ps5

July 15, 2023

[2]:
```
# Initialize Otter
import otter
grader = otter.Notebook("ps5.ipynb")
```

1 Econ 140 – Problem Set 5

Before getting started on the assignment, run the cell at the very top that imports `otter` and the cell below which will import the packages we need.

Important: As mentioned in problem set 0, if you leave this notebook alone for a while and come back, to save memory datahub will "forget" which code cells you have run, and you may need to restart your kernel and run all of the cells from the top. That includes this code cell that imports packages. If you get `<something> not defined` errors, this is because you didn't run an earlier code cell that you needed to run. It might be this cell or the `otter` cell above.

[3]:
```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

1.1 Problem 1. Instrumental Variable Estimation

Consumption of gasoline is a critical component of household expenditures, and increasingly, it is the focus intense public policy debate given the concern over greenhouse emissions. For these reasons alone economists would like to find accurate estimates of price elasticity of demand for gasoline by American consumers. The data file `gasoline.csv` contains monthly data on U.S. consumption of gasoline from 1978 to 2002.

[4]:
```
gas = pd.read_csv("gasoline.csv")
gas.head()
```

[4]:

	obs	carsales	persincome	pricegas	quantgas	transindex
0	1978:01	10.070	1756	64.8	6681.0	59.6
1	1978:02	10.450	1756	64.7	6876.0	59.7
2	1978:03	10.953	1756	64.7	7255.0	59.9
3	1978:04	11.786	1821	64.9	7202.0	60.3

```
4  1978:05    11.804        1821      65.5    7724.0        61.0
```

Question 1.a. Estimate a simple linear demand equation by regressing the quantity of gas `quantgas` consumed on the price of a gallon of gas `pricegas`. What is your estimate of the price coefficient from the OLS estimation? Remember to use robust standard errors, and to always include a constant.

[5]:
```python
xparta=sm.add_constant(gas['pricegas'])
yparta= gas['quantgas']

modelparta= sm.OLS(yparta,xparta)

resultparta= modelparta.fit(cov_type='HC1')

resultparta.summary()
```

[5]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               quantgas   R-squared:                       0.046
Model:                            OLS   Adj. R-squared:                  0.043
Method:                 Least Squares   F-statistic:                     13.84
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           0.000239
Time:                        15:04:36   Log-Likelihood:                -2356.4
No. Observations:                 296   AIC:                             4717.
Df Residuals:                     294   BIC:                             4724.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       6531.8301    223.281     29.254      0.000    6094.208    6969.453
pricegas       7.8252      2.104      3.720      0.000       3.702      11.948
==============================================================================
Omnibus:                       11.752   Durbin-Watson:                   0.191
Prob(Omnibus):                  0.003   Jarque-Bera (JB):                5.598
Skew:                           0.045   Prob(JB):                       0.0609
Kurtosis:                       2.332   Cond. No.                         696.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
"""
```

Question 1.b. Use your OLSEs to express the price elasticity of demand evaluated at the average price of gas. Does it make economic sense?

Hint: Express the price elasticity when demand is linear.

[6]:
```python
averageg= gas['pricegas'].mean()
Elasticity= (averageg*7.825)/((6531.8301)+averageg*7.825)
Elasticity
```

[6]: 0.1209758756999421

To compute the price elasticity we find the average gas price and use the price coefficient and constant to apply the Elasticity forumla. Which is simply the change in quantity demanded over the change in price. This results in a value of about .12

This means a 1% increase in price leads to a .12 decrease in quantity demanded. This makes gas inelastic. Economically one may expect a bigger effect of demand due to an increase in price. But as shown gas is an inelastic good meaning that if price increases people are not turned away from buying it. Perhaps because they need gas in their vehicles and they have no other alternatives.

Question 1.c. Now introduce per capita personal income `persincome` as a regressor in the linear demand model and re-estimate using OLS. How has your estimate of price coefficient changed?

This question is for your code, the next is for your explanation.

[7]:
```python
xpartc=sm.add_constant(gas[['pricegas','persincome']])
modelpartc=sm.OLS(gas['quantgas'], xpartc)
resultspartc=modelpartc.fit(cov_type='HC1')
resultspartc.summary()
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               quantgas   R-squared:                       0.759
Model:                            OLS   Adj. R-squared:                  0.757
Method:                 Least Squares   F-statistic:                     520.9
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           3.32e-97
Time:                        15:04:36   Log-Likelihood:                -2152.8
No. Observations:                 296   AIC:                             4312.
Df Residuals:                     293   BIC:                             4323.
Df Model:                           2
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       6632.9609    168.570     39.348      0.000    6302.569    6963.352
pricegas      -6.8606      1.361     -5.041      0.000      -9.528      -4.193
persincome     0.3188      0.010     32.050      0.000       0.299       0.338
==============================================================================
Omnibus:                        2.611   Durbin-Watson:                   0.757
Prob(Omnibus):                  0.271   Jarque-Bera (JB):                2.432
```

```
Skew:                           0.127   Prob(JB):                        0.296
Kurtosis:                       3.364   Cond. No.                     3.22e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 3.22e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 1.d. Explain.

How has your estimate of price coefficient changed?

My estimate of price coefficient has changed a lot, significantly and is now negative. From about 7 to negative 6.

Question 1.e. Do you think that the above regression suffers from omitted variable bias? If so, can you determine the sign of the bias?

Yes I do think it suffers from OMB. Because when the variable persincome is included it decreases the relevance of the original regressor. Moreover, the r squared value informs us that this regression with the new regressor included is a better fit. Meaning that it explains a greater percentage of the data. So prior, the persincome was correlated with the error term when it was omitted giving bias to orignal regressor and a worse fit overall to the regression.

Question 1.f. Give reasons why you should suspect that the gasoline price would be correlated with error term even after you introduced personal income into the regression. Evaluate the monthly sales of autos in the U.S. (carsales) serve as a good instrument for price of gas? Explain.

Possible reasons lie in the variables that we have data for but were not included in the prior regressions. For example, carsales. As long as the correlation/covariance between the explanatory variable is not zero, when not included price would be correalted with the error term. In regards to car sales, specifically when more cars are sold, the demand for gas goes up to use those cars. Thus, price would increase. Cars and gasoline are complimentary goods.

Question 1.g. Estimate the first stage of a two stage least squares estimation by regressing price of gasoline on the sales of cars. Also include personal income. Perform a test that determines whether car sales is a "strong instrument."

This question is for your code, the next is for your explanation.

[8]:
```
xpartg=sm.add_constant(gas[['carsales','persincome']])
modelpartg=sm.OLS(gas['pricegas'], xpartg)
resultspartg=modelpartg.fit(cov_type='HC1')
resultspartg.summary()
```

[8]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
```

```
Dep. Variable:               pricegas   R-squared:                       0.308
Model:                            OLS   Adj. R-squared:                  0.303
Method:                 Least Squares   F-statistic:                     43.63
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           2.61e-17
Time:                        15:04:36   Log-Likelihood:                -1245.0
No. Observations:                 296   AIC:                             2496.
Df Residuals:                     293   BIC:                             2507.
Df Model:                           2
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        162.2362     10.132     16.013      0.000     142.378     182.094
carsales      -6.3378      0.957     -6.624      0.000      -8.213      -4.463
persincome     0.0023      0.001      3.788      0.000       0.001       0.003
==============================================================================
Omnibus:                       10.733   Durbin-Watson:                   0.181
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                6.829
Skew:                           0.220   Prob(JB):                       0.0329
Kurtosis:                       2.400   Cond. No.                     5.54e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 5.54e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[9]:

