Notebook

March 15, 2021

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()`.

[3]:
```
u = np.random.normal(0, 5, 1000)
X1 = np.random.normal(0, 1, 1000)
X2 = np.exp(X1)
y = 2 + 4 * X1 - 6 * X2 + u
```

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.

[4]:
```
X1_const = sm.add_constant(X1)
model_1b_X1 = sm.OLS(y, X1_const)
results_1b_X1 = model_1b_X1.fit()
results_1b_X1.summary()
```

[4]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.233
Model:                            OLS   Adj. R-squared:                  0.232
Method:                 Least Squares   F-statistic:                     302.6
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           2.14e-59
Time:                        16:45:48   Log-Likelihood:                -3489.4
No. Observations:                1000   AIC:                             6983.
Df Residuals:                     998   BIC:                             6993.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -7.7023      0.251    -30.639      0.000      -8.196      -7.209
x1            -4.3152      0.248    -17.394      0.000      -4.802      -3.828
==============================================================================
Omnibus:                      401.992   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2648.072
Skew:                          -1.702   Prob(JB):                         0.00
Kurtosis:                      10.208   Cond. No.                         1.06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 1.c. Explain.

The R squared of X2 is signficantly stronger than X1, probably due to the fact that taking the exponetial function of X1 makes it tighter by applying a function to the curve. Another note–there is a negative correlation between the two which is rather interesting in the way it affects the curves. Thus, it would make more sense to use X2 rather than X1 for a more clear and accurate outcome.

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.

[18]:

```
X_const = sm.add_constant(np.stack([X1, X2], axis=1)) # This just puts our two␣
 ↪variables together with a const
model_1d = sm.OLS(y, X_const)
results_1d = model_1d.fit()
results_1d.summary()
```

[18]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.682
Model:                            OLS   Adj. R-squared:                  0.682
Method:                 Least Squares   F-statistic:                     1071.
Date:                Mon, 15 Mar 2021   Prob (F-statistic):          5.07e-249
Time:                        16:47:06   Log-Likelihood:                -3048.4
```

```
No. Observations:                1000   AIC:                             6103.
Df Residuals:                     997   BIC:                             6118.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.1787      0.309      7.056      0.000       1.573       2.785
x1             4.1860      0.277     15.115      0.000       3.643       4.729
x2            -6.1565      0.164    -37.572      0.000      -6.478      -5.835
==============================================================================
Omnibus:                       10.170   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.006   Jarque-Bera (JB):                6.878
Skew:                          -0.037   Prob(JB):                       0.0321
Kurtosis:                       2.600   Cond. No.                         6.28
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 1.e. Explain.

With more factors added in, the R^2 value is a lot higher than the other two previously which implies we have a much better model that reflects mutiple factors. As well, we can see that x1 now has a postive coefficient, which without doing MOLS, we would not have noticed that the x2 was causing a more negative heavy weight than x1.

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.

[19]:

```python
v = np.random.normal(0, .5, 1000)
X3 = 1 + X1 - X2 + v

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

[19]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
```

```
==============================================================================
Dep. Variable:                      y   R-squared:                       0.683
Model:                            OLS   Adj. R-squared:                  0.682
Method:                 Least Squares   F-statistic:                     714.2
Date:                Mon, 15 Mar 2021   Prob (F-statistic):          1.20e-247
Time:                        16:47:06   Log-Likelihood:                -3048.0
No. Observations:                1000   AIC:                             6104.
Df Residuals:                     996   BIC:                             6124.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.4773      0.443      5.596      0.000       1.609       3.346
x1             4.4845      0.421     10.652      0.000       3.658       5.311
x2            -6.4589      0.361    -17.906      0.000      -7.167      -5.751
x3            -0.2975      0.316     -0.941      0.347      -0.918       0.323
==============================================================================
Omnibus:                       10.367   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.006   Jarque-Bera (JB):                6.976
Skew:                          -0.036   Prob(JB):                       0.0306
Kurtosis:                       2.597   Cond. No.                         12.2
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 1.g. Explain.

It seems like when we added in the third variable X3, it did not matter as much (as the coefficent is very close to zero in comparison to X2 and X1) nor did it change the R^2 value. The standard errors also are higher than just X1 and X2 on them. Personally I believe that the X1/X2 is more effective on its own and that whatever X3 indicates is something that doesnt apply as strongly.

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.

[21]:

