Case A: Evaluating school starting age rules

Introduction

Providing good education to children is considered essential for societies. Well educated individuals are not only more productive and more innovative (e.g. Harmon, Oosterbeek and Walker, 2003), but also are less often involved in crime (Deming, 2011) and have stronger civic engagement (Hout, 2012). While these positive returns to education are widely recognized by policy makers, educational outcomes still differ substantially between countries and within-country trends in academic performance of teenagers are declining in many countries.

The PISA test assesses the performance on reading, math and science tasks of students aged 15 in different countries. The OECD first collected the PISA data in 2000 and since then the data collection has been repeated every three years. In each country at least 5,000 students take a computer-based test which takes them about two hours. Different students get different parts of the full test. After taking this test students get a one-hour questionnaire on their family background, learning habits and motivations. The goal of the OECD for collecting the PISA data was to provide information to countries that can be used to improve educational policies. Academics now also widely use the PISA data for research on educational policies, inequality of opportunities and trends in academic outcomes (e.g. Fuchs and Woessmann, 2007; Hanushek and Woessmann, 2012; and Jain and Stemper, 2023).

Assignment

In this case you will use the PISA data to investigate differences in educational outcomes between countries and within countries. You will use the PISA data for the OECD countries and consider the last three waves (2022, 2018 and 2015). This case has three parts, that each serve a different purpose. When writing your paper, it is important that you clearly distinguish which findings are correlations and which findings can be interpreted causally. In case of a causal interpretation, state the identifying assumptions and provide a justification why these assumptions hold. This assignment contains a short appendix describing the PISA data. Read this appendix carefully and pay attention to the notes at the end on plausible values and sampling and replication weights.

Part 1: Descriptive statistics

In this part you provide some descriptive statistics. Since there are many countries participating in PISA, it is important to clearly visualize the data. Furthermore, make a selection of countries that allow you to make your results insightful. The descriptive statistics are meant to provide motivation for the rest of the paper. Keep the discussion brief and to the point and make a link to the following parts in your paper.

You start with documenting that math and reading scores differ between countries and within countries over time. Next, show that also within countries there is substantial variation in outcomes. You can consider the standard deviation, decimals and quartiles, or any other measure that makes the variation in outcomes insightful. Finally, consider intergenerational mobility: focus on differences between children with high-educated and low-educated parents.

Part 2: School starting age

Educational systems differ between countries, for example in terms of class size, teaching hours, tracking, compulsory education, etc. In this part you consider the school starting age. Show how the math and reading test scores of students correlate with the mean school starting age of primary school across countries. To measure the mean school starting age use the ISCED level 1 variable in the data and compute the average for each country in each wave. Is this correlation robust to controlling for student characteristics?

Next, use the PISA data to empirically find the school starting rule for a number of countries. Only consider countries with a clear school starting rule. Compare - within countries - the students who are born just before and just after the school starting threshold. This follows Oosterbeek et al. (2021) but extends it to multiple countries. You should pool the data from the different countries in your regression analysis to find an average effect over all countries. Think carefully about whether to include country fixed effects or not. Additionally, estimate a separate effect for each country that shows a clear school starting age rule in the PISA data.

Part 3: Mechanisms, heterogeneity and robustness

The school starting age may affect how many years of education a student has followed at the time of taking the PISA tests. Furthermore, students who are born just before the school starting threshold date are always the youngest in their class while students who are born just after this threshold date are always the oldest of the class. Students who start young may also be more likely to repeat a grade. The aim of this third part is to consider possible mechanisms through which the school starting age can affect test scores of teenage students. To provide insights in mechanisms you should consider variables such as grade repetition, the age at test and the relative age within the class. In addition, you may want to explain heterogeneity in the effects of early school starting. For example, children of low-educated parents may be more at risk of poor performance when starting school too young. And the effects of early school starting may be smaller if high-quality early childhood care is provided before starting school.

For this part you should be creative in thinking about mechanisms and how to specify your econometric models. Think also carefully about which variables are outcomes and which are explanatory variables. Finally, when interpreting the results of this section take the robustness of your results into account.

Literature

Deming, D. J. (2011). Better schools, less crime?. *Quarterly Journal of Economics*, 126(4), 2063-2115.

Fuchs, T. and L. Woessmann (2007), What accounts for international differences in student performance? A re-examination using PISA Data, *Empirical Economics* 32 (2-3): 433-464.

Hanushek, E.A. and L. Woessmann (2012), Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth* 17 (4): 267-321.

Harmon, C., Oosterbeek, H., & Walker, I. (2003). The returns to education: Microeconomics. *Journal of Economic Surveys*, 17(2), 115-156.

Hout, M. (2012). Social and economic returns to college education in the United States. *Annual Review of Sociology*, 38, 379-400.

Jain, R. and S. Stemper (2023), 3G internet and human capital development, Worling Paper.

Oosterbeek H., S. ter Meulen and B. Van der Klaauw (2021), Long-term effects of school-starting-age rules, *Economics of Education Review* 84, 102144.

APPENDIX: PISA data description

You are provided a selection of variables from the PISA waves 2015, 2018 and 2022. See below for a table with a description of all variables in the dataset. The year of the data is indicated with the variable WAVE (created by us). The remainder of the variables in the datasets all come directly from the PISA repository, we only made some small adjustments to make variable values comparable across the waves.