```
resultspartg.f_test("carsales").summary()
```

[9]:

```
'<F test: F=43.8833645094151, p=1.666268858145298e-10, df_denom=293, df_num=1>'
```

Question 1.h. Explain.

The strong instrument test was an F test. The resulting F stat value was 43 which is considerably large or big. This suggests that car sales is a good instrueмtnt to use in regards to its strength.

Question 1.i. Can you suggest another instrument that is likely to be a better instrument than car sales?

A better insturment has to have a strong association with gas price (x) but not the quantity of gas consumed(y). A possible instrument could be an event such as a trade agreement betweeen oil producing countries. This would affect the price of gas but not directly the amount it would be demanded/consumed.

Question 1.j. Now perform the second stage of the TSLS estimation and report any change in the size of the coefficient on gasoline price as a result of using the instrumental variable.

Hint: `results.fittedvalues` *will give you an array of the $\hat{y}$ values.*

This question is for your code, the next is for your explanation.

[10]:
```
gas['pricegas_hat'] = resultspartg.fittedvalues
xpartj=sm.add_constant(gas[['pricegas_hat','persincome']])

modelpartj=sm.OLS(gas['quantgas'], xpartj)

resultspartj=modelpartj.fit(cov_type='HC1')

resultspartj.summary()
```

[10]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               quantgas   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     415.2
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           3.11e-86
Time:                        15:04:36   Log-Likelihood:                -2157.8
No. Observations:                 296   AIC:                             4322.
Df Residuals:                     293   BIC:                             4333.
Df Model:                           2
Covariance Type:                  HC1
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         7399.7376    402.835     18.369      0.000    6610.196    8189.279
pricegas_hat   -14.9491      3.923     -3.810      0.000     -22.639      -7.260
persincome       0.3515      0.016     21.871      0.000       0.320       0.383
==============================================================================
Omnibus:                        7.832   Durbin-Watson:                   0.794
Prob(Omnibus):                  0.020   Jarque-Bera (JB):                9.249
Skew:                           0.255   Prob(JB):                      0.00981
Kurtosis:                       3.700   Cond. No.                     7.63e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 7.63e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 1.k. Explain.

The coefficient for the price of gas is now about -14. Which is way more than it was previously at around negative 6. The R squared is also higher which indicates better fit.

Question 1.l. Is the TSLS estimate of the price coefficient statistically significant? Do you have any reason to doubt the reported values of the standard errors from the second stage? Explain.

The very low p value of 0 for the price coeffficient indicates that it is statistically significant. However, in a two stage least squares there can be suspicion about the standard errors because the estimator is biased but consistnet. So it hurts the overall accuracy of the TSLS

Question 1.m. Suppose you were instead interested in studying how the supply of gas is influenced by its price. Would you feel comfortable regressing the quantity of gas produced on its price? Why?

Yes I would feel comfortable because I think there is validity in the economic assumption of price and gas quantity. For example, at a higher price it would be belieived economically, that more gas would be produced. Moreover, this test would be ok because the OLS assumptions are satisfied.

Question 1.n. Also included in the dataset is the BLS monthly price index for consumer purchases of "transportation services" over the same sample period `transindex`. Perform TSLS estimation using this price index as an instrument. Evaluate the results of the first and second stages.

This question is for your code, the next is for your explanation.

[11]:
```python
##########FIRST



xpartn=sm.add_constant(gas[['transindex','persincome']])

modelpartn=sm.OLS(gas['pricegas'], xpartn)

resultspartn= modelpartn.fit(cov_type='HC1')
resultspartn.summary()
```

[11]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               pricegas   R-squared:                       0.342
Model:                            OLS   Adj. R-squared:                  0.338
Method:                 Least Squares   F-statistic:                     99.50
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           1.06e-33
Time:                        15:04:36   Log-Likelihood:                -1237.5
No. Observations:                 296   AIC:                             2481.
Df Residuals:                     293   BIC:                             2492.
Df Model:                           2
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         29.3495      6.830      4.297      0.000      15.964      42.735
transindex     1.1001      0.113      9.741      0.000       0.879       1.321
persincome    -0.0088      0.002     -5.704      0.000      -0.012      -0.006
==============================================================================
Omnibus:                       32.446   Durbin-Watson:                   0.051
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.866
Skew:                           0.648   Prob(JB):                     1.47e-06
Kurtosis:                       2.294   Cond. No.                     4.79e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 4.79e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[12]:

```python
#####SECOND

gas['pricegastrans_hat']=resultspartn.fittedvalues
x_1n2=sm.add_constant(gas[['pricegastrans_hat','persincome']])
model_1n2=sm.OLS(gas['quantgas'], x_1n2)
results_1n2=model_1n2.fit(cov_type='HC1')
results_1n2.summary()
```

[12]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               quantgas   R-squared:                       0.822
Model:                            OLS   Adj. R-squared:                  0.821
Method:                 Least Squares   F-statistic:                     628.6
Date:                Sat, 15 Jul 2023   Prob (F-statistic):          1.01e-106
Time:                        15:04:36   Log-Likelihood:                -2107.7
No. Observations:                 296   AIC:                             4221.
Df Residuals:                     293   BIC:                             4233.
Df Model:                           2
Covariance Type:                  HC1
================================================================================
=====
                        coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-----
const              8632.8412    269.222     32.066      0.000    8105.175
9160.507
pricegastrans_hat   -27.9567      2.730    -10.241      0.000     -33.307
-22.606
persincome            0.4040      0.014     29.901      0.000       0.378
0.430
==============================================================================
Omnibus:                        3.557   Durbin-Watson:                   1.052
Prob(Omnibus):                  0.169   Jarque-Bera (JB):                3.278
```

```
Skew:                          -0.248   Prob(JB):                        0.194
Kurtosis:                       3.140   Cond. No.                     6.77e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 6.77e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 1.o. Explain.

In the first stage, transindex is significant with a very low p value. With a coefficient value of around 1. The second stage coefficeint value for pricegastrans_hat is -27.9567 which is much more drastic (more negative) than for the values of the regressions before.

Question 1.p. Assume that you are told that at least one of the instruments above is not exogenous (it could be both). Based on your empirical results using these data, decide what you consider the "best" estimate of the price coefficient. It doesn't have to be one of the above instruments. Explain your reasoning.

The two instruments are'pricegastrans_hat','persincome' Based on the empirical data, I think it would be persincome since it has a lesser effect on the x variable. Moreover, logically I think persincome can be more at risk to be correalted with the error term than the price index in regards to the qunantity of gas.

1.2 Problem 2. Experiments

Senior management at Ctrip, China's largest travel agency, is interested in allowing their Shanghai call center employees to work from home (telecommute). Allowing telecommuting may not only reduce office rental costs but it may also lower the high attrition rates the firm was experiencing by saving the employees from long commutes. However, management is also worried that employees may be less productive if they telecommute. To determine the effects of telecommuting on productivity, Ctrip decided to run an experiment wherein participants were allowed to work from home for several days over a 9 month period. They asked employees in the airfare and hotel departments whether they would be interested in volunteering for this experiment, and not all employees agreed to participate. Each employee who volunteered for the experiment was then assigned a random share of work days over the 9 months that they must work from home. The file `ctrip.csv` contains data from all 994 employees of Ctrip.

Variable	Description	Units
personid	person ID	
age	age	years
tenure	tenure at Ctrip	months
grosswage	monthly gross salary	1000s of CNY
children	whether person has children	
bedroom	whether person has independent bedroom to work in	

Variable	Description	Units
commute	daily commute in minutes	minutes
men	whether person is male	
married	whether person is married	
volunteer	whether person volunteers for experiment (work from home)	
high_educ	tertiary education and above	
WFHShare	share of work days worked from home during experiment	
calls	average number of calls taken per week during experiment	

[13]:
```
ctrip = pd.read_csv("ctrip.csv")
ctrip.head()
```

[13]:
```
   personid   age  tenure  grosswage children bedroom  commute  men  married  \
0      3224  30.0   113.0   3.824882       no      no     40.0  1.0      0.0
1      3906  33.0    96.0   2.737547      yes     yes    180.0  0.0      1.0
2      4118  31.0    94.0   3.460380      yes      no    180.0  0.0      1.0
3      4122  30.0    94.0   4.096246       no      no    180.0  0.0      0.0
4      4164  28.0    25.0   7.253200       no     yes     65.0  0.0      1.0

