Notebook

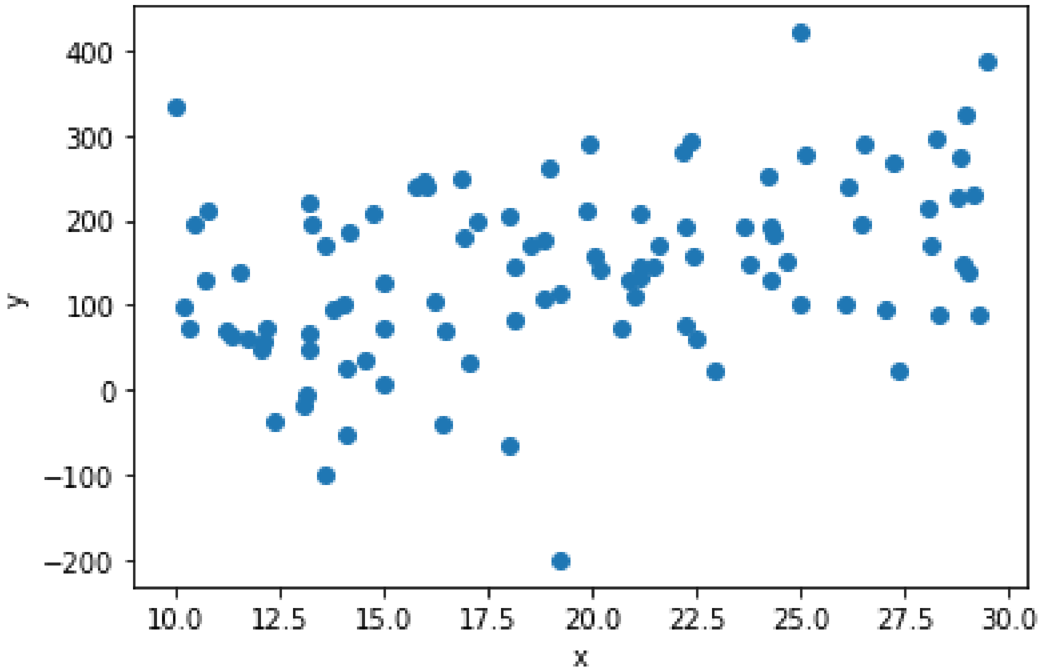
March 1, 2021

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.

[3]:
```
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.

[4]:

```python
X_1b = sm.add_constant(x)
model_1b = sm.OLS(y, X_1b)
results_1b = model_1b.fit(cov_type = 'HC1')
results_1b.summary()
```

[4]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.145
Model:                            OLS   Adj. R-squared:                  0.136
Method:                 Least Squares   F-statistic:                     16.25
Date:                Mon, 01 Mar 2021   Prob (F-statistic):           0.000110
Time:                        10:40:04   Log-Likelihood:                -598.72
No. Observations:                 100   AIC:                             1201.
Df Residuals:                      98   BIC:                             1207.
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.4191     34.850      0.270      0.787     -58.886      77.725
x1             6.8435      1.698      4.031      0.000       3.516      10.171
==============================================================================
Omnibus:                        4.297   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.117   Jarque-Bera (JB):                4.307
Skew:                          -0.252   Prob(JB):                        0.116
Kurtosis:                       3.883   Cond. No.                         71.1
==============================================================================

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

Question 1.c. Explain.

Since the R value is so low (got around .10 for R2 which tells it is about .3 roughly for R), it implies that there is little to no relation between the variables and they vary wildly. We also got a coefficent of 6 for x which is very close to the coefficent for the X variable we used. The reason for this slight difference might be the low number of a random sample and u variable that picks anywhere from the normal curve of SD of 100. Just a review of the requirements which I will refer to again throught this: 1. Mean of conditional distribution of errors ui given xi is zero, E[ui|xi] = 0 2. Random sampling, or (yi,xi)ni=1 are independently and identically distributed (i.i.d) 3. Large outliers are unlikely: 0< E[x4i]<∞and 0< E[y4i]<∞ (this is necessary for asymptotic normality only; you will need it in order to use Central Limit Theorem.)

The first requirement is not satifisied through the Prob(Omnibus), as if it was close to 1 it would indicate that BLUE is accurate and is not. Two is satified because we used random sampling to come with our y and x. Lastly, since we used u to compute error on a normal curve, it is extremely difficult to get anything within +-3 STD.

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.

The high residual (around 100) implies that the variance is very high as well, which we can confirm with our R^2 being so close to zero which tells us that there is no strong correlation between the variables and vary wildly.

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)`.

[6]:
```
sum_of_residuals = np.sum(results_1b.resid)
print("Sum of residuals: ", sum_of_residuals)
np.corrcoef(x, results_1b.resid)
```

