Economics 142
Final Project - Spring 2019
Due: Friday May 17, 12:00 pm (noon)

You must submit your project electronically via GRADESCOPE by the due date/time: 12.00 pm on Friday May 17. Your answer should take the form of a **pdf document** with answers to each question that will include narrative as well as figures, tables of regression outputs etc. In an appendix ***you must also submit all code*** that generated the figures tables and calculations.

The project will use a data set called "sample.csv" that contains information on 236,459 married couples from an earlier US Census. These couples all have at least 2 children, with a mother ("mom") of the family who is between 21 and 35 years of age and a father of the family ("dad") who worked in the previous year. The variable naming convention is that XXXXm is for the mother and XXXXd is for the dad. A list of variables with their labels and means is attached at the end of this document. Make sure that when you load in the data set you can reproduce these means!

The variables are:

demographics

- agem, aged = age in years
- educm, educd = years of completed education (from 0 to 20)
- agefstm, agefstd = age of mom and dad at the time of birth of their first kid
- blackm, blackd = dummy for black race
- whitem, whited = dummy for white race
- othracem, othraced = dummy for other race (includes Asian, Native American, etc)
- hispm, hispd = dummy for hispanic ethnicity
- wnhm, wnhd = dummy for white non-Hispanic status
- st = 2-digit state of residence (1st digit = "division") eg NY=21, CA=93 (see appendix). This has 51 values for the 50 states and DC.

work activity

- workedm, workedd = dummy for worked in past year (note workedd=1 for all obs)
- weeksm, weeksd = weeks worked last year (0-52)
- hrsweekm, hrsweekd = hours per week last year
- annhrsm, annhrsd = total annual hours last year (annhrs=weeks*hrsweek)
- earningsm, earningsd = total labor earnings last year
- famearn = earningsm+earningsd
- faminc = total family income last year
- wagem, waged = hourly wage last year = earnings/annhrs
- expm, expd = estimated potential years of work since completing school = max[0, age-educ-6]

child composition
- kidcount = total # of kids (2-6)
- morekids = indicator for having 3+ kids= 1[kidcount>2]
- boy1st = indicator that 1st kid is male
- boy2nd = indicator that 1st kid is male
- boys2 = indicator that first 2 kids are both male = boy1st*boy2nd
- girls2 = indicator that first 2 kids are both female
- samesex = indicator that first 2 kids are same sex

Part I

1. In this part you will develop and evaluate some models for having 3+ kids (i.e., having morekids=1) based on mother's age, education, and mother's age when she had her first child. Specifically, use a linear probability model (with dependent variable = morekids) to estimate 3 models:

M1: a model with 3 explanatory variables plus a constant: mother's education (educm), mother's age (agem), and mother's age at first birth (agefstm)

M2: a model with a full set of dummy variables for mother's education, plus mother's age (agem), and mother's age at first birth (agefstm) (this model has 22 covariates plus a constant)

M3: a model with a full set of dummy variables for mother's education, a full set of dummy variables for mother's age and a full set of dummy variables for mother's age at first birth (this model has 52 covariates plus a constant)

For each model, compute R-squared, adjusted R-squared, RMSE, and AIC (Akaike information criterion). Also compute the following differences in probabilities of having morekids=1 from the 3 models and the associated standard errors:

(i) the difference in probabilities for mothers age 30 vs 35

(ii) the difference in probabilities for mothers with 12 versus 16 years of education

(iii) the difference in probabilities for mothers who had a first child at age 20 versus age 25

Finally, graph the actual and predicted probabilities from M1, M2, and M3 of having morekids=1 for 35 year old mothers with 12 years of education who had their first child at ages 17, 19, ... 30.

Based on your results and any other evidence you can develop, compare the advantages and disadvantages of M1, M2 and M3. (Extra points will be rewarded for creativity in developing comparisons between the models)

2. In this part you will explore a richer set of models for morekids, and evaluate the use of samesex as a potential instrumental variable for morekids. The idea of this instrumental variable is that some families want to have both a girl and a boy and are willing to have a third kid to achieve that aim.

a) Starting from M3, estimate an extended model that also includes dad's age, dad's education, and 3 variables for mother's race/ethnicity: blackm,

hispm, othracem. (Note that with this specification the omitted group is white non-hispanic mothers). Test that the two dad variables can be excluded . Test that the 3 race/ethnicity variables can be excluded.

Hint: these are both F-tests. You should find that the dad variables are not very important but the race ethnicity variables are more important.

b) Now you want to evaluate the potential use of samesex as an "exogenous" determinant of family size.

