Notebook

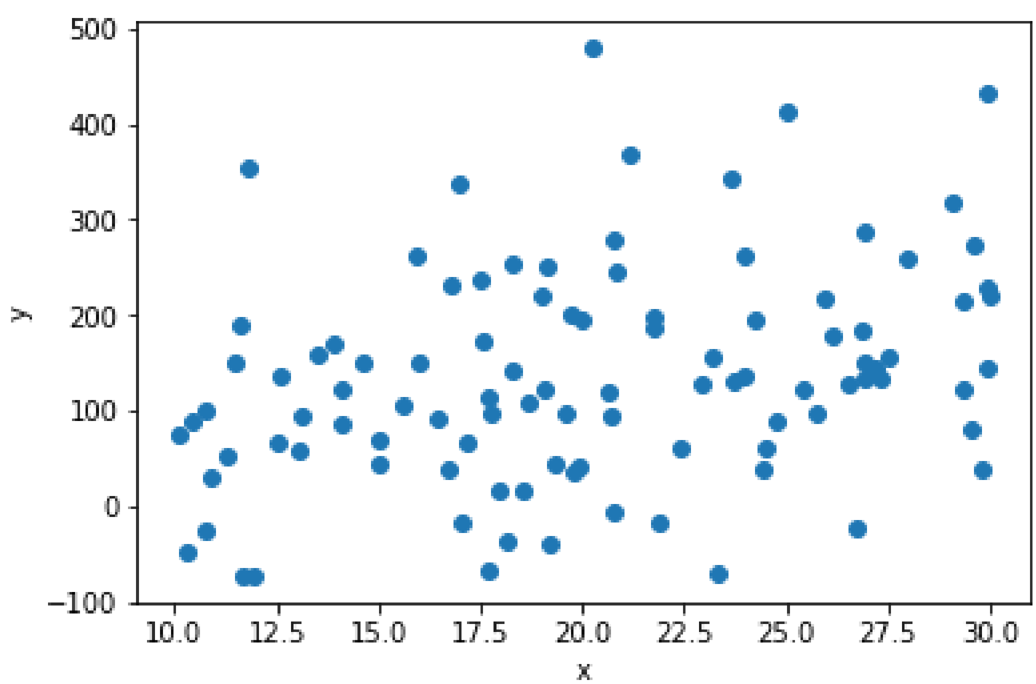
February 28, 2021

Question 1.a. Begin by specifying that there are 100 observations and generate the regressor to be $x = 10 + 20v$, where v is a uniform random variable on the unit interval. As a result, x is a random variable uniformly distributed on the interval $[10, 30]$. Next specify the dependent variable to be linearly related to this regressor according to $y = 30 + 5x + u$, where u is a random draw from a normal distribution with population mean 0 and population standard deviation 100. Then, generate a scatter plot of x and y.

Hint: You may want to check out `np.random.random_sample` to generate v. You also may want to check out `np.random.normal` to generate u.

[8]:
```
v = np.random.random_sample((100, ))
x = 10 + 20 * v
u = np.random.normal(0, 100, 100)
y = 30 + 5 * x + u

plt.scatter(x, y)
plt.xlabel("x")
plt.ylabel("y");
```



Question 1.b. Next regress y on x (calling for robust standard errors). Is each one of the three OLSE assumptions satisfied in this case? Explain why for each one. Give your assessment of how well least squares regression performs in estimating the true intercept and slope.

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

[9]:
```
X_1b = sm.add_constant(x)
model_1b = sm.OLS(y, x)
results_1b = model_1b.fit(cov_type = 'HC1')
results_1b.summary()
```

[9]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      y   R-squared (uncentered):
0.628
Model:                            OLS   Adj. R-squared (uncentered):
0.624
Method:                 Least Squares   F-statistic:
164.8
Date:                Sun, 28 Feb 2021   Prob (F-statistic):
8.52e-23
Time:                        20:30:49   Log-Likelihood:
-610.12
No. Observations:                 100   AIC:
1222.
Df Residuals:                      99   BIC:
1225.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             6.6849      0.521     12.839      0.000       5.664       7.705
==============================================================================
Omnibus:                        6.441   Durbin-Watson:                   2.264
Prob(Omnibus):                  0.040   Jarque-Bera (JB):                5.848
Skew:                           0.547   Prob(JB):                       0.0537
Kurtosis:                       3.454   Cond. No.                         1.00
==============================================================================

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

"""

Question 1.c. Explain.

