Social Capital I and II: Replication Guide

August 1, 2022

Contents

1 Overview

This replication package contains code to reproduce the results reported in Chetty et al. (2022a, b) that can be constructed using the aggregate statistics on social capital that we release publicly.

We use methods from the differential privacy literature to add noise to our aggregate social capital measures to protect privacy while maintaining a high level of statistical reliability. Estimates obtained from the publicly released data will not exactly match those reported in the published papers because of the addition of this noise and the exclusion of small cells (e.g., those with fewer than 100 low-SES and 100 high-SES Facebook users for economic connectedness measures). In practice, because the estimates we report in the papers are all population-weighted, the point estimates are very similar, but the number of observations will differ. Please see Chetty et al. (2022a), Chetty et al. (2022b), and `social_capital_codebook.pdf` for further details.

To protect users' privacy, we also only release a subset of social capital measures and population weights. As a result, only a subset of the tables and figures in the papers can be replicated using the publicly available data. In addition, some multivariate analyses in this replication package use a subset of variables used in the papers, and are thus not directly comparable.

2 Replication Instructions

1. Software requirements: Stata 17.0 and R 4.2.1. Before running the code in the replication package, please install the `readstata13` package in R using the following command: `install.packages("readstata13")` in R.

2. In the replication package, go to `Code/Paper 1` to replicate the tables and figures for Chetty et al. (2022a) and `Code/Paper 2` to replicate the tables and figures for Chetty et al. (2022b).

3. In each code subfolder, change the global `master_folder` in the respective metafile to the location of this README file on your computer.

4. In each code subfolder, run all code from their respective metafile. Section 1 of each metafile downloads the social capital data from the Humanitarian Data Exchange (HDX) data repository. The output for Chetty et al. (2022a) is stored in `Paper 1 Exhibits` and the output for Chetty et al. (2022b) is stored in `Paper 2 Exhibits`.

When executing the replication code for the first time, it may take a few minutes to install the necessary external packages. To speed up the replication code in subsequent runs, comment out the code block in each metafile that installs external packages.

3 Codebook for Publicly Available Variables

Please see `social_capital_codebook.pdf` (also available here as `README.pdf`) for a codebook describing the social capital measures constructed and released in this project. This section presents brief descriptions of other publicly available data (e.g., statistics from the Census and ACS) that are used to produce the results in the papers.

3.1 County-Level Data

`county_covariates.dta`

Variable Name	Description
county	5-digit county FIPS code.
cz	Commuting zone (1990 definition).
czname	Commuting zone name (1990 definition).
income_segregation	Income segregation defined using a Theil (entropy) index from Reardon and Bischoff (2011) using data from the ACS. See Supplementary Information A.5.1. of Chetty et al. (2022a) for details.
racial_segregation	Racial segregation constructed using a Theil H index using data from the American Community Survey (ACS). See Supplementary Information A.5.1. of Chetty et al. (2022a) for details.
pop2018	Population in 2018 from the ACS.
causal_p25_cz_cty_kr26	Raw estimate of the causal effect of growing up for an additional year in the county on earnings in adulthood, as constructed by Chetty and Hendren (2018).
causal_p25_cz_cty_kr26_se	Standard error of the raw estimate of the causal effect of growing up for an additional year in the county on earnings in adulthood, as constructed by Chetty and Hendren (2018).
teenbrth_pooled_female_p25	Teen birth rate for women with parents at 25th income percentile, from Chetty et al. (2018).
hs_pooled_pooled_p25	High school completion rate for children with parents at 25th income percentile, from Chetty et al. (2018).
kfr_pooled_pooled_p25	Mean income percentile in adulthood of a child born to parents at the 25th percentile of the income distribution, from Chetty et al. (2018).
kfr_black_pooled_p25	Mean income percentile in adulthood of a child born to parents at the 25th percentile of the income distribution for Black children, from Chetty et al. (2018).
kfr_white_pooled_p25	Mean income percentile in adulthood of a child born to parents at the 25th percentile of the income distribution for white children, from Chetty et al. (2018).
num_below_p50	Number of children with below-national-median parental household income, from Chetty et al. (2018).
kid_black_pooled_blw_p50_n	Number of Black children with below-national-median parental household income, from Chetty et al. (2018).

