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 the *American Economic Review* 2012, 102(5): 1927–1956, Dahl and Lochner (hereafter, DL) study how children's school performance depends on family income.[1] They posit the following model of the relationship

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

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

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.

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?

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.)

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-.

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

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

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

where ε_{ia} is random residual.

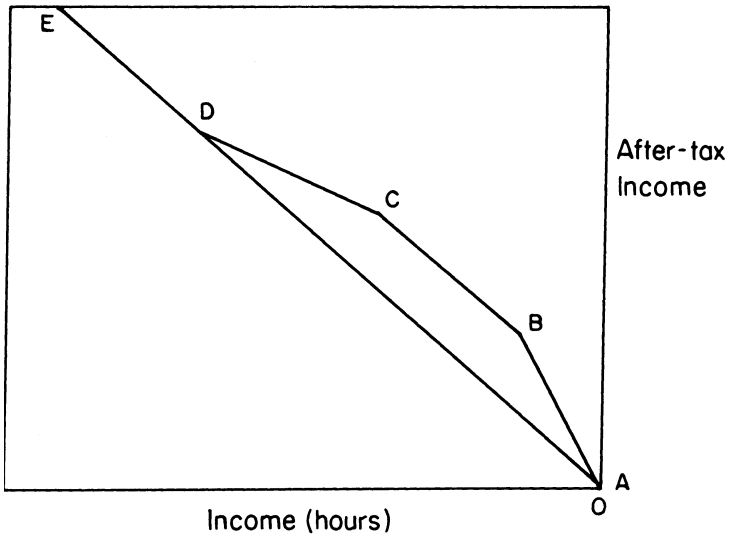
[1] Notice that the results from these estimations will not match the estimates in the paper, both because part of the data is classified, and because we have simplified their model somewhat.

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).

7. Estimate the model with fixed effects (using -xtreg, fe- in case of Stata), 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?

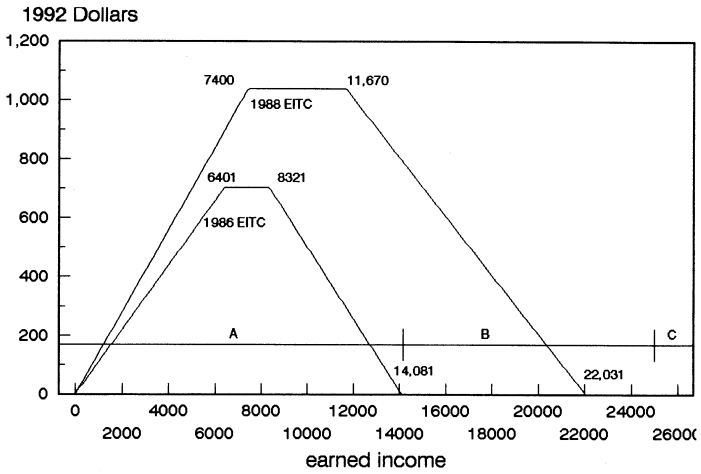
8. Why may we be worried about omitted variables bias also in the panel data models? (Hint: What is driving changes in family income?)

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

$$I_{ia} = P_{ia} + \chi_{ia}\left(P_{ia}\right) - \tau_{ia}\left(P_{ia}\right)$$

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

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 $\Delta\chi_{ia} = \chi_{ia}\left(P_{i,a}\right) - \chi_{i,a-2}\left(P_{i,a-2}\right)$ would be a better or worse instrument for ΔI_{ia}?

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

11. Should we be worried 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.

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.

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-.

Data description:

The file -DahlLochner2012AER.dta- contains *biannual* data on school performance and family income in the years 1987–1999, in addition to a number of observable characteristics of the children and their families. Each child is observed up to four times. The number of observations equals 7,280, covering 3,692 children. Because data are biannual, it will prove very useful to apply the -S2.- operator, see -help tsvarlist-.

The file includes the following variables:

	Variable	**Label**
1	id	Id
2	year	Period
3	faminc	Family income (in $1000, 2000-dollars)
4	eitc	EITC
5	eitc_sim	EITC, simulated
6	math	Mathematics
7	readingrecog	Reading recognition
8	readingcomp	Reading comprehension
9	black	Black
10	hispanic	Hispanic
11	male	Male
12	age	Age
13	agemom	Mother's age
14	ed1age23	Mother high school dropout
15	ed2age23	Mother high school graduate
16	ed3age23	Mother attended college
17	ed4age23	Mother graduated college
18	afqt	Mother's AFQT-score (normalized
19	afqt_miss	Mother's AFQT-score missing
20	married	Married
21	spouseage	Father's age
22	spouseage_miss	Father's age missing
23	famsize	No. of siblings
24	famsize_miss	No. of siblings missing
25	sib1	One sibling
26	sib3	Two or more siblings
27	laborpart	Labor participation
28	faminc_L1	Family income, 1 year previous
29	faminc_L2	Family income, 2 years previous