PHP 2410E - Assignment 2

Kevin W. McConeghy

Compiled: 2019-09-24

Contents

Introduction

This is the completed assignment 2 for the 'Medicare data' course at Brown University.
All code is stored in a Github repository, https://github.com/kmcconeghy/PHP2410E

Statement of work

This document was created solely by the author, guidance in the homework solutions was driven by class instruction, materials or prior experience. The solutions were not shared with anyone else.

Note on R markdown

This report was generated using R markdown, LaTEX, and several non-base R packages (e.g. tidyverse).

```
## Non-base packages loaded:  Scotty tidyverse rJava kableExtra
```

Assignment as written:

- Data Assignment #2
- Working with Medicare Public Use Files
- Due October 2nd, 2019

Assignment Overview

There are two data sets each containing identifying information on Medicare beneficiaries. The information includes date of birth, gender and race. The two sources of data are:

1) The Minimum Data Set, a clinical patient record assessment that is completed each time a patient is admitted to a nursing home in the US that is certified by Medicare/Medicaid. The MDS data file (in either STATA or SAS format) is likely to include multiple records per ID (unique individual) for many individuals;

2) The Medicare Enrollment Record is the individual identifier of every Medicare beneficiary. The overall file includes detailed data from Social Security as well as whether and when the beneficiary had joined a Medicare Advantage Plan and when. The current file includes only identifying information such as date of birth, gender and race. There is only one record per person per year. The ID number on the Medicare Enrollment Record is the same as that on the MDS data file.

There are two parts to this data assignment. Each is described below. There are various ways to complete these components. It is up to the student to choose how s/he goes about completing the assignment.

• Using the MDS data determine how consistent the identifying data are across records for the same persons who have multiple records. This means, among those with more than one record, what is the rate of inconsistency for date of birth, gender and race. DoB, gender and race have multiple ways in which they could disagree, in addition to calculating the rate of inconsistency, characterize the different ways (e.g. day, month or year of birth; missing categories of information in which this inconsistency manifests itself). (Note:

The Codebook for the MDS2.0 data can be found here: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/NursingHomeQualityInits/downloads/MDS20MDSAllForms.pdf)

• Match the MDS and Medicare enrollment records based upon patient ID, using the first MDS record to match for those cases that have multiple MDS records. Next, compare the DoB, gender and Race from the two different data sources. Note that DoB is necessarily measured the same way across the two data sources as is gender (although there may be differing levels of missing data). The race variables across the two data sets are coded differently so it may be important to separately estimate the degree of agreement across the categories that are comparably labeled (e.g. white vs. white; black vs. black) since the Medicare Enrollment record is known to underestimate the number of Hispanic and Asian Americans.

Assessment of Minimum Dataset Assessments

The overall objective is to better understand the validity of MDS assessments by examinining variation in patient characteristics which would typically be fixed across time (e.g. age, gender, race/ethnicity).

in-file of dataset

```
## Dataframe: MDS dataset
## Memory Size: 57 Mb  Rows: 1,481,145  Columns: 5
## id: 150,757 Missing: 0
```

