README
Replication Archive for
"The Value of a High School GPA"

Overview

This document describes the replication files that reproduce all the figures and tables in "The Value of a High School GPA". The document also provides information on data availability and the list of all datasets and variables used in the paper, followed by information on directory structure, computational requirements, and a description of all the codes used.

No table or figure can be reproduced without access to the confidential data.

Data Availability and Provenance Statements

This paper uses restricted access and public access datasets. Each of the data sources is described in further detail below, along with a detailed description of how to access the source data.

Summary of Availability

Some data cannot be made publicly available.

Details on each Data Source

1. Restricted Access Datasets from Statistics Norway

We use several restricted access administrative registries obtained from Statistics Norway. Researchers can gain access to these registries by submitting a written application to Statistics Norway. The application should include a detailed research proposal, a detailed list of datasets and variables, and the selection criteria to be used. Once received, applications must be certified to ensure that the data are processed in a manner that protects the personal integrity of all individuals included in the data. Conditional on this approval, Statistics Norway determines data access. Inquiries about access should be addressed to: Statistics Norway Postboks 2633 St. Hanshaugen, NO-0131 Oslo.

The administrative registries and variables we use in the study are the following:

- Statistics Norway's education registries (Statistisk Sentralbyrå, 2020*d*).
  - f_kurs_1974-2018: this dataset includes all courses that have been ongoing or completed in educational institutions in Norway, as well as courses completed abroad by students with grants from Norway's Loan Fund for Education.

Variable	Description
npid	Unique individual identifier, allowing matching across data sources from Statistics Norway.
lopenr_kurs	Course identifier.
tilgdato	Start date of the course.
hoved	Type of education (e.g., compulsory, university or college education).
kltrinn2000	Grade level.
kltrinn2000dato	Start date of the grade level.
nus2000	Classification of education (NUS).
uhgruppe	Grouping of higher education programs.
komp	Indicates whether the course provides academic or professional qualification.
heldel	Full-time or part-time course.
elevstatus	Student status.
skole_foretak	School identifier.
hskode	University identifier.
hskodedato	Start date of the university course.
skolekom	Municipality of the school.
eierf	Private or public school.
grunnskolepoeng	Middle school GPA.

- tab_karakter_vgskole_2003-2018: high school grades for each subject, all grade levels.

Variable	Description
npid	Unique individual identifier, allowing matching across data sources from Statistics Norway.
skolear	School year.
skule	School identifier.
videregaendepoeng	High school GPA, used for admission to the next grade level.
fagkode	Subject identifier.
stp	Teacher assessment score.
skr	Written exam grade.
mun	Oral exam grade.
fagstatus	Status of the subject (e.g., subject taken as a regular student or privately).

- fagkatalog_2020: information on high school subjects.

Variable	Description
fagkode	Subject identifier.
fagnavn	Full name of the subject.
fagtype	Type of subject (e.g., core curricula subject, elective subject).
aartimer	Number of annual instructional hours for the subject.

  - f_demo_1974-2019: this data includes all individuals who have been residents in Norway from 1970 onwards, and it reports their completed education.

Variable	Description
npid	Unique individual identifier, allowing matching across data sources from Statistics Norway.
lopenr_kurs	Course identifier.
bu_regdato	Date of completion.
bu_kltrinn	Class level.
bu	Classification of education (NUS).

- Statistics Norway's population registries (Statistisk Sentralbyrå, 2020*a*).
  - slekt_1934-2019: for all Norwegian residents, this data provides information on individuals' year and month of birth, as well as the unique identifiers of their parents.

Variable	Description
npid	Unique individual identifier, allowing matching across data sources from Statistics Norway.
bmonth	Year and month of birth.
bsex	Sex.
nmpid	Unique individual identifier of the mother.
nfpid	Unique individual identifier of the father.

- Statistics Norway's tax records (Statistisk Sentralbyrå, 2020*b*).
  - inntekt_1993-2018: for each year between 1993 and 2018, this data reports the incomes received during the calendar year for all individuals who were tax residents of Norway as of December 31st.

Variable	Description
npid	Unique individual identifier, allowing matching across data sources from Statistics Norway.
aargang	Year of observation.
wyrkinnt	Labor income: it is the sum of employee income and self-employment income.

- Statistics Norway's employment registries (Statistisk Sentralbyrå, 2020*c*).
  - regsys_2000-2019: for each year between 2000 and 2019, this dataset reports information on the main employment in November of Norwegian residents aged between 16 and 74.

Variable	Description
npid	Unique individual identifier, allowing matching across data sources from Statistics Norway.
aar	Year of observation.
pers_alder	Age.
arb_arbmark_status	Employment status (e.g., outside of the workforce, employee, self-employed).

2. Public Access Datasets

Additionally, we use one public access dataset. A copy of this dataset is in the folder "Data/Public_access". Below, we provide information on the source and content of this publicly available dataset.