(i) using the same model as in 2a, add samesex and re-estimate the model. What is the average effect of having the first two children of the same sex on the probability morekids=1?

(ii) Some people have claimed that families only care about having at least 1 son - they don't care about having daughers. Develop a test for this claim using boys2 and girls2. Do you find any significant difference in the impact of having 2 daughters versus 2 sons?

(iii) You might be concerned that the sex composition of children is not truly random (the probability of having a male baby is around 51-52% in healthy human populations, but the event of having a boy is argued by some to be random and by others to be nonrandom). To test this concern, consider a linear probability model in which you try to "explain" the event of samesex=1 using information on mother's and father's age, education, race/ethnicity, and ages when they started having children. Present a selection of models to back up the claim that sex composition is random (in this sample, at least).

3. In this part you will compare OLS and IV models for the effect of having a larger number of children on mothers' decisions to work or not. To keep the models simple we will use a relatively short list of 15 control variables:

- mothers's education, plus dummy for mother's education < 12
- dad's education, plus dummy for dad's education <12
- linear and quadratic terms in mother's age
- linear and quadratic terms in mother's age at first birth
- linear and quadratic terms in dad's age
- linear and quadratic terms in dad's age at first birth
- 3 variables for mother's race/ethnicity: blackm, hispm, othracem (most of the time mom and dad have the same race/ethnicity)

a) Start with a linear probability model for the event that mom works. Compare 2 OLS models:

- model W1: no controls. Only explanatory variable is morekids (= 1 if 3+ kids).
- model W2: all the controls listed above plus morekids

Compare the size of the effect of having morekids=1 on the probability that mom works with and without the extra controls. What do you conclude about the relationship between morekids and the residual in model W1? Hint: think about short versus long regressions and omitted variables theorem.

b) Now consider a simple causal model relating $y_i = 1[work]$ to $x_{1i} = morekids$ with no other controls. Suppose we have the instrumental variable

$z_i = samesex$. Present a table showing the estimated first stage, reduced form, and IV (or 2sls) models. Verify that the coefficient in the IV model is the ratio of the reduced form coefficient to the first stage coefficient.

c) Consider the same causal model, but with the 15 controls described above. Present a table showing the estimated first stage, reduced form, and IV (or 2sls) models from this specification. Verify that the coefficient in the IV model is the ratio of the reduced form coefficient to the first stage coefficient.

d) An econometrician claims that if an instrumental variable is truly random, then the IV estimate will be (approximately) the same whether or not you add controls to the specification. Develop a proof of this claim.

HINT: Consider the reduced form equation. Show that if z_i is orthogonal to the vector of other controls x_{Oi} then the population regression coefficient relating y_i to z_i from a univariate model (no controls) will be the same as the population regression coefficient relating y_i to z_i when the additional controls x_{Oi} are added. Now develop the same argument for the first stage.

e) Evaluate the difference between the IV estimates in parts (b) and (c) in light of the result in part (d).

HINT: the two estimates are not exactly the same. Are they very different statistically? Extra points for making this as formal as possible.

f) Based on the OLS and IV models, what do you conclude about the relationship between the "causal effect" of having extra kids on mother's decision to work and the observational comparisons between mothers with 2 or 3+ kids?

4. In this part you will compare OLS and IV models for the effect of having a larger number of children on earnings of mothers, fathers, and the family as a whole.

a) Consider 3 dependent variables: mother's earnings (earningsm); father's earnings (earningsd), their combined family earnings (famearns). For each of the 3 dependent variables you will estimate 2 OLS models – one with no controls, similar to W1, and one with 15 controls used in W2 – and 2 IV models, with and without the extra controls.

Present a table of the estimated effects of morekids on the 3 dependent variables with each row representing one of the dependent variables and 4 columns – 2 for the OLS models and 2 for the IV. Focusing on the specifications with controls, compare the OLS versus IV estimates of the effect of morekids on mother's and father's earnings. Give an explanation for why the IV estimate is *less negative* than the OLS estimate for mothers but *more negative* than the OLS for fathers.

b) Prove that if $y_{3i} = y_{1i} + y_{2i}$ and we have a single vector of controls x_i in 3 models estimated on the same sample:

$$\begin{aligned} y_{1i} &= x_i\beta_1 + u_{1i} \\ y_{2i} &= x_i\beta_2 + u_{2i} \\ y_{3i} &= x_i\beta_3 + u_{3i} \end{aligned}$$

then the OLS estimates will have the property that $\widehat{\beta}_3 = \widehat{\beta}_1 + \widehat{\beta}_2$.