   volunteer  high_educ  WFHShare  calls
0        0.0        0.0       NaN    NaN
1        1.0        0.0       0.0  342.0
2        0.0        1.0       NaN    NaN
3        0.0        0.0       NaN    NaN
4        1.0        1.0       0.0  172.0
```

Question 2.a. What percentage of employees volunteered to participate in the experiment?

Hint: Check out the `Series.value_counts()` function.

[14]:
```
counts=ctrip['volunteer'].value_counts()
percent= (counts/counts.sum())*100
percent
print(50.603622)
```

```
50.603622
```

Question 2.b.i. Use the variables `commute` as a dependent variable in a bivariate linear regression where `volunteer` is the explanatory variable.

[15]:
```
x_2bi=sm.add_constant(ctrip['volunteer'])
model_2bi=sm.OLS(ctrip['commute'], x_2bi)
results_2bi=model_2bi.fit(cov_type='HC1')
results_2bi.summary()
```

[15]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
```

```
==============================================================================
Dep. Variable:                commute   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                  0.010
Method:                 Least Squares   F-statistic:                     11.46
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           0.000739
Time:                        15:04:36   Log-Likelihood:                -5413.0
No. Observations:                 994   AIC:                         1.083e+04
Df Residuals:                     992   BIC:                         1.084e+04
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         74.4656      2.316     32.152      0.000      69.926      79.005
volunteer     12.0318      3.554      3.385      0.001       5.066      18.998
==============================================================================
Omnibus:                      122.652   Durbin-Watson:                   1.591
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              167.975
Skew:                           0.993   Prob(JB):                     3.35e-37
Kurtosis:                       3.331   Cond. No.                         2.63
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
"""
```

Question 2.b.ii. Interpret the coefficient on `volunteer` and comment on its statistical significance.

The volunteer variables shows whether person volunteers for experiment (work from home). It is a dummy variable of value zero or 1. The ceofficient value is 12.0318. This means that if a person is an employee and volunteer then their commute time is expected to increase on average about 12 minutes. Its p value of very close to zero indicates that it is statistically significant.

Question 2.c.i. Use the variable `tenure` as a dependent variable in a bivariate linear regression where `volunteer` is the explanatory variable.

[16]:
```python
x_2ci=sm.add_constant(ctrip['volunteer'])
model_2ci=sm.OLS(ctrip['tenure'], x_2ci)
results_2ci=model_2ci.fit(cov_type='HC1')
results_2ci.summary()
```

[16]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 tenure   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.006
Method:                 Least Squares   F-statistic:                     7.451
```

```
Date:                  Sat, 15 Jul 2023   Prob (F-statistic):            0.00645
Time:                          15:04:36   Log-Likelihood:                -4431.3
No. Observations:                   994   AIC:                             8867.
Df Residuals:                       992   BIC:                             8876.
Df Model:                             1
Covariance Type:                    HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         26.8422      0.972     27.624      0.000      24.938      28.747
volunteer     -3.6235      1.327     -2.730      0.006      -6.225      -1.022
==============================================================================
Omnibus:                     97.416   Durbin-Watson:                   0.099
Prob(Omnibus):                0.000   Jarque-Bera (JB):              124.805
Skew:                         0.856   Prob(JB):                     7.93e-28
Kurtosis:                     3.292   Cond. No.                         2.63
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
"""
```

Question 2.c.ii. Interpret the coefficient on `volunteer` and comment on its statistical significance.

The coefficient value for volunteer is -3.6. Its p value is very close to zero, making it statistically significant. volunteer is whether person volunteers for experiment (work from home)

If a person is an employee, then their tenure decreases on average by 3.62 years.

Question 2.d.i. Impressed by your recent econometrics training, Ctrip hires you as a consultant to analyze the results from their experiment. To begin with, you estimate a bivariate linear regression model of the productivity of workers, measured by the log of the average number of calls taken per week (call this variable `ln_calls`), on the variable `WFHShare` (work from home share).

Hint: Add the argument `missing='drop'` when constructing your OLS model to drop the missing entries.

[17]:
```
x_2di=sm.add_constant(ctrip['WFHShare'])
ctrip['ln_calls']=np.log(ctrip['calls'])
model_2di=sm.OLS(ctrip['ln_calls'], x_2di, missing='drop')
results_2di=model_2di.fit(cov_type='HC1')
results_2di.summary()
```

[17]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                ln_calls   R-squared:                       0.163
Model:                             OLS   Adj. R-squared:                  0.161
```

```
Method:                 Least Squares   F-statistic:                     142.6
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           4.23e-29
Time:                        15:04:36   Log-Likelihood:                -517.61
No. Observations:                 503   AIC:                             1039.
Df Residuals:                     501   BIC:                             1048.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.4442      0.062     87.180      0.000       5.322       5.567
WFHShare       0.9753      0.082     11.942      0.000       0.815       1.135
==============================================================================
Omnibus:                      396.980   Durbin-Watson:                   1.820
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8757.799
Skew:                          -3.269   Prob(JB):                         0.00
Kurtosis:                      22.368   Cond. No.                         4.07
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
"""
```

Question 2.d.ii. Interpret the regression coefficient on `WFHShare` in words. Is the effect statistically significant?

The WFHShare coefficient value is 0.9753. Since we took the log for dependent variable, this means that a 1% increase in WFH share would increase the average number of calls by .9753 %. The p value is zero making this coefficient statistacally significant.

Question 2.e. Has the Ctrip company achieved the ideal of a randomized controlled experiement, so that we can view the estimated effects of working from home on productivity in causal terms?

We cannot conclude causality. The people who worked from volunteered to do so. Thus, the treatment was not randomly assigned. This creates selection bias. Therefore, it is not the ideal of a randomized controlled experiement because it is not random at all.

Question 2.g.i. Create a dummy variable called `longcommute` which is equal to one if the employee has a commute of greater than or equal to 120 (i.e. 2 hours) and add it to the `ctrip` column.

