# Econometrics 3 - Problem Set 3

Pierre Pili - Isée Biglietti - Marie Gardie

16/02/2024

## 1. Data Preparation

### 1)

Table 1: PPEN97 – Scores in Years 1 and 3

| var | median | mean | min | max | sd | NAs (%) |
|---|---|---|---|---|---|---|
| score_1 | 69.8 | 69.0 | 9.1 | 99 | 12.9 | 1.1 |
| score_f_3 | 69.2 | 67.6 | 0.0 | 100 | 15.7 | 19.8 |
| score_m_3 | 67.5 | 66.0 | 0.0 | 100 | 15.4 | 19.9 |

Table 2: PSEN95 – Scores in Years 6 and 9

| var | median | mean | min | max | sd | NAs (%) |
|---|---|---|---|---|---|---|
| score_f_6 | 47.0 | 45.7 | 0.0 | 68 | 11.3 | 4.6 |
| score_m_6 | 52.0 | 50.4 | 0.0 | 78 | 13.8 | 4.8 |
| score_f_9 | 11.0 | 11.3 | 1.0 | 20 | 2.6 | 38.6 |
| score_m_9 | 11.5 | 11.3 | 0.0 | 20 | 3.4 | 38.6 |
| score_fl_9 | 12.0 | 11.6 | 0.4 | 20 | 3.3 | 38.8 |

As grades during year 9 were between 0 and 20 we rescaled them between 0 and 100 to make comparisons easier. The 1997 dataset (PPEN97) contains one variable for year 1 `score_1` and two for year 2 measuring French and Math test scores denoted `score_f_3` and `score_m_3`. The 1995 dataset (PSEN95) also contains French and Math test scores for year 6 (`score_f_6`, `score_m_6`, `score_f_9`, `score_m_9`) and one additional variable for foreign language test scores `score_fl_9`. We notice that year 9 contains more missing values with almost 40% for `score_f_9` and `score_m_9`.

### 2)

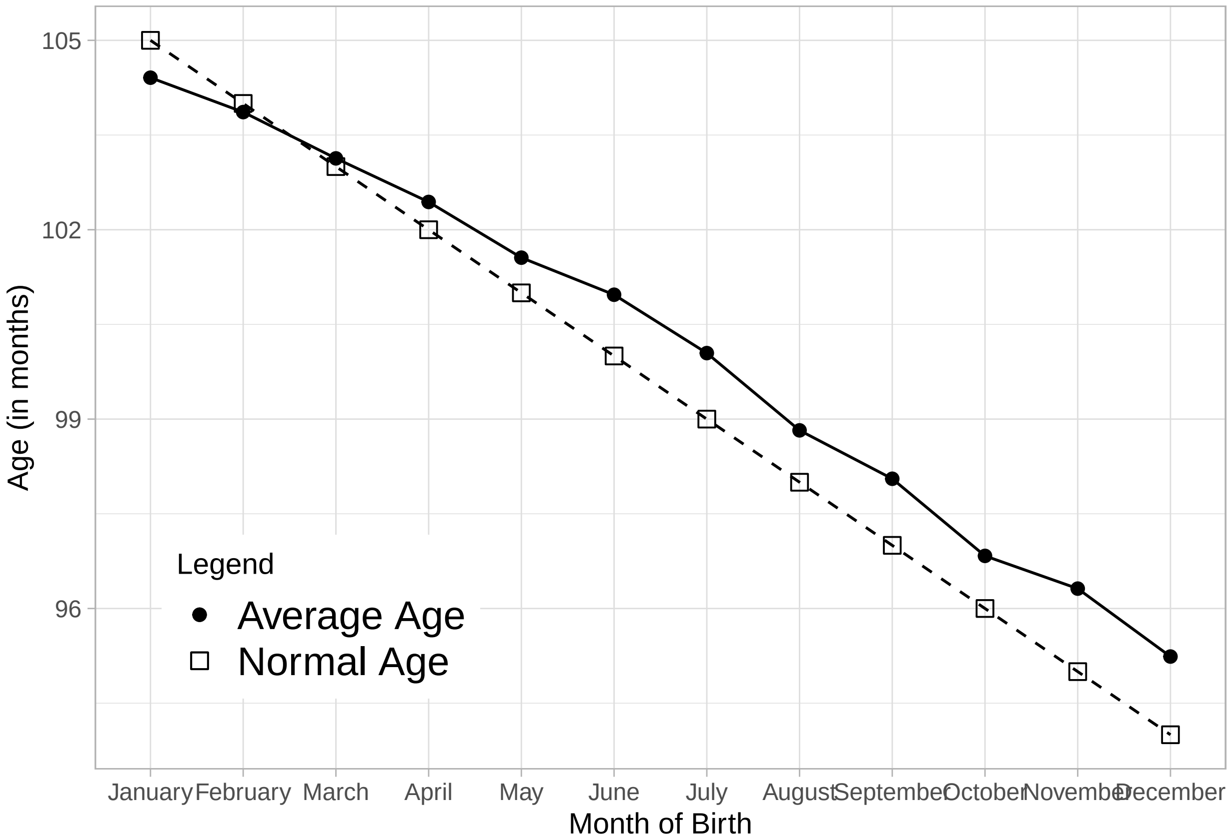
(see code)

## 2. Naïve Estimation of Month Birth Effects

### 1)

Firstly, pupil born in the beginning of the year will most likely join school a year sooner and conversely for those born at the end of the year. This is will reduce the actual age difference among each cohort. Moreover, students can be held back for a year or more which is more likely for students born at the end of the year. Both effects imply an underestimation of the true effect. Secondly, there is a source of endogeneity as social economic backrounds can explain both the month of birth and the test score of the pupil. For instance, teacher's children are more generally born in may and are more sucessful at school. There is also a measurment problem with foreign born children who sometimes get assigned a date of birth, usually in January, which leads to January born children to be over-represented. The OLS estimator will thus be biased.