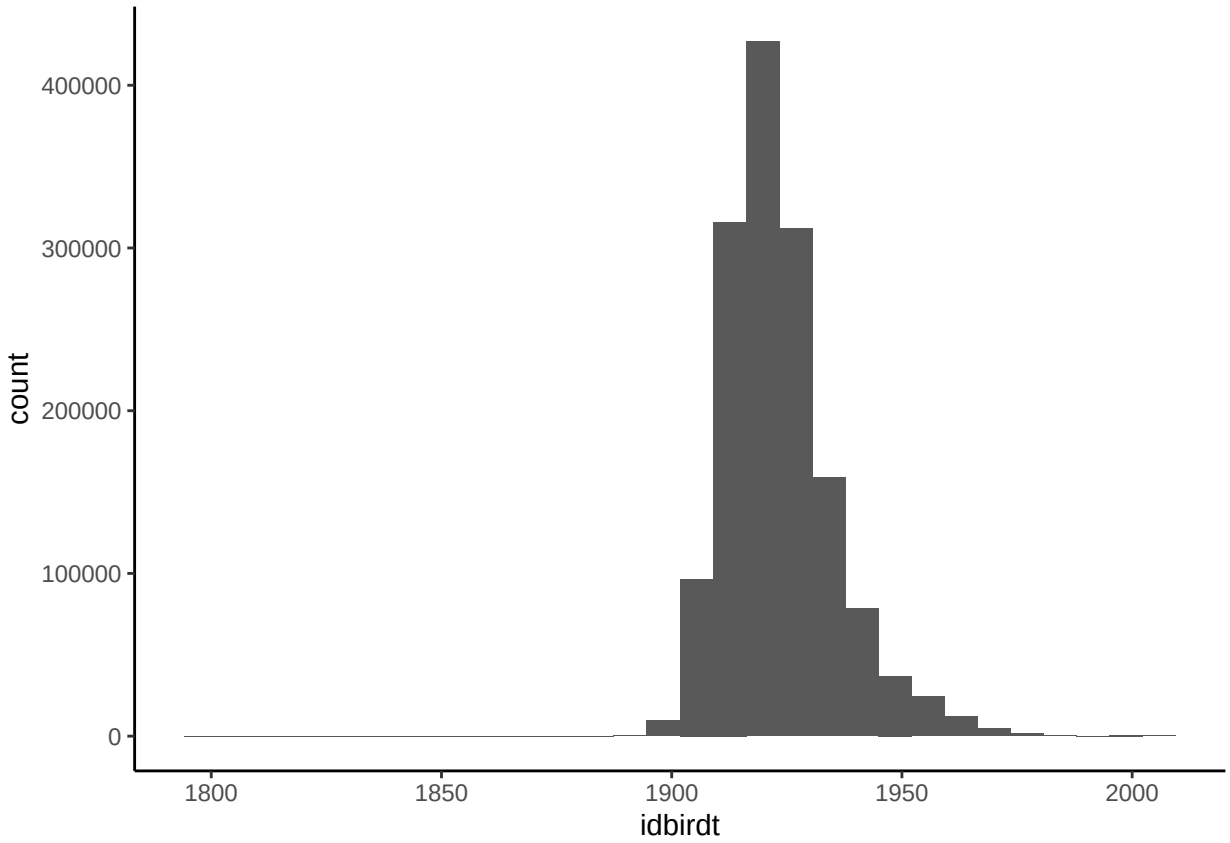
Data structure

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1481145 obs. of  5 variables:
##  $ idgendr: num  1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Gender <IN> Identification Information"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ idbirdt: Date, format: "1943-11-02" "1943-11-02" ...
##  $ idrace : num  5 5 5 5 5 5 5 5 5 5 ...
##   ..- attr(*, "label")= chr "Race/ethnicity <IN> Identification Information"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ dmdate : Date, format: "2003-09-25" "2003-12-15" ...
##  $ id     : num  1 1 1 1 1 1 1 1 2 2 ...
##   ..- attr(*, "format.stata")= chr "%12.0g"
```

Summary statistics

Date of birth

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "1800-09-17" "1915-03-16" "1921-07-08" "1923-05-02" "1928-12-29"
##         Max.         NA's
## "2008-11-19"        "302"
```



Year

```
## # A tibble: 11 x 4
##     year n_assess  n_id percent
##    <dbl>    <int> <int>   <dbl>
##  1  1998    55503 21920    3.75
##  2  1999   131571 30733    8.88
##  3  2000   131169 30109    8.86
##  4  2001   135622 30423    9.16
##  5  2002   140206 30606    9.47
##  6  2003   142857 30780    9.64
##  7  2004   144912 30872    9.78
##  8  2005   148893 31598   10.1
##  9  2006   151363 31539   10.2
## 10  2007   151430 31645   10.2
## 11  2008   147619 30992    9.97
```

Sex

```
## # A tibble: 3 x 3
##   idgendr      n percent
##     <dbl>  <int>   <dbl>
## 1       1 485132  32.8
## 2       2 995466  67.2
## 3      NA    547   0.037
```

```
## # A tibble: 6 x 3
##   idrace       n percent
##    <dbl>   <int>   <dbl>
## 1      1    5290   0.357
## 2      2   17130   1.16
## 3      3  164435  11.1
## 4      4   49427   3.34
## 5      5 1238941  83.6
## 6     NA    5922   0.4
```

Summary description

Median age date of birth is in 1921, most recent is in 2008 (highly suspicious for error), 302 missing DOB. The number of residents per year peaks in 2007. The raw datafile, 67% of assessments are female, 84% white, 11% Afr. American.

