1 Midterm - Empirical Analysis, Spring 2019

For this exercise, use the data set -DahlLochner2012AER.dta- available on Canvas. Include your code after the main text, tables and figures. Please be brief, but precise in your answers. Note that you do not have to report more in the text than is asked for.

In a recent study published in AER, Dahl and Lochner (DL) study how children's school performance depends on family income. They posit the following model of the relationship:

$$y_{ia} = \mathbf{x}_i' \boldsymbol{\alpha}_a + \mathbf{w}_{ia}' \boldsymbol{\beta} + \delta I_{ia} + u_{ia} \qquad (1)$$

where y_{ia} and I_{ia} are the performance and family income, respectively, of child i at age a; $\mathbf{x}$ and $\mathbf{w}$ are permanent and time-varying characteristics listed below, while u_{ia} reflects unobserved determinants of school performance.

Problem 1.1. There are three performance measures in the data set -math-, -readingcomp- and -readingrecog-. Create a new variable -score- as the average of these variables, and standardize it to mean equal zero and standard deviation equal one.

Solution. I implement this using Stata using the following code

```
gen score_raw = (math + readingcomp + readingrecog) / 3
egen score = std(score_raw)
replace score_raw = score
```

Problem 1.2. How much of the variation in -score- and -faminc- is coming from comparisons across individuals and how much is coming from comparisons within individuals over time?

Solution. For -score-, we obtain the following analysis of variance:

			Number of obs =		7,280
			R-squared =		0.9160

Source	SS	df	MS	F	Prob > F
Between id	6667.6768	3,691	1.8064689	10.60	0.0000
Within id	611.32316	3,588	.17037992		
Total	7279	7,279	1		

- ▷ We find that 91.59% of the variation is coming from comparisons across individuals and 8.41% of the variation is coming from comparisons within individuals.

For -famine-, we obtain the following analysis of variance:

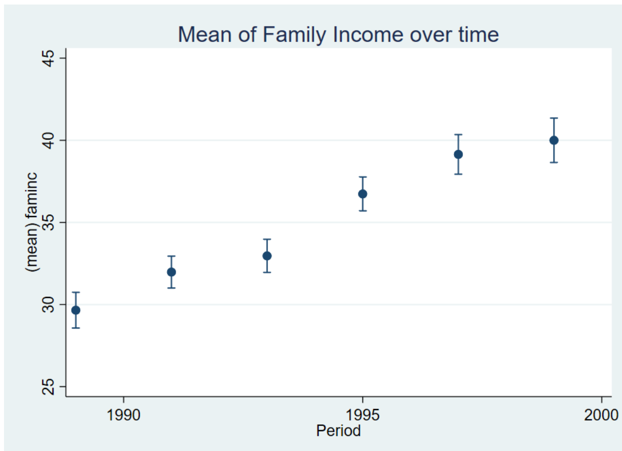
			Number of obs =		7,280
			R-squared =		0.9167

Source	SS	df	MS	F	Prob > F
Between id	2619146.5	3,691	709.60349	10.69	0.0000
Within id	238142.89	3,588	66.372044		
Total	2857289.4	7,279	392.53872		

▷ We find that 91.66% of the variation is coming from comparisons across individuals and 8.34% of the variation is coming from comparisons within individuals.

Problem 1.3. Graph the mean of -score- and -faminc- over time, and include a 95% confidence interval. (Hint: You want to graph one observation per year(try -collapse- if you use Stata). Also, you need to generate new variables for the confidence interval using the standard error of the mean.)

Solution. I implement this using Stata. We obtain the following graphs:



▷ Both the means of score and family income are increasing over time in general, but the average score dipped in the most recent period.

Problem 1.4. Estimate model (1) using OLS with -score- as the dependent variable, controlling for variables 9–26 below (i.e. -black- through -sib3-). Use robust standard errors. Interpret the coefficient on -faminc-.