Hint: First create a boolean column for `longcommute` then cast it into integers using `Series.astype(int)`.

[18]:
```python
ctrip['longcommute'] = (np.where(ctrip['commute']>= 120,True,False)).astype(int)
```

Question 2.g.ii. How would you expect that including `longcommute` as a second explanatory variable would alter the coefficient on `WFHShare` – would it increase, decrease, or stay the same? Explain.

I think since WFHShare is a random variable/ randomly assigned treatment, the covariance between these two variables should be zero. Thus, I would expect it to stay the same coefficient on WFH

Share.

Question 2.h.i. Management believes that commute (the travel time from home to office and back) is an important determinant of a worker's productivity. They have two hypotheses:

1. Employees who face a longer commute time are generally less productive than workers who have shorter commute times.
2. The effects of `WFHShare` on productivity is larger for those who face a longer commute.

Estimate a regression of `ln_calls`, with `WFHShare`, `longcommute`, and their interaction (call it `WFHShareXlongcommute`) as explanatory variables.

Hint: Once again you will need to add the argument `missing='drop'` when constructing your OLS model to drop the missing entries.

[19]:
```
ctrip['WFHShareXlongcommute']= ctrip['longcommute']*ctrip['WFHShare']
x_2hi=sm.add_constant(ctrip[['WFHShare','longcommute','WFHShareXlongcommute']])
model_2hi=sm.OLS(ctrip['ln_calls'], x_2hi, missing='drop')
results_2hi=model_2hi.fit(cov_type='HC1')
results_2hi.summary()
```

[19]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               ln_calls   R-squared:                       0.179
Model:                            OLS   Adj. R-squared:                  0.174
Method:                 Least Squares   F-statistic:                     179.4
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           6.79e-79
Time:                        15:04:36   Log-Likelihood:                -512.63
No. Observations:                 503   AIC:                             1033.
Df Residuals:                     499   BIC:                             1050.
Df Model:                           3
Covariance Type:                  HC1
================================================================================
========
                           coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
--------
const                    5.4398      0.095     57.061      0.000       5.253
5.627
WFHShare                 0.8641      0.125      6.926      0.000       0.620
1.109
longcommute              0.0162      0.103      0.158      0.875      -0.186
0.218
WFHShareXlongcommute     0.3300      0.137      2.415      0.016       0.062
0.598
==============================================================================
```

```
Omnibus:                      392.423   Durbin-Watson:                   1.841
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8736.706
Skew:                          -3.207   Prob(JB):                         0.00
Kurtosis:                      22.383   Cond. No.                         9.76
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
"""
```

Question 2.h.ii. Do your results support hypothesis (i), hypothesis (ii), both hypotheses, or neither one? Explain.

These results support hypothesis 2. (The effects of WFHShare on productivity is larger for those who face a longer commute). WFHShareXlongcommute coefficient is postivie at .3 and it is statistically significant. The effect is also greater than those who do not face long commute.

Question 2.i. If the coefficient on `longcommute` is statistically insignificant, would this lead you to drop `longcommute` from the regression model in part (h)? Explain your answer.

No. Just because it is not statistically insiginifcant does not mean it is not useful to the regression. It can give insight to the other regressor included. Moreover, to see if the variable is actually a good fit for the model it would be smart to check the r squared value for goodness of fit.

Question 2.j. Using the regression in part (h) and without estimating any other regression, write the estimated equation for the simple regression of `ln_calls` on `WFHShare` using only data for those with a commute of fewer than 120 minutes. You must show your solution to obtain full credit.

Since we only focus on short commute, the variable long commute equals to zero leaving us with:
ln_calls = 5.44 + .8641 * WFHShare

1.3 Problem 3. Natural Experiments

"Sin taxes" have not been the only way in which governments have attempted to reduce the consumption of cigarettes. In 1970, the U.S. passed a law that banned the advertising of cigarettes on radio and television. The ban took effect in 1971. The accompanying data file `cigads.csv` contains data on annual per capita consumption of tobacco measured in terms of "Annual grams of Tobacco Sold per Adult (15+)" for both the U.S. and Canada, 1968-1990 (`CIGSPC`). Also included in that file is a measure of the price of cigarettes given by the "Real Price of 20 grams Cents" for both countries (`PRICE`).

[20]:
```python
cigads = pd.read_csv("cigads.csv")
cigads.head()
```

[20]:

	YEAR	COUNTRY	CIGSPC	PRICE
0	1964	CAN	3975	128
1	1965	CAN	4095	128
2	1966	CAN	4158	127
3	1967	CAN	4168	127

```
4  1968     CAN    3971    137
```

Question 3.a. Treating the ban in cigarette advertising as a quasi-experiment, perform a differences-in-differences analysis of the effect of the ban on the consumption of tobacco. Fill in the table that indicates the conclusion of your analysis.

The top left box with work has been done for you.

[21]:
```
# Mean of annual grams of Tobacco Sold per Adult (15+) across the pre-treatment␣
↪periods in Canada
pre_period = cigads[cigads['YEAR'] <= 1970]
np.mean(pre_period[pre_period['COUNTRY'] == "CAN"]['CIGSPC'])


np.mean(pre_period[pre_period['COUNTRY']=="US"]['CIGSPC'])
postperiod=cigads[cigads['YEAR']> 1970]
np.mean(postperiod[postperiod['COUNTRY']=="CAN"]['CIGSPC'])
np.mean(postperiod[postperiod['COUNTRY']=="US"]['CIGSPC'])
```

[21]:
```
3804.05
```

	Before	After	After - Before
Canada	4043.14	3601.8	-441.34
USA	4280.71	3804.05	-476.66
USA - Canada	237.57	202.25	-35.32

Your explanation here

The difference in difference (or "double difference") estimator is defined as the difference in average outcome in the treatment group before and after treatment minus the difference in average outcome in the control group before and after treatment.

Question 3.b.i. Now create a dummy variable `post` indicating the time period whether the ban was in effect or not, plus a dummy variable `treat` for the treatment group (i.e. the U.S.) and the control group (i.e. Canada). Regress tobacco consumption on these two dummies and on the interaction between the two (you can call this `treatpost`).

Hint: Once again you will need to first create boolean columns then cast it into integers using `Series.astype(int)`.

[22]:
```
cigads['post'] = (cigads['YEAR']> 1970).astype(int)
cigads['treat'] = (cigads['COUNTRY']=="US").astype(int)
cigads['treatpost'] = cigads['treat']*cigads['post']

model_3b = sm.OLS(cigads['CIGSPC'], sm.
↪add_constant(cigads[['post','treat','treatpost']]))
results_3b = model_3b.fit(cov_type='HC1')
results_3b.summary()
```

[22]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 CIGSPC   R-squared:                       0.243
Model:                            OLS   Adj. R-squared:                  0.198
Method:                 Least Squares   F-statistic:                     13.82
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           1.09e-06
Time:                        15:04:36   Log-Likelihood:                -400.28
No. Observations:                  54   AIC:                             808.6
Df Residuals:                      50   BIC:                             816.5
Df Model:                           3
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       4043.1429     38.835    104.110      0.000    3967.027    4119.259
post        -441.3429    128.339     -3.439      0.001    -692.882    -189.803
treat        237.5714     63.652      3.732      0.000     112.815     362.328
treatpost    -35.3214    164.267     -0.215      0.830    -357.279     286.636
==============================================================================
Omnibus:                        5.878   Durbin-Watson:                   0.275
Prob(Omnibus):                  0.053   Jarque-Bera (JB):                5.770
Skew:                          -0.797   Prob(JB):                       0.0559
Kurtosis:                       2.843   Cond. No.                         9.69
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
"""
```

Question 3.b.ii. How do your results compare to your diffs-in-diffs estimator?