The assumption, that the mean of conditional distribution of errors, u, given x is zero is true because the mean of the random distribution is equal to 0 and does not rely on the value of the variable x. The random sampling is satisfied due to the use of random sampling in the code for the variable v. The assumption that large outliers are unlikely is also true because the variable u is normally distributed and the random sampling is also distributed normally.

Question 1.d. Looking at the results of this regression including the number shown above, assess how close least squares estimation is to the true variance of the error term.

As long as the OLS assumptions are met, which they are in this case, we can estimate of the true variance of the error term using the residuals from the OLS. Comparing our answers from part 1.a to the cell above, we see that they are extremely close to each other with a 0.08 difference.

Question 1.e. Generate the regression residuals and confirm they add up to zero. Also, confirm that the residuals are uncorrelated with the regressor.

Hint: The command `results_1c.resid` *will give you an array of the residuals of the regression. The function* `np.sum()` *takes an array as an argument inside the parenthases and sums all of the elements together. Remember that* `results_1c.resid` *is an array. Also, the function* `np.corrcoef()` *takes in two arrays of equal length, separated by a comma, and computes the correlation matrix of the two arrays. For example, usage might look like* `np.corrcoef(array1, array2)`*.*

[11]:
```python
sum_of_residuals = np.sum(results_1b.resid)
print("Sum of residuals: ", sum_of_residuals)
np.corrcoef(x, y)
```

```
Sum of residuals:  94.11025515255949
```

[11]:
```
array([[1.        , 0.31107695],
       [0.31107695, 1.        ]])
```

Question 1.f. Now generate the variables x and y as you did above but do it for $n = 1000$ observations. Run the regression of y on x and compare the results with the earlier case of $n = 100$. Explain the differences.

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

[12]:
```python
v_1000 = np.random.sample(1000, )
x_1000 = 10 + 20 * v_1000
u_1000 = np.random.normal(0, 100, 1000)
y_1000 = 30 + 5 * x_1000 + u_1000

X_1f = sm.add_constant(x_1000)
model_1f = sm.OLS(y_1000, x_1000)
results_1f = model_1f.fit(cov_type = 'HC1')
results_1f.summary()
```

[12]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      y   R-squared (uncentered):
0.659
Model:                            OLS   Adj. R-squared (uncentered):
0.659
Method:                 Least Squares   F-statistic:
2060.
Date:                Sun, 28 Feb 2021   Prob (F-statistic):
5.47e-245
Time:                        20:30:52   Log-Likelihood:
-6022.6
No. Observations:                1000   AIC:
1.205e+04
Df Residuals:                     999   BIC:
1.205e+04
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             6.5468      0.144     45.385      0.000       6.264       6.830
==============================================================================
Omnibus:                        1.511   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.470   Jarque-Bera (JB):                1.397
Skew:                          -0.047   Prob(JB):                        0.497
Kurtosis:                       3.157   Cond. No.                         1.00
==============================================================================