**2)**




**3)**

## 3. IV estimation of month of birth effect

**1)**

**STUVA**: Only my own assignement of assigned relative age matter for my own enrollment (if my friend has no assigned relative age, I am still going to enroll in school and pass this test, and passing the test isn't even up to me as a pupil, it's up to the school), and for my own outcome (knowing that I am assigned a relative age but my friends are not doesn't change my studdy behavior a priori).

**Random assignement**: Not as random as we want it to be as seen before. In fact, we think again about this teacher parents' matter that deprive our instrument of independence with the outcome, here test scores, as teacher parents' children are mostly born in May and are more successful at school (correlation wildly documented).

**Exclusion restriction**: The instrument must be uncorrelated with the unobservable determinants of test scores, or in other words, out of the data generating process. Again here, the teacher example show us that seasonality in birth month implies different socio-economic background for specific students, as well as grade retention can totally be a variable through which our instrument affects the output measured here.

**Relevance condition**: Our instrument must be correlated to the original variable it proxies, being here the absolute age at which the pupil took the test. The correlation is very close to one when students are young, and should gradually decrease as retention gets more common. Though we can ask ourselves about

Table 3: Naïve OLS

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

the fact that skipping classes is more frequent when young, especially in high socio-economic backgroung geographic zones (as primary school system in France is geographically based and more respected than for secondary or high schools).

**Monotonicity**: It is quite belivable that there are no defiers in this configuration, as being assigned 11 (january) will not tend to give an incentive to NOT change cohort while having a 9 would give you the incentive to skip a year. The more plausible assumption is that if you want to skip a year having been assigned 9, you will want to skip having been assigned any other number. We however expect the presence of compliers (as seen in the paper).

The instrument is not perfect, and maybe other methods could be considered as the exclusion restriction is hardly reliable here.

### 2)

The benchmark model writes :

$$s_{ig} = \alpha_g + \beta_g a_{ig} + \epsilon_{ig}$$

where, for grade level $g$ by students $i$, $s_{ig}$ identifies test score obtained, $a_{ig}$ is the absolute age, in month, at which the test is taken (the endogenous variable), $\beta_g$ is the partial impact of $a_{ig}$ on academic performance $s_{ig}$, $\epsilon_{ig}$ is the error term. Column (3) is generated based on the first stage, that writes as follow:

$$a_{ig} = \gamma_g + \delta_g z_i + \eta_g$$

where $z_i = 12 - m_i$ is the assigned relative age (our instrument), $\eta_g$ the new error term (orthogonal to $z_i), $\delta_g$ estimates the relevance of our instruments. Column (4) is generated based on the reduced form, writting:

$$s_{ig} = \lambda_g + \mu_g z_i + \nu_i g$$

Here, $\mu_g$ measures the impact of assigned relative age (instrument) on test scores (outcome), net of grade retention, early and late entry. The IV estimate writes:

$$\hat{\beta}_g^{IV} = \frac{\mu_g}{\delta_g}$$

Our IV, as always, identifies the effect of absolute age on test scores for students that started school in their right cohort (entering CP having 6 years old before the 1st of January). To put it differently, it identifies the average treatment effect on compliers, and thus correspond to a LATE.

**3)**

Table 4: First stage

| | Scores | | |
|---|---|---|---|
| | Year 1: Global score | Year 3: Math | |
| | (1) | (2) | (3) |
| Age in Months (Year 1) | 0.908*** | 0.853*** | 0.853*** |
| | (0.006) | (0.011) | (0.011) |
| Age in Months (Year 3) | 70.512*** | 95.455*** | 95.455*** |
| | (0.041) | (0.074) | (0.074) |
| Observations | 9,641 | 7,728 | 7,728 |
| $R^2$ | 0.681 | 0.424 | 0.424 |
| Adjusted $R^2$ | 0.681 | 0.424 | 0.424 |
| Residual Std. Error | 2.122 (df = 9639) | 3.385 (df = 7726) | 3.385 (df = 7726) |
| F Statistic | 20,568.520*** (df = 1; 9639) | 5,691.122*** (df = 1; 7726) | 5,691.122*** (df = 1; 7726) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 5: Reduced Form

| | Scores | | |
|---|---|---|---|
| | Year 1: Global score | Year 3: Math | Year 3: French |
| | (1) | (2) | (3) |
| Age in Months (Year 1) | 0.057*** | 0.042*** | 0.031*** |
| | (0.003) | (0.003) | (0.003) |
| Age in Months (Year 3) | −0.313*** | −0.235*** | −0.170*** |
| | (0.019) | (0.022) | (0.022) |
| Observations | 9,531 | 7,720 | 7,728 |
| $R^2$ | 0.038 | 0.021 | 0.011 |
| Adjusted $R^2$ | 0.038 | 0.021 | 0.011 |
| Residual Std. Error | 0.981 (df = 9529) | 0.990 (df = 7718) | 0.995 (df = 7726) |
| F Statistic | 373.918*** (df = 1; 9529) | 162.788*** (df = 1; 7718) | 84.436*** (df = 1; 7726) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |