Labor Economics, Spring 2021

Problem Set 1
Due Friday, March 19 in class

NOTE: For questions using Stata, please type your answers into a word processing document (using MS Word or something similar). Insert the relevant part of your Stata log file into your document, and clearly explain how your Stata output answers the question.

1. (15 points) Open the NLSY data set nlsy_S2016.csv, which contains the following 6 variables:

   wage (in dollars per hour)
   exp (years of experience)
   male (a variable that's equal to 1 for males and 0 for females)
   school (years of education)
   afqt (AFQT test score, which is very similar to an IQ test except it has a mean of zero)
   race (= 1 for non-Hispanic African-Americans, 2 for Hispanics, and 3 for non-Hispanic Caucasians)

   a. For the NLSY sample (individuals aged 14-22 in 1979), calculate the mean years of education, overall and for males and females separately. Also calculate the percent of the population that is female. [Hint: use the *summarize* command]

      See the Stata output below. Mean years of education in the full NLSY sample is 12.72. For men it's 12.51 and for women it's 12.95. The percent of the population that is female is (1-0.527) = 0.473.

```
. summarize school

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      school |      5898    12.72177    2.373274          0         20

. sum school if male==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      school |      2791    12.95199    2.295825          0         20

. sum school if male==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      school |      3107    12.51497    2.422531          3         20

. sum male

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
        male |      5898    .5267887    .4993242          0          1
```

b. Using the *tabulate* command, calculate the percentage of NLSY respondents who have 11 years of education or fewer.

20.30 percent of the NLSY respondents have 11 years of education or fewer.

```
. tab school

     school |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          2        0.03        0.03
          1 |          1        0.02        0.05
          3 |          4        0.07        0.12
          4 |          7        0.12        0.24
          5 |          5        0.08        0.32
          6 |         21        0.36        0.68
          7 |         33        0.56        1.24
          8 |        120        2.03        3.27
          9 |        243        4.12        7.39
         10 |        325        5.51       12.90
         11 |        436        7.39       20.30
         12 |      2,378       40.32       60.61
         13 |        531        9.00       69.62
         14 |        523        8.87       78.48
         15 |        235        3.98       82.47
         16 |        732       12.41       94.88
         17 |        130        2.20       97.08
         18 |         97        1.64       98.73
         19 |         37        0.63       99.36
         20 |         38        0.64      100.00
------------+-----------------------------------
      Total |      5,898      100.00
```

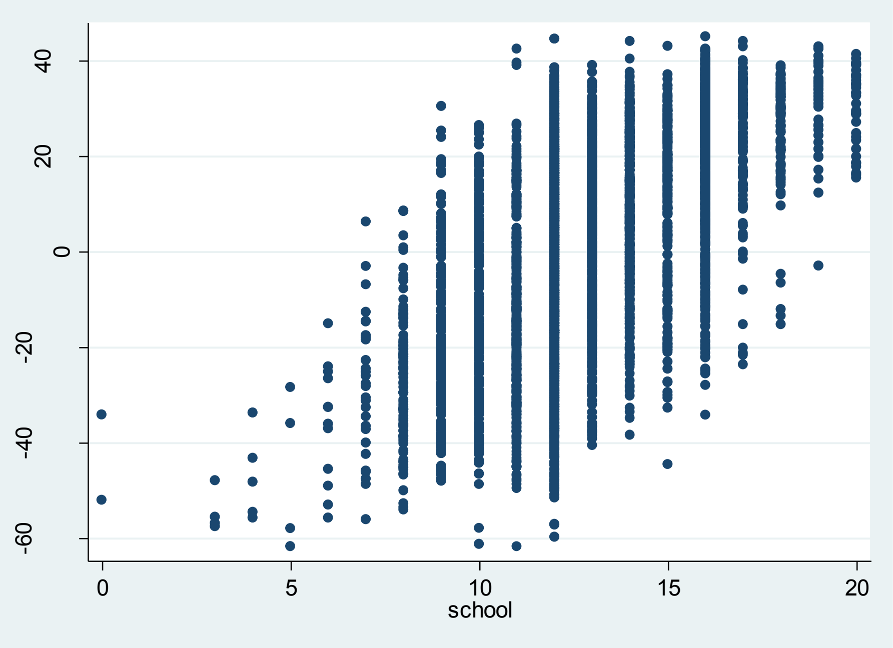
c. Again using the *tabulate* command, calculate the average AFQT score for each year of education in the sample. [Hint: *"tabulate var1, sum(var2)"* is the Stata command to get the average of a variable named "*var2*" for each value of a variable named "*var1*". You want to replace "*var1*" and "*var2*" with the names of the variables you'll use.] Does there seem to be a relationship between these two variables, and does this relationship make sense? Finally, type "*scatter afqt school"* to see a scatterplot of AFQT scores and years of education. Does the scatterplot seem to confirm the relationship you found above?

```
. tab school, sum(afqt)

            |          Summary of afqt
     school |        Mean   Std. Dev.       Freq.
```

```
------------+------------------------------------
          0 |   -43.13958   12.641853           2
          1 |  -41.279533           0           1
          3 |  -54.528173    4.467726           4
          4 |  -44.105703   9.6252849           7
          5 |   -46.99711   14.303985           5
          6 |  -38.978688   13.227395          21
          7 |  -25.970234   14.742069          33
          8 |  -26.222162   14.168921         120
          9 |  -20.459466   16.688453         243
         10 |  -14.843099    18.62481         325
         11 |  -13.600444   18.144445         436
         12 |  -3.2380301   19.139281        2378
         13 |   2.9902149   17.267753         531
         14 |   7.6569529   16.485581         523
         15 |    8.541884     17.4096         235
         16 |   19.436498   14.591826         732
         17 |    20.53743   15.007507         130
         18 |   24.267038   11.683994          97
         19 |     30.0113   9.9864336          37
         20 |   30.140038   7.8802097          38
------------+------------------------------------
      Total |  -.00514665   21.675705        5898
```