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

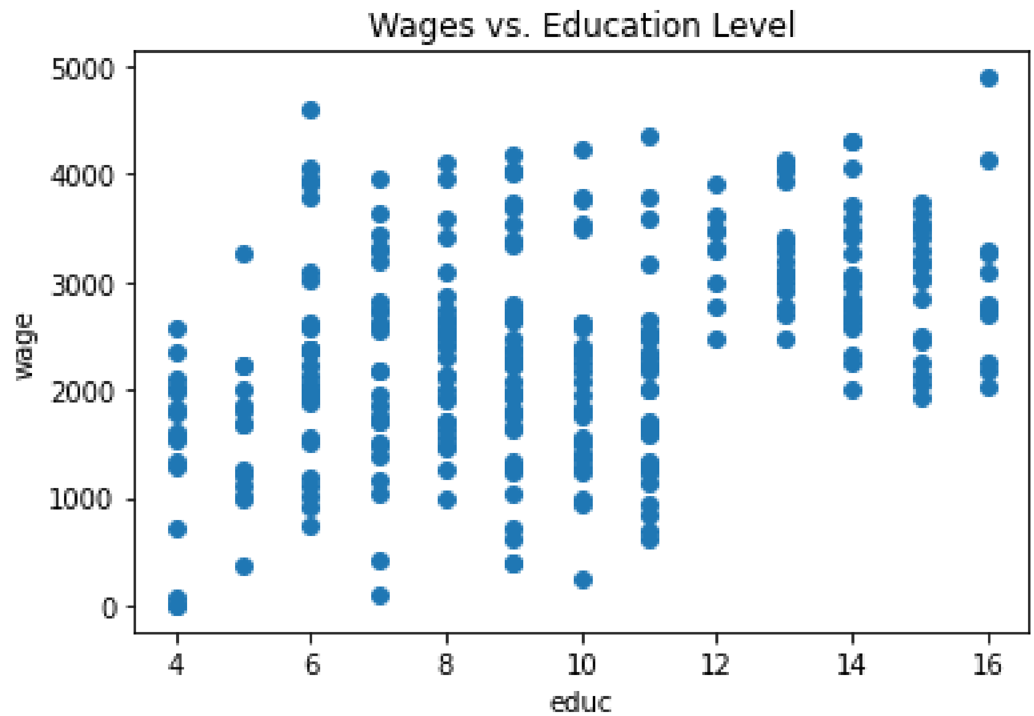
Question 1.h. Explain.

Comparing the two OLS regression results, the standard error for the 1000 observations is lower than the standard error of the 100 observations. This is due to the amount of observations for the second OLS regression result is 10 times higher than the first. Likewise, the smaller sample size of the first OLD regression result has the adjusted R-squared comes down more than the larger sample size of the second OLS regression result.

Question 2.a. Plot a scatter diagram of the average monthly wage against education level. Does it confirm your intuition? What differences do you see between individuals who did not complete high school and those that did?

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

```
[14]: plt.scatter(wages['educ'], wages['wage'])
plt.xlabel("educ")
plt.ylabel("wage")
plt.title("Wages vs. Education Level");
```



Question 2.b. Explain.

The scatter plot graph confirms the intuition that having more education correlates to having a higher average monthly wage. The differences between the individuals who did not complete highschool and those that did is that theres a much higher monthly wage.

Question 2.c. Perform an OLS regression of wages on education. Be sure to include the robust option. Give a precise interpretation of least squares estimate of the intercept and evaluate its sign, size and statistical significance. Does its value make economic sense? Do the same for the least squares estimate of the slope. Does this slope estimate confirm the scatter plot above?