Solution. I implement this using Stata.

```
Linear regression                               Number of obs     =      7,280
                                                F(18, 7261)       =     173.93
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2871
                                                Root MSE          =     .84538
```

score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
black	-.2523694	.0296207	-8.52	0.000	-.3104346	-.1943042
hispanic	-.0532542	.0319945	-1.66	0.096	-.1159728	.0094643
male	-.0398039	.0198723	-2.00	0.045	-.0787594	-.0008483
age	-.0798643	.005195	-15.37	0.000	-.090048	-.0696807
agemom	-.0046219	.0036137	-1.28	0.201	-.0117058	.002462
edlage23	0	(omitted)				
ed2age23	.1975649	.0299453	6.60	0.000	.1388635	.2562663
ed3age23	.3097689	.0373332	8.30	0.000	.2365849	.3829529
ed4age23	.3688812	.0518731	7.11	0.000	.2671949	.4705675
afqt	.2918538	.0156725	18.62	0.000	.261131	.3225765
afqt_miss	.0639619	.0605968	1.06	0.291	-.0548254	.1827492
married	-.1095284	.1067756	-1.03	0.305	-.3188396	.0997828
spouseage	.0054564	.0029109	1.87	0.061	-.0002498	.0111627
spouseage_miss	-.1352716	.48424	-0.28	0.780	-1.084523	.8139796
famsize	-.0301376	.0172245	-1.75	0.080	-.0639026	.0036274
famsize_miss	-.1914412	.0976819	-1.96	0.050	-.3829261	.0000438
sibl	.1984667	.0363853	5.45	0.000	.1271409	.2697924
sib3	-.1131375	.0212114	-5.33	0.000	-.1547181	-.0715569
faminc	.0051668	.0007034	7.35	0.000	.003788	.0065456
_cons	.9441016	.1281926	7.36	0.000	.6928067	1.195396

▷ $1000 increase in family income is associated with a 0.00517 standard deviation increase in the average score. Note that we cannot interpret this as a causal estimate for reasons discussed in the next question as well as throughout the problem set.

Problem 1.5. Do you think that the OLS estimates may be biased? Explain your answer. In which direction do you think δ is biased?

Solution. Yes, it is likely biased due to omitted variables. For example, ability may be positively correlated with both family income (if we assume ability is hereditary and income is correlated with ability) and score (since ability affects test-taking results). In this case, this would introduce an upward bias in the estimate of δ. Another similar example would be access to private tutorship due to similar series of reasoning.

We have panel data with information on school performance of each child in several years. Assume that the error term above has an individual-specific component μ_i that is fixed over time, such that

$$u_{ia} = \mu_i + \epsilon_{ia}$$

where ϵ_{ia} is random residual.

Problem 1.6. Explain how you can use the panel structure of the data to get a more reliable estimate of δ. Estimate this model using first differences for -score- and -faminc-. Include as control variables -black-, -hispanic-, -male-, -age-, -sib1-, and -sib3- (not differenced).

Solution. Recall the original model:

$$y_{ia} = \mathbf{x}_i' \boldsymbol{\alpha}_a + \mathbf{w}_{ia}' \boldsymbol{\beta} + \delta I_{ia} + u_{ia}$$

and rewrite the error term to obtain:

$$y_{ia} = \mathbf{x}_i'\boldsymbol{\alpha}_a + \mathbf{w}_{ia}'\boldsymbol{\beta} + \delta I_{ia} + (\mu_i + \epsilon_{ia})$$
$$y_{i(a+2)} = \mathbf{x}_i'\boldsymbol{\alpha}_{(a+2)} + \mathbf{w}_{i(a+2)}'\boldsymbol{\beta} + \delta I_{i(a+2)} + \left(\mu_i + \epsilon_{i(a+1)}\right)$$

which implies a regression of the following form:

$$\Delta y_{ia} = \mathbf{x}_i'\left(\boldsymbol{\alpha}_{a+2} - \boldsymbol{\alpha}_a\right) + \left(\mathbf{w}_{i(a+2)} - \mathbf{w}_{ia}\right)'\boldsymbol{\beta} + \delta\Delta I_{ia} + \underbrace{\Delta u_{ia}}_{=\Delta\epsilon_{ia}}$$

