# 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()`.

[30]:

```
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.

[31]:

```
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()
```

[31]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.314
Model:                            OLS   Adj. R-squared:                  0.313
Method:                 Least Squares   F-statistic:                     456.7
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           9.94e-84
Time:                        16:14:39   Log-Likelihood:                -3520.4
No. Observations:                1000   AIC:                             7045.
Df Residuals:                     998   BIC:                             7055.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -7.7323      0.259    -29.868      0.000      -8.240      -7.224
x1            -5.6348      0.264    -21.370      0.000      -6.152      -5.117
==============================================================================
Omnibus:                      360.796   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2058.779
Skew:                          -1.547   Prob(JB):                         0.00
Kurtosis:                       9.312   Cond. No.                         1.02
==============================================================================

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

**Question 1.c.** Explain.

Accoding to the equation given yi = 2 + 4X1i−6X2i + ui, the true parameter for  0 should be 2 and the coefficient for X1 and X2 should be 4 and -6 respectively. However, the regression model we conducted missed both of these coefficients. Being more specific, the confidence interval generated for X1 is (-6.361, -5.209), which undoubtedly exclude the true parameter 4. Similarly, the the confidence interval generated for X2 is (-6.361, -5.209), which also does not include the true parameter -6. The reason for this accuracy is the violation of the first assumption of OSL, which is the conditional mean zero errors assumption. Being more specific, the error term for X1 is −6X2i + ui. However, the expectation for this error term given X1i does not equal to 0. Similarly, the error term for X2 is 4X1i + ui, whose expectation also does not evaluate to 0 given X2i

**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.

[33]:

```
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()
```

[33]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.729
```

```
Model:                            OLS   Adj. R-squared:                  0.728
Method:                 Least Squares   F-statistic:                     1340.
Date:                Mon, 15 Mar 2021   Prob (F-statistic):          2.59e-283
Time:                        16:14:41   Log-Likelihood:                -3056.2
No. Observations:                1000   AIC:                             6118.
Df Residuals:                     997   BIC:                             6133.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.9954      0.298      6.707      0.000       1.412       2.579
x1             3.8166      0.293     13.012      0.000       3.241       4.392
x2            -6.0597      0.155    -39.065      0.000      -6.364      -5.755
==============================================================================
Omnibus:                        1.057   Durbin-Watson:                   2.050
Prob(Omnibus):                  0.590   Jarque-Bera (JB):                1.088
Skew:                          -0.016   Prob(JB):                        0.580
Kurtosis:                       2.841   Cond. No.                         6.60
==============================================================================

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

**Question 1.e.** Explain.

Compared to the OLS conducted in 1b, this OLS is much more accurate on capturing the coefficient 0, X1, and X2. The true parameter for these three terms are 2, 4, and -6. From the chart above, we can see that all three numbers are included in the confidence interval. This increment in accuration is also reflected in the increased R and the adjusted R . In this OLS, Both the R and the adjusted R2 for this OLS increases to 0.859. Compared to the R2s conducted above, this increased closer-to-1 R2 and adjusted R2indicates a comparatively stronger relationship between the independent variables and the dependent variable.

**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.

[34]:
```
v = np.random.normal(0, 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()
```

```
[34]: <class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.729
Model:                            OLS   Adj. R-squared:                  0.728
Method:                 Least Squares   F-statistic:                     892.7
Date:                Mon, 15 Mar 2021   Prob (F-statistic):          1.06e-281
Time:                        16:14:42   Log-Likelihood:                -3056.1
No. Observations:                1000   AIC:                             6120.
Df Residuals:                     996   BIC:                             6140.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.9431      0.450      4.319      0.000       1.060       2.826
x1             3.7654      0.442      8.522      0.000       2.898       4.633
x2            -6.0082      0.367    -16.367      0.000      -6.729      -5.288
x3             0.0511      0.330      0.155      0.877      -0.596       0.698
==============================================================================
Omnibus:                        1.063   Durbin-Watson:                   2.050
Prob(Omnibus):                  0.588   Jarque-Bera (JB):                1.094
Skew:                          -0.016   Prob(JB):                        0.579
Kurtosis:                       2.841   Cond. No.                         13.4
==============================================================================

