Question 1.a. Set the sample size at 1,000 and generate an error term, u_i, by randomly selecting from a normal distribution with mean 0, and standard deviation 5. Draw an explanatory variable, X_{1i}, from a standard normal distribution, $\mathcal{N}(0,1)$, and then define a second explanatory variable, X_{2i}, to be equal to $e^{X_{1i}}$ for all i. Finally, set the dependent variable to be linearly related to the two regressors plus an additive error term: $y_i = 2 + 4X_{1i} - 6X_{2i} + u_i$. Note that, by construction, the error term of this multivariate linear regression is homoskedastic.

Hint: You may want to refer to how you did this in Problem Set 2. Also, the function `np.exp()` takes a list/array of numbers and applies the exponential function to each element. This is basically the opposite funciton of `np.log()`.

```
In [4]: u = ...
        X1 = ...
        X2 = ...
        y = ...
```

Question 1.b. Regress y on X_1 with homoskedasticity-only standard errors (`statsmodels` does this by default, just don't specify a `cov_type` like we usually do to get robust errors). Do the same analysis for y and X_2. Compare the results with the true data generating process. Explain why differences arise between the population slopes and the estimated slopes, if there are any.

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

```
In [5]: X1_const = ...
        model_1b_X1 = ...
        results_1b_X1 = ...
        results_1b_X1.summary()
```

Question 1.c. Explain.

Type your answer here, replacing this text.

Question 1.d. Next, regress y on both X_1 and X_2. Compare the estimation results with those you did in part (b/c), especially the model with only the regressor X_1. Examine differences across the three regressions in terms of the coefficient estimates, their standard errors, the R^2, and the adjusted R^2.

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

```
In [8]: X_const = sm.add_constant(np.stack([X1, X2], axis=1)) # This just puts our two variables togeth
        model_1d = ...
        results_1d = ...
        results_1d.summary()
```

Question 1.e. Explain.

Type your answer here, replacing this text.

Question 1.f. Generate a third regressor: $X_{3i} = 1 + X_{1i} - X_{2i} + v_i$ where v_i is drawn from a normal distribution with mean 0 and standard deviation 0.5. Estimate the model $y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + w_i$. Compare the result with part (d/e). Do changes in OLS estimates, standard errors, the R^2, and the adjusted R_2 make sense to you? Explain why or why not.

Hint: Think about the concept of "imperfect multicollinearity".

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

```
In [9]: v = ...
        X3 = ...

        X_const_f = sm.add_constant(np.stack([X1, X2, X3], axis=1))
        model_1f = ...
        results_1f = ...
        results_1f.summary()
```

Question 1.g. Explain.

Type your answer here, replacing this text.

Question 2.a. Run a regression of `course_eval` on `beauty` using robust standard errors. What is the estimated slope? Is it statistically significant?

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

```
In [12]: y_2a = ...
         X_2a = ...
         model_2a = ...
         results_2a = ...
         results_2a.summary()
```

Question 2.b. Explain.

Type your answer here, replacing this text.

Question 2.c. Run a regression of `course_eval` on `beauty`, including some additional variables to control for the type of course and professor characteristics. In particular, include as additional regressors `intro`, `onecredit`, `female`, `minority`, and `nnenglish`. What is the estimated effect of `beauty` on `course_eval`? Does the regression in (a) suffer from important omitted variable bias (OVB)? What happens with the R^2? Based on the confidence interval from the regression, can you reject the null hypothesis that the effect of beauty is the same as in part (a)? What can you say about the effect of the new variables included?

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

```
In [13]: y_2c = ...
         X_2c = ...
         model_2c = ...
         results_2c = ...
         results_2c.summary()
```

Question 2.d. Explain.

Type your answer here, replacing this text.

Question 2.e. Estimate the coefficient on beauty for the multiple regression model in (c) using the three-step process in Appendix 6.3 (the Frisch-Waugh theorem). Verify that the three-step process yields the same estimated coefficient for beauty as that obtained in (c). Comment.

Hint: Recall that if your regression results are called ***`results`****, you could get the residuals using* ***`results.resid`****.*

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

```
In [14]: # Do the first step here (regress the outcome variable on covariates)
         course_eval = ...
         covariates = ...
         model_eval_on_covariates = ...
         results_eval = ...
         eval_residuals = ...

         # Do the second step here (regress the explanatory variable on covariates)
         beauty = ...
         model_beauty_on_covariates = ...
         results_beauty = ...
         beauty_residuals = ...

         # Do the last step here (regress the outcome variable's residuals on the explanatory variable'
         model_fw = ...
         results_fw = ...
         results_fw.summary()
```

Question 2.f. Explain.

Type your answer here, replacing this text.

Question 2.g. Professor Smith is a black male with average beauty and is a native English speaker. He teaches a three-credit upper-division course. Predict Professor Smith's course evaluation.

Type your answer here, replacing this text.

Question 3.a. What do you expect for the sign of the relationship and what mechanism can you think about to explain it?

Type your answer here, replacing this text.

Question 3.b. Run a regression of years of completed education (`yrsed`) on distance to the nearest college (`dist`), measured in tens of miles (For example, dist = 2 means that the distance is 20 miles). What is the estimated slope? Is it statistically significant? Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain.

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

```
In [16]: y_3b = ...
         X_3b = ...
         model_3b = ...
         results_3b = ...
         results_3b.summary()
```

Question 3.c. Explain.

Type your answer here, replacing this text.

Question 3.d. Now run a regression of `yrsed` on `dist`, but include some additional regressors to control for characteristics of the student, the student's family, and the local labor market. In particular, include as additional regressors: `bytest`, `female`, `black`, `hispanic`, `incomehi`, `ownhome`, `dadcoll`, `cue80`, and `stwmfg80`. What is the estimated effect of `dist` on `yrsed`? Is it substantively different from the regression in (b)? Based on this, does the regression in (b) seem to suffer from important omitted variable bias?

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

```
In [17]: y_3d = ...
         X_3d = ...
         model_3d = ...
         results_3d = ...
         results_3d.summary()
```

Question 3.e. Explain.

Type your answer here, replacing this text.

Question 3.f. The value of the coefficient on `dadcoll` is positive. What does this coefficient measure? Interpret this effect.

Type your answer here, replacing this text.

Question 3.g. Explain why `cue80` and `stwmfg80` appear in the regression. Are the signs of their estimated coefficients what you would have believed? Explain.

Type your answer here, replacing this text.

Question 3.h. Bob is a black male. His high school was 20 miles from the nearest college. His base-year composite test score (`bytest`) was 58. His family income in 1980 was \$26,000, and his family owned a home. His mother attended college, but his father did not. The unemployment rate in his county was 7.5%, and the state average manufacturing hourly wage was \$9.75. Predict Bob's years of completed schooling using the regression in (d).

Type your answer here, replacing this text.