README

This repository contains the necessary STATA code to replicate the paper “High school majors and future earnings” by Gordon Dahl, Dan-Olof Rooth and Anders Stenberg.

The Stata program “main.do” calls all of the required programs. The main.do file provides an explanation of what each of the programs do.

Step 1-5 generates data and requires about 4 hrs. Step 6-9 generates all figures, all tables and all digits mentioned in the paper and requires about 45 mins. Step 10-13 are table A9 GMM tests which require several hours.

We used Stata version 16 on Statistics Sweden’s servers for all of the data analysis. Researchers who want to replicate the analysis would need to perform the analysis on these servers; information on how to apply for access is detailed below. The analysis makes use of the following Stata packages: estout, diff, groups, rd, nrow, seq, mediation, missing, levels, qqvalue.

The repository only contains Stata program and not the underlying data as these are confidential register data. Information on how to access this data is provided in the Data Availability statement.

Data Availability

The information used in the analysis combines several Swedish administrative registers (as described in the paper). The data use is subject to the European Union’s General Data Protection Regulation(GDPR) from May 2018. The data are physically stored on computers at Statistics Sweden and, due to security considerations, the data may not be transferred to computers outside Statistics Sweden.

Researchers interested in obtaining access to the register data employed in this paper are required to submit a written proposal to gain approval from Statistics Sweden. The proposal must include a detailed description of the proposed project, its purpose, and its social contribution, as well as a description of the required datasets, variables, and analysis population. Applications can be submitted by researchers who are affiliated with Swedish institutions accepted by Statistics Sweden, or by researchers outside of Sweden who collaborate with researchers affiliated with these institutions.

MONA (Microdata Online Access) is Statistics Sweden’s platform for access to microdata. In MONA, users can process data online without the microdata ever leaving Statistics Sweden, including the specific administrative databases used in the current paper such as the LISA database containing earnings or administrative registers on applications to upper secondary school 1977-1991.

The application process is not standardized. A short request is typically attached to the project, for example stating: “The underlying population of the project is collected from the registers of the total population 1990-2014, consisting of all individuals registered as residents in Sweden above age 15. The current project aims to compare labor market outcomes of individuals with different fields of study, information is therefore collected from [a list of] various registers at Statistics Sweden.”

https://www.scb.se/en/services/ordering-data-and-statistics/ordering-microdata/mona--statistics-swedens-platform-for-access-to-microdata/

The following datasets have been used in this paper:

Statistics Sweden (2021a). Sökande/antagna till gymnasiet 1977-1991 [database]. Statistics Sweden (accessed 2018).

Statistics Sweden (2021b). Ak9-elevregistret 1988-1991 [database]. Statistics Sweden (accessed 2018).

Statistics Sweden (2021c). Income and taxation registers 1978-1989 [database]. Statistics Sweden (accessed 2018).

Statistics Sweden (2021d). Longitudinal integrated database for health insurance and labour market studies (LISA) 1990-2014 [database]. Statistics Sweden (accessed 2018).

Statistics Sweden (2021e). Multiple generation registers (*Flergenerationsregistret*) [database]. Statistics Sweden (accessed 2018).

Statistics Sweden (2021f). Population and housing census 1970, 1980 and 1985 [database]. Statistics Sweden (accessed 2018).