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

**Question 1.g.** Explain.

Compared to the OLS conducted in part e, we observe that the accuracy for every term, including coefficient, X1, X2, and X3, decreases. However, this OLS still captures the true parameter in the confidence intervals produced. Therefore, we concliude that this OLS is less accurate than that from part e but is more accurate than that from part d. However, it is worth noting that this OLS generates the greater standard error among all three OLSs. This feature should be attribute to the imperfect multicollinearity because X3 is composed as the linear combination of X1 and X2. This means that its variations can also be explained the linear combination of X1 and X2. R2 and the adjusted R2 remains the same.

**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.

[36]:

```python
y_2a = ratings['course_eval']
X_2a = ratings['beauty']
model_2a = sm.OLS(y_2a, sm.add_constant(X_2a))
results_2a = model_2a.fit(cov_type='HC1')
results_2a.summary()
```

[36]:

```
<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:                     16.94
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           4.58e-05
Time:                        16:14:44   Log-Likelihood:                -375.32
No. Observations:                 463   AIC:                             754.6
Df Residuals:                     461   BIC:                             762.9
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.9983      0.025    157.727      0.000       3.949       4.048
beauty         0.1330      0.032      4.115      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 are heteroscedasticity robust (HC1)
"""
```

**Question 2.b.** Explain.

The estimated slope here is 0.1330 and it is statistically significant on both the 0.05 level and 0.01 level. Therefore, this estimated slope may reflect a positive relationship between the beauty of the teachers and the ratings they received.

**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.

[37]:

```
y_2c = ratings['course_eval']
X_2c_beauty = ratings['beauty']
X_2c_intro = ratings['intro']
X_2c_onecredit = ratings['onecredit']
X_2c_female = ratings['female']
X_2c_minority = ratings['minority']
X_2c_nnenglish = ratings['nnenglish']
X_2c_const = sm.add_constant(np.stack([X_2c_beauty, X_2c_intro,␣
 ↪X_2c_onecredit,X_2c_female, X_2c_minority, X_2c_nnenglish], axis=1))
model_2c = sm.OLS(y_2c, X_2c_const)
results_2c = model_2c.fit(cov_type='HC1')
results_2c.summary()
```

[37]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:            course_eval   R-squared:                       0.155
Model:                            OLS   Adj. R-squared:                  0.144
Method:                 Least Squares   F-statistic:                     17.03
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           8.67e-18
Time:                        16:14:45   Log-Likelihood:                -344.85
No. Observations:                 463   AIC:                             703.7
Df Residuals:                     456   BIC:                             732.7
Df Model:                           6
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.0683      0.037    109.926      0.000       3.996       4.141
x1             0.1656      0.032      5.246      0.000       0.104       0.227
x2             0.0113      0.056      0.202      0.840      -0.099       0.121
x3             0.6345      0.108      5.871      0.000       0.423       0.846
x4            -0.1735      0.049     -3.505      0.000      -0.270      -0.076
x5            -0.1666      0.067     -2.472      0.013      -0.299      -0.034
x6            -0.2442      0.094     -2.608      0.009      -0.428      -0.061
==============================================================================
Omnibus:                       22.413   Durbin-Watson:                   1.516
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               24.406
Skew:                          -0.555   Prob(JB):                     5.02e-06
Kurtosis:                       3.179   Cond. No.                         5.81
==============================================================================

