Reproducible Quantitative Data Science - Coursework 1

Rasmus H. Klokker

2023-09-13

1 Brief background and description

Teacher mobility patterns may have structural adverse effects on pupils' academic attainment, contributing to the overall educational inequality in society. Existing research suggests that teachers with varying degrees of qualification and teaching experience are unequally distributed among schools. Evidence suggests that disadvantaged schools such as low-achieving schools, schools in impoverished areas, or schools with many low socioeconomic status (SES) and ethnic minority pupils face challenges attracting skilled teachers. Furthermore, teachers employed at these schools typically transfer to higher-achieving schools once they have acquired a certain amount of teaching experience. Thus, schools with many negative traits may end up in a vicious circle: They cannot attract or retain the teachers that could help mitigate the effects of social inequality, making these schools stay unattractive for experienced and skilled teachers. The opposite might be true for schools with positive traits, and both tendencies may further exacerbate the inequality between schools.

In this project, I therefore aim to study two overall questions:

1) How are teachers sorted across schools in terms of teacher quality?

2) How are school traits associated with the probability of attracting and retaining HQ teachers?

3) What is the causal effect of teacher sorting on pupil attainment and labor market outcomes?

2 Planning phase

During the planning phase of my Ph.D. project, although I didn't complete a full pre-registration, I created two documents with less precise data collection and analysis plans. My original grant proposal outlined expected findings, albeit not as precisely as in a pre-registration. It also detailed statistical methods, planned outcomes, predictors, and covariates for each research question. Additionally, I provided a comprehensive description of variables for analysis in the application for register data to Statistics Denmark, including criteria for defining public school teachers, students, and primary schools, as well as specifying variables for analysis.

3 Data collection and access

Data collection has been done by external organizations, primarily DST.

In addition to data from DST, I also plan to use survey data collected by the Programme for International Student Assessment(PISA) and data collected by the Danish National Agency for IT and Learning(NAIL). Data are freely available via https://www.oecd.org/pisa/data/ and https://data.stil.dk/instreghistorik/

4 Data management

Data from DST is stored on the central servers of DST and is only accessible through remote desktop. Further, the raw data supplied by DST are read-only and can thus not be manipulated without administrative access. DST also handles the backup of data.

Data from PISA and NAIL are stored locally, but the original versions of the data can be downloaded from both PISA and NAIL.

Data stored locally will be stored on the central servers of VIVE, on which backups are routinely performed. Further, I will also store backups of data on an external harddrive.

In this project, I will, unless avoidable, use open source data formats such as .csv or .feather/SQL database for storing large files.

5 Data processing and data analysis

In this project i will use open-source software, primarily R, to process and analyze data. Raw data will stay read-only, with minimal processed data saved to ease computational load.

To enhance reproducibility, I'll use three tools:

Version control: All project work, including data processing code and research documents, will be on GitHub unless it violates ethical or data protection rules.

Containers: I'll employ containers like Docker for individual paper projects to maintain a consistent computing environment and ensure reproducibility.

Rmarkdown: Rmarkdown facilitates integrating text and code, promoting version control and general reproducibility.

6 Reporting and open access

In this project i will share all code used to analyze data and produce results. All code will be shared on a GitHub repository, which also makes it possible to view the version history of all code, further adding transparency.

In this project, I aim to publish pre-prints of individual papers via the Open Science Framework or arXiv. This will increase transparency, as it will be possible to track differences between versions of submitted papers before publication and those deemed publishable. Further, publishing pre-prints helps to mitigate the "file-drawer" problem(Rosenthal 1979), i.e., the phenomenon that some results are deemed more "publishable" than others by academic journals, regardless of their validity.

Lastly, I aim to publish in journals that offer open access to published papers.

7 limitations of administrative data

In this project, I heavily rely on administrative data from DST, which presents challenges for data sharing due to its sensitive nature and legal constraints. While I can share my methodology and code to explain my conclusions, fully replicating my results outside the DST server's project folder is very challenging. This lack of data sharing significantly reduces reproducibility.

Even though the data theoretically can be obtained anyone, so long as they meet DST's requirements, the process is often tedious, requiring institutional approval and data certification. For institutions outside the EU, GDPR regulations can further complicate access.

Bilbliography

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–41. https://doi.org/10.1037/0033-2909.86.3.638.