Notebook

May 4, 2021

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.

[45]:
```
y_1a = gas['quantgas']
X_1a = gas['pricegas']
model_1a = sm.OLS(y_1a, sm.add_constant(X_1a))
results_1a = model_1a.fit(cov_type = 'HC1')
results_1a.summary()
```

[45]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           0.000239
Time:                        20:20:23   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.
==============================================================================

Warnings:
[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.

If the coefficient of `pricegas` variable is the price elasticity of demand, then it would not economically make sense because a higher price of gas would correspond with a higher amount of gas consumed. In reality, we would expect a higher price for gas would correspond to a decrease in gas consumption.

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.

[46]:
```
X_1c = sm.add_constant(gas[['pricegas', 'persincome']])
y_1c = gas['quantgas']
model_1c = sm.OLS(y_1c, X_1c)
results_1c = model_1c.fit(cov_type = 'HC1')
results_1c.summary()
```

[46]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           3.32e-97
Time:                        20:20:23   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
==============================================================================
```

```
Warnings:
[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.

We now observe that the `pricegas` coefficient is negative, indicating an invert relationship with the quantity of gas consumed, which makes more economical sense.

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?

The OLS regression from Question 1.a does suffer from omitted variable bias because by adding the variable `persincome` into the regression, it changes the linear relationship between price gas and the quantity of gas consumed from positive to negative. We determine that the sign of the bias is negative, since `persincome` is a positive value, and that the correlation between `pricegas` and `persincome` is negative.

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.

We suspect that the goasline price would be correlated with the error term because there are more variables that we could account for in our linear regression model. If there were more cars being sold, then we would expect an increase in the price of gas due to a higher demand for gas.

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.

[47]:
```python
y_1g = gas['pricegas']
X_1g = sm.add_constant(gas[['persincome', 'carsales']])
model_1g = sm.OLS(y_1g, X_1g)
results_1g = model_1g.fit(cov_type = 'HC1')
results_1g.summary()
```

[47]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           2.61e-17
Time:                        20:20:23   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
persincome     0.0023      0.001      3.788      0.000       0.001       0.003
carsales      -6.3378      0.957     -6.624      0.000      -8.213      -4.463
==============================================================================
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
==============================================================================

Warnings:
[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.
"""
```

Question 1.h. Explain.

Due to the significance in the p-value of the `carsales` variable, we conclude that `carsales` is a "strong instrument" for price of gas.

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

The `transindex` variable could be a better instrument because it encapsulates more transportation services than just car owners. By including more transportation services, we would expect the price of gas to decrease overall due to economies of scale. For example, we can fit 30 people in one bus, but can only fit at most 5 people in one car. It is more gasoline efficient to transport 30 people than 5 people.

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.

[48]:
```python
gas['pricegas_hat'] = results_1g.fittedvalues
y_1j = gas['carsales']
X_1j = sm.add_constant(gas[['persincome', 'pricegas_hat']])
model_1j = sm.OLS(y_1j, X_1j)
results_1j = model_1j.fit(cov_type = 'HC1')
results_1j.summary()
```

[48]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               carsales   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 3.496e+30
Date:                Tue, 04 May 2021   Prob (F-statistic):               0.00
Time:                        20:20:24   Log-Likelihood:                 9123.9
No. Observations:                 296   AIC:                        -1.824e+04
Df Residuals:                     293   BIC:                        -1.823e+04
Df Model:                           2
Covariance Type:                  HC1
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         25.5981   6.18e-15   4.14e+15      0.000      25.598      25.598
persincome     0.0004   4.56e-19   7.79e+14      0.000       0.000       0.000
pricegas_hat  -0.1578   6.18e-17  -2.55e+15      0.000      -0.158      -0.158
==============================================================================
Omnibus:                       50.420   Durbin-Watson:                   0.058
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               13.655
Skew:                           0.198   Prob(JB):                      0.00108
Kurtosis:                       2.025   Cond. No.                     7.63e+04
==============================================================================

Warnings:
[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.

After the second stage of the TSLS estimation, the size of the coefficient for pricegas increased from the first stage of our OLSE regression.

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 TSLS estimate of the price coefficient is statistically significant, and we do not have a reason to doubt the standard errors from the second stage because after refining our regression estimator, we expect the standard errors to decrease significantly from the first stage.

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?

Due to multiple factors affecting the quantity of gas produced, we would feel comfortable regressing the quantity of gas produced on its price. However, we would also want to include multiple regressors to have a clearer picture. Again, using a two-stage regression can help us give better results for our estimators.

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.

[49]:
```
y_1n = gas['pricegas']
X_1n = sm.add_constant(gas[['persincome', 'transindex']])
model_1n = sm.OLS(y_1n, X_1n)
results_1n = model_1n.fit(cov_type = 'HC1')
results_1n.summary()
```

[49]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           1.06e-33
Time:                        20:20:24   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
persincome    -0.0088      0.002     -5.704      0.000      -0.012      -0.006
transindex     1.1001      0.113      9.741      0.000       0.879       1.321
==============================================================================
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
==============================================================================

Warnings:
[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.
"""
```

