Capstone Project Milestone - Data Story
Foundations of Data Science Workshop

Tanvi Desai
October 2016

Hospital Inpatient Charges for top 10 DRGs in the U.S.

Background

Medicare is a federally-funded health insurance program that pays for health care (hospital and medical) expenses, typically for senior citizens aged 65 and older, adults with certain approved medical conditions or qualifying permanent disabilities. The program is administered by the Centers for Medicare and Medicaid Services (CMS), which is a division of the U.S. Department of Health and Human Services.

For payment purposes, inpatient stays for patients are divided into groups by means of a statistical classification system known as Diagnosis-related Groups(DRG). All possible diagnoses are divided into more than 20 major body system groups, which are subdivided into almost 500 groups for the purpose of Medicare reimbursement. In addition to the diagnoses involved, factors such as hospital resources necessary to treat the condition, are used to determine the DRG payment amount. For every DRG assigned to a given patient, the hospitals are paid a fixed amount for the inpatient services involved.

Client

As the agency monitoring payments to hospitals for medicare eligible inpatient expenses, the CMS would potentially be interested in knowing whether certain hospitals are overcharging for the same diagnoses as compared to others. Charges and payments may differ based on provider state, number of discharges, income groups of the patients as well as whether the hospital is a teaching hospital or not. Based on the analysis of billing and payment information, the CMS would determine whether the payment rates need to be modified for particular providers.

Data

The hospital inpatient charges dataset for the top 10 DRGs in the U.S. is owned by the US government and is freely available on Data.gov. It is also available on Kaggle.com.
The dataset consists of the following information:

- DRG Code and definition
- Provider ID and Name

- Provider Street Address
- Provider City
- Provider State
- Provider Zip Code
- Hospital Referral Region Description
- Total Discharges
- Average Covered Charges
- Average Total Payments
- Average Medicare Payments

The DRG Codes, Average Covered Charges and Average Medicare Payments are the most important variables in the dataset in order to understand the billing and payment differences across the providers. These may differ across individual providers as well as across states.

The lack of time series data and limited variables are some of the limitations of this dataset. Due to the lack of this information, we are unable to make any predictions regarding charges and payments. It also lacks information regarding different characteristics of the listed institutions such as number of people served, whether they are teaching hospitals, etc. This limits analysis of factors influencing the charges and payments.

Goal and Objectives

The goal of the project is to gain insights into the inpatient billing and payment information so as to aid the CMS in making policy decisions related to payment rates assigned per DRG for each provider.

The specific objectives of the analysis are:

1. To determine the distribution of charges for particular DRGs across states
2. Visualize the differences in charges for DRGs across states
3. To assess the differences in Medicare reimbursement for the top 10 billed DRGs across states

Data will be accessed and cleaned. Data visualization techniques will be utilized for data exploration. Clustering will be performed in order to identify patterns within the dataset. Linear regression will be performed to discover associations between charges and other variables in the dataset.

Data Wrangling

The dataset was fairly clean with no missing variables or extreme outliers. The variable names were long and were trimmed for ease of coding. The DRG variable contained both DRG code as well as definition. For example, “176 - PULMONARY EMBOLISM W/O MCC”. These we split into two separate variables, “code” and “definition”.
Three variables(Average.Covered.Charges, Average.Total.Payments, Average.Medicare. Payments) were character variables and the observations contain "$". These were converted to numeric and the "$" was removed.

Preliminary Analysis

The dataset has 163,065 observations and 11 variables.

The following table provides the summary information for the numeric variables:

Table 1: Summary information

Variable Name	**Mean**	**Minimum Value**	**Maximum Value**
Average Covered Charges	$36,134	$2,459	$929,119
Average Medicare Payments	$8,494	$1,149	$154,621
Average Total Payments	$9,707	$2,673	$156,158
Total Discharges	43	11	3383

The dataset lacks information such as provider characteristics and longitudinal data and therefore does not allow for meaningful regression or time-series analyses. Therefore, for the purposes of this project, we will focus on descriptive analysis and data visualization.