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

```
[15]: y_2c = wages['wage']
X_2c = sm.add_constant(wages['educ'])
model_2c = sm.OLS(y_2c, X_2c)
results_2c = model_2c.fit(cov_type = 'HC1')
results_2c.summary()
```

```
[15]: <class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.160
Model:                            OLS   Adj. R-squared:                  0.157
Method:                 Least Squares   F-statistic:                     70.91
Date:                Sun, 28 Feb 2021   Prob (F-statistic):           1.60e-15
Time:                        20:30:55   Log-Likelihood:                -2460.4
No. Observations:                 300   AIC:                             4925.
Df Residuals:                     298   BIC:                             4932.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       1256.1721    151.039      8.317      0.000     960.141    1552.203
educ         117.1024     13.907      8.421      0.000      89.846     144.359
==============================================================================
Omnibus:                        1.218   Durbin-Watson:                   2.068
Prob(Omnibus):                  0.544   Jarque-Bera (JB):                1.258
Skew:                           0.152   Prob(JB):                        0.533
Kurtosis:                       2.909   Cond. No.                         31.7
==============================================================================

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

Question 2.d. Explain.

The intercept size is about 1256, which is a plausible amount for an individual with a monthly wage. likewise, the intercept is a positive value which indicates that uneducated people are making about 15,000 per year. Given the p-value of 0.00, the OLS regression results does show a statistical significance. This does make economic sense that with education increasing, the amount of monthly wage would also increase. With a slope value of about 117, the slope is positive and therefore does agree with our scatter plot grap from above.

Question 2.e. List the three OLS assumptions and give a concrete example of when each of those would hold in this context. Are these assumptions plausible in this context?

(Mean distribution assumption description)

The first OLS assumption, with a mean value of 0 would hold if it was independent of other factors. In this case, it would not hold because there a variety of factors, besides education level, that would affect the dependent variable (wage). The assumption that random sampling are independently and identically distributed would hold if the jobs that are being offered don't require a level of education, but rather, skill sets that are required in the work force. In this case, the assumption is not plausible because the amount of the wage is dependent on the level of education. The assumption that outliers are unlikely can hold if there aren't individuals with a high level of education aren't working for jobs with low income or, conversely, individuals with minimal education are receiving excessively high income. In this case, it is possible that there are outliers due to extremely successful businesses.

Question 2.f. You are rightfully concerned whether education will, in fact, be rewarded in the labor market. You wonder if another year of education will yield an expected \$100 more per month (which if discounted over a typical working lifetime at say, 5%, amounts to roughly a year at Berkeley). Test the following null hypothesis: $H_0 : \beta_1 = 100$ vs $H_1 : \beta_1 \neq 100$.

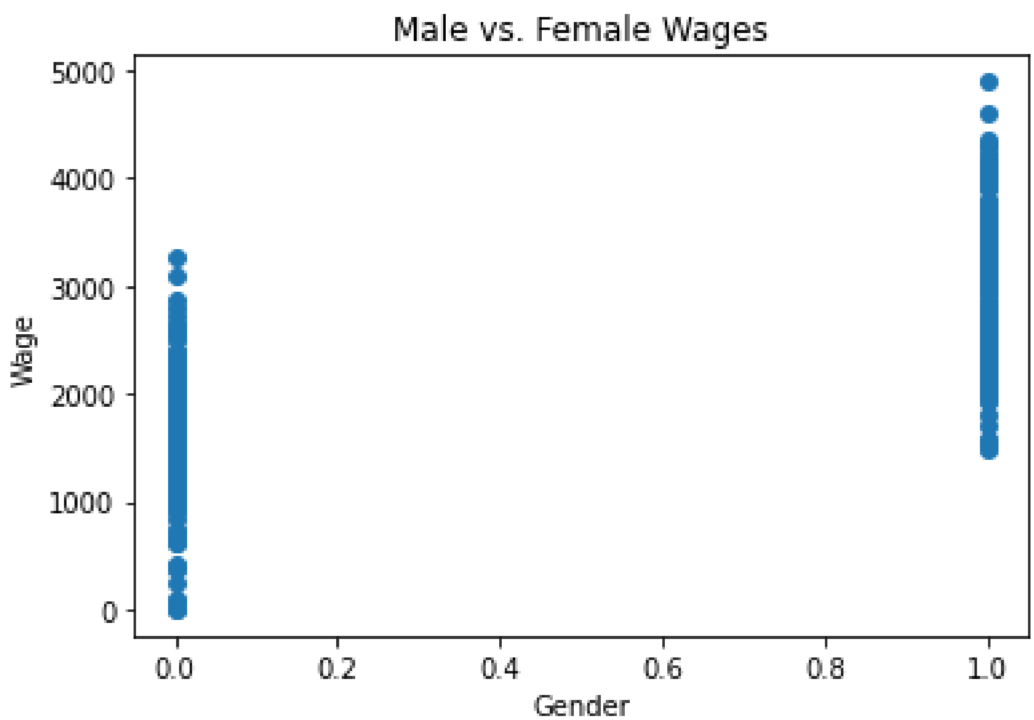
Hypothesis Test:
$H_0 : \beta_1 = 100$
$H_1 : \beta_1 \neq 100.$

We assume that the slope is 100, and that the standard error is 13.9. Given the confidence interval 2.c, we are well within the 95% confidence interval (72.76, 127.25) by multiplying the standard error by 2 and adding/subtracting with the slope of 100.

Question 2.g. Let's now return to a familiar empirical question: do men and women earn the same amount? As in part (a) above, generate a scatterplot of `wage` against the dummy variable `male`. Don't forget to label your axes! What is your answer to the question based on this graph?

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

[16]:
```
plt.scatter(wages['male'], wages['wage'])
plt.xlabel('Gender')
plt.ylabel('Wage')
plt.title('Male vs. Female Wages');
```



Question 2.h. Explain.

On average, male and women do not earn the same amount by looking at the graph, where male averages are overall higher than female averages.

Question 2.i. Run an OLS regression of `wage` on `male`. Provide a precise interpretation of the slope. Do you believe you have found evidence of wage discrimination in this data, or do you believe there is another explanation for the differences? Explain.

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

[17]:
```
y_2i = wages['wage']
X_2i = wages['male']
model_2i = sm.OLS(y_2i, X_2i)
results_2i = model_2i.fit(cov_type = 'HC1')
results_2i.summary()
```

[17]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                   wage   R-squared (uncentered):
0.755
Model:                            OLS   Adj. R-squared (uncentered):
0.754
Method:                 Least Squares   F-statistic:
3081.
Date:                Sun, 28 Feb 2021   Prob (F-statistic):
1.69e-159
Time:                        20:30:55   Log-Likelihood:
-2570.1
No. Observations:                 300   AIC:
5142.
Df Residuals:                     299   BIC:
5146.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
male        2993.1195     53.925     55.505      0.000    2887.428    3098.811
==============================================================================
Omnibus:                       39.762   Durbin-Watson:                   1.484
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               10.879
Skew:                           0.047   Prob(JB):                      0.00434
Kurtosis:                       2.072   Cond. No.                         1.00
```

```
==============================================================================

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

Question 2.j. Explain.

There is evidence that there is a correlation between higher average wages of male and lower average wages of females, but does not infer that that the gender is a causation.

Question 2.k. As we did in problem set 1, perform a t-test of a difference in wages between men and women and report the t-stat and p-value. Compare the output of that test with the regression results you got using the male dummy. To make the two results (in terms of t-stat and p-value) correspond, do you assume equal or unequal variance of men's and women's wages?

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

[18]:
```python
wages_men = wages[wages['male'] == 1]
wages_women = wages[wages['male'] == 0]