```
y_2a = ratings['course_eval']
X_2a = sm.add_constant(ratings['beauty'])
model_2a = sm.OLS(y_2a, X_2a)
results_2a = model_2a.fit()
results_2a.summary()
```

[21]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            course_eval   R-squared:                       0.036
Model:                            OLS   Adj. R-squared:                  0.034
Method:                 Least Squares   F-statistic:                     17.08
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           4.25e-05
Time:                        16:47:07   Log-Likelihood:                -375.32
No. Observations:                 463   AIC:                             754.6
Df Residuals:                     461   BIC:                             762.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.9983      0.025    157.727      0.000       3.948       4.048
beauty         0.1330      0.032      4.133      0.000       0.070       0.196
==============================================================================
Omnibus:                       15.399   Durbin-Watson:                   1.410
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               16.405
Skew:                          -0.453   Prob(JB):                     0.000274
Kurtosis:                       2.831   Cond. No.                         1.27
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 2.b. Explain.

The estimated slope seems to be .1330 which does indicated a positive correlation, however, on a scale 1-5, it does not seem to change the result of course eval as much. Because of this, I believe that beauty is in fact not a determining factor in how high course evals scores are.

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.

```
[22]: y_2c = ratings['course_eval']
      X_2c = sm.add_constant(np.stack([ratings['beauty'], ratings['female'],␣
       ↪ratings['nnenglish'], ratings['minority']], axis=1))
      model_2c = sm.OLS(y_2c, X_2c)
      results_2c = model_2c.fit()
```

```
results_2c.summary()
```

```
[22]: <class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:            course_eval   R-squared:                       0.087
Model:                            OLS   Adj. R-squared:                  0.079
Method:                 Least Squares   F-statistic:                     10.95
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           1.75e-08
Time:                        16:47:08   Log-Likelihood:                -362.61
No. Observations:                 463   AIC:                             735.2
Df Residuals:                     458   BIC:                             755.9
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.1046      0.034    121.570      0.000       4.038       4.171
x1             0.1499      0.032      4.733      0.000       0.088       0.212
x2            -0.1944      0.051     -3.822      0.000      -0.294      -0.094
x3            -0.3160      0.109     -2.910      0.004      -0.529      -0.103
x4            -0.0388      0.076     -0.513      0.608      -0.187       0.110
==============================================================================
Omnibus:                       17.214   Durbin-Watson:                   1.448
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.477
Skew:                          -0.489   Prob(JB):                     9.72e-05
Kurtosis:                       2.984   Cond. No.                         5.08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 2.d. Explain.

Beauty still falls under a low threshold of slope, as do most of the others, with 'onecredit' seeming to be one of the higher values. Age and gender also seem to have little to no effect on the data and if they do, it is negatively correlated. The good news is that our R squared value is 4 times stronger, however it is still extremely low and considered no correlation if so. I would reject the null hypothesis that beauty affect the result of the course evalutations. The effect of new variables added allow us to see different contributions that contribute to the decision of evaluations, but it seems that they typically run somewhat independant of features such as gender and age.

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.

[23]:

```
# Do the first step here (regress the outcome variable on covariates)
course_eval = ratings['course_eval']
covariates = sm.add_constant(np.stack([ratings['minority'], ratings['female'],␣
 ↪ratings['onecredit'], ratings['age']], axis=1))
model_eval_on_covariates = sm.OLS(course_eval, covariates)
results_eval = model_eval_on_covariates.fit()
eval_residuals = results_eval.resid

# Do the second step here (regress the explanatory variable on covariates)
beauty = ratings['beauty']
model_beauty_on_covariates = sm.OLS(beauty, covariates)
results_beauty = model_beauty_on_covariates.fit()
beauty_residuals = results_beauty.resid

# Do the last step here (regress the outcome variable's residuals on the␣
 ↪explanatory variable's residuals)
model_fw = sm.OLS( eval_residuals, beauty_residuals)
results_fw = model_fw.fit()
results_fw.summary()
```

[23]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      y   R-squared (uncentered):
0.051
Model:                            OLS   Adj. R-squared (uncentered):
0.049
Method:                 Least Squares   F-statistic:
24.99
Date:                Mon, 15 Mar 2021   Prob (F-statistic):
8.21e-07
Time:                        16:47:09   Log-Likelihood:
-347.36
No. Observations:                 463   AIC:
696.7
Df Residuals:                     462   BIC:
700.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
------------------------------------------------------------------------------
x1             0.1594      0.032      4.999      0.000       0.097       0.222
==============================================================================
Omnibus:                       20.756   Durbin-Watson:                   1.505
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.470
Skew:                          -0.536   Prob(JB):                     1.32e-05
Kurtosis:                       3.122   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 2.f. Explain.

Both values are in fact extremely low from what I can gather. The reason for doing this is to ensure that no other variable cross checks with that variable so we know more about the actual values and strength they have on the outcome.

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.

4.1 + -0.0388(1) = 4.06 roughly. (Given that NNEgnlish = 0, Avg beauty = 0, minority = 1, intro = 0, onecredit= 0)

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

I think the distance will be positive, as one goes to college away from home they focus better as well as do not have to worry about transportation and food as it is provided.

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.

