

Replication files for

Bingley, Cappellari and Tatsiramos

Family, Community and Long-Term Socioeconomic Inequality: Evidence from Siblings and Youth Peers

The data used in this paper are drawn from Danish administrative registers and are confidential. Our access is not unique and others can gain similar access by following a procedure described by Statistics Denmark at <http://www.dst.dk/en/TilSalg/Forskningsservice.aspx>. Physically the micro data are located on specific computers at Statistics Denmark and may not be transferred to computers outside Statistics Denmark due to data security considerations.

Researchers who wish to analyse the micro data for replication purposes need to submit a written application to Statistics Denmark with a detailed research proposal describing the goals and methods of the project, a detailed list of variables, and the selection criteria to be used. Researchers are allowed to use these data if their research project is approved by Statistics Denmark and if they are affiliated with a research institution accepted by Statistics Denmark. If a researcher at a university or other research institution outside Denmark wishes to use the micro data, this may be accomplished by visiting a Danish research institution or by cooperating with researchers working in Denmark.

For researchers wishing to analyse the micro data for replication purposes, we will provide guidance with regard to project approval at Statistics Denmark. All of the results of the paper are obtained from the second and fourth moments of the data, which we provide and thus can be used for replication without access to the micro data.

All estimations are carried out on the VIVE server at Statistics Denmark (DST) using STATA v14.2 MP4 on an Intel E5-2697 CPU with two 2.7 GHz processors.

All reported computing times in master do files are approximate because there are about 45 users on the server and changes in congestion affect computing times.

A. Overall structure of analysis

The analysis has 4 steps:

1. Extraction of source data from DST registers.
2. Extraction of estimating sample from source data and generation of relevant variables, creating dataset with estimating sample, either a long-form panel or a cross-section (in the case of education).
3. Creation of datasets of moments from the estimating sample.
4. Model estimation and post-estimation using data sets of moments.

B. Description of relevant of datasets

The source data are described in source-data-list.txt.

The estimating data for earnings have the following variables:

pnr	individual identifier
familyid	family identifier
border	birth order
yob	year of birth
schoolid	school identifier
parish	parish identifier
sameschool	1(siblings attending same school at 15)
sameparish	1(siblings living in same parish at 15)

same_ps	1(siblings attending same school at 15 or living in same parish at 15)
earnings	annual labor earnings in real terms
logearn	log() of earnings
cohort	3-year birth cohort
year	
age	
ageco	imputed age based on 3-year cohort of birth
prime_ind	1(first observation on an individual)
couple	1(person belongs to a sibling couple)

The estimating data for unemployment is as above, but has an unemployment indicator (see text) in place of earnings.

The estimating data for education are obtained from the earnings panel, keeping one observation per person and adding educational attainment.

Description of datasets of moments

There are 74 datasets that are stored in \C_minimum_distance. The folder has 2 subdirectories \Men and \Women containing the same statistics for each gender.

These subdirectories each have three subdirectories \earnings, \education, \unemployment containing statistics for each of the variables described below.

Each dataset contains second and fourth moments of length of education, annual earnings, or degree of unemployment. Because the nature of the variables is different, the statistics for each variable are organized slightly differently.

\education contains one dataset. The dataset contains variables with the following naming convention, denoting what statistic is calculated or what sample the statistic is calculated for:

1. m covariance of outcomes between year t and year s
2. bobt type of the person with outcome in year t:
bobt=1 if the person is a singleton, first born sibling or community peer
bob2=2 if the person is a second born sibling
3. bobs type of the person with outcome in year s:
bobs=1 if the person is a singleton or first-born sibling
bobs=2 if the person is a second born sibling
bobs=3 if the person is a community peer
4. cohoot birth cohort of the outcome in year t
5. cohos birth cohort of the outcome in year s
6. same indicator for siblings sharing the community
7. dcoYYYY (with YYYY=1966 to 1984) - binary indicator for cohoot=YYYY and cohos=YYYY and bobt=bobs
8. dev the square root of the diagonal elements of the fourth moment matrix
9. idev 1/dev

\unemployment contains six datasets. Each dataset contains variables with the following naming convention, denoting what statistics are calculated or what sample the statistic is calculated for:

1. m covariance of outcomes between year t and year s
2. yt year t
3. ys year s

4. bobt type of the person with outcome in year t:
bobt=1 if the person is a singleton, first born sibling or community peer
bob2=2 if the person is a second born sibling
5. bobs type of the person with outcome in year s:
bobs=1 if the person is a singleton or first-born sibling
bobs=2 if the person is a second born sibling
bobs=3 if the person is a community peer
6. cohort birth cohort of the outcome in year t
7. cohos birth cohort of the outcome in year s
8. same indicator for siblings sharing the community
9. cell the number of pairs used to compute m
10. v is the diagonal elements of the fourth moments matrix