- Using the MDS data determine how consistent the identifying data are across records for the same persons who have multiple records. This means, among those with more than one record, what is the rate of inconsistency for date of birth, gender and race. DoB, gender and race have multiple ways in which they could disagree, in addition to calculating the rate of inconsistency, characterize the different ways (e.g. day, month or year of birth; missing categories of information in which this inconsistency manifests itself).

Basic strategy: The MDS dataset will be grouped by ID, a series of flags will be generated to identify disparate records among those variables expect to be fixed. We reformat gender to 0/1 (1=Male), and race into dummary categories (e.g. rc_white). Then the findings will be summarized in a set of tables.

Consistency of data across assessments

	n_gendr	n_race	n_birdt	n_birmonyr
	1:148556	1:145706	1:143430	1:146002
	2: 2191	2: 4841	2: 6863	2: 4537
	3: 10	3: 194	3: 434	3: 210
	NA	4: 15	4: 27	4: 7
	NA	5: 1	5: 2	7: 1
	NA	NA	8: 1	NA

	n_gendr	n_race	n_birdt	n_birmonyr
	0: 32	0: 477	0: 10	1:146002
	1:148822	1:146607	1:143600	2: 4537
	2: 1903	2: 3550	2: 6706	3: 210
	NA	3: 115	3: 416	4: 7
	NA	4: 8	4: 22	7: 1
	NA	NA	5: 2	NA
	NA	NA	8: 1	NA

By individual a value of 0 denotes no entries, a value of 2 means two distinct values for that variable (within person), and so on. There are a few missing entries but it's not the primary concern. While most only have one value, many have multiple values for the same variable (which is incoherent) and would need to be adjudicated in an analysis. In the extreme, one individual has 8 different birth dates across assessments. Birth dates may vary by a few days, or only be entered as the first of the month etc. If you recode the birth date to the month-year you may be able to collapse conflicting birthdates into the same category of month-year (`birmonyr`). This resolves about 3000 conflicted birthdates but still leaves many conflicts.

Assessment of Medicare Enrollment File

in-file of dataset

```
## Dataframe: Denominator dataset
## Memory Size: 182 Mb  Rows: 4,768,016  Columns: 5
## id: 671,591 Missing: 0
```