[50]:
```
gas['pricegas_hat'] = results_1n.fittedvalues
y_1n_2 = gas['transindex']
```

```python
X_1n_2 = sm.add_constant(gas[['persincome', 'pricegas_hat']])
model_1n_2 = sm.OLS(y_1n_2, X_1n_2)
results_1n_2 = model_1n_2.fit(cov_type = 'HC1')
results_1n_2.summary()
```

[50]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:             transindex   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 3.222e+31
Date:                Tue, 04 May 2021   Prob (F-statistic):               0.00
Time:                        20:20:24   Log-Likelihood:                 8658.1
No. Observations:                 296   AIC:                        -1.731e+04
Df Residuals:                     293   BIC:                        -1.730e+04
Df Model:                           2
Covariance Type:                  HC1
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -26.6785   3.84e-14  -6.94e+14      0.000     -26.679     -26.679
persincome       0.0080   3.05e-18   2.63e+15      0.000       0.008       0.008
pricegas_hat     0.9090   4.23e-16   2.15e+15      0.000       0.909       0.909
==============================================================================
Omnibus:                       35.354   Durbin-Watson:                   0.075
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.357
Skew:                           0.443   Prob(JB):                     0.000103
Kurtosis:                       2.160   Cond. No.                     6.77e+04
==============================================================================

Warnings:
[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, we see that the price index is a strong instrument for estimating the transportation services index. In stage two, we used the refined estimator, which is `pricegas_hat`, to regress on `transindex` and we see that our estimators have very small standard errors. This indicates the strength of our regression results, and proves that using the two-stage approach helps us become more confident in our OLS analysis.

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.

Due to the relatively larger coefficient on pricegas_hat (0.9) on our regression analysis for transindex compared to carsales pricegas_hat coefficient (-0.16), we would consider that the better estimate of the price coefficient would be `transindex`.

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

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

[52]:
```python
ctrip['volunteer'].value_counts()
percentage = 503 / len(ctrip['volunteer'])
percentage
```

[52]:
```
0.506036217303823
```

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

[53]:
```python
y_2bi = ctrip['commute']
X_2bi = sm.add_constant(ctrip['volunteer'])
model_2bi = sm.OLS(y_2bi, X_2bi)
results_2bi = model_2bi.fit(cov_type = 'HC1')
results_2bi.summary()
```

