PHP 2410E - Assignment 1

Kevin W. McConeghy

Compiled: 2019-09-18

Contents

Introduction

This is the completed assignment 1 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 #1
- Working with Medicare Public Use Files
- Due September 19th, 2019

Assignment Overview

> CMS is committed to increasing access to its Medicare claims data through the release of de-identified data files available for public use. The first phase in this effort is the release of the 5% sample Public Use Files for a variety of Medicare claim types for the periods 2006-2014. These files are available to researchers as free downloads in CSV and/or Excel format, depending upon the year. They contain non-identifiable claim-specific information and are within the public domain.
>
> Of paramount importance in the release of Public Use Files is the protection of beneficiary confidentiality. To that end all directly identifiable information has been removed. Moreover, other potentially identifying variables, which might cause identification by themselves or enable it in combination with other variables, have either been removed from the files or their values re-coded. See the general documentation file for each claim type for specific information concerning de-identification and variable values.
>
> The files can be find here: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/BSAPUFS/index.html
>
> Each file has its own documentation describing file layout and variable values, as well as program code for creating SAS datasets. Click on the link in the left menu for the specific PUF to access documentation and download instructions.
>
> Specification of Data Assignment There are five possible PUF files on the CMS web site. Each is described below. Select at least one and do the following:

Assigment 1.1 Infile the data

1. Read the data into a SAS or STATA analysis file using the format statement provided;

Public Use Files

For this assignment the inpatient file was chosen and downloaded.

https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/BSAPUFS/Downloads/2008_BSA_Inpatient_Claims_PUF.zip

Read in SAS infile statement for labels

CMS provides a SAS proc format statement for working with data, this is read into R and used to relabel the raw '.csv' file in R.

Basic description of file

```
## Dataframe: Raw Inpatient PUF for 2008, downloaded from CMS
## Memory Size: 67 Mb  Rows: 588,415  Columns: 8
## IP_CLM_ID: 588,415 Missing: 0
```

The file is one row per unique claim ID.

Comparison to reported statistics from website

The CMS report on this data was downloaded in PDF format from the web and read into R for comparison.

Extract table results from PDF report

Table 1. CMS reported inpatient use rates by gender and enrollment

V1	Months of Enrollment	Under 65	65 - 69	70 - 74	75 - 79	80 - 84	85 and older	Total
Female	12 months	21.14%	12.59%	15.20%	19.19%	23.09%	27.54%	19.29%
Female	Less than 12 months(2)	10.81%	5.85%	22.11%	33.39%	44.18%	50.49%	19.42%
Male	12 months	17.92%	12.64%	15.34%	18.95%	22.97%	26.92%	17.76%
Male	Less than 12 months(2)	11.52%	6.51%	24.67%	38.57%	47.94%	55.32%	18.18%
Total		17.93%	10.86%	15.80%	20.26%	25.01%	30.77%	18.64%

Table 2. Sample table on gender from codebook

Variable Value	Formatted Value	Frequency	Frequency (%)
1	Male	258,217	43.883%
2	Female	330,198	56.117%

Raw data-file comparison for sex

```
## BENE_SEX_IDENT_CD
```

```
## .
##      1      2 
## 258217 330198
```

The file loaded into R appears to be consistent with reported tables from CMS.

Assignment 1.2 Summary statistics

2. Run summary statistics on all variables (except the ID number); this is either a frequency distribution for a nominal or ordinal variable and means, standard deviations and percentiles for the continuous variables like expenditures, etc.

Summary Statistics

First some data-formatting; rename all variables to lower string, factor categories, set the codes to character strings vs. integers.