ttest_2k = stats.ttest_ind(wages_men['wage'], wages_women['wage'], equal_var =␣
 ↪True)

tstat_2k = ttest_2k.statistic
pval_2k = ttest_2k.pvalue

print("t-stat: {}".format(tstat_2k))
print("p-value: {}".format(pval_2k))
```

```
t-stat: 17.435055261853524
p-value: 2.0502180343257218e-47
```

Question 2.l. Explain.

We assume that they are equal variance because the variance between male and female wages is relatively similar.

Question 3.a. What is contained in the error term? Provide a couple of examples. Do you think that the first OLS assumption is plausible in this context?

The error term contains the distance between the predicted value from the regression line and the actual value of observed prices. For example, the grapes that were harvested one year ago would have a predicted price of 100 dollars. However, the true price was a 110 dollars, making the error term u_i 10 dollars. The OLS assumption is plausible because if grapes were harvested today, then there wouldn't be much of a difference between the prices of wine.

Question 3.b. Suppose you estimate your model via OLS and you obtain the following estimated coefficients (standard errors are reported in parenthesis), with $R^2 = 0.77$:

$$price_i = \underset{(2.57)}{1.75} + \underset{(1.02)}{5.5}\, vintage_i + \hat{u}_i$$

Interpret the regression coefficients.

If the grapes were harvested today, then their price would be 1.75 dollars, and would increase by 5.5 dollars per year that they were harvested.

Question 3.c. Comment on the R^2. Given this statistic what can you infer about causality in the relationship of prices and vintage?

With a R^2 value of 0.77, there is a high correlation between the prices of wine and the time that the grapes were harvested (in years). However, we cannot determine a causality based on a correlation.

Question 3.d. Predict the fitted value of price of a bottle whose grapes were harvested ten years ago, and that for a bottle harvested nine years ago; then compute the difference between the two values.

$$price_i = 1.75 + 5.5vintage_i + \hat{u}_i$$

$$1.75 + 5.5 * 10 + 0 = 56.75$$

$$1.75 + 5.5 * 9 + 0 = 51.25$$

$$56.75 - 51.25 = 5.5$$

Question 3.e. Derive the marginal effect of the increase in one year in vintage on price. Do you get the same result as in part (d)? Why? Explain.

The marginal effect would equal to 5.5, which is the same answer as part (d) because taking the derivative of the linear regression line gives us how much the price increases due to an increase in a single vintage year.

Question 3.f. Using the results above, give a 95% confidence interval for the difference in average price for a ten year bottle vs a five year bottle. Can you reject the null hypothesis that this difference is \$40?

null hypothesis: the difference is equal to 40 dollars. alternative hypothesis: the difference is not equal to 40 dollars.

$$StandardError = 1.02 * 5 + 2.57 = 7.67$$

$$Averagedifference = 27.5$$

$$1.96 * 7.67 = 15.03$$

$$ConfidenceInterval = (12.47, 42.53)$$

In this case, we fail to reject the null hypothesis.

Question 4.a. Since we want to see what happens to the share of expenditures spent on food, create the variable `foodshare` = `foodpq/totexppq`. Run a regression of food share on family size. What is the interpretation of the estimated coefficient on family size? Is it statistically and economically significant? Do your findings support the theory that large families can enjoy economies of scale (e.g., house, TV, etc.) and allocate more of their expenses to food?

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

[20]:
```
ces['foodshare'] = ces['foodpq'] / ces['totexppq']
y_4a = ces['foodshare']
X_4a = sm.add_constant(ces['fam_size'])
model_4a = sm.OLS(y_4a, X_4a)
results_4a = model_4a.fit(cov_type = 'HC1')
results_4a.summary()
```

[20]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:              foodshare   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     4.394
Date:                Sun, 28 Feb 2021   Prob (F-statistic):             0.0363
Time:                        20:30:57   Log-Likelihood:                 898.39
No. Observations:                1000   AIC:                            -1793.
Df Residuals:                     998   BIC:                            -1783.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1654      0.007     25.034      0.000       0.152       0.178
fam_size       0.0047      0.002      2.096      0.036       0.000       0.009
==============================================================================
Omnibus:                      347.206   Durbin-Watson:                   2.027
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1606.241
Skew:                           1.557   Prob(JB):                         0.00
Kurtosis:                       8.372   Cond. No.                         6.10
==============================================================================

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

Question 4.b. Explain.

For every increase in family size by 1, there will be an expected increase in foodshare by 0.0047. From an economical standpoint, this is significant because as family size increases, the amount of expenditure going towards food increases. With a p-value of 0.036, this is also statistically significant because it goes beyond the 5 percent confidence level. It is possible that economies of scale can be used to explain why an increase in family size leads to an increase in food expenditure.

Question 4.c. What is the predicted share of expenditures spent on food for a single mother with two kids?

$$0.0047 * 3 + 0.1654 = 0.1795$$

Question 4.d. Now regress food share on the logarithm of family size. Do the regression results differ? How does the interpretation of the coefficient on log family size differ from the prior regression?

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

[21]:
```
ces['log_fam_size'] = np.log(ces['fam_size'])
y_4d = ces['foodshare']
