# Homework 2: Submission 3

**Research Methods, Spring 2024**

Leila Mulveny

Homework 2: Repository

## Enrollment Data

Answer the following based on the enrollment data:

1. How many hospitals filed more than one report in the same year? Show your answer as a line graph of the number of hospitals over time.

After first identifying hospitals that filed more than one report in the same year by summarizing the dataset to count the number of distinct provider numbers for which duplicates exist, I created a summary dataframe that groups the data by fiscal year (fyear) and calculates the count of distinct hospitals for each year. The number of hospitals that filed more than one report in the same year can be visualized in Figure 1.

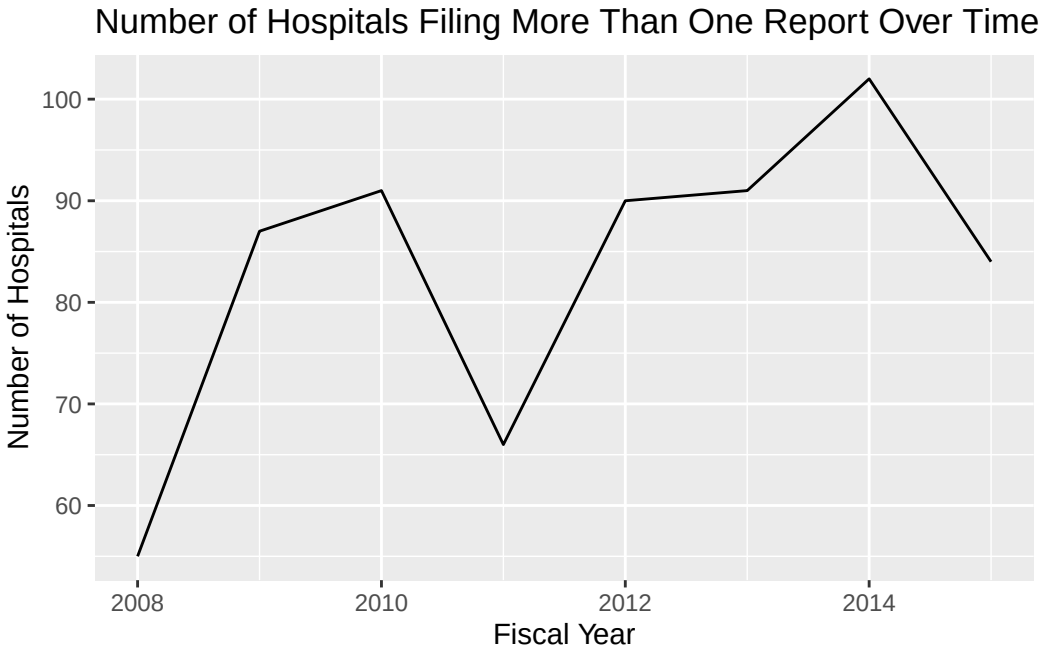


Figure 1: Number of Hospitals Over Time

2. After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

After removing/combining multiple reports, there exist 6747 unique hospital IDs in the data.

3.What is the distribution of total charges (tot_charges in the data) in each year? Show your results with a "violin" plot, with charges on the y-axis and years on the x-axis.

For this question I ended up filtering out outliers in the data to get a more presentable violinplot (without filtering the distributions were stretched too thin, making the plot difficult to interpret). I dropped data for the years 2007 and 2016, as they contained a lot of outliers, and don't pertain to the rest of the assignment. I then grouped the data by year and filtered out extreme values by identifying the 1st and 99th percentiles of the observations, and then removing them. Additionaly, I had to change my x-axis data to as.factor(year) so that each year would be separated as a 'factor' instead of grouped together.

The distribution of total charges in each year can be visualized in: Figure 2.