Data structure:

```
## 'data.frame':    588415 obs. of  7 variables:
##  $ ip_clm_base_drg_cd  : Factor w/ 311 levels "\"Heart transplant or implant of heart assist system\
##  $ ip_clm_icd9_prcdr_cd: Factor w/ 100 levels "'Not elsewhere classified'",..: 32 NA 55 NA 71 46 NA
##  $ ip_clm_days_cd      : Factor w/ 4 levels "'1 day'","'2-3 days'",..: 4 2 4 2 1 2 2 4 2 3 ...
##  $ ip_drg_quint_pmt_avg: int  86240 3447 34878 3007 3352 2690 5234 2713 9143 23354 ...
##  $ ip_drg_quint_pmt_cd : Factor w/ 5 levels "1","2","3","4",..: 4 2 5 2 2 1 3 2 5 5 ...
##  $ gender              : Factor w/ 2 levels "Male","Female": 2 2 1 2 2 1 1 2 1 2 ...
##  $ age_cat             : Factor w/ 6 levels "'Under  65 '",..: 4 5 1 2 2 1 3 2 1 3 ...
```

Tables 3. Gender

gender	n	percent
Male	258217	43.88
Female	330198	56.12

Tables 4. Age Category

age_cat	n	percent
'Under 65 '	116080	19.73
'65 - 69 '	77597	13.19
'70 - 74 '	86205	14.65
'75 - 79 '	91487	15.55
'80 - 84 '	94759	16.10
'85 & Older'	122287	20.78

Tables 5. Length of stay

ip_clm_days_cd	n	percent
'1 day'	76025	12.92
'2-3 days'	261419	44.43
'4-7 days'	122073	20.75
'8 or more days'	128898	21.91

Table 6. DRG hospitalization categories - Top 10 Codes

ip_clm_base_drg_cd	n	percent
"Heart failure & shock"	29374	4.99
"Simple pneumonia & pleurisy"	24317	4.13
"Major joint replacement or reattachment of lower extremity"	23111	3.93
"Chronic obstructive pulmonary disease"	22865	3.89
"Psychoses"	21248	3.61
"Septicemia w/o MV 96+ hours"	17904	3.04
"Rehabilitation"	17219	2.93
"Esophagitis, gastroent & misc digest disorders"	15226	2.59
"Cardiac arrhythmia & conduction disorders"	14695	2.50
"Kidney & urinary tract infections"	14127	2.40

Table 7. Procedures with DR - Top 10 Codes

ip_clm_icd9_prcdr_cd	n	percent
NA	276546	47.00
'Joint repair'	33100	5.63
'Other nonoperative proc'	29267	4.97
'Intest incis/excis/anast'	27553	4.68
'Other heart/pericard ops'	21613	3.67
'Vessel inc/excis/occlus'	21350	3.63
'Other ops on vessels'	19759	3.36
'Not elsewhere classified'	18535	3.15
'Non-op intubat & irrigat'	13530	2.30
'Other dx radiology'	11402	1.94