Warnings:
```

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

**Question 2.d.** Explain.

The estimated effect of beauty on course_eval increases from 0.1330 to 0.1656. However, we cannot reject the hypothesis that the effect of beauty is the same as in part (a) because the confidence interval conducted in c for beauty is (0.104, 0.227), which includes 0.1330. According to this result, we may conclude that the regression in (a) does seemingly suffers from important omitted variable bias (OVB). Among all other factors, x3, which is whether this class is taken as a one credit class, has the greatest weight. This means that instructors who have students taking classes in one credit would receive a higher rating on average. Approaching these coefficient for factors through a similar way, we find some outstanding results. On average, female instructors receive lower ratings than male, indicated by the negative sign in front of the coefficient. Similarly, instructors who is minority or non-English speakers have the same pattern compared with majority or English speaking instructors. On class level, instructors who teach introductary level class on average have higher rating. Incorporating all these factors into our model, we find that our R2 increases from 0.036 to 0.156. This means that the including all these factors make our OLS more accurate.

**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.

[38]:

```
# Do the first step here (regress the outcome variable on covariates)
course_eval = ratings['course_eval']
covariates = sm.add_constant(np.stack([X_2c_intro, X_2c_onecredit,␣
 ↪X_2c_female,X_2c_minority, X_2c_nnenglish], axis=1))
model_eval_on_covariates = sm.OLS(course_eval, covariates)
results_eval = model_eval_on_covariates.fit(cov_type='HC1')
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(cov_type='HC1')
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(cov_type='HC1')
results_fw.summary()
```

```
[38]: <class 'statsmodels.iolib.summary.Summary'>
      """
                                       OLS Regression Results
      ================================================================================
      =======
      Dep. Variable:                      y   R-squared (uncentered):
      0.060
      Model:                            OLS   Adj. R-squared (uncentered):
      0.058
      Method:                 Least Squares   F-statistic:
      27.88
      Date:                Mon, 15 Mar 2021   Prob (F-statistic):
      1.99e-07
      Time:                        16:14:46   Log-Likelihood:
      -344.85
      No. Observations:                 463   AIC:
      691.7
      Df Residuals:                     462   BIC:
      695.8
      Df Model:                           1
      Covariance Type:                  HC1
      ==============================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
      ------------------------------------------------------------------------------
      x1             0.1656      0.031      5.280      0.000       0.104       0.227
      ==============================================================================
      Omnibus:                       22.413   Durbin-Watson:                   1.516
      Prob(Omnibus):                  0.000   Jarque-Bera (JB):               24.406
      Skew:                          -0.555   Prob(JB):                     5.02e-06
      Kurtosis:                       3.179   Cond. No.                         1.00
      ==============================================================================

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

**Question 2.f.** Explain.

Estimating the coefficient on beauty for the multiple regression model using the Frisch-Waugh theorem, we could potentially isolate/minimize the effect of other factors, such as nnenglish, minority, onecredit, etc, which is counted as the marginal effect in this model and get the coefficient just for the factor beauty. The result we get using the Frisch-Waugh theorem is 0.1656, which is exactly the same as part(c)

**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.

4.0683 + -0.1666 * 1

minority = 1 female = 0 nnenglish = 0 onecredit = 0 intro = 0 beauty = average = 0 Therefore, predicted rating = 4.0683 + -0.1666 * 1 = 3.9017

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

I expect the sign of relationship to be negative. As on overage, the closer students's from high school to nearest four-year college, the larger the number of completed years of education.

**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.

[40]:

```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(cov_type = "HC1")
results_3b.summary()
```

[40]:

```
<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:                     29.83
Date:                Mon, 15 Mar 2021   Prob (F-statistic):           5.01e-08
Time:                        16:14:51   Log-Likelihood:                -7632.2
No. Observations:                3796   AIC:                         1.527e+04
Df Residuals:                    3794   BIC:                         1.528e+04
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.9559      0.038    369.093      0.000      13.882      14.030
dist          -0.0734      0.013     -5.462      0.000      -0.100      -0.047
==============================================================================
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 are heteroscedasticity robust (HC1)
```

```
"""
```

**Question 3.c.** Explain.

The estimated slope is -0.0734, it is statistically significant at 1% level. According to $R^2$, distance to college explain a small fraction of the variance in educational attainment across individuals.