```
Sum of residuals:  1.4068746168049984e-12
```

[6]:
```
array([[1.00000000e+00, 2.36289801e-16],
       [2.36289801e-16, 1.00000000e+00]])
```

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.

[7]:
```
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_1f)
results_1f = model_1f.fit(cov_type = 'HC1')
results_1f.summary()
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.108
Model:                            OLS   Adj. R-squared:                  0.107
Method:                 Least Squares   F-statistic:                     115.5
Date:                Mon, 01 Mar 2021   Prob (F-statistic):           1.44e-25
Time:                        10:40:07   Log-Likelihood:                -6033.5
No. Observations:                1000   AIC:                         1.207e+04
Df Residuals:                     998   BIC:                         1.208e+04
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.8792     12.124      0.567      0.570     -16.884      30.643
x1             6.0983      0.567     10.747      0.000       4.986       7.210
==============================================================================
Omnibus:                        4.475   Durbin-Watson:                   1.972
Prob(Omnibus):                  0.107   Jarque-Bera (JB):                3.698
Skew:                          -0.051   Prob(JB):                        0.157
Kurtosis:                       2.720   Cond. No.                         77.3
==============================================================================

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

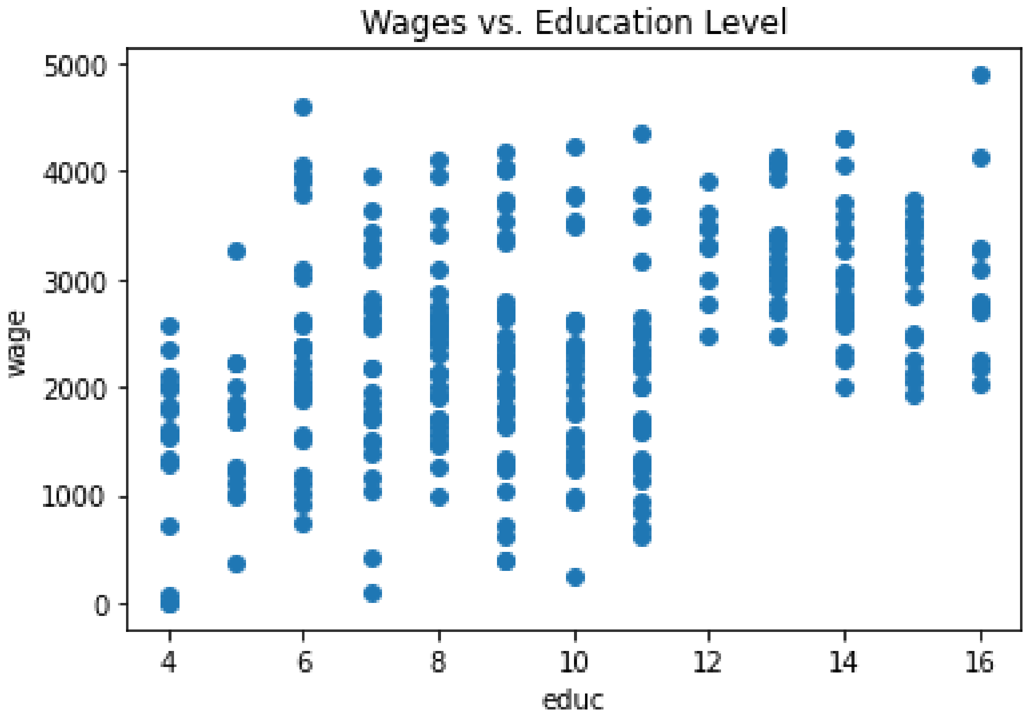
Question 1.h. Explain.