Therefore, we can obtain a more reliable estimate of δ using first-differences. In implementing this regression, we assume that $\mathbf{w}$'s do not vary with age as we are told to include control variables not differenced. Running this in Stata, we obtain:

```
. reg score_d faminc_d black hispanic male age sib1 sib3, robust

Linear regression                               Number of obs     =      3,445
                                                F(7, 3437)        =       3.05
                                                Prob > F          =     0.0034
                                                R-squared         =     0.0061
                                                Root MSE          =      .5414

------------------------------------------------------------------------------
             |               Robust
     score_d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    faminc_d |   .0010299   .0009071     1.14   0.256    -.0007486    .0028085
       black |  -.0720681   .0209105    -3.45   0.001    -.1130663   -.0310699
    hispanic |  -.0002503   .0286675    -0.01   0.993    -.0564573    .0559568
        male |   .0331323   .0185027     1.79   0.073    -.0031452    .0694097
         age |   .0068856   .0059166     1.16   0.245    -.0047148     .018486
        sib1 |   .0085307   .0336405     0.25   0.800    -.0574267     .074488
        sib3 |   .0360153   .0195716     1.84   0.066    -.0023579    .0743884
       _cons |  -.1943313   .0741695    -2.62   0.009     -.339752   -.0489107
------------------------------------------------------------------------------
```

▷ We find that a \$1,000 increase in family income is associated with a 0.0010299 standard deviation increase in the average score. The estimate, however, is not significant.

Problem 1.7. Estimate the model with fixed effects including the same controls. Why does Stata exclude the variables -black-, -hispanic-, and -male? How would you interpret the coefficient on these variables in the model in first differences?

Solution. We obtain the following results:

```
Fixed-effects (within) regression               Number of obs     =      7,280
Group variable: id                              Number of groups  =      3,692

R-sq:                                           Obs per group:
     within  = 0.0657                                         min =          1
     between = 0.0905                                         avg =        2.0
     overall = 0.0667                                         max =          4

                                                F(4,3584)         =      63.04
corr(u_i, Xb)  = 0.1590                         Prob > F          =     0.0000

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      faminc |   .0001726    .000828     0.21   0.835    -.0014508     .001796
       black |          0  (omitted)
    hispanic |          0  (omitted)
        male |          0  (omitted)
         age |  -.0521711   .0033484   -15.58   0.000    -.0587362   -.0456061
        sib1 |   .0897478   .0692599     1.30   0.195     -.046045    .2255406
        sib3 |    .013605   .0440047     0.31   0.757    -.0726717    .0998817
       _cons |   .5672455   .0493478    11.49   0.000     .4704928    .6639981
-------------+----------------------------------------------------------------
     sigma_u |  .97093666
     sigma_e |  .39919725
         rho |  .85540164   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(3691, 3584) = 9.26                 Prob > F = 0.0000
```

▷ The three variables drop out due to collinearity because once you know the ID of the person, you can perfectly determine the values for -black-, -hispanic-, and -male. They do not vary over time and are fixed for each individual.

▷ Since we are asked to interpret coefficient on these variables (black, hispanic, and male) in the model with first differences:

* Over a general two-year time period, blacks experience additional decrease in test scores of 0.0720681 times its standard deviation than non-blacks do over a general two-year time period.
* Over a general two-year time period, hispanics experience additional decrease in test scores of 0.00025 times its standard deviation than non-hispanics do over a general two-year time period.
* Over a general two-year time period, males experience additional increasein test scores of 0.03313 times its standard deviation than females do over a general two-year time period.

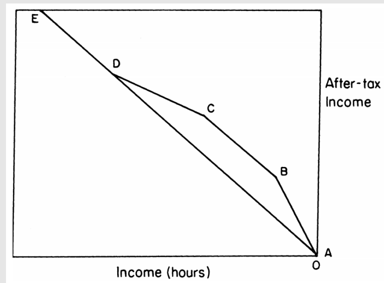
Problem 1.8. Why may we be worried about omitted variables bias also in the panel data models? (Hint: What is driving changes in family income?)