`kid_white_pooled_blw_p50_n`	Number of white children with below-national-median parental household income, from Chetty et al. (2018).
`kfr_26_pooled_pooled_p25`	Mean income percentile at 26 years old of a child born to parents at the 25th percentile of the income distribution, from Chetty et al. (2018).
`gini99_simple`	The raw Gini coefficient estimated using tax data minus the income share of the top 1%, from Chetty et al. (2018). See Supplementary Information A.5.1. of Chetty et al. (2022a) for details.
`jobs_total_5mi_2015`	Jobs within 5 miles in 2015, from Chetty et al. (2018).
`job_growth_2004_2013`	Job growth rate 2004–2013, from Chetty et al. (2018).
`emp2000`	Employment rate in 2000, from Chetty et al. (2018).
`hhinc_mean2010`	Mean household income in 2010, from Chetty et al. (2018).
`singleparent_share2000`	Share of single parents in 2000, from Chetty et al. (2018).
`gsmn_math_g3_2013`	Mean 3rd grade math score in 2013, from Chetty et al. (2018).
`frac_coll_plus2000`	Share of college graduates in 2000, from Chetty et al. (2018).
`poor_share2000`	Share below poverty line in 2000, from Chetty et al. (2018).
`nonpoor_share2000`	Share above poverty line in 2000, from Chetty et al. (2018).
`share_white2000`	Share white (2000), from Chetty et al. (2018).
`share_black2000`	Share Black (2000), from Chetty et al. (2018).
`share_asian2000`	Share Asian (2000), from Chetty et al. (2018).
`share_hisp2000`	Share Hispanic (2000), from Chetty et al. (2018).
`share_white2017`	Share white (2017), from Chetty et al. (2018).
`share_black2017`	Share Black (2017), from Chetty et al. (2018).
`share_hisp2017`	Share Hispanic (2017), from Chetty et al. (2018).
`med_hhinc2000`	Median household income (2000), from the ACS.
`med_hhinc2017`	Median household income in 2017, from the ACS.
`pop2000`	Population in 2000 from the ACS.

`chetty_2016_JAMA_table_11.dta`

Please see the codebook for Table 11 of Chetty et al. (2016) for details. This codebook is available on the Opportunity Insights website.

3.2 ZIP Code-Level Data

`zip_covariates.dta`

Variable Name	Description
`zip`	5-digit ZIP code tabulation area code.
`county`	5-digit county FIPS code.
`cz`	Commuting zone (1990 definition).
`czname`	Commuting zone name (1990 definition).
`pop2018`	Population in 2018 from the ACS.
`kfr_pooled_pooled_p25`	Mean income percentile in adulthood of a child born to parents at the 25th percentile of the income distribution, from Chetty et al. (2018).
`kfr_black_pooled_p25`	Mean income percentile in adulthood of a child born to parents at the 25th percentile of the income distribution for Black children, from Chetty et al. (2018).
`kfr_white_pooled_p25`	Mean income percentile in adulthood of a child born to parents at the 25th percentile of the income distribution for white children, from Chetty et al. (2018).
`num_below_p50`	Number of children with below-national-median parental household income, from Chetty et al. (2018).
`kid_black_pooled_blw_p50_n`	Number of Black children with below-national-median parental household income, from Chetty et al. (2018).
`kid_white_pooled_blw_p50_n`	Number of white children with below-national-median parental household income, from Chetty et al. (2018).
`kfr_26_pooled_pooled_p25`	Mean income percentile at 26 years old of a child born to parents at the 25th percentile of the income distribution, from Chetty et al. (2018).
`jobs_total_5mi_2015`	Jobs within 5 miles in 2015, from Chetty et al. (2018).
`job_growth_2004_2013`	Job growth rate 2004–2013, from Chetty et al. (2018).

emp2000	Employment rate in 2000, from Chetty et al. (2018).
hhinc_mean2010	Mean household income in 2010, from Chetty et al. (2018).
hhinc_mean2000	Mean household income in 2000, from Chetty et al. (2018).
singleparent_share2000	Share of single parents in 2000, from Chetty et al. (2018).
gsmn_math_g3_2013	Mean 3rd grade math score in 2013, from Chetty et al. (2018).
frac_coll_plus2000	Share of college graduates in 2000, from Chetty et al. (2018).
poor_share2000	Share below poverty line in 2000, from Chetty et al. (2018).
nonpoor_share2000	Share above poverty line in 2000, from Chetty et al. (2018).
share_white2000	Share white (2000), from Chetty et al. (2018).
share_black2000	Share Black (2000), from Chetty et al. (2018).
share_asian2000	Share Asian (2000), from Chetty et al. (2018).
share_hisp2000	Share Hispanic (2000), from Chetty et al. (2018).
share_white_2018	Share white (2018), from the ACS.
share_black_2018	Share Black (2018), from the ACS.
share_hispanic_2018	Share Hispanic (2018), from the ACS.
share_asian_2018	Share Asian (2018), from the ACS.
share_hawaii_2018	Share Hawaiian (2018), from the ACS.
share_natam_2018	Share Native American (2018), from the ACS.
med_hhinc2000	Median household income (2000), from the ACS.
med_inc_2018	Median household income in 2018, from the ACS.
pop_2000_n	Population in 2000 from the ACS.