c) Use your result in (b) to show that the estimated OLS effect of morekids on mother earnings and the estimated OLS effect of morekids on father earnings will add up to the effect on family earnings. Verify that this is true for the OLS models with and without extra controls. Considering the effect of morekids on *total family earnings*, what share is driven by the effect on mothers versus fathers?

d) Show that the result in part (b) is also true for IV estimates estimated in the same sample using the same instrumental variable. Verify that this is true for the IV models with and without extra controls. Taking the IV estimate of the effect of morekids on total family earnings, what share is driven by the effect on mothers versus fathers?

e) Looking at your overall sample of families with 2+ kids, and assuming that samesex is randomly assigned and that there are no "defier" families:

(i) what is the fraction of "always takers" (AT) who always have another kid after having two of the same sex

(ii) what is the fraction of "never takers" (NT) who have only two kids regardless of the sex composition of the first two

(iii) what is the fraction of "compliers" (C) who have 3+ kids when their first 2 are of the same sex but not otherwise

(iv) compare the fractions of AT/NT/C for four subgroups of families, based on mother's education: less than 12 years, equal to 12 years, between 13 and 15 years, and 16+ years of education.

f) using a model with controls and the technique discussed in class, calculate the following means for the *overall set of compliers*:

- mean fraction of C moms with <12 years of schooling
- mean education of C moms
- mean age of C moms at first birth
- fraction of C moms who had a first birth before age 21
- fraction of C moms who are hispanic
- fraction of C moms who are black
- fraction of C moms who are white non-hispanic
- fraction of C moms from Utah (state=87)

Compare these with the means for all families, and comment on the differences.

Part II

In this part you will investigate the wages of fathers of different ages who live in different states. Your dependent variable will be the natural log of the average hourly wage rate ($log(waged)$). You will focus on men with 16 years of education (i.e. a BA or BS and no advanced degree). **Before you start**, select the subsample of fathers with 16 years of education in families with rv<0.75 (n=20,923) as the estimation sample and the subsample of fathers with 16 years of education in families with rv>=0.75 (n=7,065) as the holdout sample.

1. Estimate a simple OLS model on the estimation sample that includes age (*aged*), a set of dummies for the state of residence, and interactions of the state of residence with age. Since there are 51 states (including DC), this model has a constant, age, 50 state dummies, and 50 interactions of age with state – a total of 101 covariates in addition to the constant. Use this model to predict earnings of men age 26 in each state, and men of age 35 in each state.

2. Using k-fold cross validation within the estimation sample, estimate a lasso version of the same model, and use this model to predict earnings of men age 26 in each state, and men of age 35 in each state.

3. Based on the two models, what is the "best" (highest log wage) state for young men (age 26) with a college degree. What is the best state for men who are age 35?

4. Now use the two models (OLS and lasso) to predict log wages of men in the *holdout sample*. Calculate the RMSE for the two models at each age (i.e., for men of age a in the holdout sample, calculate $RSS(a) = \sum_{i \in a}(y_i - \widehat{y}_i)^2$ where $\widehat{y}_i$ is the prediction from either OLS or lasso). Graph the RSS values for each age and comment on the relative accuracy of the OLS versus lasso predictions.

```
Contains data from sample.dta
  obs:       236,459
 vars:            41                          27 APR 2019 14:30
 size:    21,754,228
-------------------------------------------------------------------------------
-------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
-------------------------------------
kidcount        byte    %8.0g                 count of kids in household
boy1st          byte    %8.0g                 first birth boy
boy2nd          byte    %8.0g                 second birth boy
boys2           byte    %8.0g                 first two births boys
girls2          byte    %8.0g                 first two births girls
samesex         byte    %8.0g                 first two kids are of same sex
morekids        byte    %8.0g                 had more than 2 kids
blackm          byte    %8.0g                 =1 of black
hispm           byte    %8.0g                 =1 if hispanic
whitem          byte    %8.0g                 mother race=white
othracem        byte    %8.0g                 =1 if other race (white is ref)
blackd          byte    %8.0g                 =1 of black
hispd           byte    %8.0g                 =1 if hispanic
whited          byte    %8.0g                 dad race=white
othraced        byte    %8.0g                 =1 if other race (white is ref)
educm           byte    %8.0g                 moms educ
educd           byte    %8.0g                 dads educ
agefstm         byte    %8.0g                 age of mom at first birth
agefstd         byte    %8.0g                 age of dad when kid first born
workedm         byte    %8.0g                 mom worked last year
workedd         byte    %8.0g                 dad worked last year
hrsweekd        byte    %8.0g                 hours of work per week, dad
hrsweekm        byte    %8.0g                 hours of work per week, mom
annhrsm         int     %8.0g                 annual hours mom
annhrsd         int     %8.0g                 annual hours dad
wnhm            byte    %8.0g                 white non-hispanic mom
wnhd            byte    %8.0g                 white non-hispanic dad
earningsm       double  %12.0g                earnings of mother
earningsd       double  %12.0g                earnings of father
faminc          double  %12.0g                family income
famearn         double  %12.0g                total earnings mom+dad
agem            byte    %8.0g                 age of mom
aged            byte    %8.0g                 age of dad
weeksm          byte    %8.0g                 weeks worked of mom
weeksd          byte    %8.0g                 weeks worked of dad
wagem           double  %12.0g                hrly wa mom (if working)
waged           double  %12.0g                hrly wage dad
expm            byte    %8.0g                 potential experience mom
expd            byte    %8.0g                 potential experience dad
st              byte    %8.0g                 state code
rv              double  %12.0g                uniform rv
-------------------------------------------------------------------------------
-------------------------------------
Sorted by:


. sum

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