The results are fairly similar, howecer they follow the line a lot better but still are weakly clustered around this line, probably due to the SD of 100, which will reflect that if so.

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.

[9]:
```python
plt.scatter(wages['educ'], wages['wage'])
plt.xlabel("educ")
plt.ylabel("wage")
plt.title("Wages vs. Education Level");
```



Question 2.b. Explain.

It generally confirms my suspicions however not in the way I expected. First of all, 4 years of education to 6 years (indicating the change from a bachelors (4 years) to masters (6 years), the masters has a larger range of education. After that, from 7-11 years of education, the range varies and there is barely any difference, however then we get to 12+ years of education where the range is smaller, but the average is higher. Jobs that require this much education are limited to but not including medical doctors and professors, which recieve MD and PhDs, which require more education many of which require time and reseach in the field they plan on going into, therefore making their skills more marketable. Sadly, after getting a masters, money seems pretty much the same until you get the PhD or MD, so that will be difficult.

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.

[10]:

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

```
print(np.mean(y_2c))
print((np.std(y_2c)/np.std(x))*(.16**(.5)))
results_2c.summary()
```

```
2382.6971551
117.03933238160823
```

```
[10]: <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:                Mon, 01 Mar 2021   Prob (F-statistic):           1.60e-15
Time:                        10:40:09   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 first number given is the y intercept and the second is the slope. Both are positive which is expected as you cannot make negative money when working a job, even if you have no formal education. This number tells us that if we are just out of high school, we will make a little under 2.5k per month. The slope is positive and for every year of education we are expected to make a little over 100 dollars more. Statistically however, due to the nature of how spread the values are, there is high error and this will not accurately make an assessment of how much you will make over time.

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?

1. Mean of conditional distribution of errors ui given xi is zero, E[ui|xi] = 0
2. Random sampling, or (yi,xi)ni=1 are independently and identically distributed (i.i.d)
3. Large outliers are unlikely: 0< E[x4i]<∞and 0< E[y4i]<∞ (this is necessary for asymptotic normality only; you will need it in order to use Central Limit Theorem.)

First will only hold if the only factor contributing to the amount of money made is education. Second will hold up if the sample is taken randomly and from the same population. Third will hold up if it follows the normal curve and we happen to sample the correct part of the population.

In this context, it makes sense that they would be plausible in this context, however the data is too variable to give us an accurate number. Included, with the First one, we know that not only does education play into pay, but the prior experience and job market matters as well. For instance, I could work at Starbucks for a little over minimum wage with a bachelors degree. With a Data Science major, I wouldnt want to do this as I have oppurtunity to make more, however some people find themselves in this postision which will account for the lower end of the pay range.

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 $H_1 : \beta_1 \neq 100$.

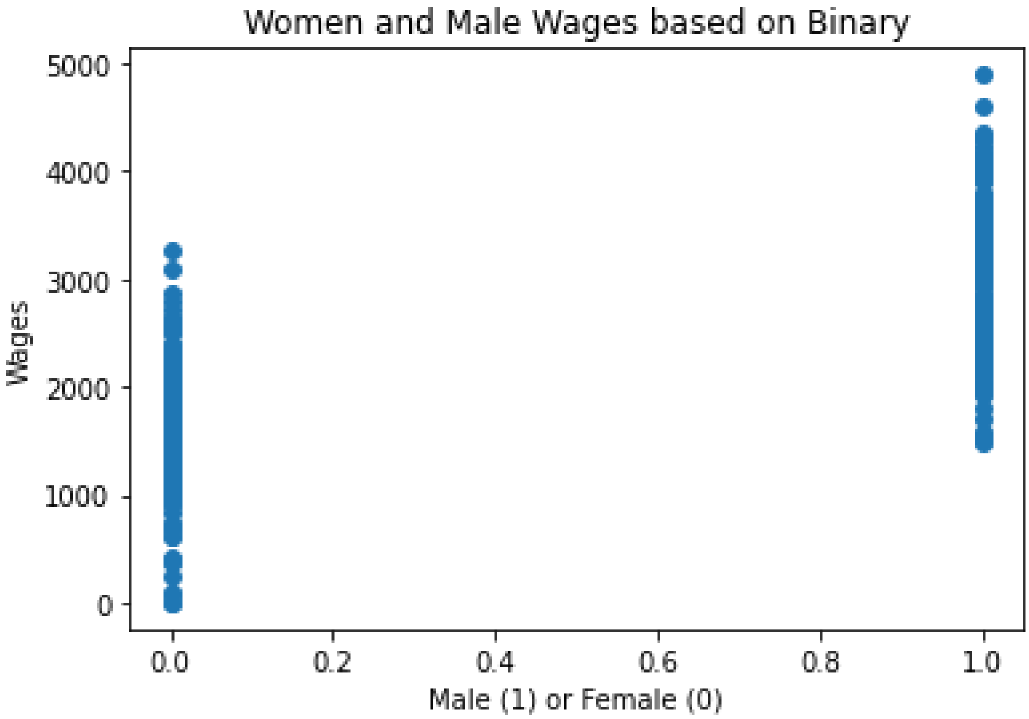
From the result in 2c, we can see in the [0.025 0.975] interval, 100 is included, so we fail to reject the null at 5% because of this fact. Since it will follow a normal distribution, it will be very unlikely that 100 would be in 2.5% on each side as the STD doesnt account for this.

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.

[11]:

```python
plt.scatter(wages['male'], wages['wage'])
plt.xlabel("Male (1) or Female (0)")
plt.ylabel("Wages")
plt.title("Women and Male Wages based on Binary");
```



Question 2.h. Explain.

From what I see, other than this is not the proper way to use a scatter plot (would use a box plot or even a histogram to compare but besides that) the range of women wages and male wages vary signficantly. The range of male go between 5000 and 1500 while womens goes between 0 and 3300. Although this doesnt imply much (for instance, our random sample mightve just reflected a less educated group of women or the area we sampled in happen to have more housewives/stay at home mothers than the other). This can be indicative of a wage disparity, but without ruling out these other confounding variables, we can never be sure.

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.

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

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

```
==============================================================================
Dep. Variable:                   male   R-squared:                       0.253
Model:                            OLS   Adj. R-squared:                  0.250
Method:                 Least Squares   F-statistic:                     240.9
Date:                Mon, 01 Mar 2021   Prob (F-statistic):           3.13e-40
Time:                        10:40:12   Log-Likelihood:                -172.10
No. Observations:                 300   AIC:                             348.2
Df Residuals:                     298   BIC:                             355.6
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1736      0.062     -2.792      0.005      -0.295      -0.052
educ           0.0759      0.005     15.522      0.000       0.066       0.085
==============================================================================
Omnibus:                      209.104   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               19.312
Skew:                           0.053   Prob(JB):                     6.40e-05
Kurtosis:                       1.762   Cond. No.                         31.7
==============================================================================

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

Question 2.j. Explain.

The values it gives us for the values in between are double while we have int (rather boolean in respect to ‘if male’) which gives us data that is insignificant in a way, since we cannot find a linear assosiation with these values. It does have a postisive R value (~.5) however this is not enough to confirm that there is a higher chance of men being paid more than women.

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.

[13]:

```
wages_men = wages[wages['male'] == 1]['wage']
wages_women = wages[wages['male'] == 0]['wage']

ttest_2k = stats.ttest_ind(wages_men, wages_women, equal_var = False)

tstat_2k = ttest_2k.statistic
pval_2k = ttest_2k.pvalue

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