Solution. Omitted variable bias will be an issue here if the omitted variables are correlated with our regressor (ΔI_{ia}). Specifically, we are concerned about the possibility that changes in unobserved factors affecting child development $(\Delta\epsilon_{ia})$ are correlated with changes in family income (ΔI_{ia}) and with changes in score (Δy_{ia}).

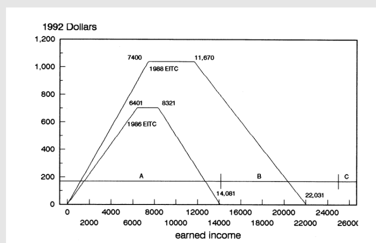
One example may be environmental factors that may impede child's development as well as decrease productivity, yielding lower family income. Another example may be economic expansion, which may increase the change in family income as well as improve the quality of education, thereby increasing the average score.

This concern is legitimate for both the first-differences and the fixed effects model. In the first-differences, we are subtracting subsequent observations; in the fixed-effects, we are de-meaning for the specified group. For both cases, omitted variables is a legitimate estimate.

The Earned Income Tax Credit (EITC) is a major US transfer program that provides direct transfers to working families depending on their income and the number of children. The following figure shows how the EITC changes the budget constraint:



While the EITC and other tax schedules do not generally vary with the child's age in any given year, they do sometimes change over time (that is: with the age of the child a). The following figure illustrates this for the 1986 and 1988 EITC in the US:



Total net family income is therefore given by

$$I_{ia} = P_{ia} + \chi_{ia} P_{ia} - \tau_{ia} P_{ia}$$

where P_{ia} is family income prior to taxes and transfers, and χ_{ia} and τ_{ia} are the EITC and tax schedules respectively.

Problem 1.9. Explain why

$$\Delta\chi_{ia}\left(P_{i,a-1}\right) = \chi_{ia}\left(P_{i,a-1}\right) - \chi_{i,a-2}\left(P_{i,a-2}\right)$$

may be an instrument for ΔI_{ia}. Do you think

$$\Delta\chi_{ia} = \chi_{ia}\left(P_{ia}\right) - \chi_{i,a-2}\left(P_{i,a-2}\right)$$

would be a better or worse instrument for ΔI_{ia}?

Solution. The instrument is based on the observation that low- and middle-income families benefited substantially from expansions of the EITC in the late-1980s and mid-1990s, whereas higher-income families did not. So to the extent that income affects achievement, we should be able to observe relative improvements in the test scores of children from families that benefit most from these EITC expansions. Ultimately, we want the instrument to capture only the changes in I_{ia} deriving from changes in EITC and avoid incorporating general changes in family income. We verify the conditions for ΔI_{ia} to be a valid instrument:

- ▷ Random assignment: The policy change was exogenous, so it is unlikely that people were able to select into the treatment.

- ▷ Exclusion: This is likely satisfied if we use $\Delta\chi_{ia}(P_{i,a-1})$ but not if we use $\Delta\chi_{ia}$. The reason is that I_{ia} is by construction correlated to P_{ia} so $\Delta\chi_{ia}$ will be endogenous for the same reason we argued I_{ia} is endogenous. By letting $\chi_{ia}(P_{i,a-1})$, we can get away from this problem.
- ▷ Relevance: $\Delta\chi_{ia}(P_{i,a-1})$ needs to be correlated with the instrumented variable, I_{ia}. We show in question 11 that this is indeed the case, verified by the first-stage regression.

Essentially, we are only exploiting variation in EITC income due to government changes in EITC schedules over time and not due to changes in family structure. Instead, suppose we used:

$$\Delta\chi_{ia} = \chi_{ia}(P_{ia}) - \chi_{i,a-2}(P_{i,a-2})$$

as our instrument. In this case, the proposed instrument depends on P_{ia}. Then the exclusion restriction may not be satisfied since P_{ia} may be correlated with macro trends (as discussed in the previous section) which may affect I_{ia} directly. Therefore, this will be a worse instrument.