\earnings contains 30 datasets. Each dataset contains variables with the following naming convention, denoting what statistics are calculated or what sample the statistic is calculated for:

1. m covariance of outcomes between year t and year s
2. yt year t
3. t yt - 1990
4. ys year s
5. s ys - 1990
6. bobt type of the person with outcome in year t:
bobt=1 if the person is a singleton, first born sibling or community peer
bob2=2 if the person is a second born sibling
7. bobs type of the person with outcome in year s:
bobs=1 if the person is a singleton or first-born sibling
bobs=2 if the person is a second born sibling
bobs=3 if the person is a community peer
8. cohort birth cohort of the outcome in year t
9. cohos birth cohort of the outcome in year s
10. aget age of the outcome in year t
11. ages age of the outcome in year s
12. same indicator for siblings sharing the community
13. drY (with Y=0 to 24) - binary indicator for yt=Y+1990
14. dcY (with Y=0 to 24) - binary indicator for ys=Y+1990
15. dcoYYYY (with YYYY=1966 to 1984) - binary indicator for cohort=YYYY and cohos=YYYY and bobt= bobs
16. dd binary indicator for yt=ys
17. tds |yt-ys|
18. dev* columns of the matrix of deviations of cross products from sample moments, which are used to accumulate the fourth moment matrix V
19. varia diag(V), i.e., the diagonal elements of the matrix V
20. stdev varia^{.5}, i.e., the square root of the diagonal elements of the matrix V

dev* requires more explanation because the * further indicates what sample the deviation refers to. * has the format _AA_BBBBCCCC_DDDDEEEE, where AA refers to bobt and bobs, BBBB is the first cohort-year of comparison (cohort), CCCC is the second cohort-year of comparison (cohos), DDDD is the first calendar year of comparison (yt), EEEE is the second calendar year of comparison (ys). For example, dev_12_19661972_20032010 refers to siblings (that is bobt=1 and bobs=2) comparing the 1966 with 1972

cohort (for first and second sibling respectively) and the 2003 with 2010 calendar year (for first and second sibling respectively).

C. Do files

We provide all the do files for the analysis, grouped into three distinct folders.

1. \A_from_raw_data_to_moments. These files perform steps one to three of the analysis (see Section A above) and operate on confidential microdata.

Extraction of source data from DST registers is implemented by:

community-data-creation-1.do

community-data-creation-2.do

Creation of estimating sample and creation of datasets of moments are implemented by:

\earnings\master_earnings_moments.do

\education\master_edu.do

\unemployment\master_unem.do

The do files should be run by gender after selecting the appropriate gender-specific command lines within the do files named community-data-creation*.do and inspection*.do as indicated in the do files themselves.

2. \B_simulation_tab1_fig1. To replicate Table 1 and Figure 1 that are produced using the confidential microdata we simulate synthetic data with a structure analogous to the real data. The simulations are implemented by running:

master-simulation.do

Note that the file fig1sim.do, that is called by master-simulation.do for replicating Figure 1, uses the STATA command “spmap”, which works with all recent versions of STATA with the exception of version 15.0. Users of version 15.0 should substitute the “spmap” with the “grmap” command within fig1sim.do.

3. \C_minimum_distance. All remaining figures and tables in the paper or appendix are produce using second and forth moments, which we provide for the purposes of replication. This is step four of the analysis (see Section A above) that is implemented by running:

\Gender\earnings\master_earnings_models.do

\Gender\education\master_edu.do

\Gender\unemployment\master_unem.do

for Gender = Men, Women

combine_graphs.do should be run afterwards to compile Figures 2 to 10 from gender-specific figures.

The mapping between notation in the paper and parameter names in the Minimum Distance do files is as follows:

$a_{ij}, j=1,2$	$\sigma_{\omega 24j}^2$
$b_{ij}, j=1,2$	$\sigma_{\xi j}^2$
af	$\sigma_{\mu F}^2$
ac	$\sigma_{\mu C}^2$
afc	$\sigma_{\mu FC}^2$
bf	$\sigma_{\gamma F}^2$

bc	$\sigma_{\gamma C}^2$
abf	$\sigma_{\mu \gamma F}^2$
abc	$\sigma_{\mu \gamma C}^2$
b3j, j=1,2	$\sigma_{\varepsilon j}^2$
b3_age_knot	Age spline
b4j, j=1,2	ρ_j
b3_12	σ_f
b30_1j, j=1,2	σ_{24j}^2
b3_13	λ
pwt	π_{1990+t}
tw	η_{1990+t}
pco	δ

The log files of Minimum Distance estimates report an initial regression output of the STATA `nl`-routine for non-linear least squares that has unadjusted standard errors. This is followed by a print out of the same point estimates but accompanied by standard errors that are adjusted as explained in the Appendix. The log files end with the decomposition of the sibling correlation coefficient.