```python
print("p-value: {}".format(pval_2k))
```

```
t-stat: 17.558368459097736
p-value: 1.6265962860910122e-47
```

Question 2.l. Explain.

Now that we have compared the values in a simple hypothesis test to tell us at a 1% confidence level, even without that we have at least 17 standard deviations away from the value being similar, with a way less than a 1% chance of it happening within the population randomly. Assuming these factors arent caused by a random sample of the population having less pay for the women or the population has a higher likelyhood of having housewives or women who do not pursue higher paying jobs as a norm, we can assume that women are in fact paid less than men. (But then again we have to test specific job fields, specific education levels, and other factors to confirm rather than just wages because as I mentioned with education and wage, its not the only factor contruibuting to pay.)

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?

In the error term, we can consider the fact that some years, harvest isnt good so therefore the grapes arent that good as well as other factors such as if a lack of supply exist, it might drive up the price especially if its a well known wine. Last thing it may account for will be the fact that different markets have different prices. The same wine sold to me in San Francisco may be a couple dollars more expensive in some small town in Kansas for instance. In this context, knowing that OLS E[ui|xi] = 0 would be satisfied because over those many bottles it will level out as time goes on.

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.

I said for example 50 dollars instead of using a dollar sign because it would use LaTex which is unintened and messes with the formatting of the document.

The first coeffcient is if the grapes were just harvested (basically grape juice), the starting price would be around 1.75 dollars with error of about 3 dollars and for every year that the wine is aged, it will raise about 5.5 dollars, with an error of 1 dollar.

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

R^2 gives us a lot of information as the correlation R is a little over .8 and implies that the wine prices tend to follow this trendline very well. It also tells us that we are fitted fairly close to the line as well with the exception of a few, which could be result of the u factor.

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.

The value would be roughly 56.75 dollars for the 10 year wine and 51.25 dollars for the 9 year wine. The difference is the slope, which is 5.50 dollars.

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.

Yes. Every year it will raise 5.50 dollars. This is because without accounting for the value of the errors, this will tend to be the difference between the values.

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?

After crunching some numbesr we find the difference can be between 11.50 and 42.50 dollars through taking the lower end error and upper end error of the change and it is very much within range to be 40 dollars, however it will be somewhat unlikely, as the average tends to be closer to $27.5 dollars.

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.

[19]:

```python
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')
print(np.mean(ces['foodshare']))
results_4a.summary()
```

```
0.177105215511823
```

[19]:

```
<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:                Mon, 01 Mar 2021   Prob (F-statistic):             0.0363
Time:                        10:48:43   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.

The estimated coefficent on family size is very low, in fact I am given 0.0047 which tells me its not as significant. It is economically or statistically not significant since you can have a large family, but if you make a lot of money, everyone is fed well, and you can have a small family and no one working. Due to this finding, we can assume larger families if they make the money to support feeding their family they can enjoy luxuries.

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

0.1654 + 0.0047 * 3 = .1795, which shows that the predicted share of expenditures is a little above average. (the first number shown on 4a)

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]:

```python
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:                Mon, 01 Mar 2021   Prob (F-statistic):              0.135
Time:                        10:50:22   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 results are similar, and there was little to no change in the coefficient and intercept. If anything, it is a little more than fam_size, although the R^2 is less. Regardless, the estimation is not accurate enough for us to predict anything.

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.

There can be a trend, however the R^2 tells us that the values are extremely and loosely fit around the line of best fit that we created, and thus it cannot be a reliable method to determine family size and food share are related, at least linerly.

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.

[35]:
```
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['fam_size'])
model_4g =  sm.OLS(y_4g, X_4g)
results_4g = model_4g.fit(cov_type = 'HC1')
results_4g.summary()
```

[35]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              foodshare   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.8087
Date:                Mon, 01 Mar 2021   Prob (F-statistic):              0.369
Time:                        11:05:38   Log-Likelihood:                 445.57
No. Observations:                 532   AIC:                            -887.1
Df Residuals:                     530   BIC:                            -878.6
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2177      0.010     21.179      0.000       0.198       0.238
fam_size      -0.0027      0.003     -0.899      0.369      -0.008       0.003
==============================================================================
Omnibus:                      198.101   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              899.278
Skew:                           1.617   Prob(JB):                    5.30e-196
Kurtosis:                       8.487   Cond. No.                         7.18
==============================================================================

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

Question 4.h. Explain.

We get something very similar as the ones before, however here we see that the coefficent for fam size is now negative which implies as we add more family members the less the food share is per person. This is different than what we saw before since they limit the per capita expenditure to 3000 and as we have less money for more and more people we are unable to earn more money to make up for that loss.

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.

[36]:

```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()
```

[36]:

```
<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:                Mon, 01 Mar 2021   Prob (F-statistic):           1.71e-21
Time:                        11:06:26   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.

This makes a lot more sense and follows our trend we saw before with the coefficent of family size being negative and decreasing as each member is added per capita expendeture. Although the R^2 is not strong enough to tell us it fits this line perfectly, we can estimate and consider the error that confounding variables contribute as well (which would be working individuals, age of individuals (as children and elderly tend to eat less than adults))