zip_income_bins_and_median_income_by_age_14_18.dta

Variable Name	Description
	Note: All variables in this dataset are from the 2014-2018 ACS
zip	5-digit ZIP code tabulation area code
ajzle019	No. of householders 25 to 44 years

ajzle020	No. of householders 25 to 44 years with less than $10,000 household income
ajzle021	No. of householders 25 to 44 years with $10,000-$14,999 household income
ajzle022	No. of householders 25 to 44 years with $15,000-$19,999 household income
ajzle023	No. of householders 25 to 44 years with $20,000-$24,999 household income
ajzle024	No. of householders 25 to 44 years with $25,000-$29,999 household income
ajzle025	No. of householders 25 to 44 years with $30,000-$34,999 household income
ajzle026	No. of householders 25 to 44 years with $35,000-$39,999 household income
ajzle027	No. of householders 25 to 44 years with $40,000-$44,999 household income
ajzle028	No. of householders 25 to 44 years with $45,000-$49,999 household income
ajzle029	No. of householders 25 to 44 years with $50,000-$59,999 household income
ajzle030	No. of householders 25 to 44 years with $60,000-$74,999 household income
ajzle031	No. of householders 25 to 44 years with $75,000-$99,999 household income
ajzle032	No. of householders 25 to 44 years with $100,000-$124,999 household income
ajzle033	No. of householders 25 to 44 years with $125,000-$149,999 household income
ajzle034	No. of householders 25 to 44 years with $150,000-$199,999 household income
ajzle035	No. of householders 25 to 44 years with $200,000 or more household income
ajzme003	Median household income for householders 25 to 44 years

3.3 High School Data

`nces_highschools.dta`

Variable Name	Description
	Note: All variables in this dataset are from the National Center for Education Statistics.
`nces_school_id`	12-digit NCES school ID
`zip`	5-digit ZIP code tabulation area code.
`county`	5-digit county FIPS code.
`cz`	Commuting zone.
`school_name`	School name.
`students_9_to_12`	Number of students in the school (grades 9 to 12).
`frac_students_fr_red_lunch_5yr_m`	Fraction of students eligible for free or reduced lunch.
`frac_students_female`	Share female.
`frac_students_native`	Share Native American.
`frac_students_white`	Share white.
`frac_students_black`	Share Black.
`frac_students_hispanic`	Share Hispanic.
`frac_students_asian`	Share Asian.

`crdc_nces_collapse_hs.dta`

Variable Name	Description
	Note: All variables in this dataset are from the National Center for Education Statistics (Civil Rights Data Collection).
`nces_id`	12-digit NCES school ID
`tot_enr_m`	Total enrolment (male)
`tot_enr_f`	Total enrolment (female)
`tot_gtenr_m`	Total enrolment (male) in gifted and talented programs
`tot_gtenr_f`	Total enrolment (female) in gifted and talented programs

`tot_apenr_m`	Total enrolment (male) in Advanced Placement programs
`tot_apenr_f`	Total enrolment (female) in Advanced Placement programs

3.4 College Data

Datasets:

1. `chetty_2020b_QJE_table_3.dta`, from Chetty et al. (2020). The codebook for this dataset is available on the Opportunity Insights website.

2. `chetty_2020b_QJE_opeid_superopeid_crosswalk.dta`, from Chetty et al. (2020). The codebook for this dataset is available on the Opportunity Insights website.

3. `college_characteristics.dta`, from Chetty et al. (2020). The codebook for this dataset is available on the Opportunity Insights website.

References

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018). The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility. Working Paper 25147, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2020). Income segregation and intergenerational mobility across colleges in the united states. *The Quarterly Journal of Economics 135*(3), 1567–1633.

Chetty, R. and N. Hendren (2018). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics 133*(3), 1163–1228.

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebel, N. Hendren, R. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, D. Johnston, M. Koenen, E. Laguna-Muggenberg, F. Mudekereza, T. Rutter, N. Thor, W. Townsend, R. Zhang, M. Bailey, P. Barberá, M. Bhole, and N. Wernerfelt (2022a). Social capital I: Measurement and associations with economic mobility. *Nature 608*(7921), 108–121.

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebel, N. Hendren, R. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, D. Johnston, M. Koenen, E. Laguna-Muggenberg, F. Mudekereza, T. Rutter, N. Thor, W. Townsend, R. Zhang, M. Bailey, P. Barberá, M. Bhole, and N. Wernerfelt (2022b). Social capital II: Determinants of economic connectedness. *Nature 608*(7921), 122–134.

Chetty, R., M. Stepner, S. Abraham, S. Lin, B. Scuderi, N. Turner, A. Bergeron, and D. Cutler (2016). The association between income and life expectancy in the united states, 2001-2014. *Jama 315*(16), 1750–1766.

Reardon, S. F. and K. Bischoff (2011). Income inequality and income segregation. *American Journal of Sociology 116*(4), 1092–1153.