Replication Data and Code for "Empirical Decomposition of the IV–OLS Gap with Heterogeneous and Nonlinear Effects"

Shoya Ishimaru*

Overview

This document describes the data and the programs for replicating the empirical results in the paper. The analyses of the data are performed on Stata. Figures are produced using R.

College Cost Instrument with Geographic Variation (Section 5.1)

The instructions for replicating the results in Section 5.1 of the paper are as follows. See also Online Appendix D.1 for details.

Data Sources

The paper uses the data from the National Longitudinal Survey of Youth of 1979 (NLSY79). The main part of the data is publicly available from the Bureau of Labor Statistics (BLS). The construction of some geographic variables, however, requires access to the restricted-use geocode data. The restricted-use data cannot be posted here, but the programs to process the data are provided. Information about colleges is from the Higher Education General Information Survey (HEGIS) data compiled by Kling (2001), which is available from the data archive of the paper at the *Journal of Business and Economics Statistics* website. Local earnings data are from the county-level economic accounts by the Bureau of Economic Analysis. State unemployment rates are from the BLS for the years 1976–81 and from the 1977 Statistical Abstract of the United States for the years 1974–75. The definition of urbanized areas, which is used to construct urban status variable, is from the U.S. Census Bureau. CPI–U for inflation adjustment is from the Federal Reserve Bank of Minneapolis.

Data and Script Files for Replication

The following structured list explains data and script files under the folder "College_IV" in this replication package. The numbers in the list indicate the order in which script files should be executed.

- public (for cleaning of public-use data of the NLSY79)

*Hitotsubashi University, Department of Economics (email: shoya.ishimaru@r.hit-u.ac.jp).

  - data (stores .dta files downloaded from the NLS Investigator and NLSY79 tag files)
  - code
    - clean (stores scripts to be called by clean.do)
    1. clean.do (the master script for cleaning the data)
  - processed (stores .dta files processed by clean.do)
  - export
    - basic_variables.dta (the final product of clean.do; should be manually copied to restricted/import)

- restricted (for the main analysis using geocode data)
  - import
    - basic_variables.dta (a copy imported from public/export)
    - cpi.dta (CPI data)
    - geo_info.dta (a copy imported from geographic/export)
  - code
    2. clean.do (cleans geocode data)
    3. merge.do (puts everything together)
    4. descriptive.do (computes descriptive stats for Table A.4)
    4. decom_base.do (decomposes the IV–OLS gap for Table 1)
    4. weight.do (computes covariate and treatment-level weights for Table 2 and Figure 1)
    4. first_stage.do (computes first-stage coefficients for Table A.5)
    4. decom_alt.do (decomposes the IV–OLS gap for Table A.6)
    4. weight_alt.do (computes covariate and treatment-level weights for Table A.7)
    5. graph.r (creates a line plot for Figure 1)
  - processed (an empty folder to store processed files)
  - result (stores results)

- geographic (for cleaning county data)
  - data
    - BEA_income.dta, BEA_VA.dta (local earnings data)

    - county_coordinates.dta (longitudes and latitudes of counties for computing distance)[1]
    - county_ua.dta (urbanized area definition from the 1990 Census)
    - county1980.dta (list of counties in the 1980 Census)
    - state_unemployment.dta (state unemployment data)
  - code
    1. county.do (the master script for constructing the county data)
  - export
    - geo_info.dta (the final product of county.do; should be manually copied to restricted/import)
  - HEGIS (stores files to compile HEGIS data)[2]

RD-Based Compulsory Schooling Instrument (Section 5.2)

The instructions for replicating the results in Section 5.2 of the paper are as follows. See also Online Appendix D.2 for details.

Data Sources

The paper uses the data from the British General Household Surveys (GHS). The raw data sets are available through the U.K. data archive (https://www.data-archive.ac.uk/). To ensure comparability of the regression analyses, I use the compiled data sets provided by the replication package for Oreopoulos (2006), which reflect the corrections made in Oreopoulos (2008).

Data and Script Files for Replication

The following structured list explains data and script files under the folder "CSL_RD" in this replication package. The numbers in the list indicate the order in which script files should be executed.

- data

[1]Downloaded from Geocorr Applications: https://mcdc.missouri.edu/applications/geocorr.html

[2]These files are originally from the data archive of Kling (2001). I am grateful to Jeffrey R. Kling for generously granting me permission to redistribute them.