The coeffficient is equal to the diffs-in-diffs estimator and so is the post value to the estimator for the Canada row. In looking at the treatpost coefficient we would not reject the null for it, in other words it is not significant. It is the only variable which is not significant.

Question 3.c.i. Finally, recognizing that price does also affect consumption, you introduce the price variable into the regression in (b).

[23]:
```
model_3c = sm.OLS(cigads['CIGSPC'], sm.
 ↪add_constant(cigads[['post','treat','treatpost','PRICE']]))



results_3c = model_3c.fit(cov_type='HC1')
```

```python
results_3c.summary()
```

[23]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 CIGSPC   R-squared:                       0.854
Model:                            OLS   Adj. R-squared:                  0.842
Method:                 Least Squares   F-statistic:                     72.98
Date:                Sat, 15 Jul 2023   Prob (F-statistic):           5.03e-20
Time:                        15:04:36   Log-Likelihood:                -355.80
No. Observations:                  54   AIC:                             721.6
Df Residuals:                      49   BIC:                             731.5
Df Model:                           4
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       5599.8931    124.292     45.054      0.000    5356.285    5843.501
post        -191.9745     42.022     -4.568      0.000    -274.336    -109.613
treat        -60.8905     54.984     -1.107      0.268    -168.656      46.875
treatpost   -259.1679     83.122     -3.118      0.002    -422.083     -96.252
PRICE        -11.8706      0.926    -12.812      0.000     -13.687     -10.055
==============================================================================
Omnibus:                        2.758   Durbin-Watson:                   0.402
Prob(Omnibus):                  0.252   Jarque-Bera (JB):                2.656
Skew:                           0.500   Prob(JB):                        0.265
Kurtosis:                       2.575   Cond. No.                         906.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
"""
```

Question 3.c.ii. Report your results and compare to those from (b).

Including price shows its coefficient as negative. An inrease in price would lead to a decrease in consumption per capita. The treatment variable is not significant at the 5% level since the p value is .26

Question 3.d. Why would you expect that the price of a pack of cigarettes might be correlated with the error term? Note that some economists have argued that the advertising ban reduced competition among cigarette makers by eliminating one dimension on which they compete for customers, which in turn led to higher prices.

The price may correlated with the error term because there are external facotrs that affect the demand of cigarretes but the regression does account for because it is focused only on per capita consumption as demand and determinant for price. For example, anti cigrette campaigns can affect the demand to smoke, and therefore would impact the price.

1.4 Problem 4. Regression Discontinuity

The data set `rd.csv` contains student level data for 65,535 students who finished high school and were eligible to enter college. In the specific country where the data orginate (Chile), students write a standardized test at the end of high school, called the PSU test. Their scores on this test, plus high school GPA, determine which colleges they can get into. Students who score at least 475 points on the PSU test are also eligible for a loan from the government for college costs, while students who score less than 475 points cannot receive the loan. In this exercise we will use regression discontinuity methods to analyze the effect of the loan program on the probability of college entry.

Variable	Description
psu	PSU test score (ranges from 300 to 700)
over475	1=PSU score is 475 or higher
entercollege	1=student entered college
hsgpa	high school GPA (ranges from 0 to 70)
privatehs	1=student went to privatre high school
hidad	1=father has more than a high school education
himom	1=mother has more than a high school education

[24]:
```
rd = pd.read_csv("rd.csv")
rd.head()
```

[24]:

	hsgpa	psu	entercollege	privatehs	hidad	himom	over475
0	60	396.0	0	0	0	0	0
1	65	402.5	0	0	0	0	0
2	55	485.0	0	0	0	0	1
3	0	461.5	0	0	0	0	0
4	62	394.0	0	0	0	0	0

Question 4.a. Construct the average values of `entercollege`, `hsgpa`, `privatehs`, `hidad`, `himom` for each integer value of `psu` (e.g., get the averages for scores from 300 to 300.99, and assign them to the "300" bucket; then get the averages for scores from 301 to 301.99 and assign them to the "301" bucket, etc.). This is sometimes called "collapsing" the data to integer cells. This is a bit tricky, so we provide the commands for you below.

[25]:
```
rd['psu_integer'] = np.floor(rd['psu'])
rd_temp = rd.groupby('psu_integer').agg(['mean']).reset_index()

rd_collapsed = pd.DataFrame()
rd_collapsed['psu_integer'] = rd_temp['psu_integer']
rd_collapsed['hsgpa'] = rd_temp['hsgpa']['mean']
rd_collapsed['psu'] = rd_temp['psu']['mean']
rd_collapsed['entercollege'] = rd_temp['entercollege']['mean']
rd_collapsed['privatehs'] = rd_temp['privatehs']['mean']
```

```python
rd_collapsed['hidad'] = rd_temp['hidad']['mean']
rd_collapsed['himom'] = rd_temp['himom']['mean']
rd_collapsed['over475'] = rd_temp['over475']['mean']
rd_collapsed.head()
```

[25]:

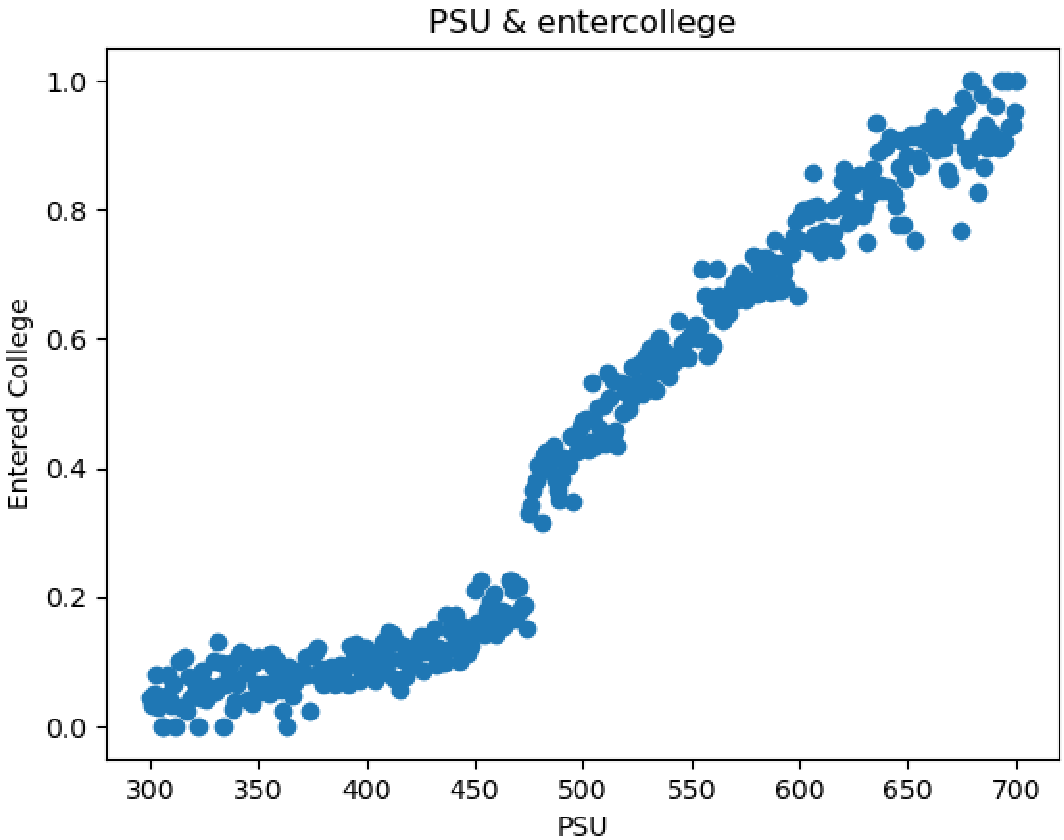
```
   psu_integer      hsgpa         psu  entercollege  privatehs     hidad  \
0        300.0  53.136364  300.113636      0.045455   0.000000  0.045455
1        301.0  50.677419  301.290323      0.032258   0.000000  0.064516
2        302.0  49.833333  302.133333      0.050000   0.016667  0.016667
3        303.0  53.657895  303.223684      0.078947   0.000000  0.026316
4        304.0  51.057143  304.214286      0.028571   0.000000  0.028571

