Documentation for CenSoc Dataset

Monica Alexander

March 2018

Contents

1 Introduction

The 'CenSoc' project involves producing a dataset which contains records from the full-count 1940 US census to the social security deaths masterfile. The resulting CenSoc dataset provides researchers with a tool for studying mortality inequalities in the US and how conditions have changed over time.

This document outlines the data sources and methods used to produce the initial CenSoc dataset. The results of the matching process and relative characteristics of the matched and unmatched records are also described.

1.1 The `censoc` R package

The CenSoc project has an accompanying R package `censoc`, which contains supporting R code and documentation used to create and work with the CenSoc dataset. The package can be found here: `https://github.com/MJAlexander/censoc`.

2 Data sources

2.1 Census

The original 1940 census records were digitized by Ancestry.com and are made available through the Minnesota Population Center (MPC). The MPC provides a de-identified version of the complete count census as part

of the IPUMS-USA website: `https://usa.ipums.org/usa/complete_count.shtml`. However, name and other identifying information is not available from the IPUMS website.

Access to the resticted 1940 census data was granted by agreement between UC Berkeley and MPC. The data are encrypted and can only be accessed through computers or servers on the Berkeley demography network.

2.2 Social Security Deaths

The Social Security Administration (SSA) has kept a record of all deaths that have been reported to the SSA since 1962 in their death master file (SSDM). The SSDM is used by financial and government agencies to match records and prevent identity fraud.

The SSDM is considered a public document under the Freedom of Information Act. Monthly and weekly updates of the file are sold by the National Technical Information Service of the U.S. Department of Commerce.

In our initial match, we obtained the SSDM files through the wesbite `http://ssdmf.info/download.html`. This wesbite appears to be run by an individual who has bought a copy of the SSDM and decided to distribute it freely. The 2013 version of the SSDM is available here: `http://cancelthesefunerals.com/`.

Note: XX need to work out whether we are using the 2011 or 2013 version. Based on dates of deaths in the CenSoc data, it looks like probably 2011. Also need to confirm that I should be restricting the latest date of death.

2.2.1 Data completeness

Completeness of death reporting in the SSDM is lower pre-1970s, as a substantial proportion of the population did not pay into the social security system. Deaths are more likely to be reported at older ages, when are more likely to be receiving social security benefits. In addition, previous studies have found death coverage in the SSDM is larger for men and for the white population. Previous studies suggest more than 90% completeness of deaths over age 65 reported in the SSDM since 1975, compared to vital statistics sources. Our own comparison with HMD suggests XX% completeness of the SSDM data after 1975.

3 Data preparation

Prior to matching the two datasets, the relevant variables are cleaned to ensure the that matches are not affected by issues like trailing white space, punctuation, differences in case, etc. The cleaning steps for each dataset are described below.

3.1 Census

For the Census dataset, the following releveant variables are retained:

- `SERIAL40`: 1940 census serial number
- `PERNUM`: person number in household
- `STATEFIP`: US state code
- `NUMPREC40`: total size of household
- `AGE`: age of respondent
- `SEX`: sex of respondent
- `NAMELAST`: first name of respondent (and middle name, if applicable)
- `NAMEFRST`: last name of respondent

Age, sex and name are required to match the census data with the social security deaths data. The first four variables (`SERIAL40`, `PERNUM`, `STATEFIP`, `NUMPREC40`) are needed to merge the CenSoc dataset with the IPUMS census data at a later stage.

The following pre-processing steps are done:

1. Convert all name strings to upper case.
2. Remove the middle name. The original first name variable contains both first and middle name. The name string is split and only the first name is used for matching.
3. Remove rows where either the first or last name are just question marks or blank.
4. Create a match key by concatenating last name, first name and age.
5. Subset the data to only include males. At this stage, we are not attempting to match the female population.

Each of these steps is done using the `load_census` function within the `censoc` package.

3.2 Social Security Deaths

For the social security deaths files, there are three raw files in which rows contain a continuous string of characters. For each of the three files, each row is split into social security number, last name, first name, middle initial, date of death and date of birth. The three files are then binded together to create one large file. (XX Josh: where did you get the other variable info from)?

The following pre-processing steps are done:

1. Remove any trailing white space from first and last names
2. Split up date of birth and date of death to get day, month and year of birth and death.
3. Calculate age of person at census. The age is calculated based on knowing the date of birth and that the 1940 census was run in April.
4. Remove any deaths where the date of birth is missing
5. Remove any deaths of people who born after 1940
6. Remove any deaths before 1975
7. Create a match key by concatenating last name, first name and census age.

Each of these steps is done using the `load_socsec_deaths` function.

4 Match Method

Currently, the two datasets are matched based on exact matches of first name, last name and age. For example, a match key could be `ALEXANDERMONICA31`. Census records that have a key that is not found in the social security deaths database are not matched. Any records with a duplicate match key are removed. In future, alternative match methods will be investigated, for example, accounting for name variants and spelling mistakes, simulating matched datasets based on duplicate keys, etc.

The match is done using the `create_censoc` function. Each state is initially matched separately. The specific steps are:

1. Load in the cleaned census and social security datasets.
2. Remove any duplicate keys.
3. Merge the two datasets based on key.

4.1 Creating the national matched dataset

Due to file size, the matching step is done separately for each census file in each US state. The resulting national dataset is created by:

1. Loading in and binding all state matched files.
2. Removing all rows that have duplicated keys.

This can be done using the `create_national_censoc` function.

5 Resulting Dataset

A total of 7,564,451 individual males were matched across the census and SSDM to create the CenSoc dataset. As the 1940 full count census had 66,093,146 males, this corresponds to a raw match rate of 11.4%. A total of 43,881,719 males in the census had unqiue keys; as such the match rate on unique keys was 17.2%.

5.1 Match rates by group

The raw match rates differ marked by cohort/age at census. As the table below illustrates, match rates are highest for 15-40 year olds. This corresponds to cohorts 1900-1925.

census_age	match_rate	match_rate_unique
0-4	9.1	14.4
10-14	14.5	22.7
15-19	17.0	26.3
20-24	18.2	27.4
25-29	18.0	26.8
30-34	16.6	24.8
35-39	13.9	20.7
40-44	10.8	16.0
45-49	7.4	11.0
50-54	4.2	6.2
55-59	1.7	2.6
5-9	11.6	18.4
60-64	0.4	0.7
65-69	0.1	0.1
70-74	0.0	0.0
75+	0.0	0.0

XX Need to add more characteristics

5.2 Accounting for mortality

Note that these raw match rates do not take into consideration mortality. Some individuals would have died before 1975, and some are still alive after 2005, and so do not appear in the SSDM. Thus we would never expect to get match rates of 100% given we only observe a truncated window of deaths.

Based on overall US mortality reported in the HMD. . . XX TODO. . .

5.3 Characteristics of matched versus unmatched data

6 Merging with public-use IPUMS file

The CenSoc dataset contains information about individaul's birth date (and implied at the 1940 census), death date and state of residence in 1940. The dataset does not contain any other information on the characteristics of the individuals; however, by using the unique identifier information contained in the CenSoc dataset, the data can be merged back to the 1940 census available on the IPUMS-USA website.

For instructions on how to obtain full count census data from IPUMS, see documentation here: `https://github.com/MJAlexander/censoc/blob/master/documentation/ipums_document.pdf`.