  - combined general household survey.dta (GHS data file complied by Oreopoulos, 2006)[3]

- code

  1. clean.do (clean the data and aggregate them by cohort, year, and schooling level)
  2. decom_base.do (decomposes the IV–OLS gap for Table 1)
  2. weight.do (computes covariate and treatment-level weights for Table 3 and Figure 1)
  2. appendix.do (the master script for performing all additional analyses in Appendix C.2)
  3. graph.r (creates line plots for Figure 1 and Figure A.1)

  - appendix (stores scripts to be called by appendix.do)
    * decom_alt.do (for the second row of Table A.9)
    * decom_NI.do (for the third row of Table A.9)
    * decom_NI_alt.do (for the fourth row of Table A.9)
    * weight_NI.do (computes covariate and treatment-level weights for Table A.10 and Figure A.1)

- processed (stores the data cleaned by clean.do)

- result (stores results)

Compulsory Schooling Instrument with DID Variation (Section 5.3)

The instructions for replicating the results in Section 5.3 of the paper are as follows. See also Online Appendix D.3 for details.

Data Sources

The paper use the 1960–80 U.S. Census data compiled by IPUMS (Ruggles et al., 2020). The compulsory schooling laws (CSL) data are from the data archive of Acemoglu and Angrist (2000). Some analyses in Online Appendix D.3 use the CSL data from Stephens and Yang (2014).

[3]This file is originally from the data archive of Oreopoulos (2006). I redistribute it under Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/), following the license statement in the archive.

Data and Script Files for Replication

The following structured list explains data and script files under the folder "CSL_DID" in this replication package. The numbers in the list indicate the order in which script files should be executed.

- data
  - CSL.dta (the CSL data file from Acemoglu and Angrist, 2000)[4]
  - schooling_laws_sy.dta (the CSL data file from Stephens and Yang, 2014)[5]
  - region_division_state.dta (a file that associates fips state codes with census regions)
  - Census.dta (the census data file)
- code
  1. clean.do (clean the data and aggregate them by cohort, year, and schooling level)
  2. decom_base.do (decomposes the IV–OLS gap for Table 1)
  2. weight.do (computes covariate and treatment-level weights for Table 4 and Figure 1)
  2. appendix.do (the master script for performing all additional analyses in Appendix C.3)
  3. graph.r (creates line plots for Figure 1 and Figure A.2)
  - appendix (stores scripts to be called by appendix.do)
    - first_stage.do (for Table A.12)
    - decom_ca.do (for the second row in Table A.13)
    - decom_rs.do (for the third row in Table A.13)
    - decom_qob.do (for the fourth row in Table A.13)
    - weight_ca.do (for the "CA IV" column in Table A.14)
    - weight_rs.do (for the "RS IV" column in Table A.14)
    - weight_qob.do (for the "QOB IV" column in Table A.14)

[4]This file is originally from the data archive (https://doi.org/10.7910/DVN/ZKEY9J) of Acemoglu and Angrist (2000), which is under CC0 Public Domain Dedication. I converted it into a dta format for Stata.

[5]This file is originally from the data archive of Stephens and Yang (2014). I redistribute it under Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/), following the license statement in the archive.

- processed (stores the data cleaned by clean.do)
- result (stores results)

Monte Carlo Analysis (Online Appendix C.6)

The folder "Monte_Carlo" provides the simulation code for the Monte Carlo analysis in Appendix C.6. The following structured list explains data and script files. The numbers in the list indicate the order in which script files should be executed.

- input
  - parameter.csv (specifies the parameters of the simulated model)
- code
  1. MCmain.do (runs the simulation)
  2. summary.do (summarizes the results)
- results (stores the results generated by MCmain.do)

References

Acemoglu, D., Angrist, J. 2000. How large are human-capital externalities? Evidence from compulsory schooling laws. NBER Macroeconomics Annual. 15, 9–59.

Kling, J. R. 2001. Interpreting instrumental variables estimates of the returns to schooling. Journal of Business & Economic Statistics. 19 (3), 358–364.

Oreopoulos, P. 2006. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. American Economic Review. 96 (1), 152–175.

Oreopoulos, P. 2008. Estimating average and local average treatment effects of education when compulsory schooling laws really matter: Corrigendum. American Economic Review.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., Sobek, M. 2020. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS.

Stephens, J., Melvin, Yang, D.-Y. 2014. Compulsory education and the benefits of schooling. American Economic Review. 104 (6), 1777–92.