Problem set 5

Due on Thursday November 10, 2022 (by 11:59 PM EST)

Note: No credit will be given if you report only the final answers without showing formulas and calculations when appropriate. This applies to both theoretical and empirical questions. For the empirical questions, make sure to submit the R scripts and output on Latte. No credit will be given if the R output is missing.

Problem 1

Consider a wage equation for U.S. working married women,

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + u_i$$

where exper and exper^2 denotes years (and years squared) of work experience, while educ stands for years of education. We have good reasons to believe that exper and exper^2 are exogenous variables. On the other hand, the error term u is thought to be correlated with educ because of omitted ability, as well as other factors, such as quality of education and family background.

(a) Suppose that we can collect data on husband, mother, and father education - huseduc motheduc and fatheduc - and that we want to use these three variables as instruments for the endogenous variable educ. Write down the first stage equation that you should run in this case.

(b) Using data on 428 working married women in the US, you want to test whether the 3 instruments you have been considering are relevant. For that, you have just obtained the output of the following two regressions. Perform an F test (it's ok to assume errors are homoscedastic!) to look into this question. Based on the results of you test, would you conclude that the instruments are relevant or not?

Regression 1

$$\text{educ}_i = 12.36936 + 0.564919\, exper_i - 0.0019043\, exper_i^2 \qquad R^2 = 0.0049$$

Regression 2

$$\text{educ}_i = 5.538311 + 0.374977\, exper_i - 0.0006002\, exper_i^2 + 0.1141532\, motheduc_i + 0.1060801\, fatheduc_i + 0.3752548\, huseduc_i \quad R^2 = 0.4286$$

Problem 2

You are looking into the effect of education on fertility in developing countries using data for women in Botswana. The description of the variables available in the data, summary statistics for the data, and the results from three regressions are given below. Use this output to answer the following questions.

a. First consider Regression 1. (i) Interpret the coefficient on educ in words. (ii) If 100 women receive another year of education each, how many fewer children are they expected to have as a group?

b. Now compare Regression 2 to Regression 1. Which specification is preferable? Why? Be sure to consider both economic and statistical factors in your explanation.

c. The regressions all use "robust standard errors." What does that mean? Why are such errors being used?

d. The variable educ is unlikely to be exogenous. Explain why not.

e. You consider using the variable *frsthalf*, which is a dummy variable equal to 1 if the woman was born during the first six months of the year and 0 otherwise, as an instrument for *educ*. Is *frsthalf* a good instrument for *educ*? In answering this question, consider both logical arguments and the results of Regression 3. (Hint: Consider a situation where children enter school once they have reached age 5, but parents often make daughters drop out of school when they turn 14. Assume that school starts in August for everyone.)

```
  obs:         4,361
 vars:            27                          17 Aug 1999 15:26
 size:       139,552 (86.3% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
mnthborn        byte   %8.0g                  month woman born
yearborn        byte   %8.0g                  year woman born
age             byte   %8.0g                  age in years
electric        byte   %8.0g                  =1 if has electricity
radio           byte   %8.0g                  =1 if has radio
tv              byte   %8.0g                  =1 if has tv
bicycle         byte   %8.0g                  =1 if has bicycle
educ            byte   %8.0g                  years of education
ceb             byte   %8.0g                  children ever born
agefbrth        byte   %8.0g                  age at first birth
children        byte   %8.0g                  number of living children
knowmeth        byte   %8.0g                  =1 if know about birth control
usemeth         byte   %8.0g                  =1 if ever use birth control
monthfm         byte   %8.0g                  month of first marriage
yearfm          byte   %8.0g                  year of first marriage
agefm           byte   %8.0g                  age at first marriage
idlnchld        byte   %8.0g                  'ideal' number of children
heduc           byte   %8.0g                  husband's years of education
agesq           int    %8.0g                  age^2
urban           byte   %8.0g                  =1 if live in urban area


urb_educ        byte   %8.0g                 urban*educ
spirit          byte   %9.0g                 =1 if religion == spirit
protest         byte   %9.0g                 =1 if religion == protestant
catholic        byte   %9.0g                 =1 if religion == catholic
frsthalf        byte   %9.0g                 =1 if mnthborn <= 6
educ0           byte   %9.0g                 =1 if educ == 0
evermarr        byte   %9.0g                 =1 if ever married
-------------------------------------------------------------------------------
```

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mnthborn	4361	6.331346	3.323333	1	12
yearborn	4361	60.43362	8.682723	38	73
age	4361	27.40518	8.685233	15	49
electric	4358	.1402019	.3472363	0	1
radio	4359	.7017665	.457535	0	1
tv	4359	.0929112	.2903413	0	1
bicycle	4358	.2758146	.4469751	0	1
educ	4361	5.855996	3.927075	0	20
ceb	4361	2.441642	2.406861	0	13
agefbrth	3273	19.0113	3.092333	10	38
children	4361	2.267828	2.222032	0	13
knowmeth	4354	.9632522	.1881636	0	1
usemeth	4290	.5776224	.4939956	0	1
monthfm	2079	6.270322	3.619943	1	12
yearfm	2079	76.91246	7.760183	50	88
agefm	2079	20.68639	5.002383	10	46
idlnchld	4241	4.615892	2.219303	0	20
heduc	1956	5.144683	4.803028	0	20
agesq	4361	826.46	526.9232	225	2401
urban	4361	.5166246	.4997808	0	1
urb_educ	4361	3.469158	4.294228	0	20
spirit	4361	.4221509	.493959	0	1
protest	4361	.2277001	.4193961	0	1
catholic	4361	.1024994	.3033387	0	1
frsthalf	4361	.5404724	.4984164	0	1
educ0	4361	.2077505	.4057437	0	1
evermarr	4361	.4767255	.4995153	0	1