```
    kidcount |    236,459    2.486533    .7285427          2          6
      boy1st |    236,459     .514715    .4997845          0          1
      boy2nd |    236,459    .5128796    .4998351          0          1
       boys2 |    236,459    .2664352    .4420954          0          1
      girls2 |    236,459    .2388406     .426376          0          1
-------------+---------------------------------------------------------
     samesex |    236,459    .5052758    .4999732          0          1
    morekids |    236,459    .3731006    .4836295          0          1
      blackm |    236,459    .0497126    .2173511          0          1
       hispm |    236,459    .0250953    .1564148          0          1
      whitem |    236,459    .8967263    .3043167          0          1
-------------+---------------------------------------------------------
    othracem |    236,459    .0284658    .1662998          0          1
      blackd |    236,459    .0509687    .2199342          0          1
       hispd |    236,459     .026013    .1591741          0          1
      whited |    236,459    .8967474    .3042891          0          1
    othraced |    236,459    .0262709    .1599403          0          1
-------------+---------------------------------------------------------
       educm |    236,459    12.42263    2.392778          0         20
       educd |    236,459    13.04468    3.077283          0         20
     agefstm |    236,459    20.84279    2.911087         15         33
     agefstd |    236,459    23.78704    3.553351         15         43
     workedm |    236,459    .5341391    .4988342          0          1
-------------+---------------------------------------------------------
     workedd |    236,459           1           0          1          1
    hrsweekd |    236,459    44.20936    9.805291          1         99
    hrsweekm |    236,459    16.78046    18.19839          0         99
     annhrsm |    236,459    630.9714    817.0996          0       5148
     annhrsd |    236,459    2185.041    585.3178          3       5148
-------------+---------------------------------------------------------
        wnhm |    236,459    .8967263    .3043167          0          1
        wnhd |    236,459    .8967474    .3042891          0          1
   earningsm |    236,459    10140.96    15276.55          0    99661.8
   earningsd |    236,459    65221.29    33047.16   4167.707   241899.9
      faminc |    236,459    78612.67    36846.42   190.2273   259400.8
-------------+---------------------------------------------------------
     famearn |    236,459    75362.25    35141.78   4167.707   314774.3
        agem |    236,459    30.36734    3.374119         21         35
        aged |    236,459    32.80577    4.127869         21         45
      weeksm |    236,459    19.17827    21.85144          0         52
      weeksd |    236,459    49.26945    7.041211          1         52
-------------+---------------------------------------------------------
       wagem |    126,302    18.83854    48.85837          0   10377.95
       waged |    236,459    32.83272     71.6352   .8348771   19604.94
        expm |    236,459    11.94482    3.692718          0         29
        expd |    236,459    13.76122    4.933979          0         39
          st |    236,459    51.53003    25.11601         11         95
-------------+---------------------------------------------------------
          rv |    236,459    .4999056    .2886722   5.49e-08   .9999987
```

```
. tab kidcount morekids

  count of |
   kids in | had more than 2 kids
 household |         0          1 |     Total
-----------+----------------------+----------
         2 |   148,236          0 |   148,236
         3 |         0     66,952 |    66,952
         4 |         0     16,564 |    16,564
```