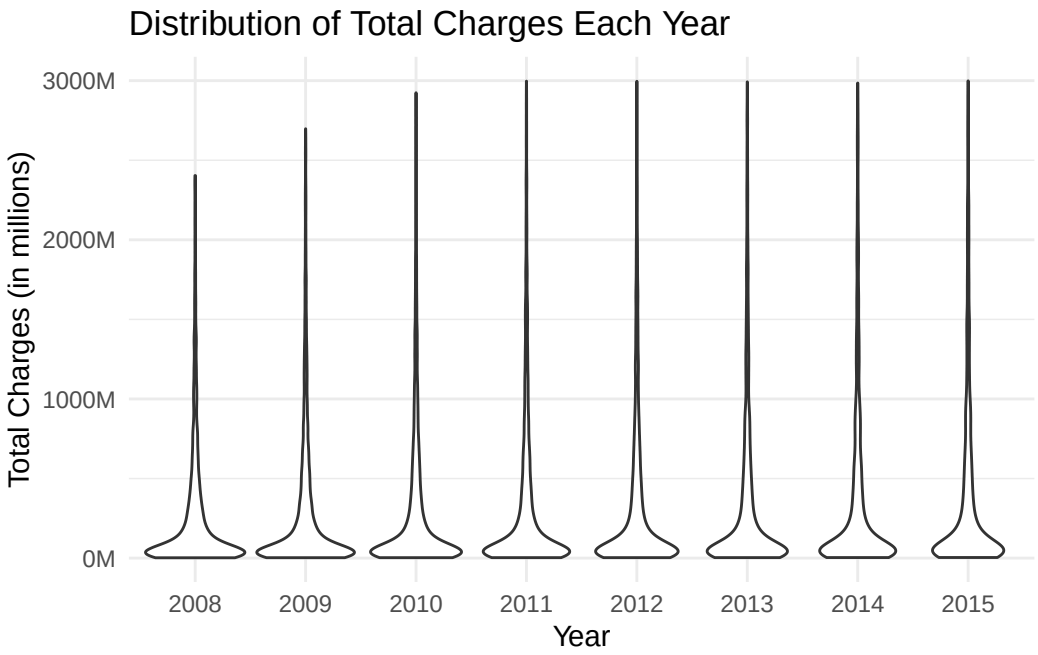


Figure 2: Distribution of Total Charges by Year

4. What is the distribution of estimated prices in each year? Again present your results with a violin plot, and recall our formula for estimating prices from class. Be sure to do something about outliers and/or negative prices in the data.

After defining 'price' and removing outliers and negative prices in the data, the distribution of estimated prices in each year can be visualized in Figure 3.