Problem 1.10. In the data, $\chi_{ia}(P_{ia}) =$ eitc and $\chi_{ia}(P_{i,a-1}) =$ eitc_sim. Estimate the model in first differences (as in 6 above) using $\Delta\chi_{ia}(P_{i,a-1})$ as an instrument.

Solution. We will estimate the following model:

$$\Delta y_{ia} = \mathbf{x}_i'\boldsymbol{\alpha}_a + \mathbf{w}_{ia}'\boldsymbol{\beta} + \delta\Delta I_{ia} + \Delta\epsilon_{ia}$$

and use $\Delta\chi_{ia}(P_{i,a-1})$ as our instrument for ΔI_{ia}. We obtain the following results:

```
Instrumental variables (2SLS) regression          Number of obs     =      3,445
                                                  Wald chi2(7)      =      21.43
                                                  Prob > chi2       =     0.0032
                                                  R-squared         =          .
                                                  Root MSE          =     .55142
```

score_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
faminc_d	.0113712	.0076042	1.50	0.135	-.0035327	.0262751
black	-.0721035	.0212255	-3.40	0.001	-.1137048	-.0305021
hispanic	.0057273	.0292529	0.20	0.845	-.0516072	.0630619
male	.0317945	.0188753	1.68	0.092	-.0052005	.0687895
age	.00807	.00607	1.33	0.184	-.0038271	.019967
sib1	.0030757	.0339889	0.09	0.928	-.0635413	.0696926
sib3	.0343756	.0200056	1.72	0.086	-.0048346	.0735859
_cons	-.2173203	.0771009	-2.82	0.005	-.3684353	-.0662054

```
Instrumented:  faminc_d
Instruments:   black hispanic male age sib1 sib3 inst_d
```

Problem 1.11. Should we be worried about $\Delta\chi_{ia}(P_{i,a-1})$ being a weak instrument?

Solution. We can test for relevance by regressing -faminc- on the instrument.

```
. reg faminc_d inst_d black hispanic male age sib1 sib3

      Source |       SS           df       MS      Number of obs   =     3,445
-------------+----------------------------------   F(7, 3437)      =      7.34
       Model |  5526.88529         7  789.555042   Prob > F        =    0.0000
    Residual |  369734.024     3,437  107.574636   R-squared       =    0.0147
-------------+----------------------------------   Adj R-squared   =    0.0127
       Total |   375260.91     3,444  108.960775   Root MSE        =    10.372

------------------------------------------------------------------------------
    faminc_d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      inst_d |     2.5266   .3638152     6.94   0.000     1.813284    3.239916
       black |  -.5973676   .4110474    -1.45   0.146    -1.403289    .2085543
    hispanic |  -.6928083   .5226354    -1.33   0.185    -1.717516     .331899
        male |   .1661264   .3540098     0.47   0.639    -.5279645    .8602174
         age |  -.1384597   .1143552    -1.21   0.226    -.3626707    .0857513
        sib1 |   .7175039   .6304075     1.14   0.255    -.5185073    1.953515
        sib3 |   .0827349   .3759061     0.22   0.826    -.6542869    .8197568
       _cons |   2.284736   1.428687     1.60   0.110    -.5164246    5.085897
------------------------------------------------------------------------------
```

▷ Since the coefficient is significant and sizable, we are not concerned about $\Delta\chi_{ia}\left(P_{i,a-1}\right)$ being a weak instrument.

We may be worried that also $P_{i,a-1}$ is endogenous, since it may be associated with $P_{i,a}$ by e.g. serially correlated shocks. By including in our IV model flexible controls for $P_{i,a-1}$, we may more plausibly incorporate in our instrument only the changes in I_{ia} deriving from changes in EITC, and avoid incorporating general changes in family income.

Problem 1.12. Reestimate the IV model in 10 above, including as control variables the dummy -laborpart- and a fifth-order polynomial in -faminc_L1-. Compare the estimates to those you got above.