**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.

[41]:
```
y_3d = dist['yrsed']
X_3d = sm.add_constant(dist[['dist', 'bytest', 'female', 'black', 'hispanic',␣
 ↪'incomehi', 'ownhome', 'dadcoll', 'cue80', 'stwmfg80']])
model_3d = sm.OLS(y_3d, X_3d)
results_3d = model_3d.fit(cov_type = "HC1")
results_3d.summary()
```

[41]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  yrsed   R-squared:                       0.279
Model:                            OLS   Adj. R-squared:                  0.277
Method:                 Least Squares   F-statistic:                     197.7
Date:                Mon, 15 Mar 2021   Prob (F-statistic):               0.00
Time:                        16:14:52   Log-Likelihood:                -7025.9
No. Observations:                3796   AIC:                         1.407e+04
Df Residuals:                    3785   BIC:                         1.414e+04
Df Model:                          10
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.8275      0.241     36.583      0.000       8.355       9.300
dist          -0.0315      0.012     -2.705      0.007      -0.054      -0.009
bytest         0.0938      0.003     31.479      0.000       0.088       0.100
female         0.1454      0.050      2.885      0.004       0.047       0.244
black          0.3680      0.068      5.449      0.000       0.236       0.500
hispanic       0.3985      0.074      5.394      0.000       0.254       0.543
incomehi       0.3952      0.062      6.382      0.000       0.274       0.517
ownhome        0.1521      0.065      2.343      0.019       0.025       0.279
dadcoll        0.6961      0.071      9.838      0.000       0.557       0.835
cue80          0.0232      0.009      2.493      0.013       0.005       0.041
```

```
stwmfg80      -0.0518      0.020     -2.632      0.008      -0.090      -0.013
==============================================================================
Omnibus:                      118.266   Durbin-Watson:                   1.924
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               97.867
Skew:                           0.320   Prob(JB):                     5.60e-22
Kurtosis:                       2.543   Cond. No.                         539.
==============================================================================

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

**Question 3.e.** Explain.

The estimated effect of dist on yrsed is -0.0315, it's statistically significant at 1% level. $R^2$ increased dramatically compared with part(a), and coefficients also decreased a lot. The regression in (b) did seem to suffer from important omitted variable bias.

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

Holding other variables constant, on overage, students with dadcoll = 1 (which means their father did go to college) have about 0.6961 more years of education compared with those with dadcoll = 0 (which means their father did not go to college).

**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.

Those reflect the opportunity cost of going to college. An increase in cue80 will lead to increase in years of schooling, as it is harder to find jobs. An increase in cue80 will lead to decrease in years of education.

**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).

-0.0315 * 2 + 0.0938 * 58 + 0.1454 * 0 + 0.3680 * 1 + 0.3985 * 0 + 0.3952 * 1 + 0.1521 * 1 + 0.6961 * 0 + 0.0232 * 7.5 - 0.0518 * 9.75 + 8.8275 = 14.79

**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)?

Because there might be racial and gender discrimination issues, thus it will be more reliable to use people of a single race and gender.

**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.*

On average, people with 9 years of education earns 0.227 less in log(hourly wages) compared with those with 12 years of education, while holding all other variables constant.

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

14th year of education might corresponds to those students graduating from a junior college(two-year college), and 15th year of education might corresponds to those who could not finish college and got dropped off. Compared with 15th year, employers may prefer 14th year.

**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?

Now: 0.052 + 0.083 = 0.135. After: 0.178 + 0.245 = 0.423. 0.423 - 0.135 = 0.288. Thus, there will be approximately a 28.8% increase.

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

Complete a professional degree other than PhD. Marginal effect of professional degress is 0.289, while marginal effect of PhD degree is 0.067.

**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.*

Null hypothesis: all diploma effect in (4) are zero. Alternative: at least one in (4) is not zero.

F = ((0.154-0.147)/8) / ((1-0.154)/(8957-28-1)) = 9.234

9.234 > 1.93, we reject the null.