PHP 2410E - Assignment 2

Kevin W. McConeghy

Compiled: 2019-09-27

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).

Dataset

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

Data structure

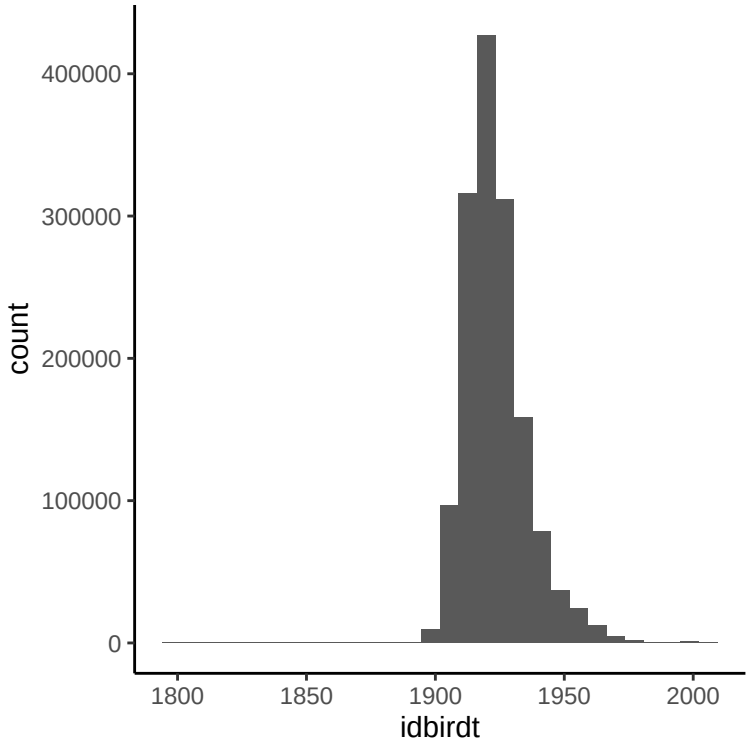
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1481145 obs. of  5 variables:
##  $ idgendr: num  1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Gender <IN> Identification Information"
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ idbirdt: Date, format: "1943-11-02" "1943-11-02" ...
##  $ 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"
```

Figure 1. Histogram of MDS date of birth



Year

year	No. assessments	No. persons	%
1998	55503	21920	3.7
1999	131571	30733	8.9
2000	131169	30109	8.9
2001	135622	30423	9.2
2002	140206	30606	9.5
2003	142857	30780	9.6
2004	144912	30872	9.8
2005	148893	31598	10.1
2006	151363	31539	10.2
2007	151430	31645	10.2
2008	147619	30992	10.0

Sex

idgendr	No. assessments	percent
1	485132	32.8
2	995466	67.2
	547	0.0

Where 1 is male and 2 is female.

idrace	No. assessments	%
1	5290	0.4
2	17130	1.2
3	164435	11.1
4	49427	3.3
5	1238941	83.6
	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: We recode race according to the denominator definitions. The MDS dataset will be grouped by ID, a series of flags will be generated to identify disparate records among those variables expected to be fixed. Then the findings will be summarized in a set of tables.

Reformat

Consistency of data across assessments

	Dup. MDS gender	Dup. MDS race cat.	Dup. MDS race recode	Dup. MDS race white	Dup. MDS race black	
	0: 32	0: 477	1:145706	0: 477	0: 477	
	1:148822	1:146607	2: 4841	1:146900	1:148589	
	2: 1903	2: 3550	3: 194	2: 3380	2: 1691	
		3: 115	4: 15			
		4: 8	5: 1			

Discussion

By individual a value of 0 denotes no entries, a value of 1 or more means 1 or more distinct values for that variable (within person). 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

Dataset

