Problem set 3: Replication and discussion of an IV

Luc Behaghel (luc.behaghel@psemail.eu), Pietro Geuna (pietro.geuna@psemail.eu)
Econometrics 3 - APE
Paris School of Economics

February 2025

```
Questions containing (*) at the end do not necessarily have to be handed in.
However, students are expected to think about them in advance and discuss and
present them in class.  Submit to metrics2324@gmail.com
```

```
Students are required to hand in the answers to the Problem Set as well as the
complete R script that implements the replication.
```

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.[1] 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 in the course website.

1 Data preparation

1. The first step in every empirical analysis should be to know the data we are working with. Open the PPEN97 and PSEN95 datasets in R and get familiar with variables' names and

[1] These datasets are provided for pedagogical purposes only. Please do not share or use in other contexts than this problem set.

coding. We will work with test scores for years 1 and 3 from PPEN97, and 6 and 9 from PSEN95. Provide basic summary statistics of the test scores variables for each of these years and comment.

2. Next, you need to have a ready database with the variables to be used in the analysis.[2] First, make sure that binary indicators are in a 0-1 format (otherwise convert them). Second, you need to generate the two age-related variables used in the paper: the pupil's age at test (measured in months), and the "assigned relative age".
   Hint 1: You may want to check the class of each variable before proceeding (see `?class`*). In cases of character variables, the* `substr()` *function may be of help.*
   Hint 2: In order to measure the pupil's age at the test, you need to take into account the date at which the exam is taken.[3]
   Hint 3: You might find useful to have a look at the `mutate()` *function from the dplyr package.*

2 Naïve estimation of month of birth effects

The goal of the paper is to measure the relationship between pupil performance and date of birth while providing evidence on the underlying mechanisms of such effect.

1. What are the main empirical challenges when trying to estimate the effect of age on educational attainment by running a simple OLS of month of birth on test scores within school grade?

2. We will now replicate Figure 7(a), that plots theoretical versus observed age differences, using the PPEN97 data.[4] Beware that you will need to generate the "normal" age (measured in months) for pupils in Year 3 (cf. Hint 2). What can you conclude from it? Relate it to your answer to question 1.

3. Column 2 from Table 1 in the paper reports the results of the naïve estimation of the impact of observed age at test (in months) on test scores using OLS. Replicate the first three rows, corresponding to Year 1 and Year 3, using the PPEN97 (there is a small difference in the

[2] These are: test scores, age (in months), assigned relative age, a dummy for gender, a dummy for French nationality, the occupation of the pupil's parents, the number of students in the class, a dummy for public school, school department, a dummy for schools located in a priority education zone, size of the agglomeration.

[3] October 1997 for Year 1 pupils; October 1999 for Year 3 pupils who were not held back in Year 1 or 2, or October 2000 if they were held back once. For Year 6 pupils, exam is taken on October 1995, and Year 9 test scores correspond to a continuous assessment of the junior high school certificate, either in 1999, 2000 or 2001 depending on whether the pupil was held back or not. We will consider June as the month of assessment.

[4] ERRATA: Second paragraph in the "Theoretical vs. observed age differences" in pp. 19: The solid line refers to age at the exam date for Year 3 pupils (in October 1999), not Year 1. Likewise, the note to Figure 7(a), should read "The normal age at which pupils sit the Year 3 test is the age they would have had in October 1999 had they started primary schooling at the normal age (i.e. born in 1991) without repeating a year".

number of observations, which will result in differences in the third decimal).

Hint 4: Remember to normalize the scores so that they have a mean of zero and a standard deviation of one.
Hint 5: In order to use the same sample of pupils across specifications, the author defines a "regression sample" for each year containing all pupils with nonmissing information in all the regression variables for that particular year (cf. footnote 2), and runs all the regressions on the specific subsample.

4. Replicate now rows 4-8 from column 2 (Table 1), using the PSEN95. (*)
5. Interpret the results. Why would you expect these OLS estimates to be biased?

3 IV estimation of month of birth effects

1. The author proposes using the "assigned relative age" as an instrument for age in the test score equation. Explain how this instrument meets the necessary conditions to be valid.
2. Write down the statistical model behind columns 3 and 4, and derive the IV estimator. What does it identify? On which population?
3. Taking into account the hints previously mentioned, replicate the results in columns 3 to 7 using the PPEN97 (rows 1-3) and export them to a properly formatted table.
   Hint 6: For categorical variables, you should transform them into separate dummies to be included in the regression. Check the `factor()` *function in R.*
   Hint 7: To export results in formatted tables, have a look at the `stargazer` *package.*
   Hint 8: For IV regressions, use the `ivreg()` *function from the AER package*
4. Now repeat exercise 3 for rows 4-8 using PSEN95.(*) *Hint: You might want to check the use of loops in R.*
5. Comment on the results. What can you conclude from the comparison between OLS and IV estimates? Is there a bias? If yes, in which direction?

4 Heterogeneity analysis (*)

Table 2 studies how age affects performance across different subpopulations. We will explore heterogeneous effects by socioeconomic origin (columns 3 and 4) for Year 1 pupils (first row only).[5]

[5]The author defines *privileged background* as pupils in households where the household head is self-employed, professionals, managers, or intermediary occupations (codes 21-23, 31-35,37-38,42-48 of the `pcschef97` variable).

1. Estimate the impact of month of birth on test scores in Year 1, using the assigned relative age as an instrument, on the two separate subsamples (privileged and underprivileged), and comment on the results. Do not add any additional controls for now.

2. Now estimate the same effect but on the entire sample, interacting both the endogenous variable and the instrument with a privileged background indicator. Compare your results with those from question 1. What can you conclude? How can you test if the age effects are different for pupils from different socioeconomic origin?

3. Add the same group of control variables used in the previous exercise, and repeat exercises 1 and 2 (at this point you should obtain the same results as row 1 columns 3 and 4 in Table 2 of the paper). Do your conclusions hold?

4. Finally, interact all your controls with the privileged background indicator. How do the results change? Why?

Household heads who are farmers, employees, manual workers, unemployed or economically inactive (i.e. all other codes) are considered from *underprivileged background*. See footnotes 20 and 21 in the paper.