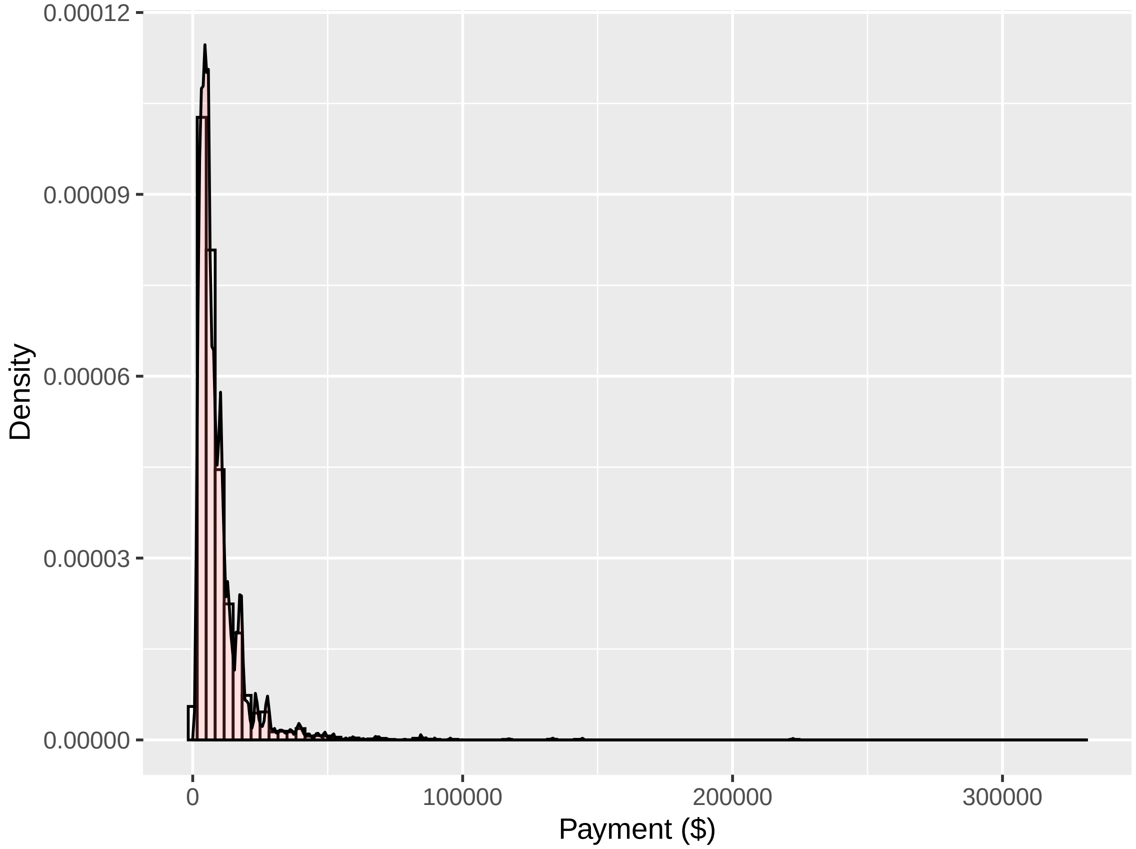
DRG Payment

```
## Variable: ip_drg_quint_pmt_avg
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      SD
##       0    4008    6352    9313   10760  329467   10483
```

Payment distribution

Figure 1. Distribution of Payments among Hospitalized Medicare Beneficiaries



Assignment 1.3 Dependent variable:

3. Compare some possible dependent variable that characterizes the utilization event (length of stay in days, expenditures, cost or some other analytic variable) by age categories of your choosing and sex.

We will use the `IP_DRG_QUINT_PMT_AVG` variable. Per the data dictionary:

> Average Medicare total claim payment amount of the quintile for the payments (of a particular DRG) in the 100% Inpatient claims data for 2008.

Restated this is the average payment for a person of equal sex, age, DRG category who is in the same quintile of payment distribution as the person in this limited file. So a reasonable approximation of the actual payment for that person.

Table 8. Average and standard deviation for DRG payments by gender, age

Gender	Age category	Mean ($)	SD ($)
Male	’Under 65 ’	9,632	11,809
Male	’65 - 69 ’	10,650	13,035
Male	’70 - 74 ’	10,519	12,032
Male	’75 - 79 ’	10,413	11,806
Male	’80 - 84 ’	9,745	10,830
Male	’85 & Older’	8,774	8,755
Female	’Under 65 ’	8,750	10,333
Female	’65 - 69 ’	9,343	10,640
Female	’70 - 74 ’	9,458	10,401
Female	’75 - 79 ’	9,285	10,154
Female	’80 - 84 ’	8,874	9,603
Female	’85 & Older’	7,981	7,384

Figure 2-4. Payments (mean, log, and 95% upper threshold)