- inflation: for each year between 1930 and 2018, this data reports the consumer price index from Statistics Norway with 2015 as the base year. We collected this data on the website: https://www.ssb.no/en/priser-og-prisindekser/konsumpriser/statistikk/konsumprisindeksen (Table 1), last accessed on August 24$^{\text{th}}$, 2023.

Dataset list

Overview of Datasets

Dataset	Description	Source	Notes	Provided
f_kurs_1974-2018	Education registry	Statistics Norway	Confidential	No
tab_karakter_vgskole_2003-2018	Education registry	Statistics Norway	Confidential	No
fagkatalog_2020	Education registry	Statistics Norway	Confidential	No
f_demo_1974-2019	Education registry	Statistics Norway	Confidential	No
slekt_1934-2019	Population registry	Statistics Norway	Confidential	No
inntekt_1993-2018	Tax records	Statistics Norway	Confidential	No
regsys_2000-2019	Employment registry	Statistics Norway	Confidential	No
inflation	Consumer price index	Statistics Norway	Public	Yes

Computational requirements

The code was last run on a server with two AMD EPYC 7452 32-Core Processors, totaling 64 cores and 128 threads, on an x86_64 architecture. The server operates on Red Hat Enterprise Linux 8.7 (Ootpa).

The code was run with Stata/MP (version 16.1).

Replicators should expect the code for data preparation to finish in about 1 hour and 40 minutes, and the code generating the tables and figures of the paper to finish in about 45 minutes.

Software Requirements

The code was run with Stata/MP (version 16.1), with the following packages:

- egenmore
- keeporder
- grstyle
- colorpalette
- winsor2
- texsave
- esttab
- ivolsdec
- pdslasso
- ritest
- boottest

Description of programs/code

Our project folder for replication has the following structure:

Folder	Content
Restat_replication/Programs/	All codes to replicate the figures and tables of the paper.
Restat_replication/Programs/Data_construction/	Codes to prepare the final datasets used in the study.
Restat_replication/Programs/Graphs/	Codes to generate the figures of the paper.
Restat_replication/Programs/Tables/	Codes to generate the tables of the paper.

Folder	Content
Restat_replication/Data/Public_access/	Public access datasets listed in the previous section.
Restat_replication/Outputs/	Figures and tables produced by the programs of the replication package.
Restat_replication/Log/	Logs of the programs.
/data/prosjekt/.alfa/	All confidential datasets are stored in a separate folder on a secure server. This folder is not included in the replication package. Researchers willing to replicate the results of the study should submit a written application to Statistics Norway to gain access to the confidential data and variables listed in the previous section. Once access has been granted, they should store the data in a similar folder before running the codes provided in the replication package.

Description of codes:

- The program "Restat_replication/Programs/master.do" calls the following do files in order to prepare the data and generate all the figures and tables in the paper:
  - "Restat_replication/Programs/Data_construction/data1.do" to "Restat_replication/Programs/Data_construction/data5.do": prepares the data.
  - "Restat_replication/Programs/Graphs/distributions_correlations.do": generates Figure A1.
  - "Restat_replication/Programs/Graphs/monotonicity.do": generates Figure A2.
  - "Restat_replication/Programs/Tables/BC.do": generates Table A2.
  - "Restat_replication/Programs/Tables/main.do": generates Tables 1, 3, A1, A4, A5.
  - "Restat_replication/Programs/Tables/heterogeneity.do": generates Tables 2, A6, A7.
  - "Restat_replication/Programs/Tables/monotonicity.do": generates Table A9.

– "Restat_replication/Programs/Tables/robustness.do": generates Table A3.

– "Restat_replication/Programs/Tables/IV_OLS_decomposition.do": generates Table A8.

Instructions to Replicators

Download the folder Restat_replication. Maintain the folder structure and folder names.

Edit Restat_replication/Programs/master.do to adjust the default paths (in particular, adjust the path to the restricted access data once access to has been granted).

Once access to the restricted datasets has been granted, run Restat_replication/Programs/master.do to recreate the figures in the paper. Without access to the restricted access datasets, no figures or tables can be reproduced.

References

Statistisk Sentralbyrå. 2020*a*. "Data om folkemengden eller familier og husholdninger, innvandring, og andre demografiske opplysninger." https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/befolkning, (last accessed on December 15$^{\text{th}}$, 2023).

Statistisk Sentralbyrå. 2020*b*. "Data om inntekter, formue, barnetrygd eller pensjoner." https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/inntekt, (last accessed on December 15$^{\text{th}}$, 2023).

Statistisk Sentralbyrå. 2020*c*. "Data om personers tilknytning til arbeidsmarkedet, slik som for eksempel næring, arbeidssted, arbeidstid, yrke eller arbeidsledighet." https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/arbeidsmarked, (last accessed on December 15$^{\text{th}}$, 2023).

Statistisk Sentralbyrå. 2020*d*. "Data om personers utdanningsnivå og utdanningsretning, karakterer og resultater." https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/utdanning, (last accessed on December 15$^{\text{th}}$, 2023).