      himom  over475
0  0.045455      0.0
1  0.032258      0.0
2  0.050000      0.0
3  0.026316      0.0
4  0.057143      0.0
```

Question 4.b. Generate plots of the average values of `entercollege`, `hsgpa`, `privatehs`, `hidad`, `himom` (from 4.a) as a function of `psu` (be sure to label your axes and give each plot a title). You should see a jump in `entercollege` at 475 points, but relatively smooth values of the other variables. The following cell is for your code.

[26]:

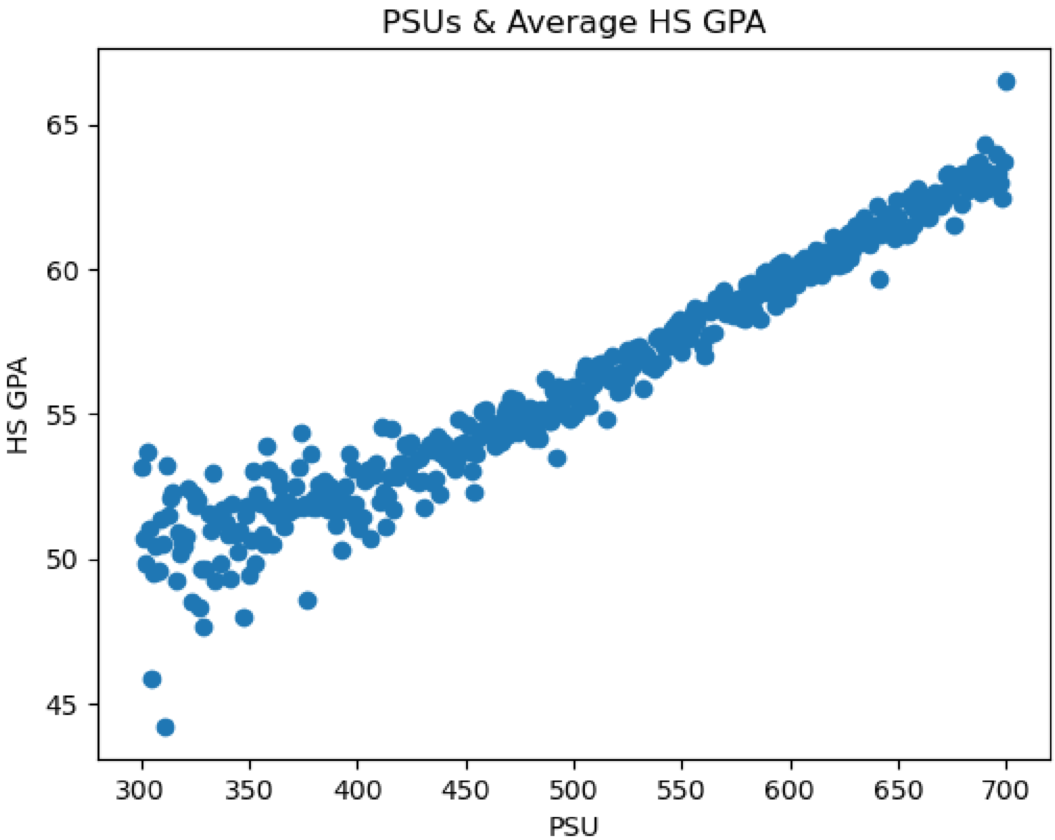
```python
plt.scatter(rd_collapsed['psu_integer'], rd_collapsed['entercollege'])

plt.ylabel('Entered College')