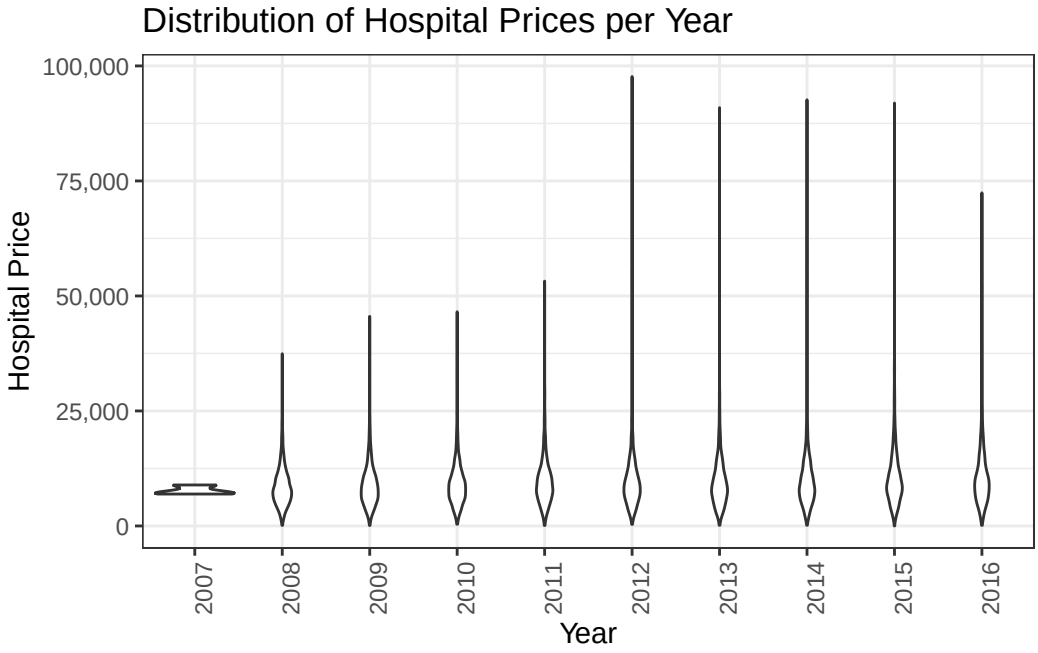


Figure 3: Estimated Prices in Each Year (Excluding Outliers)

## Estimate ATEs

For the rest of the assignment, you should include only observations in 2012. So we are now dealing with cross-sectional data in which some hospitals are penalized and some are not. Please also define penalty as whether the sum of the HRRP and HVBP amounts are negative (i.e., a net penalty under the two programs).

5. Calculate the average price among penalized versus non-penalized hospitals.

The mean price among penalized hospitals is $9896.3084978. The mean price among non-penalized hospitals is $9560.4132266. The difference between mean prices for penalized versus non-penalized hospitals is $335.8952712.

6. Split hospitals into quartiles based on bed size. To do this, create 4 new indicator variables, where each variable is set to 1 if the hospital's bed size falls into the relevant quartile. Provide a table of the average price among treated/control groups for each quartile.

| Quartile | Average Price (Treated) | Average Price (Control) |
|:---:|:---:|:---:|
| 1 | 8319 | 7684 |
| 2 | 8691 | 8511 |
| 3 | 10127 | 9857 |
| 4 | 12068 | 12356 |

7. Find the average treatment effect using each of the following estimators, and present your results in a single table:

Nearest neighbor matching (1-to-1) with inverse variance distance based on quartiles of bed size Nearest neighbor matching (1-to-1) with Mahalanobis distance based on quartiles of bed size Inverse propensity weighting, where the propensity scores are based on quartiles of bed size Simple linear regression, adjusting for quartiles of bed size using dummy variables and appropriate interactions as discussed in class

Results:

| Estimator | ATE |
|:---|:---:|
| Nearest Neighbor (Inverse Variance Distance) | 190.1 |
| Nearest Neighbor (Mahalanobis Distance) | 190.1 |
| Inverse Propensity Weighting | 190.1 |
| Simple Linear Regression | 190.1 |

8. With these different treatment effect estimators, are the results similar, identical, very different?

The results are identical across all four separate methods, yeilding a positive average treatment effect estimator of about 190.

9. Do you think you've estimated a causal effect of the penalty? Why or why not? (just a couple of sentences)

The estimated ATE of 190.1416 across all estimators may not adequately represent a causal effect of the penalty. The estimation did not fully account for potential biases introduced by unobserved factors influencing both the imposition of penalties and hospital prices. If bed size is indeed correlated with both the treatment (penalties) and the outcome (hospital prices), it introduces the possibility of endogeneity or reverse causality. For example, larger hospitals (larger bed size) may have higher prices not solely because of the imposition of penalties but also because they provide better-quality care or offer specialized services, which naturally command higher prices.

10. Briefly describe your experience working with these data (just a few sentences). Tell me one thing you learned and one thing that really aggravated or surprised you.

I still think that the ungroup() function is one of the key things I learned from this assignment just because I didn't realize the effect that groupby() was having on my code (my dataset) throughout the assignment. It is just another thing that I am now going to be aware of going forward, hopefully making debuging future assignments easier.

One of the more aggravating parts of this assignment, for me, was in working with the violin plots for questions 3 and 4. I don't think I fully understood just what the violin plot was doing and therefore displaying, so I did not realize where I could improve my code to generate a more presentable plot. After last class I definitely understand how the violin plot represents distributions, so I was therefore able to go back into my code and remove the outliers that were skewing my results.