The average afqt score for each year of education in the sample is shown in the Stata output above. The relationship between year of education and afqt score is positive: students with more years of education have higher mean afqt scores, on average. It makes sense because it is arguably less costly for students who have higher afqt scores to achieve higher education.



The scatterplot of afqt score for each year of education in the sample is shown in the Stata output above. It is consistent with the relationship we found: students who have higher afqt scores tend to have more education.

d. Calculate the average logarithm of wages for two different groups – those who attended some college (*school*>12) and those who never attended college (*school*<=12). To do this, first generate the log wage variable by typing "*gen lwage= log(wage)*", and then generate a variable called "*college*" that's equal to one if the person went to college and zero otherwise by typing "*gen college = (school>12)*".

The average log wage for college attendees is 2.00, and for those who never went to college it is 1.68, as you can see from the Stata output below:

```
. gen college = school>12
. tab college, sum(lwage)
```

```
            |          Summary of lwage
    college |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   1.6758912   .65630401        3575
          1 |   2.0001521   .70869537        2323
------------+------------------------------------
      Total |   1.8036054   .69564887        5898
```

e. Test the hypothesis that the population mean of *lwage* for those who never went to college is equal to the population mean of *lwage* for those who did go to college (type "*ttest lwage, by(college)*", and note that there are two "t"s in "ttest"). Use a significance level α = 0.01.

I put the Stata command for the t-test below. I end up with a p-value of 0.0000 (it's the middle p-value shown below), which is less than any alpha I would choose. I chose alpha = 0.01, but any typical alpha like 0.01, 0.05, or 0.10 would give the same answer.

```
. ttest lwage, by(college)
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   3,575    1.675891    .0109766     .656304     1.65437    1.697412
       1 |   2,323    2.000152     .014704    .7086954    1.971318    2.028986
---------+--------------------------------------------------------------------
combined |   5,898    1.803605    .0090581    .6956489    1.785848    1.821363
---------+--------------------------------------------------------------------
    diff |           -.3242609     .018053               -.3596513   -.2888705
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -17.9616
Ho: diff = 0                                     degrees of freedom =     5896

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

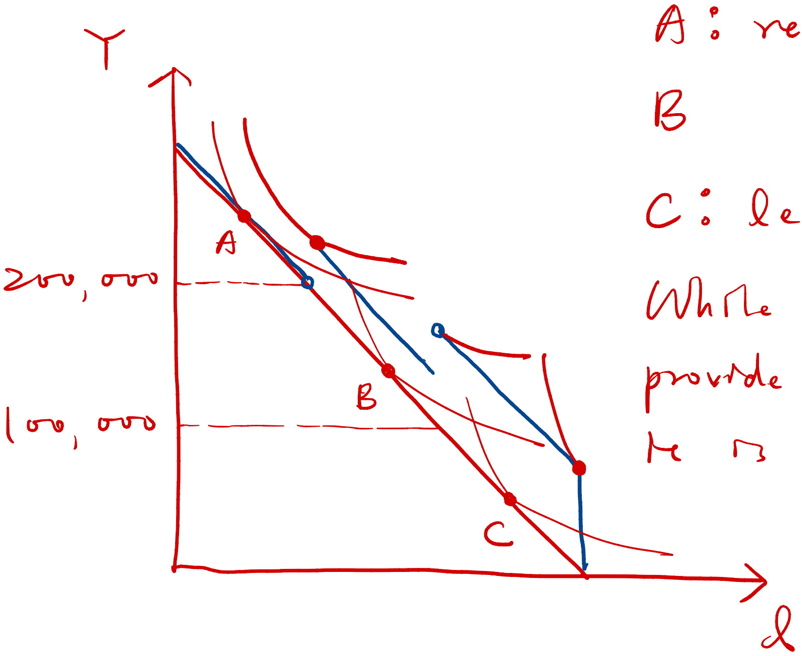
2. (5 points) Listen to the following episode of Freakonomics Radio:

https://freakonomics.com/podcast/covid-19-cities/

In this episode, Harvard economist, Ed Glaeser, talked about his preferred model to provide income support during a pandemic:

"Now, there's a question as to why make it unconditional. I think my preferred model is one in which everyone gets the checks, but let's say you come around to next year's tax season, if you're earning more than $200,000 and you spent that check, you're going to pay 100 percent tax on that check. If you're earning between one and $200,000, you're going to spend 50 percent. But if your earnings are less than $100,000, then the money's free and clear."

What is your prediction of the effect of his proposal on labor supply? Use labor supply model to explain.



A: reduce their earning to $200,000

B $100,000

C: leave labor force

While the proposed policy will provide income support to workers, He is predicted to reduce labor supply.