Regression 1

```
. regress children educ age agesq, robust

Regression with robust standard errors                Number of obs =    4361
                                                      F(  3,  4357) = 1922.00
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5687
                                                      Root MSE      =  1.4597

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0905755   .0060483   -14.98   0.000    -.1024332   -.0787178
         age |   .3324486   .0192071    17.31   0.000     .2947929    .3701043
       agesq |  -.0026308    .000352    -7.47   0.000    -.0033209   -.0019408
       _cons |  -4.138307   .2436211   -16.99   0.000    -4.615928   -3.660685
------------------------------------------------------------------------------
```

Regression 2

```
. regress children educ age agesq electric tv bicycle, robust

Regression with robust standard errors                Number of obs =    4356
                                                      F(  6,  4349) =  972.51
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5761
                                                      Root MSE      =  1.4478

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0767093   .0063442   -12.09   0.000    -.0891471   -.0642714
         age |   .3402038   .0191306    17.78   0.000      .302698    .3777096
       agesq |  -.0027081   .0003497    -7.74   0.000    -.0033937   -.0020225
    electric |  -.3027293   .0743809    -4.07   0.000    -.4485538   -.1569047
          tv |  -.2531443   .0826522    -3.06   0.002    -.4151846   -.0911039
     bicycle |    .317895   .0489639     6.49   0.000     .2219008    .4138892
       _cons |  -4.389784   .2444385   -17.96   0.000    -4.869008    -3.91056
------------------------------------------------------------------------------
```

Regression 3

```
. regress educ age agesq frsthalf, robust

Regression with robust standard errors                Number of obs =    4361
                                                      F(  3,  4357) =  201.72
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1077
                                                      Root MSE      =   3.711

------------------------------------------------------------------------------
             |               Robust
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1079504   .0402228    -2.68   0.007    -.1868076   -.0290932
       agesq |  -.0005056   .0006802    -0.74   0.457    -.0018392    .000828
    frsthalf |  -.8522854   .1132665    -7.52   0.000    -1.074345   -.6302254
       _cons |   9.692864   .5414317    17.90   0.000     8.631383    10.75435
------------------------------------------------------------------------------
```

Problem 3

Earnings functions, whereby the log of earnings is regressed on years of education, years of on the job training, and individual characteristics, have been studied for a variety of reasons. Some studies have focused on the returns to education, others on discrimination, union non-union differentials, etc. For all these studies, a major concern has been the fact that ability should enter as a determinant of earnings, but that it is close to impossible to measure and therefore represents an omitted variable.

Assume that the coefficient on years of education is the parameter of interest. Given that education is positively correlated to ability, since, for example, more able students attract scholarships and hence receive more years of education, the OLS estimator for the returns to education could be upward biased. To overcome this problem, various authors have used instrumental variable estimation techniques. For each of the instruments potential instruments listed below briefly discuss instrument validity.

(a) The individual's postal zip code.
(b) The individual's IQ or test score on a work related exam.
(c) Years of education for the individual's mother or father.
(d) Number of siblings the individual has.

Problem 4 (empirical)

Can television inform people about public affairs? It is a tricky question because those who watch a public-affair-oriented tv are well informed individual to begin with. Political scientists Bethany Albertson and Adria Lawrence in 2009 conducted a field experiment in which they randomly assigned people to treatment and control groups. Those assigned to the treatment group were told to watch a specific television broadcast about affirmative action and that they would be interviewed about what they had seen. Those in the control group were not told about the television program but were told that they would be interviewed again at a later time. The program they studied aired in California prior to vote on Proposition 209, a controversial proposition relating to affirmative action (more information here).

For this question, you will be using the data posted on Latte in the NewsStudy.RData file. Information on the key variables is below:

Variable name	Description
ReadNews	Political news reading habits (never = 1 to every day = 7)
PoliticalInterest	Interest in political affairs (not interested = 1 to very interested = 4)
Education	Education level (eighth grade or less = 1 to advanced graduate degree = 13)
TreatmentGroup	Assigned to watch program (treatment = 1; control = 0)
WatchProgram	Actually watched program (watched = 1, did not watch = 0)
InformationLevel	Information about Proposition 209 prior to election (none = 1 to great deal = 4)

a. Estimate an OLS regression model in which the information the respondent has about Proposition 209 is the dependent variable and whether the person watched the program is the independent variable. Comment on the results, especially whether and how they may be biased.
b. Re-estimate the model in part a but now include measures of political interest, newspaper reading, and education. Are the results different? Have you been able to defeat endogeneity?
c. Why might the assignment variable be a good instrument for watching the program? Please run any test that you can use to answer this question.
d. Estimate a 2SLS model from using the assignment to the treatment group as an instrument for whether a given respondent watched the program. Include the additional variables from part b in this new model. Compare the first stage results to results in part c. Are they similar? Are they identical? (hint: compare sample sizes)
e. What do the 2SLS results suggest about the effect of watching the program on information levels? Compare the results to those in part b. Have you been able to defeat endogeneity now?