Nhu Vo

Lab #4

1. Run a simple regression, with at least two Xs in it, and interpret your results. (Did the results fit your expectations? Why? Why not?)

2. Add an interaction term to that model that you think might moderate the original relationship between X1 and X2. Explain why you think an interaction might be present and in what direction it would work. Explain your results. Did it work out? Yes? No?

3. Extra Credit: Plot the relationship found in the interaction.

```
from __future__ import division
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import os
import matplotlib.pyplot as plt
```

I'm using the ARDA Data Archive again

```
from google.colab import files
uploaded = files.upload()

import io
g = pd.read_csv(io.BytesIO(uploaded['GSS.2006.csv']))

g.head()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving GSS.2006.csv to GSS.2006.csv
```

	vpsu	vstrat	adults	ballot	dateintv	famgen	form	formwt	gender1	hompop	...	away7	gender14	old14	relate
0	1	1957	1	3	316	2	1	1	2	3	...	NaN	NaN	NaN	N
1	1	1957	2	2	630	1	2	1	2	2	...	NaN	NaN	NaN	N
2	1	1957	2	2	314	2	1	1	2	2	...	NaN	NaN	NaN	N
3	1	1957	1	1	313	1	2	1	2	1	...	NaN	NaN	NaN	N
4	1	1957	3	1	322	2	2	1	2	3	...	NaN	NaN	NaN	N

5 rows × 1261 columns

1. Run a simple regression, and interpret your results. (Did the results fit your expectations? Why? Why not?)

Does one's race (White, Black and Others) and their spouse' level of education (less than highschool, high school, junior college, bachelor, and graduate) affect how many children they ever have (0-5+)? I expect that the less educated their spouse is, the more children they're going to have.

Independent Variable 1: RACE

1) White

2) Black

3) Other

Independent Variable 2: SPDEG

0) Less than high school

1) High school

2) Junior college

3) Bachelor

4) Graduate

Dependent Variable: CHILDS

0) 0

1) 1

2) 2

3) 3-4

4) 5+

Here is the model:

```
lm1 = smf.ols(formula = 'childs ~ spdeg + C(race)', data = g). fit()
print (lm1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 childs   R-squared:                       0.042
Model:                            OLS   Adj. R-squared:                  0.040
Method:                 Least Squares   F-statistic:                     20.56
Date:                Fri, 23 Jun 2023   Prob (F-statistic):           4.88e-13
Time:                        22:57:20   Log-Likelihood:                -2532.1
No. Observations:                1398   AIC:                             5072.
Df Residuals:                    1394   BIC:                             5093.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        2.5249      0.072     34.901      0.000       2.383       2.667
C(race)[T.2]     0.3819      0.150      2.541      0.011       0.087       0.677
C(race)[T.3]     0.1899      0.116      1.630      0.103      -0.039       0.418
spdeg           -0.2169      0.032     -6.715      0.000      -0.280      -0.154
==============================================================================
Omnibus:                      135.189   Durbin-Watson:                   1.832
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              192.262
Skew:                           0.741   Prob(JB):                     1.78e-42
Kurtosis:                       4.050   Cond. No.                         8.65
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

intercept = 2.5249

The intercept represents the estimated mean value of the dependent variable (childs) when all independent variables are zero.

C(race)[T.2] coef = 0.3819

compared to people of White race, being Black is associated with an increase of 0.3819 in number of children, holding other variables constant.

C(race)[T.3] coef = 0.1899

compared to people of White race, being Others race is associated with an increase of 0.18 in number of children, holding other variables constant. However, this coefficient is not statistically significant at the 0.05 significance level (P>|t| = 0.103).

spdeg = -0.2169

The coefficient for spdeg indicates that a one-unit increase in the independent variable spdeg is associated with a decrease of 0.2169 in the average value of the dependent variable number of chidlren, holding other variables constant. This coefficient is statistically significant (P>|t| < 0.001). This results fit my expectation because the higher the spouse education, the lower the number of children, ie. they're more educated on contraceptives

2. Add an interaction term to that model that you think might moderate the original relationship between X1 and X2. Explain why you think an interaction might be present and in what direction it would work. Explain your results. Did it work out? Yes? No?