Solution. Now we construct a modified model:

$$\Delta y_{ia} = \mathbf{x}_i'\boldsymbol{\alpha}_a + \mathbf{w}_{ia}'\boldsymbol{\beta} + \delta\Delta I_{ia} + \Phi\left(P_{i,a-1}\right) + \Delta\epsilon_{ia}$$

where $\Phi\left(P_{i,a-1}\right)$ represents a flexible controls for $P_{i,a-1}$. Adding the dummy -laborpart- and a fifth-order polynomial in -faminc_L1-, we have the following estimates:

```
Instrumental variables (2SLS) regression          Number of obs   =      3,445
                                                  Wald chi2(13)   =      14.90
                                                  Prob > chi2     =     0.3138
                                                  R-squared       =          .
                                                  Root MSE        =     .77961

------------------------------------------------------------------------------
             |               Robust
     score_d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    faminc_d |   .0568364   .0361477     1.57   0.116    -.0140119    .1276847
       black |   .0351855    .070458     0.50   0.618    -.1029095    .1732806
    hispanic |   .0681735   .0589273     1.16   0.247     -.047322    .1836689
        male |   .0231602   .0276339     0.84   0.402    -.0310014    .0773217
         age |   .0187135   .0113947     1.64   0.101    -.0036197    .0410467
        sib1 |    .042029   .0492368     0.85   0.393    -.0544733    .1385313
        sib3 |   .0057888   .0346476     0.17   0.867    -.0621192    .0736969
   laborpart |  -.1984162   .1189039    -1.67   0.095    -.4314636    .0346311
   faminc_L1 |   .0730289   .0710664     1.03   0.304    -.0662587    .2123165
 faminc_L1_2 |  -.0038857   .0040168    -0.97   0.333    -.0117585    .0039872
 faminc_L1_3 |   .0001094   .0001091     1.00   0.316    -.0001044    .0003233
 faminc_L1_4 |  -1.49e-06   1.38e-06    -1.08   0.281    -4.20e-06    1.22e-06
 faminc_L1_5 |   7.76e-09   6.61e-09     1.17   0.240    -5.20e-09    2.07e-08
       _cons |   -.897299   .5800846    -1.55   0.122    -2.034244     .239646
------------------------------------------------------------------------------
Instrumented:  faminc_d
Instruments:   black hispanic male age sib1 sib3 laborpart faminc_L1
               faminc_L1_2 faminc_L1_3 faminc_L1_4 faminc_L1_5 inst_d
```

▷ Compared to the results from part 10, the point estimate is much larger (about five times as large). It still remains insignificant, however.

Problem 1.13. Using this final model, create a loop that estimates the model repeatedly, setting as the dependent variable one of the test-score variables: -score-, -math-, -readingcomp-, and -readingrecog-.

Solution. We obtain the following estimates for each dependent variable.

▷ Using -score- as dependent variable:

Instrumental variables (2SLS) regression

Number of obs	=	3,445
Wald chi2(13)	=	14.90
Prob > chi2	=	0.3138
R-squared	=	.
Root MSE	=	.77961

score_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
faminc_d	.0568364	.0361477	1.57	0.116	-.0140119	.1276847
black	.0351855	.070458	0.50	0.618	-.1029095	.1732806
hispanic	.0681735	.0589273	1.16	0.247	-.047322	.1836689
male	.0231602	.0276339	0.84	0.402	-.0310014	.0773217
age	.0187135	.0113947	1.64	0.101	-.0036197	.0410467
sib1	.042029	.0492368	0.85	0.393	-.0544733	.1385313
sib3	.0057888	.0346476	0.17	0.867	-.0621192	.0736969
laborpart	-.1984162	.1189039	-1.67	0.095	-.4314636	.0346311
faminc_L1	.0730289	.0710664	1.03	0.304	-.0662587	.2123165
faminc_L1_2	-.0038857	.0040168	-0.97	0.333	-.0117585	.0039872
faminc_L1_3	.0001094	.0001091	1.00	0.316	-.0001044	.0003233
faminc_L1_4	-1.49e-06	1.38e-06	-1.08	0.281	-4.20e-06	1.22e-06
faminc_L1_5	7.76e-09	6.61e-09	1.17	0.240	-5.20e-09	2.07e-08
_cons	-.897299	.5800846	-1.55	0.122	-2.034244	.239646

