README: "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major"

Data Availability Statements

Many of the data used in Bleemer and Mehta (2020a) are confidential, but may be obtained with Data Use Agreements with the Office of the Registrar of the University of California, Santa Cruz (UC-CHP, 2020), the Student Experience in the Research University Consortium (SERU, 2019), the National Student Clearinghouse (NSC, 2019), and the California Employment Development Department (EDD, 2019). Researchers interested in access to the data are strongly suggested to contact the authors (bleemer@berkeley.edu) and/or the University of California, Berkeley's Center for Studies in Higher Education (cshe@berkeley.edu), as CSHE's own Data Use Agreements permit data redistribution in some cases. Otherwise, researchers should contact:

1. The Registrar of the University of California, Santa Cruz, who is currently Tchad Sanger: cpsanger@ucsc.edu. Note that data are only available to researchers affiliated with UC Santa Cruz.
2. The Center for Studies in Higher Education at UC Berkeley (https://cshe.berkeley.edu/seru/about-seru/seru-data-guidelines), which manages data collection for the SERU Consortium. Contact Gregg Thomson: gthomson@berkeley.edu.
3. The National Student Clearinghouse Research Center (https://nscresearchcenter.org/contactus/). Note that data are only available to researchers affiliated with UC Santa Cruz and have an associated fee.
4. The California Employment Development Department's Labor Market Information Customized Data Services (https://www.labormarketinfo.edd.ca.gov/resources/lmi-custom-data-services.html), who can be reached at 916-262-2162. Note that data are only available to researchers affiliated with the University of California and have an associated fee.

It can take months to negotiate data use agreements and gain access to the data.

All publicly available data used in Bleemer and Mehta (2020a) have been deposited in the AEA Data and Code Repository openicpsr-126941 (Bleemer and Mehta, 2020b). These include a subset of data from IPUMS USA (Ruggles *et al*, 2020). IPUMS USA does not currently provide the ability to store or reference custom extracts, but allows for redistribution for the purpose of replication (see https://ipums.org/about/terms). The archive contains the extracted data and codebook. The data citation has the full URL.

Datafile and Codebook: `Data/Raw/usa_00075.dat and Data/Raw/usa_00075.xml`

They also include data from:

1. The U.S. Bureau of Labor Statistics (BLS, 2018) All Urban Consumers Consumer Price Index; see https://www.bls.gov/cpi/data.htm.

   Datafiles: `Data/Raw/CPI.csv` and `Data/Raw/CPI_California.csv`

2. The Individual Income Tax Statistics - ZIP Code Data from the Internal Revenue Service's SOI Tax Stats (IRS, 2018); see https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi.

   Datafiles: `Data/Raw/Zipcode_Income/ZIP_Code_YYYY.dta`, for YYYY from 1998 to 2017.

3. The U.S. Census Bureau (Census, 2019) 2017 NAICS Structure; see https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Structure.xlsx

   Datafile: `Data/Raw/NAICS_Codes.xlsx`

Finally, we also include crosswalks of our own creation from state name to FIPS code, two-digit NAICS codes to industries, and University of California, Santa Cruz major codes to American Community Survey major codes (Bleemer and Mehta, 2020b). The first two crosswalks are based on information from the U.S. Census (Census, 2019); the last was generated at the discretion of the authors, and is reported in Table A-7 of Bleemer and Mehta (2020a).