X_4d = sm.add_constant(ces['log_fam_size'])
model_4d = sm.OLS(y_4d, X_4d)
results_4d = model_4d.fit(cov_type = 'HC1')
results_4d.summary()
```

[21]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:              foodshare   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     2.240
Date:                Sun, 28 Feb 2021   Prob (F-statistic):              0.135
Time:                        20:30:58   Log-Likelihood:                 897.09
No. Observations:                1000   AIC:                            -1790.
Df Residuals:                     998   BIC:                            -1780.
Df Model:                           1
Covariance Type:                  HC1
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.1708      0.006     30.594      0.000       0.160       0.182
log_fam_size     0.0086      0.006      1.497      0.134      -0.003       0.020
==============================================================================
Omnibus:                      347.526   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1613.669
Skew:                           1.557   Prob(JB):                         0.00
Kurtosis:                       8.388   Cond. No.                         2.83
```

```
==============================================================================

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

Question 4.e. Explain.

The coefficient of log-family size is almost twice as large as the prior linear regression line, which infers that with every increase in the family size, the log-food shares would have roughly twice the effect.

Question 4.f. The R^2 is pretty small for both of the above regressions. Does this cast doubt on whether there is a relationship between family size and food share? Explain.

Because R^2 is a measure of good fit of the relationship between the variables, there is some doubt between the relationship of family size and food share due to the extremely small R^2 values.

Question 4.g. The theory applies in particular to poor households whose food expenses are at a bare minimum. Rerun the same regression for families who expenditure per capita are less than \$3,000. Does that change your answer to the previous question?

Hint: First you may need to create a new per capita expenditure variable.

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

[22]:
```python
ces['exp_pc'] = ces['totexppq'] / ces['fam_size']
ces_3000 = ces[ces['exp_pc'] < 3000]
y_4g = ces_3000['foodshare']
X_4g = sm.add_constant(ces_3000['log_fam_size'])
model_4g = sm.OLS(y_4g, X_4g)
results_4g = model_4g.fit(cov_type = 'HC1')
results_4g.summary()
```

[22]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:              foodshare   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     2.202
Date:                Sun, 28 Feb 2021   Prob (F-statistic):              0.138
Time:                        20:31:00   Log-Likelihood:                 446.53
No. Observations:                 532   AIC:                            -889.1
Df Residuals:                     530   BIC:                            -880.5
Df Model:                           1
Covariance Type:                  HC1
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
```

```
const             0.2211      0.010     23.106      0.000       0.202       0.240
log_fam_size     -0.0125      0.008     -1.484      0.138      -0.029       0.004
==============================================================================
Omnibus:                      194.553   Durbin-Watson:                   2.086
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              866.124
Skew:                           1.592   Prob(JB):                    8.39e-189
Kurtosis:                       8.379   Cond. No.                         3.23
==============================================================================

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

Question 4.h. Explain.

This further suggests that there is no correlation between the family size and foodshare. Instead, it suggests that there's a possibility of an omitted variable.

Question 4.i. Now regress expenditure per capita on family size and interpret the coefficient. What does this tell you about the validity of your former results?

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

[23]:
```python
y_4i = ces['exp_pc']
X_4i = sm.add_constant(ces['fam_size'])
model_4i = sm.OLS(y_4i, X_4i)
results_4i = model_4i.fit(cov_type = 'HC1')
results_4i.summary()
```

[23]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 exp_pc   R-squared:                       0.049
Model:                            OLS   Adj. R-squared:                  0.048
Method:                 Least Squares   F-statistic:                     94.96
Date:                Sun, 28 Feb 2021   Prob (F-statistic):           1.71e-21
Time:                        20:31:00   Log-Likelihood:                -9936.1
No. Observations:                1000   AIC:                         1.988e+04
Df Residuals:                     998   BIC:                         1.989e+04
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       6129.5860    304.938     20.101      0.000    5531.918    6727.254
fam_size    -749.0452     76.868     -9.745      0.000    -899.704    -598.386
==============================================================================
Omnibus:                     1014.214   Durbin-Watson:                   1.954
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            56891.090
Skew:                           4.750   Prob(JB):                         0.00
Kurtosis:                      38.709   Cond. No.                         6.10
==============================================================================

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

Question 4.j. Explain.

As family size increases, the amount of food expenditure per capita decreases. As a result, it is possible that we are experiencing economies of scale with an increasing family size.