IPPS Hospital Charge Data

Dan Matthews

The dataset I am working with shows the hospital charges for the 100 most frequently billed discharges. It includes data for over 3,000 hospitals in the United States and categorizes the procedures by Diagnosis Related Group codes, or DRG's. Procedures for a given DRG are expected to use a similar amount of hospital resources. Each row describes the charges and payments for a DRG at a particular provider. There are data for 100 different DRG's.

The main column of interest in this dataset is the "Mean Covered Charges" as that is what the hospital charged on average for the given category of procedures. There is also a "Mean Total Payments" column, which is the payments made by Medicare to the hospital, and while I found that on average these payments were around 20-30% of the charged cost, in some cases the mean of payments were actually more than the mean cost. This occurred in less than 1% of the rows in the dataset. For the purposes of this analysis I focused mainly on the Mean Covered Charges, as that seemed to vary dramatically depending on the hospital.

To get a better handle on what is actually contained in the dataset I looked at properties such as the number of hospitals and number of procedures. Each provider is designated by a provider ID so I selected all unique ID's and found there were 3,337 hospitals. I also found there were 100 different DRG categories, as it said in the description of the dataset. The address of each hospital is given in the data, so I was also curious about the number of procedures reported both by hospital and by state. I constructed a 2-D grid that shows the number of each procedure in each state. I used this to make bar graphs of the total number of each procedure nationwide and the number of procedures in each state. These are the first and second graphs in the notebook.

I was also curious about how the Mean Covered Charges compared to the Mean Total Payments, so I chose one procedure and made a scatter plot of these two quantities and found that while the Mean Total Payments is somewhat consistent for all providers, the Mean Covered Charges varied wildly depending on the hospital. This is shown in the third graph in the notebook. Next I looked at how the Mean Covered Charges varied for different procedures. My first thought was to look at the median value of the reported Mean Covered Charges nationwide for each procedure. The next graph in the notebook shows a bar graph with this median value for each procedure, as well as the median of the reported Mean Payments.

After looking at the median value for each procedure I wanted to see how the charges reported by each hospital compared to this national median. For each line in the dataset I calculated the fractional difference between the reported charge and the national median. I calculated the fractional difference to put all of the numbers on somewhat equal footing, since a given difference can have a very different significance depending on the size of the total cost. For each provider I calculated the average fractional difference in cost from the national median over all procedures. In effect this shows whether a provider tends to charge above or below the national median for all procedures. I made a scatter plot showing this quantity as a function of Provider ID, which is shown at the end of the notebook. It turns out that the points tended to cluster according to their Provider ID, and I found this was because the Provider ID actually gives an indication of what state the provider is located in. For instance the Provider ID's for providers in Alabama are all around 10000, Alaska's are around 20000, and so on. I labeled each cluster by state in the final plot in the notebook.

Overall this is an interesting dataset. The main point I took away is that on average the cost of a given procedure does seem to vary depending on the state the provider is located in, as can be seen in the final two plots of the notebook. For instance the cost in California and New Jersey seems to be significantly higher than in other states. As a next step I would like to make a better visualization of this

tendency, such as a chloropleth map of the different states.