Datafiles: `Data/Raw/State_to_FIPS.csv`, `Data/Raw/NAICS2_to_Names.csv`, and `Data/Raw/UCSC_Majors_to_ACS.csv`

Dataset list

1. The following datasets were provided by the University of California, Santa Cruz Office of the Registrar. They are confidential and are not provided in the data repository.
   (a) `UC-CHP File 1 - Student Info prior to F99.xlsx`
   (b) `UC-CHP File 1 - Student Info F99-F11.xlsx`
   (c) `UC-CHP File 1 - Student Info W12-S19.xlsx`
   (d) `UC-CHP File 2 - Educational Info prior to F99.xlsx`
   (e) `UC-CHP File 2 - Educational Info F99-F11.xlsx`
   (f) `UC-CHP File 2 - Educational Info W12-S19.xlsx`
   (g) `UC-CHP File 4 - Courses prior to F99 - Revised.xlsx`
   (h) `UC-CHP File 4 - Courses W00-F03 - Revised.xlsx`
   (i) `UC-CHP File 4A - Courses_Primary Instuctors F99-F11.csv`

(j) `UC-CHP File 4A - Courses_Primary Instructors W12-S19.xlsx`

(k) `UC-CHP File 4B - Courses_Section Instructors F99-F11.xlsx`

(l) `UC-CHP File 4B - Courses_Section Instructors W12-S19.xlsx`

2. The following datasets were provided by the Student Experience in the Research University organization at the Center for Studies in Higher Education at the University of California, Berkeley. They are confidential and are not provided in the data repository.

   (a) `SERU_ID_UCSC.csv`

   (b) `SERU_Data_UCSC.csv`

3. The following datasets were provided by the National Student Clearinghouse. They are confidential and are not provided in the data repository.

   (a) `NSC_Data_UCSC.csv`

4. The following datasets were provided by the California Employment Development Department. They are confidential and are not provided in the data repository.

   (a) `EDD_Data_UCSC.csv`

5. The following datasets were obtained from IPUMS. They are provided in the data repository.

   (a) `Data/Raw/usa_00075.dat`

   (b) `Data/Raw/usa_00075.xml`

6. The following datasets were derived from the datasets obtained from IPUMS (see `0_Initiation_Programs.R`). They are provided in the data repository.

   (a) `Data/Derived/ACS_Data_UCSCEcon.Rda`

   (b) `Data/Derived/ACS_Data_UCSCEcon_NAICS.Rda`

   (c) `Data/Derived/ACS_Data_UCSCEcon_NAICS_AllYears.Rda`

   (d) `Data/Derived/ACS_Industries_INCWAGE.Rda`

   (e) `Data/Derived/ACS_Majors_INCWAGE.Rda`

7. The following datasets were obtained from the Bureau of Labor Statistics. They are provided in the data repository.

   (a) `Data/Raw/CPI.csv`

   (b) `Data/Raw/CPI_California.csv`

8. The following datasets were obtained from the IRS. They are provided in the data repository.

(a) `Data/Raw/ZIP_Code_YYYY.dta`, where YYYY is 1998, 2001, 2004, 2005, 2006, 2007, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, and 2017.

9. The following datasets were obtained from the U.S. Census Bureau. They are provided in the data repository.
   (a) `Data/Raw/NAICS_Codes.xlsx`
10. The following datasets were created by the authors. They are provided in the data repository.
    (a) `Data/Raw/State_to_FIPS.csv`
    (b) `Data/Raw/NAICS2_to_Names.csv`
    (c) `Data/Raw/UCSC_Majors_to_ACS.csv`

Computational requirements

Software Requirements

- R 3.5.1
  - `readstata13` (0.9.2)
  - `rdd` (0.57)
  - `lfe` (2.8-5)
  - `glmnet` (4.0)
  - `ipumsr` (0.4.4)
  - `spatstat` (1.64-1)
  - `colorspace` (1.4-1)
  - `RDHonest` (0.3.2)
  - `plotrix` (3.7-8)
  - `plyr` (1.8.6)
  - `dplyr` (1.0.0)
  - `readxl` (1.3.1)
  - `readr` (1.3.1)
  - The file "`0_Initiation_Programs.R`" will install all dependencies (latest version), and should be run once prior to running other programs.

Description of programs

- The program `0_Initiation_Programs.R` will install and load all required packages, load requisite functions, and load functions the categorize majors into disciplines.