[25]:

```python
y_3b = dist['yrsed']
X_3b = sm.add_constant(dist['dist'])
model_3b = sm.OLS(y_3b, X_3b)
results_3b = model_3b.fit()
results_3b.summary()
```

[25]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  yrsed   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.007
```

```
Method:                 Least Squares   F-statistic:                     28.48
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           1.00e-07
Time:                        16:47:10   Log-Likelihood:                -7632.2
No. Observations:                3796   AIC:                         1.527e+04
Df Residuals:                    3794   BIC:                         1.528e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.9559      0.038    369.945      0.000      13.882      14.030
dist          -0.0734      0.014     -5.336      0.000      -0.100      -0.046
==============================================================================
Omnibus:                     7187.794   Durbin-Watson:                   1.769
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              361.676
Skew:                           0.410   Prob(JB):                     2.90e-79
Kurtosis:                       1.729   Cond. No.                         3.73
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 3.c. Explain.

Since I got a value of -0.0734 it seems that distance does not have as big of an effect on the years of school a student chooses to pursue. In fact the R2 for the OLS is extrememly low and is very unviable for prediction.

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.

[26]:
```
y_3d = dist['yrsed']
X_3d = sm.add_constant(np.stack([dist['dist'], dist['female'],␣
 ↪dist['hispanic'], dist['black']], axis=1))
model_3d = sm.OLS(y_3d, X_3d)
results_3d = model_3d.fit()
results_3d.summary()
```

[26]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
```

```
==============================================================================
Dep. Variable:                  yrsed   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.021
Method:                 Least Squares   F-statistic:                     21.14
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           2.95e-17
Time:                        16:47:11   Log-Likelihood:                -7604.5
No. Observations:                3796   AIC:                         1.522e+04
Df Residuals:                    3791   BIC:                         1.525e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         14.1082      0.054    260.962      0.000      14.002      14.214
x1            -0.0830      0.014     -6.048      0.000      -0.110      -0.056
x2             0.0057      0.059      0.098      0.922      -0.109       0.121
x3            -0.2050      0.083     -2.458      0.014      -0.369      -0.042
x4            -0.5614      0.076     -7.396      0.000      -0.710      -0.413
==============================================================================
Omnibus:                     5005.538   Durbin-Watson:                   1.776
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              338.878
Skew:                           0.387   Prob(JB):                     2.59e-74
Kurtosis:                       1.757   Cond. No.                         9.03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 3.e. Explain.

This is experiencing the same issue and ommitted variables are definitly causing some conundrums here. I will agree that some valuables including debt, workload, and others will affect how and when the student decides to go to school rather than not.

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

The coefficent measures the effect that the individuals dad went to college. Typically those with college degrees tends to get better education to use for more office-like jobs rather than blue collar jobs that most out of high school acheive. By having this known throughout your life there might also be an expectation to complete college as well in ones household.

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.

The unemployment rate and the paywage is very important to keep in mind as it will allow us to get a better picture, however the coeffiencts, since all the same will add to 0 as they add in no additional detail that we didnt alreadt know.

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).

```
[27]: 14.1082 + (20) * -0.0830 + (1)*-0.5614
```

```
[27]: 11.8868
```

Question 4.a. Why do you think Jaeger and Page estimate their model using only people of a single race and gender (in this particular case the sample consists of white males)?

By taking out any confounding variables such as race and gender, they can ensure they recieve results that pertain just to that group of people and how they would end up faring due to their specific groups.

Question 4.b. Look at column (3) of the table. In words, interpret the coefficient on the dummy variable "9".

Hint: Note that "12" is the omitted category.

12 I will assume is our baseline reference for each, so if you are 9 years (assuming high schooler freshman) you understandably would make less than you would if you were graduated.

Question 4.c. Why do you think the effect of the 14th year of education is larger than that of the 15th?

The 14th year I assume is an assosiates degree which implies that if you pass that point you will have credentials to get a better job than one who is in their 3rd year of college on the way to their bachelors degree.

Question 4.d. Now look at column (4). Think about a student who is currently a senior. What is the average difference in the student's wage now and the one that the student could get at the end of the year following graduation?

The difference in wage is .02. Since it is positive, it is a better rate than you would just finishing.

Question 4.e. Based on the results presented in this column, would you rather choose to complete a PhD or a professional degree? Explain.

I would prefer to go for the PhD since it has .5 as the coefficient while professional degree has .27 which is not nearly as high.

Question 4.f. Using the results from columns (3) and (4), how would you test the presence of a "diploma effect"? Carry out the test at a 5% significance level.

Hint: You may find some of the information you need in the footnote of the table.

Straight up, I have like 4 minutes to submit this and did not have enough time to complete this. I have 3 midterms this week all due within 2 days so I have been stressed to the moon. I accept the L on this assignment but it kinda do hurt you know.