Instrumented: faminc_d
Instruments: black hispanic male age sib1 sib3 laborpart faminc_L1 faminc_L1_2 faminc_L1_3 faminc_L1_4 faminc_L1_5 inst_d

▷ Using -math- as dependent variable:

Instrumental variables (2SLS) regression

Number of obs	=	3,445
Wald chi2(13)	=	18.98
Prob > chi2	=	0.1238
R-squared	=	.
Root MSE	=	.80023

math_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
faminc_d	.0156222	.0373226	0.42	0.676	-.0575288	.0887732
black	-.0215768	.072233	-0.30	0.765	-.1631509	.1199973
hispanic	.004982	.0603307	0.08	0.934	-.113264	.1232279
male	.015157	.0281364	0.54	0.590	-.0399893	.0703032
age	-.0242366	.0116973	-2.07	0.038	-.0471629	-.0013104
sib1	-.0047992	.0533983	-0.09	0.928	-.1094579	.0998595
sib3	.0291927	.0343121	0.85	0.395	-.0380578	.0964433
laborpart	-.0969815	.1226859	-0.79	0.429	-.3374414	.1434785
faminc_L1	.0090187	.0740802	0.12	0.903	-.1361758	.1542131
faminc_L1_2	-.0002302	.0042105	-0.05	0.956	-.0084826	.0080223
faminc_L1_3	7.13e-06	.0001147	0.06	0.950	-.0002176	.0002319
faminc_L1_4	-1.53e-07	1.46e-06	-0.11	0.916	-3.01e-06	2.70e-06
faminc_L1_5	1.13e-09	6.97e-09	0.16	0.872	-1.25e-08	1.48e-08
_cons	.1729458	.5991561	0.29	0.773	-1.001379	1.34727

Instrumented: faminc_d
Instruments: black hispanic male age sib1 sib3 laborpart faminc_L1 faminc_L1_2 faminc_L1_3 faminc_L1_4 faminc_L1_5 inst_d

▷ Using -readingcomp- as dependent variable:

readingcom~d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
faminc_d	.1035712	.056624	1.83	0.067	-.0074097	.2145521
black	.11028	.1123766	0.98	0.326	-.1099741	.3305342
hispanic	.0861164	.0967558	0.89	0.373	-.1035215	.2757543
male	.0426549	.0453679	0.94	0.347	-.0462646	.1315744
age	.0451407	.0184709	2.44	0.015	.0089384	.0813429
sib1	.0855219	.0808045	1.06	0.290	-.0728521	.2438959
sib3	.0117355	.0557384	0.21	0.833	-.0975098	.1209807
laborpart	-.2910437	.1887374	-1.54	0.123	-.6609622	.0788747
faminc_L1	.1362923	.1094578	1.25	0.213	-.078241	.3508256
faminc_L1_2	-.0074337	.0062315	-1.19	0.233	-.0196472	.0047799
faminc_L1_3	.0002097	.0001708	1.23	0.220	-.0001251	.0005445
faminc_L1_4	-2.82e-06	2.18e-06	-1.29	0.197	-7.10e-06	1.47e-06
faminc_L1_5	1.43e-08	1.05e-08	1.36	0.173	-6.26e-09	3.49e-08
_cons	-1.845645	.8990838	-2.05	0.040	-3.607817	-.0834731