- The program `1_Construct_Data.R` has four sections, each respectively constructing and cleaning the UCSC Registrar data, NSC data, EDD data, and IPUMS data. This file also constructs Figure A-18(a).
- The program `2_Clean_Data.R` further cleans the UCSC Registrar data and merges the respective data sources into a single student database for analysis. The file also constructs Figure A-2.
- The program `3_Analyze_Data.R` conducts all analysis presented in Bleemer and Mehta (2020a), including the production of all other figures and tables in the paper.

Memory and Runtime Requirements

The code was last run on a 12-core Intel server with Windows Version 10.0.14393. Computation took less than 12 hours.

Instructions

- Download public programs and data files and place them in a directory, adding an additional folder called `Figures`. Place secure data in a separate directory. Update these two directories in the first two lines of `0_Initiation_Programs.R`.
- Run programs in order: 0, 1, 2, and then 3. All tables and figures will be saved in the `Figures` directory.

List of tables and programs

Figure or Table #	Program	Line #	Output file	Conf. Data?
Table 1	3	36-51	SummaryStats.csv	Y
Figure 1	3	79-248	RD_Economics.png	Y
Figure 2	3	79-248	RD_wage_sum_1718.png	Y
Figure 3	3	477-500	RDIV_TimeTrend.png	Y
Figure 4(a)	3	79-248	RD_Grad_IncNSC.png	Y
Figure 4(b)	3	79-248	RD_GradSch.png	Y
Figure 4(c)	3	79-248	RD_GPA_Overall_FE.png	Y
Figure 5(a)	3	79-248	RD_Intend_Bus_SJ_Outliers.png	Y
Figure 5(b)	3	79-248	RD_FIREAcc.png	Y
Figure 5(c)	3	79-248	RD_Industry_Mean1718.png	Y
Figure 6	3	256-445	CounterfactualMajorFigure.png	Y

References

Bleemer, Zachary and Aashish Mehta. 2020a. "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major". Manuscript.

Bleemer, Zachary and Aashish Mehta. 2020b. "Replication Data for: Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major." *American Economic Association [publisher], Inter-university Consortium for Political and Social Research[distributor].* https://doi.org/10.3886/E126941V1.

United States Bureau of Labor Statistics (BLS). 2019. "All Urban Consumers (Current Series) Consumer Price Index - CPI [dataset]" *Bureau of Labor Statistics.* https://www.bls.gov/cpi/data.htm. Accessed December 2019.

California Employment Development Department (EDD). 2019. "California Labor Market Information Customized Data [dataset]" *State of California.* Accessed April 2019.

United States Census Bureau (Census). 2019. "2017 NAICS Structure with Change Indicator [dataset]" *U.S. Census Bureau.* https://www.census.gov/eos/www/naics/downloadables/downloadables.html. Accessed December 2019.

Internal Revenue Service (IRS). 2018. "SOI Tax Stats - Individual Income Tax Statistics - ZIP Code Data (SOI)" *Internal Revenue Service.* https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi. Accessed January 2018.

National Student Clearinghouse (NSC). 2019. "StudentTracker Database [dataset]" *National Student Clearinghouse.* Accessed January 2019.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. "Integrated Public Use Microdata Series (IPUMS) USA: Version 10.0 [dataset]". Minneapolis, MN: *Minnesota Population Center, IPUMS.* https://doi.org/10.18128/D010.V10.0. Accessed March 2020.

Student Experience in the Research University (SERU). 2019. "University of California, Santa Cruz UCUES Survey Responses [dataset]," *University of California.* Accessed December 2019.

University of California ClioMetric History Project (UC-CHP). 2020. "University of California, Santa Cruz Campus Transcript Database [dataset]," *University of California ClioMetric History Project.* Accessed February 2020.

Acknowledgements

This README file is based on the Social Science Data Editors' template; see https://social-science-data-editors.github.io/guidance/template-

README.html. Some content on this page was copied from Hindawi.