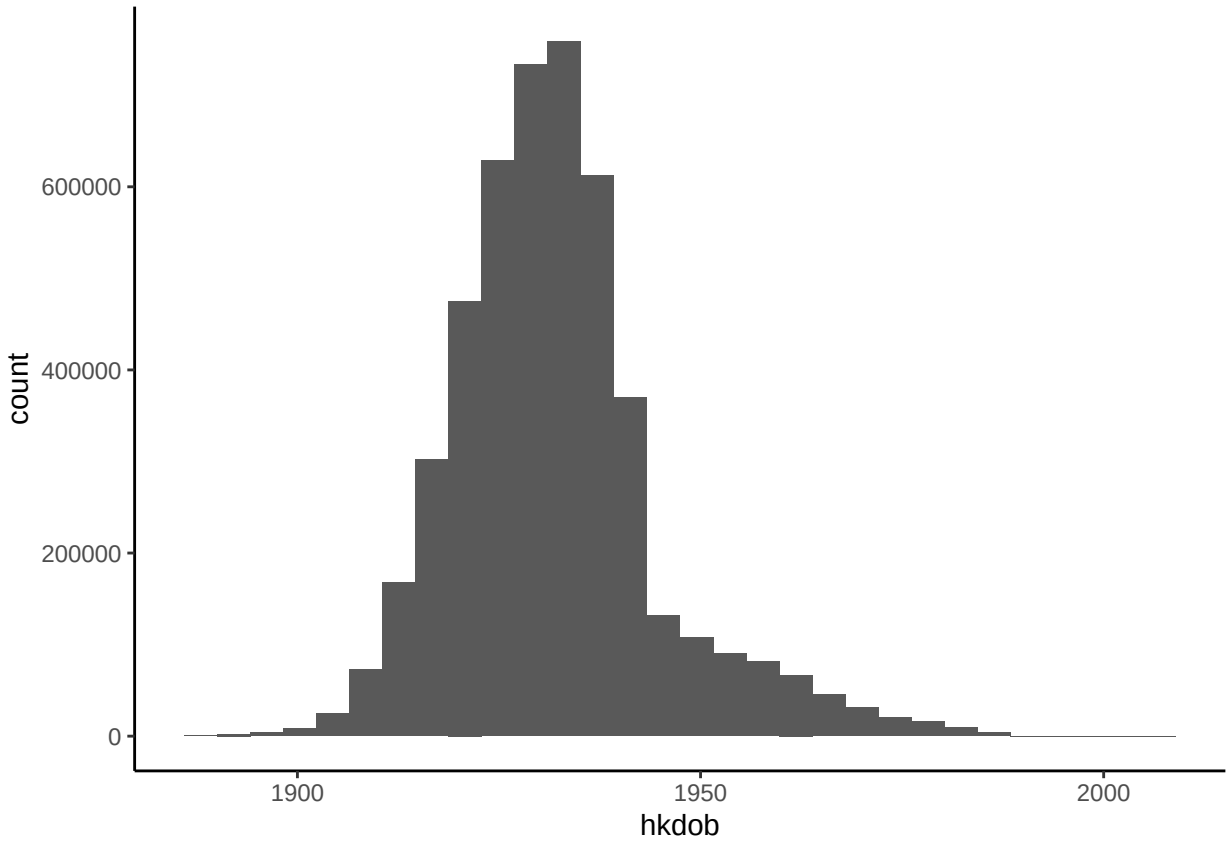
Data structure

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    4768016 obs. of  5 variables:
##  $ id    : num  1 2 3 3 3 3 4 4 4 5 ...
##   ..- attr(*, "format.stata")= chr "%12.0g"
##  $ hkyear: num  2008 2008 1998 1999 2000 ...
##   ..- attr(*, "label")= chr "Denom: Reference year (year of data)"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ hkdob : Date, format: "1943-11-02" "1947-02-15" ...
##  $ hksex : num  1 1 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Denom: Beneficiary sex (1=M 2=F & impute <65=M,65+=F)"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ hkrace: num  1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Denom: Bene race (0=Unk 1=Whi 2=Bla 3=Oth 4=Asn 5=His 6=Nat)"
##   ..- attr(*, "format.stata")= chr "%8.0g"
```

Summary statistics

Date of birth

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "1888-01-10" "1923-10-13" "1930-11-06" "1931-11-16" "1937-08-22"
##         Max.
## "2006-11-28"
```



Year

hkyear	n_recs	n_id	percent
1998	403999	403999	27.276
1999	408188	408188	27.559
2000	412847	412847	27.874
2001	417321	417321	28.176
2002	422779	422779	28.544
2003	429255	429255	28.981
2004	436078	436078	29.442
2005	444430	444430	30.006
2006	453160	453160	30.595
2007	464059	464059	31.331
2008	475900	475900	32.131

Importantly, there are no duplicates within year. Each enrollee should have one record per year.

Year cross-over

n_yrs	n_benes	percent
1	52666	3.556
2	53245	3.595
3	49758	3.359
4	47558	3.211
5	45446	3.068
6	43802	2.957
7	40704	2.748
8	39152	2.643
9	37761	2.549
10	35170	2.375
11	226329	15.281

15% of the sample exists across all years (n_yrs=11).

Sex

hksex	n	percent
1	2093428	141.338
2	2674588	180.576

hkrace	n	percent
0	10160	0.686
1	4019074	271.349
2	464395	31.354
3	66939	4.519
4	76473	5.163
5	112571	7.600
6	18404	1.243

Summary description

Median age date of birth is in 1930s (i.e. the denominator file represents younger person than the MDS assessments). The years span 1998 - 2008, with no duplicates within year. most recent is in 2008 (highly suspicious for error), 302 missing DOB. The number of residents per year peaks in 2007. The raw datafile, 67% of assessments are female, 84% white, 11% Afr. American.

> Match the MDS and Medicare enrollment records based upon patient ID, using the first MDS record to match for those cases that have multiple MDS records. Next, compare the DoB, gender and Race from the two different data sources. Note that DoB is necessarily measured the same way across the two data sources as is gender (although there may be differing levels of missing data). The race variables across the two data sets are coded differently so it may be important to separately estimate the degree of agreement across the categories that are comparably labeled (e.g. white vs. white; black vs. black) since the Medicare Enrollment record is known to underestimate the number of Hispanic and Asian Americans.

Session Info

Thank you for taking the time to review my work!

```
##  setting  value
##  version  R version 3.6.1 (2019-07-05)
```

```
##  os       Windows 10 x64
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  English_United States.1252
##  ctype    English_United States.1252
##  tz       America/New_York
##  date     2019-09-24
```