```
## Dataframe: Denominator dataset
```

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

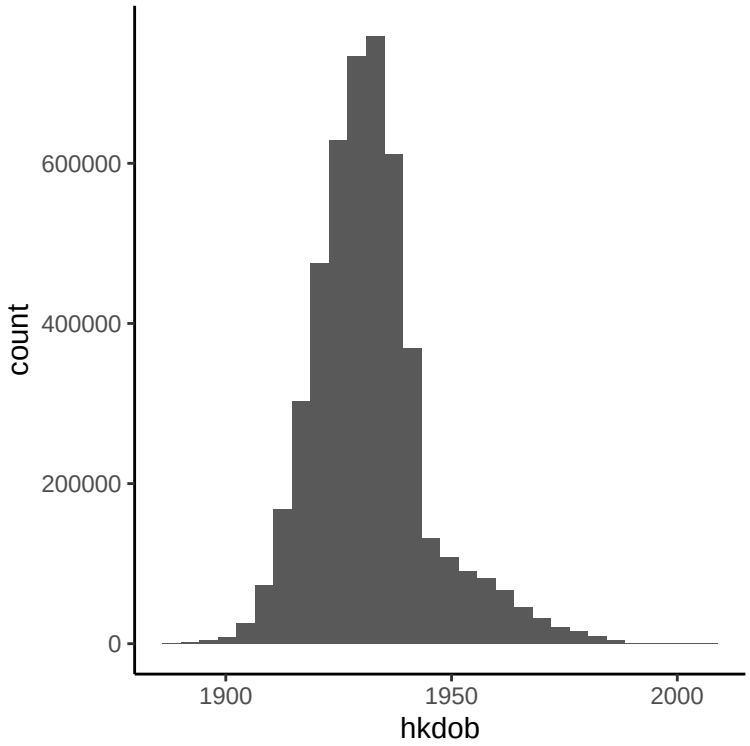
Data structure

```
## 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

Figure 2. Denominator file 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	No. records	No. persons	%
1998	403,999	403,999	8.47
1999	408,188	408,188	8.56
2000	412,847	412,847	8.66
2001	417,321	417,321	8.75
2002	422,779	422,779	8.87
2003	429,255	429,255	9.00
2004	436,078	436,078	9.15
2005	444,430	444,430	9.32
2006	453,160	453,160	9.50
2007	464,059	464,059	9.73
2008	475,900	475,900	9.98

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

Year cross-over

No. years / person	No. persons	%
1	52,666	1.10
2	53,245	1.12
3	49,758	1.04
4	47,558	1.00
5	45,446	0.95
6	43,802	0.92
7	40,704	0.85
8	39,152	0.82
9	37,761	0.79
10	35,170	0.74
11	226,329	4.75

15% of the sample exists across all 11 years.

Sex

hksex	No. records	%
1	2,093,428	43.91
2	2,674,588	56.09

Note: 1 is male, 2 is female

hkrace	No. records	%
0	10,160	0.21
1	4,019,074	84.29
2	464,395	9.74
3	66,939	1.40
4	76,473	1.60
5	112,571	2.36
6	18,404	0.39

Consistency of data across enrollment years

	No. records	Dup. Denom. Gender	Dup. Denom. Race	Dup. Denom. DOB
	11 :226329	1:671591	1:671591	1:671591
	2 : 53245			
	1 : 52666			
	3 : 49758			
	4 : 47558			
	5 : 45446			
	(Other):196589			

Discussion. Sex, race, and date of birth do not vary within ID number (across enrollment years).

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. In the raw file, 67% of assessments are female, 84% white, 11% Afr. American.

Matching MDS and Medicare file

> 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.

Although not explicility said above, each beneficiary may have many records in the enrollment file. Particular instruction for joining wasn't given on this dimension, however as shown above the person characteristics do not vary by year so not a practical issue.

Prepare for matching

MDS record, first row per person

```
## Dataframe: limited to first row by person, year
## Memory Size: 10 Mb  Rows: 150,757  Columns: 11
## id: 150,757 Missing: 0
```

Match 1:many join of first MDS assessment with denom. file

```
## N rows, MDS file: 150757 N rows, Denom file: 4768016
```

```
## N IDs, MDS file: 150757 N IDs, Denom file: 671591
```

```
## Dataframe: Inner join MDS / Denom.
## Memory Size: 15 Mb  Rows: 150,757  Columns: 15
## id: 150,757 Missing: 0
```

All ID's in the MDS file had at least one matching record in the enrollment file.