[53]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           0.000739
Time:                        20:20:24   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
==============================================================================
```

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

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

If somebody is a volunteer, then their commute would be approximately 12 minutes longer than someone who was not a volunteer. Given a p-value of 0.001 in the `volunteer` variable, 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.

[54]:
```
y_2ci = ctrip['tenure']
X_2ci = sm.add_constant(ctrip['volunteer'])
model_2ci = sm.OLS(y_2ci, X_2ci)
results_2ci = model_2ci.fit(cov_type = 'HC1')
results_2ci.summary()
```

[54]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):            0.00645
Time:                        20:20:24   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
==============================================================================

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

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

The coefficient on the variable `volunteer` indicates that tenure is approximately 4 months less for volunteers than non-volunteers. Given a p-value of 0.006, the `volunteer` variable in this regression is also statistically significant.

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.

[55]:
```python
ctrip['ln_calls'] = np.log(ctrip['calls'])
ctrip['ln_calls']
y_2di = ctrip['WFHShare']
X_2di = sm.add_constant(ctrip['ln_calls'])
model_2di = sm.OLS(y_2di, X_2di, missing='drop')
results_2di = model_2di.fit(cov_type = 'HC1')
results_2di.summary()
```

[55]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               WFHShare   R-squared:                       0.163
Model:                            OLS   Adj. R-squared:                  0.161
Method:                 Least Squares   F-statistic:                     35.02
Date:                Tue, 04 May 2021   Prob (F-statistic):           6.06e-09
Time:                        20:20:25   Log-Likelihood:                -73.987
No. Observations:                 503   AIC:                             152.0
Df Residuals:                     501   BIC:                             160.4
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.5098      0.168     -3.026      0.002      -0.840      -0.180
ln_calls       0.1671      0.028      5.918      0.000       0.112       0.223
==============================================================================
Omnibus:                       47.844   Durbin-Watson:                   1.808
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               15.722
Skew:                          -0.107   Prob(JB):                     0.000385
Kurtosis:                       2.161   Cond. No.                         49.3
==============================================================================

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

"""

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

We expect the share of days worked at home to increase by approximately 0.1671 if we increased the log calls by one unit. The p-value suggests that this effect is statistically 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?

It appears that the experiement did not take into account whether somebody volunteered to work from home when deciding how many hours they would work from home. The Ctrip has achieved a randomized controlled experiment, in which that there is a correlation on the effects on working from home on productivity. However, they did not achieve the objective of determining a causal effect on working from home on productivity.

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)`.

[56]:
```
ctrip['longcommute'] = (ctrip['commute'] >= 120).astype(int)
ctrip['longcommute']
```

[56]:
```
0      0
1      1
2      1
3      1
4      0
      ..
989    0
990    0
991    1
992    0
993    1
Name: longcommute, Length: 994, dtype: int64
```

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.

We expect that the `longcommute` variable to increase the coefficient on `WFHShare` because people with longer commutes are more incentivized to work at home instead of commuting to work.

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.

[57]:
```
ctrip['WFHShareXlongcommute'] = ctrip['WFHShare'] * ctrip['longcommute']
y_2hi = ctrip['ln_calls']
X_2hi = sm.add_constant(ctrip[['WFHShare', 'longcommute',␣
↪'WFHShareXlongcommute']])
model_2hi = sm.OLS(y_2hi, X_2hi, missing = 'drop')
results_2hi = model_2hi.fit(cov_type = 'HC1')
results_2hi.summary()
```

[57]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           6.79e-79
Time:                        20:20:27   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
==============================================================================
```

```
Warnings:
[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.

We do not support the first hypothesis because the `longcommute` variable is a positive coefficient, which infers that the longer commute times are associated with more productivity at home. We also do not support the second hypothesis because the variable `WFHShare` has a larger, positive coefficient than the variable `WFHShareXlongcommute`.

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.

We would drop the `longcommute` variable because it adds very little significance to the linear regression model, and adds unnecessary other variables to the model.

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.

ln_calls $= 5.4398 + 0.8641 * WFHShare + 0.0162 * longcommute + 0.3300 * WFHShareXlongcommute$

ln_calls $= 5.4398 + 0.8641 * WFHShare + 0.0162 * 0 + 0.3300 * 0$

ln_calls $= 5.4398 + WFHShare * 0.8641$

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.

[59]:
```python
# Mean of annual grams of Tobacco Sold per Adult (15+) across the pre-treatment
↪periods in Canada
pre_period = cigads[cigads['YEAR'] <= 1970]
print(np.mean(pre_period[pre_period['COUNTRY'] == "CAN"]['CIGSPC']))
# Canada After
pre_period = cigads[cigads['YEAR'] > 1970]
print(np.mean(pre_period[pre_period['COUNTRY'] == "CAN"]['CIGSPC']))
# USA Before
pre_period = cigads[cigads['YEAR'] <= 1970]
print(np.mean(pre_period[pre_period['COUNTRY'] == "US"]['CIGSPC']))
# USA After
pre_period = cigads[cigads['YEAR'] > 1970]
print(np.mean(pre_period[pre_period['COUNTRY'] == "US"]['CIGSPC']))
```

```
4043.1428571428573
3601.8
```

```
4280.714285714285
3804.05
```

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)`*]().

[60]:
```
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()
```

[60]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           1.09e-06
Time:                        20:20:30   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
==============================================================================

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

```
"""
```

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

When comparing diffs-in-diffs estimator from Question 3.a and 3.b OLS regression results, we see that the estimators are extremely close to each other.

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

[61]:
```python
model_3c = sm.OLS(cigads['CIGSPC'], sm.add_constant(cigads[['post', 'treat',␣
↪'treatpost', 'PRICE']]))
results_3c = model_3c.fit(cov_type = 'HC1')
results_3c.summary()
```

[61]:
```
<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:                Tue, 04 May 2021   Prob (F-statistic):           5.03e-20
Time:                        20:20:31   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.
==============================================================================

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

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

After including the `PRICE` variable into our linear regression, we see that the coefficient on `treatpost` decreases siginificantly from approximately -35 to about -259.

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` variable is correlated with the error term because higher prices would decrease the demand on the consumption of cigarettes, which incentivizes people to smoke less cigarettes i.e The Law of Demand. Perhaps the error term also encompasses other variables, such as advertising which would also be correlated with the price of cigarettes.