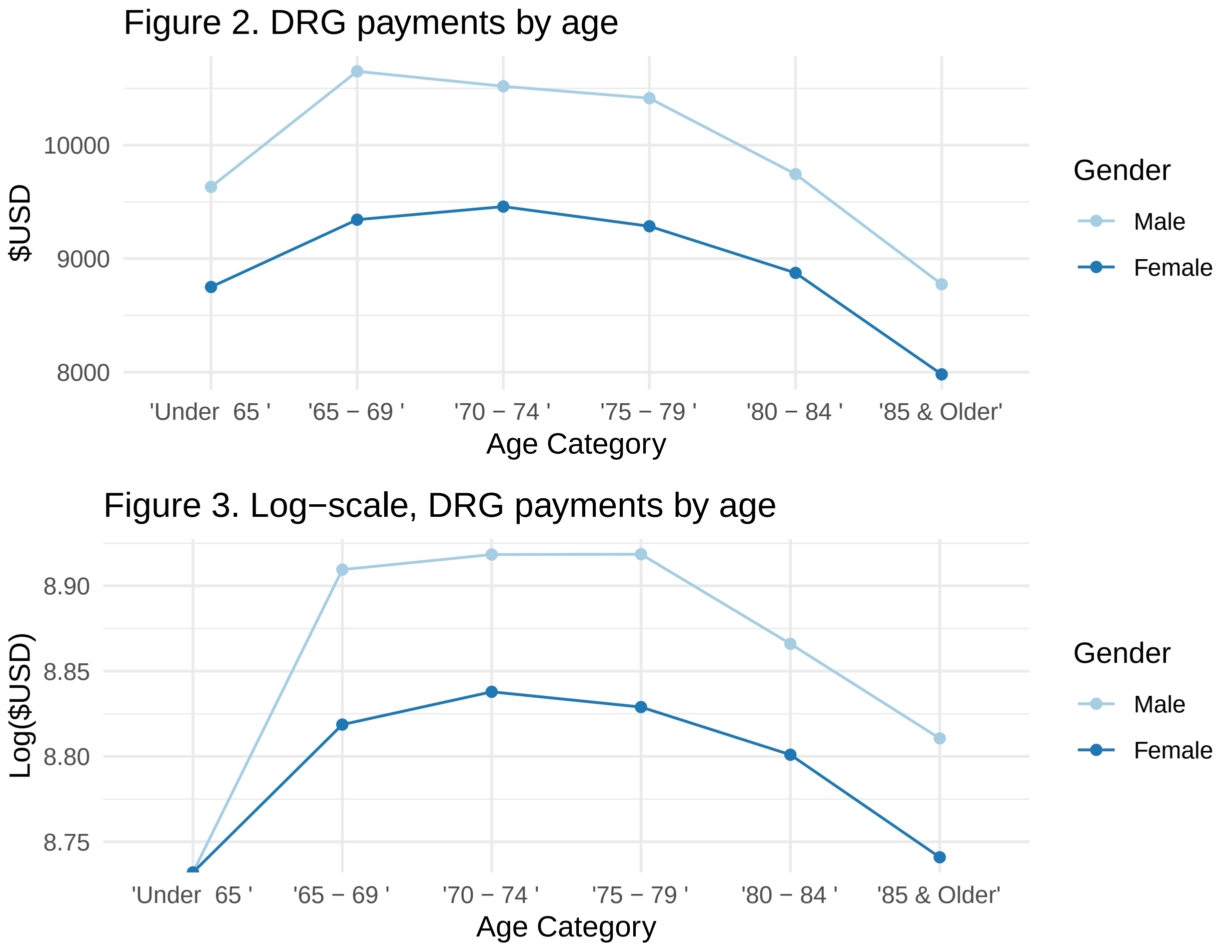
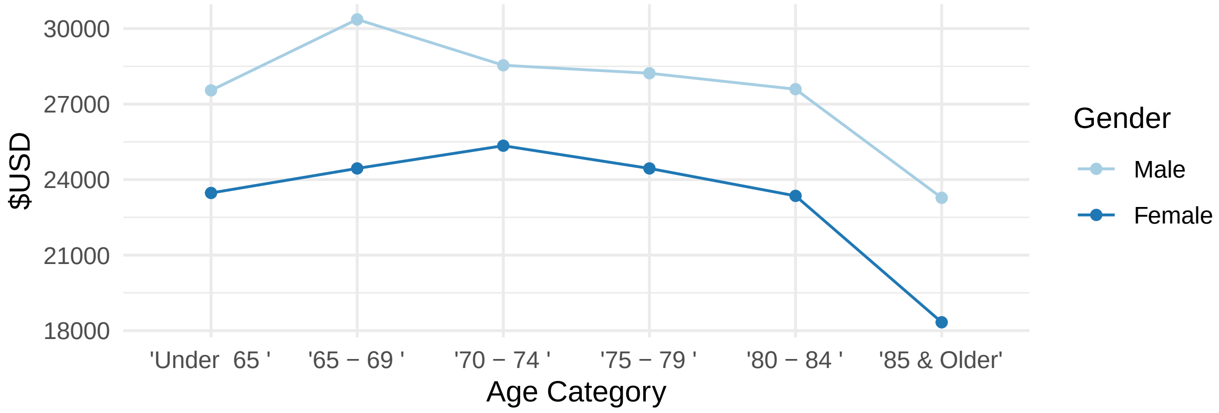