```
         5 |         0      3,863 |     3,863
         6 |         0        844 |       844
-----------+----------------------+----------
     Total |   148,236     88,223 |   236,459
```

Comarisons of subsamples

```
. sum waged educd aged expd

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       waged |    236,459    32.83272     71.6352   .8348771   19604.94
       educd |    236,459    13.04468    3.077283          0         20
        aged |    236,459    32.80577    4.127869         21         45
        expd |    236,459    13.76122    4.933979          0         39
```

Estimation Sample for Part II

```
. sum waged educd aged expd if rv<0.75 & educd==16

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       waged |     20,923    38.50751    97.68986   1.335803   10378.19
       educd |     20,923          16           0         16         16
        aged |     20,923    33.30244    3.440303         21         45
        expd |     20,923    11.30254    3.439975          0         23
```

Hold out Sample for Part II

```
. sum waged educd aged expd if rv>=.75 & educd==16

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       waged |      7,065    37.80498    50.25596   1.567213   3432.738
       educd |      7,065          16           0         16         16
        aged |      7,065    33.26525    3.442681         21         45
        expd |      7,065    11.26539    3.442198          0         23
```

State Codes

```
. tab st

 state code |      Freq.     Percent        Cum.
------------+-----------------------------------
         11 |      1,430        0.60        0.60
         12 |      1,137        0.48        1.09
         13 |        682        0.29        1.37
         14 |      5,635        2.38        3.76
         15 |        997        0.42        4.18
         16 |      3,202        1.35        5.53
         21 |     15,884        6.72       12.25
         22 |      7,315        3.09       15.34
```

```
         23 |     12,403        5.25       20.59
         31 |     12,416        5.25       25.84
         32 |      6,680        2.83       28.67
         33 |     11,586        4.90       33.56
         34 |     10,697        4.52       38.09
         35 |      5,750        2.43       40.52
         41 |      5,008        2.12       42.64
         42 |      3,617        1.53       44.17
         43 |      5,198        2.20       46.37
         44 |        863        0.36       46.73
         45 |        777        0.33       47.06
         46 |      1,908        0.81       47.87
         47 |      2,767        1.17       49.04
         51 |        572        0.24       49.28
         52 |      3,871        1.64       50.92
         53 |        138        0.06       50.97
         54 |      5,397        2.28       53.26
         55 |      2,447        1.03       54.29
         56 |      5,998        2.54       56.83
         57 |      3,316        1.40       58.23
         58 |      5,599        2.37       60.60
         59 |      7,085        3.00       63.59
         61 |      4,393        1.86       65.45
         62 |      4,955        2.10       67.55
         63 |      4,152        1.76       69.30
         64 |      2,448        1.04       70.34
         71 |      2,437        1.03       71.37
         72 |      4,519        1.91       73.28
         73 |      3,387        1.43       74.71
         74 |     16,055        6.79       81.50
         81 |        982        0.42       81.92
         82 |      1,319        0.56       82.48
         83 |        694        0.29       82.77
         84 |      3,296        1.39       84.16
         85 |      1,411        0.60       84.76
         86 |      2,814        1.19       85.95
         87 |      2,487        1.05       87.00
         88 |        693        0.29       87.30
         91 |      4,562        1.93       89.22
         92 |      3,029        1.28       90.51
         93 |     21,022        8.89       99.40
         94 |        466        0.20       99.59
         95 |        963        0.41      100.00
------------+-----------------------------------
      Total |    236,459      100.00

. exit, clear
```

Census/CPS State Codes and State Names

63	Alabama
94	Alaska
86	Arizona
71	Arkansas
93	California
84	Colorado
16	Connecticut
51	Delaware
53	District of Columbia
59	Florida
58	Georgia
95	Hawaii
82	Idaho
33	Illinois
32	Indiana
42	Iowa
47	Kansas
61	Kentucky
72	Louisiana
11	Maine
52	Maryland
14	Massachusetts
34	Michigan
41	Minnesota
64	Mississippi
43	Missouri
81	Montana
46	Nebraska
88	Nevada
12	New Hampshire
22	New Jersey
85	New Mexico
21	New York
56	North Carolina
44	North Dakota
31	Ohio
73	Oklahoma
92	Oregon
23	Pennsylvania
15	Rhode Island
57	South Carolina
45	South Dakota
62	Tennessee
74	Texas
87	Utah
13	Vermont
54	Virginia
91	Washington
55	West Virginia
35	Wisconsin
83	Wyoming