plt.title("PSU & entercollege")
plt.xlabel("PSU")
```

[26]: `Text(0.5, 0, 'PSU')`



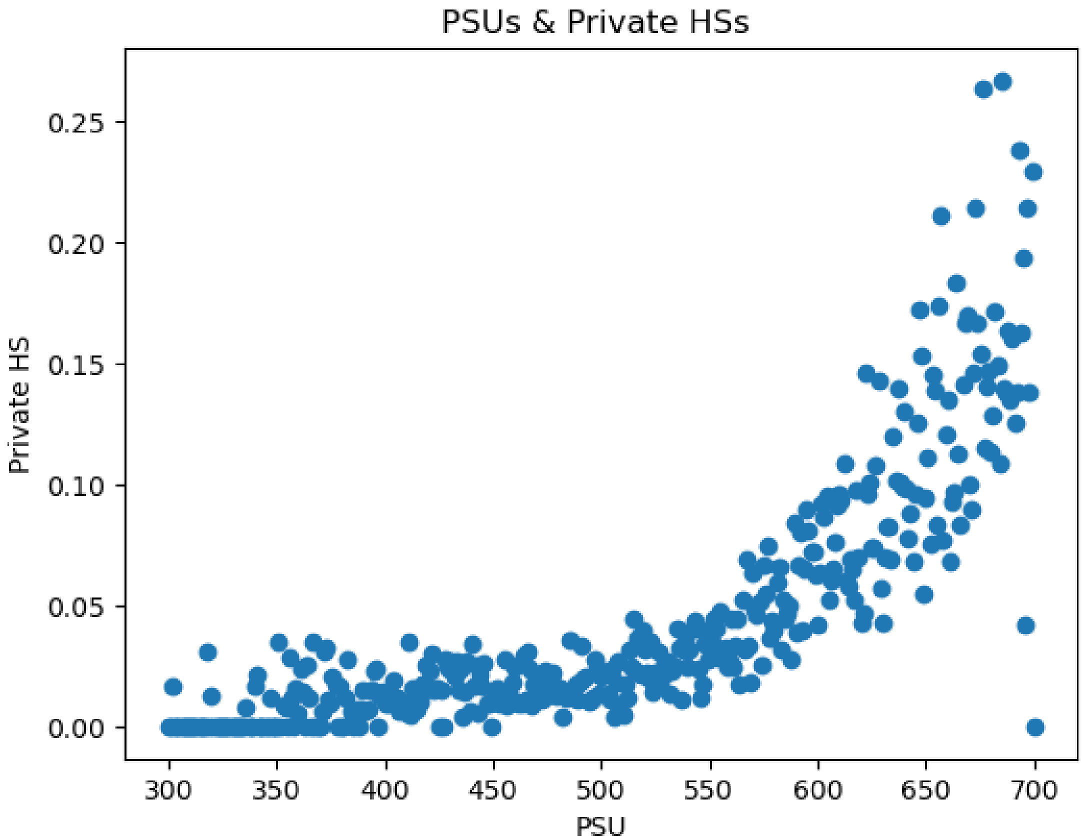
[27]:
```
plt.scatter(rd_collapsed['psu_integer'], rd_collapsed['hsgpa'])
plt.ylabel('HS GPA')
plt.xlabel('PSU')
plt.title('PSUs & Average HS GPA')
```

[27]: Text(0.5, 1.0, 'PSUs & Average HS GPA')



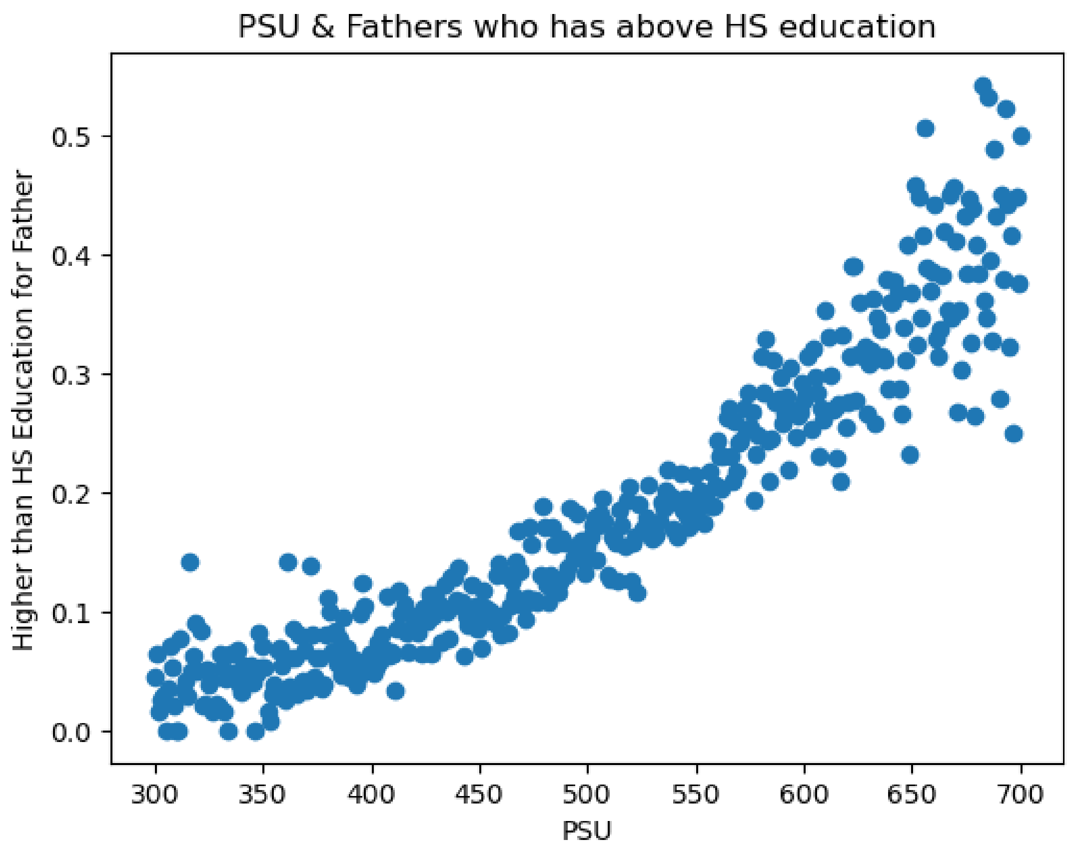
[28]:
```
plt.scatter(rd_collapsed['psu_integer'], rd_collapsed['privatehs'])
plt.ylabel('Private HS')
plt.xlabel('PSU')
plt.title('PSUs & Private HSs ')
```

[28]: `Text(0.5, 1.0, 'PSUs & Private HSs ')`



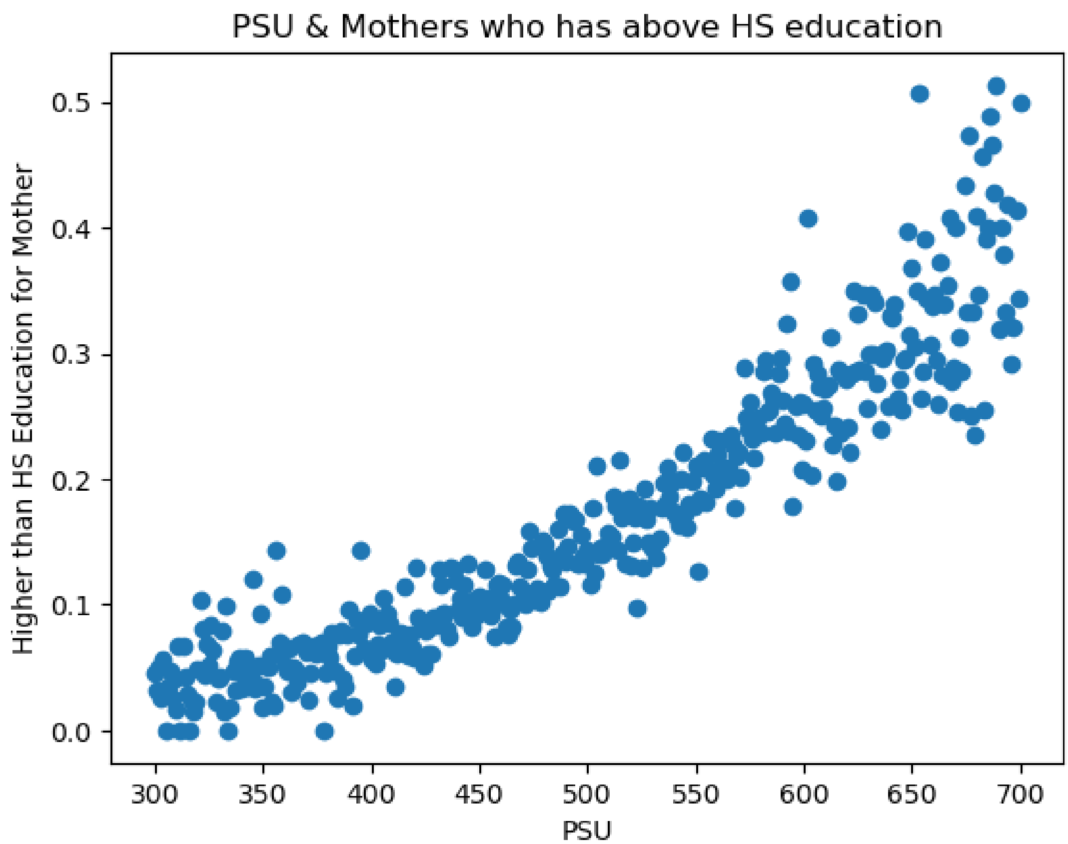
[29]:
```
plt.scatter(rd_collapsed['psu_integer'], rd_collapsed['hidad'])
plt.ylabel('Higher than HS Education for Father')
plt.xlabel('PSU')
plt.title('PSU & Fathers who has above HS education')
```

[29]:
```
Text(0.5, 1.0, 'PSU & Fathers who has above HS education')
```



[30]:
```
plt.scatter(rd_collapsed['psu_integer'], rd_collapsed['himom'])
plt.ylabel('Higher than HS Education for Mother')
plt.xlabel('PSU')
plt.title('PSU & Mothers who has above HS education')
```

[30]: Text(0.5, 1.0, 'PSU & Mothers who has above HS education')



Question 4.c. Next you will fit *local linear* regressions using different bandwidths. To do this you will regress one of the dependent variables Y_i on the following independent variables: `constant`, `psu`,`over475` and $\tilde{psu} = psu - 475$, i.e., you will fit the model

$Y_i = \beta_0 + \beta_1 over475_i + \beta_2 \tilde{psu}_i + \delta_3(\tilde{psu}_i \cdot over475_i) + u_i$

Interpret the coefficients of this regression model.

[31]:
```
rd_collapsed['psu_squiggle']=rd_collapsed['psu']-arr