Figure 2. DRG payments by age

Figure 3. Log−scale, DRG payments by age

Figure 4. 95th percentile for DRG payments by age



Assignment 1.4 Summary Description

4. Write up the results of what you've done in no more than 3 paragraphs, referring to summary tables associated with task 2 or 3 above. This write up should, among other things, comment on the level of skew and variation in the analytic variables like length of stay or expenditures.

Introduction and methods

This analysis was an exploratory exercise conducted as part of a course on understanding and analyzing Medicare data for research purposes. A limited datafile which contained information on Medicare beneficiaries' hospitalizations in 2008 was downloaded from an online repository maintained by ResDAC, the data analysis center for the Center for Medicare and Medicaid Services (CMS). The file was formatted, summarized, and the relationship between avergae DRG-based payments, age and gender was evaluated in an informal exercise. The primary research aim was to evaluate how payment rates vary by age, gender and variability of the data. The comma-separated file was downloaded directly from the CMS website, formatted and edited for exploratory data analysis. Categorical variables were tabulated, while expenditures were evaluated using a 7-statistic summary and density histogram.

Results

The inpatient sample is 56% female, the median age cannot be reported for privacy purposes but the most common age groups are those under 65 (19.7%) and those 85 and older (20.8%). Most hospitalizations occur over 2-3 days, and the top 10 DRG conditions are listed in Table 6. These include heart failure (5.0%), pneumonia (4.1%), joint replacement (3.9%) and Chronic Obstructive Pulmonary Disease (COPD, 3.9%). Approximately half of inpatient admissions do not have a primary ICD9 procedure code, the most common are joint repairs (5.6%), and other cardiovascular, abdominal or non-specific surgical procedures. The limited file truncates the procedure code at two digits (i.e. detail is limited). The inpatient file reports payments for each individual according to average payments of like individuals (gender, age, procedure, actual payment). The mean payment is $9,313, median $6,352, suggesting a right skewed distribution which is confirmed by the histogram (Figure 1). Table 8 shows the mean payments by gender and age category. The average expenditure is lowest for those youngest and oldest in the sample, with men generally higher than women. The log-scale follows a similar relationship, while 'high-utilizers', those in the 95th percentile, seem less variable by age.

Discussion

The sample represents *hospitalized* Medicare beneficiaries so may not be representative of the whole Medicare cohort or those with managed care plans. The results reveal several important points. Men generally have higher DRG-based payments than women, and those in either age extreme have lower DRG payments then those 65-85. This could be explained by rates of important DRG categories and surgeries which are higher in men and those <85 years such as myocardial infarcts, heart failure or COPD. Younger and older individuals may have different reasons for lower costs; e.g. shorter stays, different DRG groups (particularly less surgical procedures in the oldest cohort). These hypotheses are testable but would require a more in depth evaluation of the interplay between age, gender, and particular diagnosis groups to understand the importance of underlying condition and its prevalence in the general Medicare cohort.

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     2019-09-18
```