Problem Set 3: Replication and Discussion of an IV

Valentin Auplat, Tom Hamburger, Solal Godechot

February 2025

Education and Date of Birth Effects

In the 2011 paper "Academic performance, Educational Trajectories and the Persistence of Date of Birth Effects. Evidence from France", Julien Grenet studies the impact of date of birth on educational and labor market outcomes, exploiting institutional features of the French educational system. Read in detail sections 1, 2, 3, and 6.1 from the paper to understand the setting, the empirical strategy, and his main results.

In this problem set, we will replicate some of his results, using data from the Panel Primaire de l'Éducation nationale 1997 (PPEN97) and the Panel Secondaire de l'Éducation nationale 1995 (PSEN95), which are uploaded to the course website together with the paper. These datasets are provided for pedagogical purposes only. Please do not share or use them in other contexts than this problem set. You can check the "pupil datasets" and "test scores" subsections from the data section in the paper for details on both data sources. The data has been partially cleaned, and a variables' dictionary (in English) is provided on the course website.

Data Preparation

1. Open the PPEN97 and PSEN95 datasets in R and get familiar with variables' names and coding. Provide basic summary statistics of the test scores variables for each year and comment.

Table 1: Summary Statistics of score variables from PPEN97

Variable	Mean	SD	Min	Max	Missing_values
scglob	68.99529	12.90255	9.090909	98.97959	110
score_f_3	67.62938	15.68777	0.000000	100.00000	1913
score_m_3	66.03594	15.42584	0.000000	100.00000	1921

Table 2: Summary Statistics of score variables from PSEN95

Variable	Mean	SD	Min	Max	Missing_values
brevfra1	11.30173	2.592899	1.0	20	6873
brevlv11	11.59653	3.254623	0.4	20	6923
brevmat1	11.28627	3.437818	0.0	20	6875
franel	45.73405	11.314201	0.0	68	817
mathel	50.35004	13.791457	0.0	78	856

We don't see much differences between the scores from a year to another. The global score of year 1 in the data set PPEN97 is equivalent to the scores and male and female students in year 3 (3% difference), and standard error are as well. We note that the averages presented here are above the average of the distribution. For year 6 and year 9, we note a slight decrease in the standard deviations of the scores, and most importantly a decrease in the average grade compared to years 1 and 3. However, the distribution of grades should be standardized as they go from 0 to 20 or from 0 top 100 depending on the year. Section 4 precises that they are, but it is not reflected in the data set. In addition, we can note that a bit less than a third of the values of some variables from both data sets are missing. Therefore, we can suspect there might be endogeneity in our analysis because the absence of values is unlikely to be randomly distributed.

2. Prepare the data set with necessary variables:
   - Convert binary indicators into 0-1 format.
   - Generate the pupil's age at test (in months) and the "assigned relative age."

```
PPEN97 <- PPEN97 %>%
  mutate_at(vars(sexe, public_1, public_3), funs( . - 1)) %>%
  mutate_at(vars(prior_area_1:prior_area_3), funs(ifelse(.==2,0,.))) %>%
  mutate_at(vars(nati), funs(ifelse(.==100,0,1))) %>%
  mutate_at(vars(pcspere_1:pcsmere_3, size_area_1:dep_3), funs(as.character(.))) %>%
  mutate(age_exam_1 = (97-anai)*12+(10-mnai)) %>%
  mutate(age_exam_3 = (yr_3-(1900+anai))*12+(10-mnai)) %>%
  mutate(ass_rel_age = 12 - mnai)

PSEN95 <- PSEN95 %>%
  mutate_at(vars(sexe, public_6, public_9), funs(as.numeric(.))) %>%
  mutate_at(vars(sexe, public_6, public_9), funs( . - 1)) %>%
  mutate_at(vars(prior_area_6, prior_area_9), funs(ifelse(.==2,0,.))) %>%
  mutate_at(vars(nateleve), funs(ifelse(.==100,0,1))) %>%
  mutate_at(vars(pcspere_6,pcsmere_6,pcspere_9,pcsmere_9,size_area_6,size_area_9,dep_6,dep_9), funs(as.
  mutate(mnai = as.numeric(substr(datenai, 3, 4))) %>%
  mutate(anai = as.numeric(substr(datenai, 5, 6)))  %>%
  mutate(age_exam_6 = (yr_6-(1900+anai))*12+(10-mnai)) %>%
  mutate(age_exam_9 = (yr_9-(1900+anai))*12+(6-mnai)) %>%
  mutate(ass_rel_age = 12 - mnai)
```