Some of the questions can directly be linked to questions from the questionnaire, which can be found here for 2022:
https://www.oecd.org/pisa/data/2022database/CY8_202111_QST_MS_STQ_CBA_NoNotes.pdf

The remaining variables are derived from (multiple) questions in the questionnaire, to read how this was done and on which questions they are based you can search for the variable in this document (again for wave 2022):
https://www.oecd.org/pisa/data/pisa2022technicalreport/PISA-2022-Technical-Report-Ch-19-PISA-Scaling-Procedures-Construct-Validation-Context-Questionnaire-Data.pdf

NOTE 1: The Pisa dataset makes use of so-called plausible values. These plausible values are generated through multiple imputations based upon pupils' answers to the sub-set of test questions they were randomly assigned and their responses to the background questionnaires. **Analysis of such plausible values is a bit more complicated, but for the sake of brevity, you may just take the average score per student on each domain as an outcome variable for your analyses.**

NOTE 2: The Pisa dataset is based on a sample, and not on the whole population of 15-year olds. This means that you will have to use the provided sampling weight (W_FSTUWT). Technically, you would need to use another set of 80 replicate weights to get correct standard errors. **However, for sake of brevity, we have omitted these and you may ignore this issue.**

If you want to dive deeper into these two issues, see here:
https://www.oecd.org/pisa/data/httpoecdorgpisadatabase-instructions.htm
At this link you can also find software packages that directly execute all steps necessary to deal with both plausible values and the sampling weights. **If you want to, you are allowed to use these packages, but whether you use them or not will not count towards the grading of your assignment.**

Description of variables

name	Variable lable
CNT	Country code 3-character
CNTSCHID	Intl. School ID
CNTSTUID	Intl. Student ID
CYC	PISA Assessment Cycle (2 digits + 2 character Assessment type - MS/FT)
STRATUM	Stratum ID 5-character (cnt + original stratum ID)
OECD	OECD country
ST001D01T	Student International Grade (Derived)
ST003D02T	Student (Standardized) Birth - Month
ST003D03T	Student (Standardized) Birth -Year
ST004D01T	Student (Standardized) Gender
ST230Q01JA	How many siblings (including brothers, sisters, step-brothers, and step-sisters)
ST021Q01TA	How old were you when you arrived in [country of test]?
ST126Q01TA	How old were you when you started [ISCED 1]: Years
OCOD1	ISCO-08 Occupation code - Mother
OCOD2	ISCO-08 Occupation code - Father
OCOD3	ISCO-08 Occupation code - Self
PROGN	Unique national study programme code

AGE	Students' age
GRADE	Grade compared to modal grade in country
IMMIG	Index on immigrant background (OECD definition)
REPEAT	Grade repetition
EXPECEDU	Highest expected educational level
DURECEC	Duration in early childhood education and care
BSMJ	Expected occupation status (free response)- 4 digits
MISCED	Mother's level of education (ISCED)
FISCED	Father's level of education (ISCED)
PAREDINT	Index highest parental education (international years of schooling scale)
BMMJ1	Mother's occupational status (ISEI) based on 4-digit human coded ISCO
BFMJ2	Father's occupational status (ISEI) based on 4-digit human coded ISCO
HISEI	Highest parental occupational status (ISEI) based on 4-digit human coded ISCO
ICTRES	ICT Resources (WLE)
HOMEPOS	Home possessions (WLE)
ESCS	Index of economic, social and cultural status
W_FSTUWT	FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT WEIGHT
PV1MATH	Plausible Value 1 in Mathematics
PV2MATH	Plausible Value 2 in Mathematics
PV3MATH	Plausible Value 3 in Mathematics
PV4MATH	Plausible Value 4 in Mathematics
PV5MATH	Plausible Value 5 in Mathematics
PV6MATH	Plausible Value 6 in Mathematics
PV7MATH	Plausible Value 7 in Mathematics
PV8MATH	Plausible Value 8 in Mathematics
PV9MATH	Plausible Value 9 in Mathematics
PV10MATH	Plausible Value 10 in Mathematics
PV1READ	Plausible Value 1 in Reading
PV2READ	Plausible Value 2 in Reading
PV3READ	Plausible Value 3 in Reading
PV4READ	Plausible Value 4 in Reading
PV5READ	Plausible Value 5 in Reading
PV6READ	Plausible Value 6 in Reading
PV7READ	Plausible Value 7 in Reading
PV8READ	Plausible Value 8 in Reading
PV9READ	Plausible Value 9 in Reading
PV10READ	Plausible Value 10 in Reading
PV1SCIE	Plausible Value 1 in Science
PV2SCIE	Plausible Value 2 in Science
PV3SCIE	Plausible Value 3 in Science
PV4SCIE	Plausible Value 4 in Science
PV5SCIE	Plausible Value 5 in Science
PV6SCIE	Plausible Value 6 in Science
PV7SCIE	Plausible Value 7 in Science
PV8SCIE	Plausible Value 8 in Science
PV9SCIE	Plausible Value 9 in Science
PV10SCIE	Plausible Value 10 in Science
WAVE	