```
lm2 = smf.ols(formula = 'childs ~ spdeg * C(race)', data = g). fit()
print (lm2.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                 childs   R-squared:                       0.044
Model:                            OLS   Adj. R-squared:                  0.041
Method:                 Least Squares   F-statistic:                     12.84
Date:                Fri, 23 Jun 2023   Prob (F-statistic):           3.14e-12
Time:                        22:57:25   Log-Likelihood:                -2530.8
No. Observations:                1398   AIC:                             5074.
Df Residuals:                    1392   BIC:                             5105.
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept                2.4801      0.078     31.930      0.000       2.328       2.633
C(race)[T.2]             0.5389      0.265      2.033      0.042       0.019       1.059
C(race)[T.3]             0.3554      0.160      2.219      0.027       0.041       0.670
spdeg                   -0.1915      0.036     -5.305      0.000      -0.262      -0.121
spdeg:C(race)[T.2]      -0.1133      0.168     -0.675      0.500      -0.443       0.216
spdeg:C(race)[T.3]      -0.1312      0.088     -1.486      0.138      -0.305       0.042
==============================================================================
Omnibus:                      133.279   Durbin-Watson:                   1.838
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              188.987
Skew:                           0.734   Prob(JB):                     9.16e-42
Kurtosis:                       4.043   Cond. No.                         17.2
==============================================================================

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

⊛ intercept (when all the X's = 0) = 2.4801

the number of children a White individual would have if their spouse education = 0

⊛ C(race)[T.2] coef = 0.5389

- individuals who are Black would have 0.54 more children than a White individual when their spouse's education = 0

⊛ C(race)[T.3] coeff = 0.3554

- individuals of Other races category would have 0.35 more children than a White individual when their spouse's education = 0

⊛ spdeg = -0.1915 (Slope for White individuals)

- For individuals of White race, as their spouse education degree increase, they have 0.19 less children

⊛ spdeg:C(race)[T.2] = -0.1133 (Slope for Black individuals).

- This is an interaction term between spdeg and Black.

- The coefficient of -0.1133 suggests that the effect of spdeg on the dependent variable is modified when race = 2.

- For individuals who are Black, their number of children goes down by 0.11 units more steeply than their White counterparts for every degree level their spouse attained.

⊛ spdeg: C(race)[T.3] = -0.1312 (Slope for individuals of Other race)

- This is an interaction term between spdeg and Other race.

- The coefficient of -0.1312 suggests that the effect of spdeg on the dependent variable is modified when race = 3.

- For individuals of Other race, their number of children goes down by 0.13 units more steeply than their White counterparts for every degree level their spouse attained .

⊛ Interaction Analysis

- The last 2 bottoms are the interactions (spdeg:C(race)[T.2] & spdeg:C(race)[T.3]).

- The relationship is not statistically significant. This means that adding the interaction between the 2 independent variables did not provide any meaningful relationship.

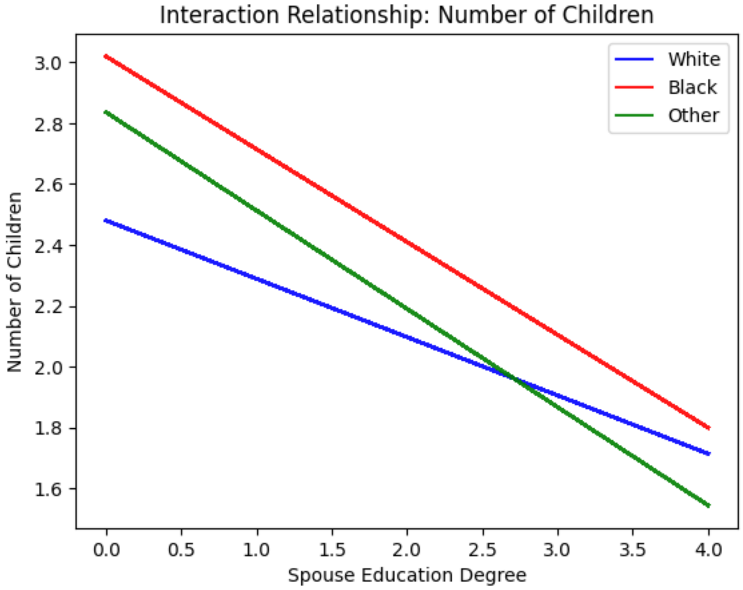
3. Extra Credit: Plot the relationship found in the interaction.

```
import matplotlib.pyplot as plt

# Plotting the interaction relationship
plt.plot(g["spdeg"], lm2.params[0] + lm2.params[3] * g["spdeg"], 'b', label='White', alpha=0.9)
plt.plot(g["spdeg"], lm2.params[0] + lm2.params[1] + (lm2.params[3] + lm2.params[4]) * g["spdeg"], 'r', label='Black', alpha=0.9)
plt.plot(g["spdeg"], lm2.params[0] + lm2.params[2] + (lm2.params[3] + lm2.params[5]) * g["spdeg"], 'g', label='Other', alpha=0.9)

# Adding labels and title to the plot
plt.title("Interaction Relationship: Number of Children")
plt.xlabel("Spouse Education Degree")  # Updated x-axis label
plt.ylabel("Number of Children")

# Displaying the legend and plot
plt.legend()
plt.show()
```



This plot shows what I predicted initially, despite not being statistically significant in the interaction model lm2. For all of the 3 races category, the higher the spouses' education degree, the less children they have

● ×