Naïve Estimation of Month of Birth Effects

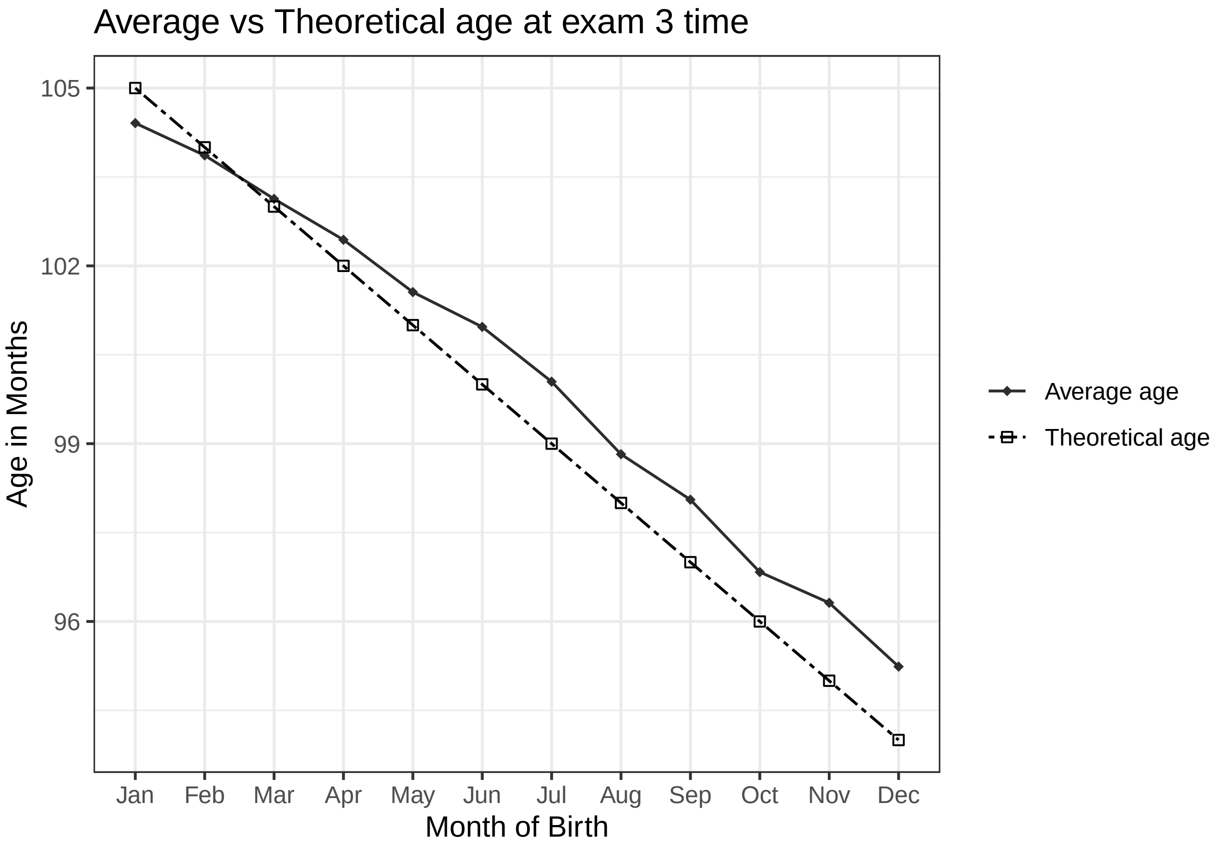
1. Discuss the main empirical challenges of estimating the effect of age on educational attainment using OLS.

What we are interested in in the first place is the effect of age on academic success. For a naive estimation to be telling about the month of birth effect, we would need the relationship between pupils' age and their time spent at school to be the same for each pupil. Otherwise, the age differences in an observed cohort (defined as pupils attending school at the same time) would not be telling at all because some behaviors would diminish it artificially. Actually, we are likely to observe this for two reasons according to the author:

First, pupils born at the beginning of the year are more likely to begin their schooling a year in advance, whereas pupils born at the end of the year are more likely to start school one year late. This is because parents decide to enroll their children at school by comparing their age with the age of their potential classmates. And they try to make them match. In addition, the children born at the end of the year are more likely to be held back a year, and children born at the beginning of the year are more likely to skip a grade. These phenomena diminish the actual age difference in a cohort, which would lead us to under estimate the effect of age on academic success.