Comparison of MDS and denominator data

Comparison of Gender

```
##         hksex     1     2
## idgendr
## 1           53892   829
## 2            1626 94295
## NA             42    73
```

Row percentages

```
##        hksex
## idgendr     1     2
##    1    0.985 0.015
##    2    0.017 0.983
##    <NA> 0.365 0.635
```

The MDS variable 'idgendr' is missing in a few cases, and is discordant in about 1.5% of cases. 1.5% where the MDS says male and enrollment says female, and 1.7% where MDS says female and enrollment says male. Note: 'NA' means missing.

Comparison of Race

```
##           hkrace      0      1      2      3      4      5      6
## rc_recode
## 0                     1    745    106     11     35     23      5
## 1                   395 126272   1039    378    279    317    182
## 2                    44    499  12762     68     12     33     18
## 4                    14    160     36    514   1110     12      2
## 5                    18   2862    120    222     23   1873     15
## 6                     1    273     32     30     15      9    192
```

Note: The MDS race/ethnicity variable has been recoded to match categories with the enrollment file race variable. MDS missing was recoded to 0, unknown. The denominator file has a category, 'other' with no comparable category in the MDS assessment. 0 - Unknown, 1 - White, 2 - Black, 3 - Other, 4 - Asian, 5 - Hispanic, 6 - Native American.

Row percentages

```
##          hkrace
## rc_recode     0     1     2     3     4     5     6
##         0 0.001 0.805 0.114 0.012 0.038 0.025 0.005
##         1 0.003 0.980 0.008 0.003 0.002 0.002 0.001
```

```
##         2 0.003 0.037 0.950 0.005 0.001 0.002 0.001
##         4 0.008 0.087 0.019 0.278 0.601 0.006 0.001
##         5 0.004 0.558 0.023 0.043 0.004 0.365 0.003
##         6 0.002 0.495 0.058 0.054 0.027 0.016 0.348
```

There is 98% agreement between the MDS and denominator on white race (category 1), 95% on black race (cat. 2), 60% agreement on asian race (cat. 4), and poor agreement on hispanic or native american race (cat. 5, 6). Many of the individuals coded as hispanic or native american, are coded white in the denominator file.

Comparison of Date of Birth

Match by exact date of birth

chk_dob	No. records	%
0	14,284	9.47
1	136,413	90.49
	60	0.04

The date of birth exactly matches between MDS and enrollment in 90% of cases.

Match by same month-year of birth

chk_mob	No. records	%
0	8,345	5.54
1	142,352	94.42
	60	0.04

The month and year of birth exactly matches in 94% of cases, increasing match rate over exact date.

Evaluate a systematic vs. random error in date-match

Understanding the dates do not match in a small subset of cases, one question would be are the unmatched dates randomly different, or does one date consistently run over- under- the other.

MDS DOB > Denom. DOB	No. records	%
0	143,113	94.93
1	7,584	5.03
	60	0.04

MDS DOB < Denom. DOB	No. records	%
0	143,997	95.52
1	6,700	4.44
	60	0.04

4.4% of MDS DOBs are less than the denominator file DOB, and 5% are more than the denominator DOB. So if the date of birth from CMS enrollment records is considered more valid, the MDS may be biased towards moderately younger (i.e. later) dates of birth than the actual.

Summary discussion

In this assignment, two limited data files, 1) Nursing home minimum dataset assessments and 2) The Centers for Medicaid and Medicare Services enrollment file ('denominator') were summarized and joined.

The primary objective was to compare how consistent reported gender/sex, race and date of birth are between the two files. To simplify the exercise the first asssesment by nursing home resident was used. The results reported above show general agreement between the two files, but some notable discrepancies for minority race categories, particularly hispanics, asians and native-americans. Dates of birth are consistent for 90% of cases outright, and the match can be improved by allowing matches by month-year instead. Overall the MDS assessments are imperfect but in reasonable agreement with the denominator file.

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     2019-09-27
```