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.

```
In [ ]: v = np.random.random_sample(...)
        x = ...
        u = np.random.normal(..., ..., ...)
        y = ...

        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.

```
In [ ]: X_1b = sm.add_constant(...)
        model_1b = sm.OLS(..., ...)
        results_1b = model_1b.fit(...)
        results_1b.summary()
```

Question 1.c. Explain.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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)`.*

```
In [ ]: sum_of_residuals = np.sum(...)
        print("Sum of residuals: ", sum_of_residuals)
        np.corrcoef(..., ...)
```

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.

```
In [ ]: v_1000 = np.random.sample(...)
        x_1000 = ...
        u_1000 = np.random.normal(..., ..., ...)
        y_1000 = ...

        X_1f = ...
        model_1f = sm.OLS(..., ...)
        results_1f = model_1f.fit(...)
        results_1f.summary()
```

Question 1.h. Explain.

Type your answer here, replacing this text.

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.

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

Question 2.b. Explain.

Type your answer here, replacing this text.

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.

```
In [ ]: y_2c = ...
        X_2c = sm.add_constant(...)
        model_2c = sm.OLS(..., ...)
        results_2c = ...
        results_2c.summary()
```

Question 2.d. Explain.

Type your answer here, replacing this text.

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?

Type your answer here, replacing this text.

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 $H_1 : \beta_1 \neq 100$.

Type your answer here, replacing this text.

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.

```
In [ ]: plt.scatter(..., ...)
        plt.xlabel(...)
        plt.ylabel(...)
        plt.title(...);
```

Question 2.h. Explain.

Type your answer here, replacing this text.

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.

```
In [ ]: y_2i = ...
        X_2i = ...
        model_2i = ...
        results_2i = ...
        results_2i.summary()
```

Question 2.j. Explain.

Type your answer here, replacing this text.

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.

```
In [ ]: wages_men = ...
        wages_women = ...

        ttest_2k = stats.ttest_ind(..., ..., ...)

        tstat_2k = ttest_2k.statistic
        pval_2k = ttest_2k.pvalue

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

Question 2.l. Explain.

Type your answer here, replacing this text.

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?

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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?

Type your answer here, replacing this text.

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.

```
In [ ]: ces['foodshare'] = ...
        y_4a = ...
        X_4a = ...
        model_4a = ...
        results_4a = ...
        results_4a.summary()
```

Question 4.b. Explain.

Type your answer here, replacing this text.

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

Type your answer here, replacing this text.

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.

```
In [ ]: ces['log_fam_size'] = ...
        y_4d = ...
        X_4d = ...
        model_4d = ...
        results_4d = ...
        results_4d.summary()
```

Question 4.e. Explain.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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.

```
In [ ]: ces['exp_pc'] = ...
        ces_3000 = ...
        y_4g = ...
        X_4g = ...
        model_4g = ...
        results_4g = ...
        results_4g.summary()
```

Question 4.h. Explain.

Type your answer here, replacing this text.

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.

```
In [ ]: y_4i = ...
        X_4i = ...
        model_4i = ...
        results_4i = ...
        results_4i.summary()
```

Question 4.j. Explain.

Type your answer here, replacing this text.