This is not the only issue, because these strategies from the parents are likely to introduce endogeneity. For example, only few parents implement strategies to control the month of birth of their children (to optimize their career path, or the schooling of the children), which would affect the age difference within a class and could be related to the children's socio-economic background. The latter, which would be captured in the residuals of a naive OLS regression, would be correlated with the outcome, and the month of birth variable, introducing endogeneity.

2. Replicate Figure 7(a) from the paper, plotting theoretical versus observed age differences using the PPEN97 data. What do you conclude? Relate it to the previous question.



3. Replicate column 2, rows 1-3 from Table 1 in the paper using the PPEN97 dataset.

	Scores		
	Year 1 Exam	Year 3 Exam French	Year 3 Exam Maths
	(1)	(2)	(3)
Age in Months (Year 1 Exam)	0.030***		
	(0.003)		
Age in Months (Year 3 Exam)		−0.012***	−0.011***
		(0.003)	(0.003)
Constant	−2.248***	1.248***	1.100***
	(0.205)	(0.255)	(0.257)
Observations	9,531	7,728	7,663
R^2	0.012	0.003	0.002
Adjusted R^2	0.012	0.003	0.002
Residual Std. Error	0.994 (df = 9529)	0.999 (df = 7726)	0.998 (df = 7661)
F Statistic	120.006*** (df = 1; 9529)	23.913*** (df = 1; 7726)	18.354*** (df = 1; 7661)
Note:			*p<0.1; **p<0.05; ***p<0.01

4. Replicate rows 4-8 using the PSEN95 dataset (*).

	Scores	
	Year 6 french exam	Year 6 math exam
	(1)	(2)
Age in Months (Year 6 Exam)	-0.056^{***}	-0.056^{***}
	(0.001)	(0.001)
Age in Months (Year 9 Exam)	7.754^{***}	7.755^{***}
	(0.142)	(0.143)
Observations	17,003	16,964
R^2	0.149	0.148
Adjusted R^2	0.149	0.148
Residual Std. Error	0.923 (df = 17001)	0.923 (df = 16962)
F Statistic	$2{,}973.209^{***}$ (df = 1; 17001)	$2{,}954.924^{***}$ (df = 1; 16962)
Note:		*p<0.1; **p<0.05; ***p<0.01

	Scores	
	Year 9 french exam	Year 9 math exam
	(1)	(2)
Age in Months (Year 6 Exam)	-0.049^{***}	-0.051^{***}
	(0.001)	(0.001)
Age in Months (Year 9 Exam)	8.893^{***}	9.283^{***}
	(0.230)	(0.229)
Observations	10,947	10,945
R^2	0.120	0.131
Adjusted R^2	0.120	0.131
Residual Std. Error	0.938 (df = 10945)	0.932 (df = 10943)
F Statistic	$1{,}496.150^{***}$ (df = 1; 10945)	$1{,}650.045^{***}$ (df = 1; 10943)
Note:		*p<0.1; **p<0.05; ***p<0.01

	Scores
	Year 9 LV exam
Age in Months (Year 9 Exam)	-0.051^{***}
	(0.001)
Constant	9.248^{***}
	(0.230)
Observations	10,897
R^2	0.130
Adjusted R^2	0.130
Residual Std. Error	0.933 (df = 10895)
F Statistic	$1{,}625.986^{***}$ (df = 1; 10895)
Note:	*p<0.1; **p<0.05; ***p<0.01

5. Interpret the results. Discuss potential biases in OLS estimates.

We see that the naive OLS estimation is not accurate and cannot be trusted there as it provides contradictory results. They go against the robust assumption that older pupils get higher grades than younger ones in a same cohort (around minus 0.05 points of normalized score per month for year 9 and year 6 exams, and minus 0.01 for the year 3 exam). And most importantly because if we trust the naive OLS, we have to conclude that the pupils' age has an opposite effect for the first year exam (plus 0.03 points of normalized score per additional month)! Therefore, we have to recognize that the naive OLS probably provides biased results, and does not allow to reach the conclusions the author aims for. We have to think about ways to work around potential endogeneity (such as students held up remaining in the higher age range and scoring low).

IV Estimation of Month of Birth Effects

1. Explain how "assigned relative age" is a valid instrument for age in the test score equation.

For an IV instrument to be valid, it needs to fulfill the relevance condition and the exclusion restriction.

Firstly, as $z_i = 12 - m_i$ (m_i being the month of birth), the relevance condition is fulfilled because the large majority of children enter school at the normal age. Therefore, the correlation between pupils' age and their relative assigned age must be close to one. As school years pass by, the correlation should decrease, as more and more pupils will skip or retake a school year (but this is marginal). We can note that the instrumental variable score will be higher of pupils born in January, and minimized for pupils born in December.

Secondly, if we assume that m_i is randomly distributed, then the exclusion condition is fulfilled. But the author mentions it is not necessarily the case. As m_i might not be random, we might not assume that the relative assigned age is not correlated with other determinants of academic success. We have already discussed the fact that some parents, belonging to specific socio-economic classes, might choose the month of birth of their children, which directly impacts the instrument, and might be correlated with children's socio-economic background characteristics determining academic and professional success.

2. Write down the statistical model for columns 3 and 4 in Table 1 and derive the IV estimator. What does it identify?

The third column of the table corresponds to the result of the first-stage regression which we can write as:

$$a_{ig} = \gamma_g + \delta_g z_i + \eta_{ig} = \gamma_g + \delta_g(12 - m_i) + \eta_{ig}$$

Where a_{ig} is the absolute age when the test is taken, and z_i is the instrument (assigned relative age). The first stage measures the independent variations of the absolute age through the relative assigned age (if we assume that both the relevance and exclusion conditions hold). The link between both variables is $\delta_g = \frac{Cov(a_{ig};z_i)}{V(z_i)}$.

Once we have this, we can run the second-stage regression to get the same results as in column 4. We can write the regression model (or reduced form) as:

$$s_{ig} = \lambda_g + \mu_g z_i + \nu_{ig} = \lambda_g + \mu_g(12 - m_i) + \nu_{ig}$$

Where s_{ig} is test score obtained in grade level g by pupil i. μ_g measures the impact of relative age on test scores (still assuming our 2 conditions hold).

We can compute the IV estimator as:

$$\beta_{IV} = \frac{\frac{Cov(s_g;z)}{V(s)}}{\frac{Cov(a_g;z)}{V(z)}} = \frac{Cov(s_g;z)}{Cov(a_g;z)} = \frac{\mu_g}{\delta_g} = LATE$$

We recognize the Local Average Treatment Effect in this IV estimate (as long as the two conditions and the no-defiers assumption hold), i.e. the average treatment effect on compliers, i.e. the influence of the exogenous variations of absolute age on test scores for pupils that have started school normally (no advance or delay).

3. Replicate columns 3-7 of Table 1 using PPEN97 data and format the results in a table.

	First Stage	Reduced Form	IV	IV with controls	IV pupils born in Jan. or Dec.
Year.1..Global.score	0.908	0.057	0.063	0.06	0.056
SE	(0.006)	(0.003)	(0.003)	(0.003)	(0.005)
t.stat	[143.417]	[19.337]	[18.943]	[20.082]	[11.976]
Year.3..Maths	0.853	0.042	0.05	0.049	0.047
SE.1	(0.011)	(0.003)	(0.004)	(0.004)	(0.006)
t.stat.1	[75.44]	[12.759]	[12.166]	[13.078]	[7.77]
Year.3..French	0.853	0.031	0.036	0.037	0.036
SE.2	(0.011)	(0.003)	(0.004)	(0.004)	(0.006)
t.stat.2	[75.44]	[9.189]	[8.948]	[10.205]	[6.171]

5. Compare OLS and IV estimates. Discuss the direction of potential bias.

All the coefficients of the IV estimations are statistically significant at the 1% risk level. Being a month older increases the Year 1 score by 0.063 units. We see that this effect is persistent but less and less important over time, falling to 0.05 units for math exam in Year 3, and to 0.036 units for french exam in Year 3.

We see that the coefficients we get from the IV models are very different from the ones we get with the naive OLS. They have the same sign for the first year, but the OLS estimators become negative for year 3 whereas as the IV estimators remain positive. Therefore, we can conclude that there must be a downward bias in the OLS estimates. This bias is due to the increase in the difference between the observed age of pupils and their normal age for their grade. Because there is an increasing number of skippers and retainers over time, the absolute age differences shrink, and therefore the OLS regression of test scores on month of birth "would under estimate the true impact of absolute age differences on pupils performance" according to the author.

We are comforted in this analysis as the IV models used seem robust. All of them provide almost the same estimates, and all have low variance. In particular, we see that adding controls and keeping only pupils born in January and December don't change the estimates and their significance too much. Therefore, we could consider that our framework is robust to the violations of the exclusion condition we were concerned about.