rd_collapsed['psu_over475']=rd_collapsed['psu_squiggle']*rd_collapsed['over475']

X_4c=sm.add_constant(rd_collapsed[['over475','psu_squiggle','psu_over475']])

model_4c=sm.OLS(rd_collapsed['entercollege'], X_4c)


results_4c=model_4c.fit(cov_type='HC1')
```

```
results_4c.summary()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[31], line 1
----> 1 rd_collapsed['psu_squiggle']=rd_collapsed['psu']-arr
      3␣
 ↪rd_collapsed['psu_over475']=rd_collapsed['psu_squiggle']*rd_collapsed['over475']
      5 X_4c=sm.
 ↪add_constant(rd_collapsed[['over475','psu_squiggle','psu_over475']])

NameError: name 'arr' is not defined
```

The constant is .1677. At zero, the entercollege value is .1677 over475 0.2271 psu_squiggle 0.0008 psu_over475 0.0019

All of these coefficients are statistically significant.

The R squared is high showing that the regression is good fit. ie. a high percentage of the data is accounted for by the regresssion.

In intrepereting the coefficients, one can see for the over475 coef for example, that if someone is over a 475 scoret then they have a 0.2271 average point increase in entering college.

Question 4.d. Using the "collapsed" data from part 4.a, which has one observation per integer value of `psu_integer`, and a bandwidth of 10 on each side of the 475 cutoff, fit the model for each of the dependent variables $Y_i = entercollege$, $Y_i = hsgpa$,, $Y_i = hidad$, $Y_i = himom$ (i.e., you are fitting four separate models here). The following cell is for your code.

Hint: This means that you fit the regression models to the collapsed data for the subset of data with $465 \leq psu\_integer \leq 485$. *This data set will have 21 observations – 10 observations for scores less than 475 and 11 observations for scores of 475 or higher.*

[ ]:
```
rd_collapsed1=rd_collapsed[rd_collapsed['psu_integer']<= 485]
rd_collapsed1=rd_collapsed1[rd_collapsed1['psu_integer']>= 465]
X_4di=sm.add_constant(rd_collapsed1[['over475','psu_squiggle','psu_over475']])
model_4di=sm.OLS(rd_collapsed1['entercollege'], X_4di)
results_4di=model_4di.fit(cov_type='HC1')

X_4dii=sm.add_constant(rd_collapsed1[['over475','psu_squiggle','psu_over475']])
model_4dii=sm.OLS(rd_collapsed1['hsgpa'], X_4dii)
results_4dii=model_4dii.fit(cov_type='HC1')

X_4diii=sm.add_constant(rd_collapsed1[['over475','psu_squiggle','psu_over475']])

model_4diii=sm.OLS(rd_collapsed1['hidad'], X_4diii)
results_4diii=model_4diii.fit(cov_type='HC1')

X_4div=sm.add_constant(rd_collapsed1[['over475','psu_squiggle','psu_over475']])
```

```
model_4div=sm.OLS(rd_collapsed1['hidad'], X_4div)
results_4div=model_4div.fit(cov_type='HC1')



results_4dii.summary()
```

Question 4.e. Repeat part 4.d using a bandwidth of 20 points. Do you find that the estimated jumps are similar for all four dependent variables as with a bandwidth of 10?

The first cell is for your code, the second cell is for your question answer.

[ ]:
```
rd_collapsed2=rd_collapsed[rd_collapsed['psu_integer']<= 495]
rd_collapsed2=rd_collapsed2[rd_collapsed2['psu_integer']>= 455]

X_4ei=sm.add_constant(rd_collapsed2[['over475','psu_squiggle','psu_over475']])
model_4ei=sm.OLS(rd_collapsed2['entercollege'], X_4ei)
results_4ei=model_4ei.fit(cov_type='HC1')

X_4eii=sm.add_constant(rd_collapsed2[['over475','psu_squiggle','psu_over475']])
model_4eii=sm.OLS(rd_collapsed2['hsgpa'], X_4eii)
results_4eii=model_4eii.fit(cov_type='HC1')

X_4eiii=sm.add_constant(rd_collapsed2[['over475','psu_squiggle','psu_over475']])
model_4eiii=sm.OLS(rd_collapsed2['hidad'], X_4eiii)
results_4eiii=model_4eiii.fit(cov_type='HC1')

X_4eiv=sm.add_constant(rd_collapsed2[['over475','psu_squiggle','psu_over475']])
model_4eiv=sm.OLS(rd_collapsed2['hidad'], X_4eiv)
results_4eiv=model_4eiv.fit(cov_type='HC1')



results_4dii.summary()
```

The estimate jumps are indeed similar for the dependent variables with a bandwidth of 10.

Question 4.f. For every bandwidth from 5 to 50, develop a plot to show the estimate of β_1 when the dependent variable Y_i is *entercollege*. The following cell is for your code.

[ ]:
```
results_4eii.params[1]
b=[]





for i in np.arange(5,51):
```

```
    rd_collapsed1=rd_collapsed[rd_collapsed['psu_integer']<= 475 +i]
    rd_collapsed1=rd_collapsed2[rd_collapsed2['psu_integer']>= 475 -i]

    X_4ei=sm.
 ↪add_constant(rd_collapsed1[['over475','psu_squiggle','psu_over475']])
    model_4ei=sm.OLS(rd_collapsed1['entercollege'], X_4ei)
    results_4ei=model_4ei.fit(cov_type='HC1')
    temp=results_4eii.params[1]
    b.append(temp)



plt.scatter(range(5,51), b_)
plt.ylabel('beta1 Coeff for over 475)')
plt.xlabel('Band width')
plt.title('Beta dif bandwidths')
```

1.5 Submission

Make sure you have run all cells in your notebook in order before running the cell below, so that all images/graphs appear in the output. The cell below will generate a zip file for you to submit. **Please save before exporting!**

[ ]:
```
# Save your notebook first, then run this cell to export your submission.
grader.export()
```