age              0.3833      0.058      6.584      0.000       0.269       0.497
bachelor         0.0202      2.398      0.008      0.993      -4.681       4.721
bachelor:age     0.2610      0.081      3.241      0.001       0.103       0.419
==============================================================================
Omnibus:                     2042.778   Durbin-Watson:                   1.942
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6342.505
Skew:                           1.405   Prob(JB):                         0.00
Kurtosis:                       6.545   Cond. No.                         840.
==============================================================================

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

The model I would use is:
ahe = 4.3139 + 0.3833 * **age** + 0.0202 * **bachelor** - 0.2610 * **(bachelor * age)**
This what basically states is that just the fact of having a bachelor degree gives you an additional increase in the *age*, besides that, a change in the *age* gives a higher change in *ahe* for those who hold a bachelor degree than for those who do not have a bachelor degree.

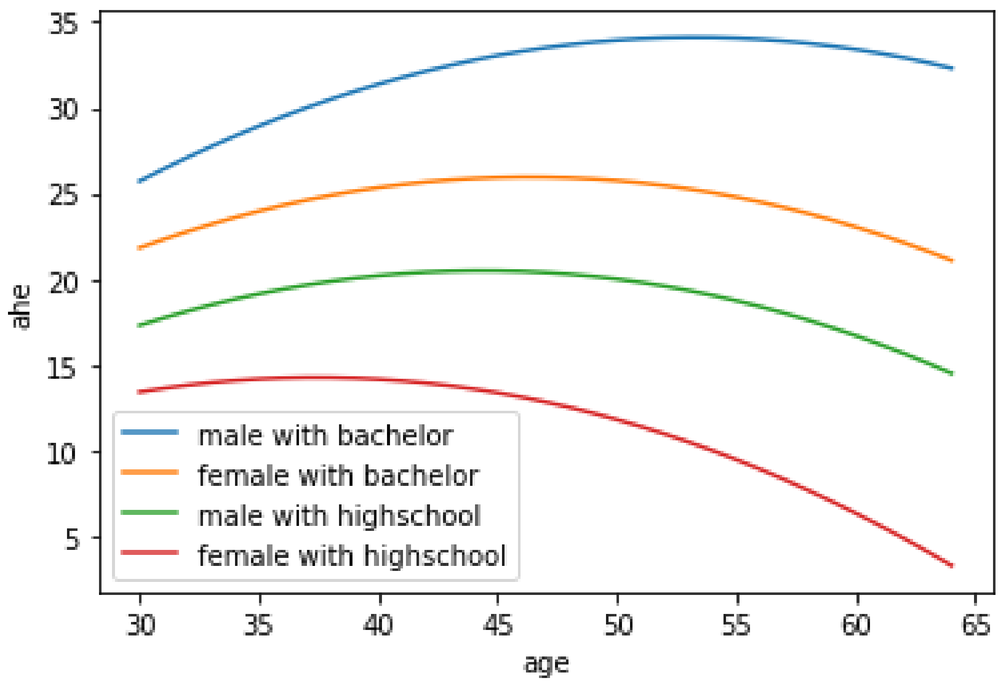
E8.2 - l

[11]:
```
sampleL = ols("ahe ~ age + age2 + female + bachelor + (age * female) + (age *␣
 ↪bachelor)", data = dataExcel).fit()
print(sampleL.params)
plt.plot([x for x in range(30, 65)], [(-9.9924 + 1.3750 * x -0.0155 * x ** 2 +␣
 ↪0.122186 + 0.276004 * x) for x in range(30, 65)], label="male with bachelor")
plt.plot([x for x in range(30, 65)], [(-9.9924 + 1.3750 * x -0.0155 * x ** 2 +␣
 ↪2.585777 +  0.122186 - 0.215476 * x +  0.276004 * x) for x in range(30,␣
 ↪65)], label="female with bachelor")
plt.plot([x for x in range(30, 65)], [(-9.9924 + 1.3750 * x -0.0155 * x ** 2)␣
 ↪for x in range(30, 65)], label="male with highschool")
plt.plot([x for x in range(30, 65)], [(-9.9924 + 1.3750 * x -0.0155 * x ** 2 +␣
 ↪2.585777 - 0.215476 * x) for x in range(30, 65)], label="female with␣
 ↪highschool")
plt.legend(loc='lower left')
plt.xlabel('age')
```

```
plt.ylabel('ahe')
```

```
Intercept      -9.992386
age             1.375001
age2           -0.015531
female          2.585777
bachelor        0.122186
age:female     -0.215476
age:bachelor    0.276004
dtype: float64
```

[11]: Text(0, 0.5, 'ahe')



This new model I propose basically summarizes everything we have done so far and it is:

ahe = - 9.992386 + 1.375001 * **age** - 0.015531 * **age2** + 2.585777 * **female** + 0.122186 * **bachelor** - 0.215476 * **(age * female)** + 0.276004 * **(age * bachelor)**
(I think collinearity is not a problem because the duplicate variables as **age** and **age2** or **age** and **bachelor * age** are exactly the same and therefore changes in one of the duplicates also applies for the second variable, in other words, they do not have collinearity by chance but because the variables are the same)

As we see in the plot, male with a bachelor degree is the combination that gives the maximum expected *ahe* and it is also the one that expects the biggest *ahe* change given one more *year* before peaking the expected maximum.
Just behind the male with a bachelor degree we find the female with a bachelor degree being the

second combination that expects the second highest *ahe*.
It is also interesting that a change in *age* for a female with a bachelor degree expects almost the same change in the *ahe* for a male with a highschool.
Just behind female with a bachelor degree, a male with highschool expects the third biggest *ahe*.
Finally a female with highschool expects the worst *ahe* in comparisson with the remaining combinations.