▷ Using -readingrecog- as dependent variable:

readingrec~d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
faminc_d	.0276106	.0321791	0.86	0.391	-.0354594	.0906805
black	.0021783	.0614186	0.04	0.972	-.1182	.1225565
hispanic	.0849884	.0504264	1.69	0.092	-.0138454	.1838223
male	.002009	.023632	0.09	0.932	-.0443089	.048327
age	.0274315	.0100527	2.73	0.006	.0077286	.0471343
sib1	.027835	.0428026	0.65	0.515	-.0560566	.1117265
sib3	-.025976	.0291986	-0.89	0.374	-.0832042	.0312521
laborpart	-.1244686	.1054807	-1.18	0.238	-.331207	.0822698
faminc_L1	.0433171	.0627532	0.69	0.490	-.0796769	.1663111
faminc_L1_2	-.0023725	.0035749	-0.66	0.507	-.0093791	.004634
faminc_L1_3	.0000658	.0000975	0.68	0.500	-.0001253	.000257
faminc_L1_4	-8.83e-07	1.24e-06	-0.71	0.476	-3.31e-06	1.55e-06
faminc_L1_5	4.59e-09	5.93e-09	0.77	0.439	-7.04e-09	1.62e-08
_cons	-.6449549	.5067719	-1.27	0.203	-1.63821	.3482997

In all of these cases, we find that faminc_d is not a significant predictor of these individual test scores.

```
/*

Empirical Analysis III - MIDTERM
Simon Sangmin Oh
University of Chicago, Booth School of Business

*/

sysuse auto, clear
set scheme s2color

* -------------------------
* Import Data
use "C:\Users\Simon Oh\Dropbox\7. PHD\Year 1\Empirical Analysis III\Exams\Take-Home Midterm\DahlLochner2012AER.dta"

* -------------------------
* Q1 - Describe the data
gen score_raw = (math + readingcomp + readingrecog) / 3
egen score = std(score_raw)
replace score_raw = score

* -------------------------
* Q2 - Describe the data
loneway score id
loneway faminc id

* -------------------------
* Q3 - Graphing the means over time including 95% confidence interval
* 1. Score
preserve
collapse (mean) mean_score = score (semean) semean_score = score, by (year)
serrbar mean_score semean_score year, scale(1.96) title("Mean of score over time")
restore

* 2. Faminc
preserve
collapse (mean) mean_faminc = faminc (semean) semean_faminc = faminc, by (year)
serrbar mean_faminc semean_faminc year, scale(1.96) title("Mean of Family Income over time")
restore

* -------------------------
* Q4 - Estimate model using score as dependent variable
reg score faminc black hispanic male age agemom ed1age23 ed2age23 ed3age23 ed4age23 ///
          afqt afqt_miss married spouseage spouseage_miss famsize famsize_miss sib1 sib3 ///
          , robust

* -------------------------
* Q6 - First Differences
xtset id year
gen score_d = score - L2.score
gen faminc_d = faminc - L2.faminc

reg score_d faminc_d black hispanic male age sib1 sib3, robust

* -------------------------
* Q7 - Fixed Effects
xtreg score faminc black hispanic male age sib1 sib3, i(id) fe robust

* -------------------------
* Q10 - First Differences with Instruments
gen inst_d = eitc_sim - L2.eitc
ivregress 2sls score_d (faminc_d = inst_d) black hispanic male age sib1 sib3, robust

* -------------------------
* Q11 - Test Relevance
reg faminc_d inst_d black hispanic male age sib1 sib3

* -------------------------
* Q10 - Full Model
```

```
gen faminc_L1_2 = faminc_L1 * faminc_L1
gen faminc_L1_3 = faminc_L1_2 * faminc_L1
gen faminc_L1_4 = faminc_L1_3 * faminc_L1
gen faminc_L1_5 = faminc_L1_4 * faminc_L1
ivregress 2sls score_d (faminc_d = inst_d) black hispanic male age sib1 sib3 laborpart
faminc_L1 faminc_L1_2 faminc_L1_3 faminc_L1_4 faminc_L1_5, robust

* -------------------------
* Q13 - Create a a loop
gen math_d = math - L2.math
gen readingcomp_d = readingcomp - L2.readingcomp
gen readingrecog_d = readingrecog - L2.readingrecog
foreach var of varlist score math readingcomp readingrecog {
ivregress 2sls `var'_d (faminc_d = inst_d) black hispanic male age sib1 sib3 laborpart
faminc_L1 faminc_L1_2 faminc_L1_3 faminc_L1_4 faminc_L1_5, robust
}
```