Instructions for Reading the Data

* The dataset and programs For the Angrist and Evans paper are in this sub-folder of the MIT link that has been circulated earlier

<https://economics.mit.edu/faculty/angrist/data1/data/angev98>

* From the zip data file on the webpage extract the 1980 SAS dataset
  + m\_d\_806
* Read in the data using python code. You will get a data set that looks like this:

In [**12**]: df.head(5)

Out[**12**]:

   STATE  SEXK   AGEK QTRBKID  RACEK SPANISH BIRTHPLK SCHOOLK  ... WEEKSD HOURSD  INCOME1D  INCOME2D AWEEK79D AHOUR79D AINC1D AINC2D

0  b'01'  b'1'  b'09'    b'3'  b'01'    b'0'   b'001'    b'1'  ...  b'52'  b'40'  b'28005'  b'00000'     b'0'     b'0'   b'0'   b'0'

1  b'01'  b'1'  b'08'    b'3'  b'01'    b'0'   b'001'    b'2'  ...  b'52'  b'72'  b'00000'  b'10005'     b'0'     b'0'   b'0'   b'3'

2  b'01'  b'0'  b'09'    b'1'  b'01'    b'0'   b'001'    b'1'  ...  b'16'  b'48'  b'00000'  b'16005'     b'0'     b'0'   b'2'   b'0'

3  b'01'  b'0'  b'05'    b'2'  b'01'    b'0'   b'001'    b'1'  ...    NaN    NaN       NaN       NaN      NaN      NaN    NaN    NaN

4  b'01'  b'1'  b'11'    b'1'  b'01'    b'0'   b'001'    b'1'  ...  b'32'  b'40'  b'09925'  b'00000'     b'0'     b'0'   b'3'   b'3'

There are 85 columns which have names

df.keys()

Out[**13**]:

Index(['STATE', 'SEXK', 'AGEK', 'QTRBKID', 'RACEK', 'SPANISH', 'BIRTHPLK',

       'SCHOOLK', 'GRADE', 'FINGRADE', 'ASEX', 'AAGE', 'AQTRBRTH', 'ABIRTHPL',

       'ASCHOOL', 'AYEARSCH', 'AFINGRAD', 'YOBK', 'AGEQK', 'SEX2ND', 'RACE2ND',

       'SPAN2ND', 'SCH2ND', 'GRADE2ND', 'FIN2ND', 'ASEX2ND', 'AAGE2ND',

       'AQTR2ND', 'ASCH2ND', 'AYEAR2ND', 'AFIN2ND', 'AGEQ2ND', 'SEX3RD',

       'AGEQ3RD', 'SEX4TH', 'AGEQ4TH', 'SEX5TH', 'AGEQ5TH', 'TWIN1ST',

       'TRIPLET', 'KIDCOUNT', 'FAMINC', 'AGEM', 'QTRBTHM', 'MARITAL', 'RACEM',

       'SPANISHM', 'FERT', 'TIMESMAR', 'AGEMAR', 'QTRMAR', 'WIDOW', 'GRADEM',

       'FINGRADM', 'CLASSM', 'WEEKSM', 'HOURSM', 'INCOME1M', 'INCOME2M',

       'INCOME6M', 'POVERTY', 'AFERTIL', 'AWEEK79M', 'AHOUR79M', 'AINC1M',

       'AINC2M', 'YOBM', 'AGED', 'QTRBTHD', 'RACED', 'SPANISHD', 'TIMEMARD',

       'AGEMARD', 'QTRMARD', 'GRADED', 'FINGRADD', 'CLASSD', 'WEEKSD',

       'HOURSD', 'INCOME1D', 'INCOME2D', 'AWEEK79D', 'AHOUR79D', 'AINC1D',

       'AINC2D'],

      dtype='object')

* The SAS code for setting up the variables (Y, X and W) are in a zip file on the same webpage. Below are extracts from the code that are relevant for the data that is needed for this analysis. Comments are starred (\*). Create these variables using corresponding code in python. Instructions for the Project, course material and the Angrist and Evans paper should provide guidance and context for identifying Y, X and W and doing the analysis. Adapt indicator variables as need for the ML exercise.

\* construct the sexes of 1st two kids;

boy1st=(sexk=0);

boy2nd=(sex2nd=0);

boys2=((sexk=0) and (sex2nd=0));

girls2=((sexk=1) and (sex2nd=1));

samesex=((boys2=1) or (girls2=1));

morekids=kidcount>2;

length girls2 boys2 morekids samesex boy1st boy2nd 3;

label boy1st='first birth boy';

label samesex='first two kids are of same sex';

label morekids='had more than 2 kids';

label boys2='first two births boys';

label girls2='first two births girls';

\* construct race indicators for mom and dad;

blackm=racem='02';

hispm=racem='12';

whitem=racem='01';

othracem=1-blackm-hispm-whitem;

length blackm othracem whitem hispm 3;

blackd=raced='02';

hispd=raced='12';

whited=raced='01';

othraced=1-blackd-hispd-whited;

length blackm othracem whitem hispm 3;

label blackm='=1 of black';

label hispm='=1 if hispanic';

label othracem='=1 if other race (white is ref)';

label blackd='=1 of black';

label hispd='=1 if hispanic';

label othraced='=1 if other race (white is ref)';

\* educ stuff;

if fingradm=2 or fingradm=1 then educm=gradem-2;

else educm=gradem-3;

educm=max(0,educm);

label educm = 'moms educ';

hsgrad=(educm=12);

hsormore=(educm=>12);

moreths=(educm>12);

\* get ages of mom and dad;

\* plus ages when 1st kid was born;

agem1=agem\*1;

aged1=aged\*1;

length agem1 aged1 3;

label agem1='age in years of mom';

label aged1='age in years of dad';

if qtrbthd=0 then yobd=80-aged;

else yobd=79-aged;

ageqm=4\*(80-yobm)-qtrbthm-1;

ageqd=4\*(80-yobd)-qtrbthd;

agefstm=int((ageqm-ageqk)/4);

agefstd=int((ageqd-ageqk)/4);

length agefstd agefstm 3;

label agefstd='age of dad when kid first born';

label agefstm='age of mom when kid first born';

\* calculate mom and dad labor supply variables;

weeksm1=weeksm\*1;

weeksd1=weeksd\*1;

workedm=weeksm>0;

workedd=weeksd>0;

hourswd=hoursd\*1;

hourswm=hoursm\*1;

incomed=income1d+max(0,income2d);

incomem=income1m+max(0,income2m);

\*\* deflate wages \*\*;

incomem=incomem\*2.099173554;

incomed=incomed\*2.099173554;

faminc1=faminc\*1;

faminc1=faminc1\*2.099173554;

famincl=log(max(faminc1,1));

nonmomi=faminc1-income1m\*2.099173554;;

nonmomil=log(max(1,nonmomi));

label weeksm1='weeks worked mom';

label workedm='mom worked last year';

label weeksd1='weeks worked dad';

label workedd='dad worked last year';

label famincl='log family income';label agefstm='age of mom at first birth';

label nonmomi='income not generated by mom';

label incomed='dads labor income';

label incomem='moms labor income';

label faminc1='family income';

label hourswd='hours of work per week, dad';

label hourswm='hours of work per week, mom';

length incomem incomed faminc1 nonmomi 4

workedm workedd weeksm1 weeksd1 hourswm hourswd 3 ;