Homework 2

Research Methods, Spring 2025

Answer Key

My answers to the homework questions are described below. As with the first homework assignment, note that my analysis is in a separate `R` script. My analysis file is available here.

Summarize the 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.

As discussed in the ReadMe file for the HCRIS GitHub repository, there are a few reasons why hospitals may submit more than one report in the same year. We try to deal with this in the final data, but it's still good to get a sense of how often this occurs. Figure 1 presents a line graph of the number of hospitals submitting duplicative reports each year.

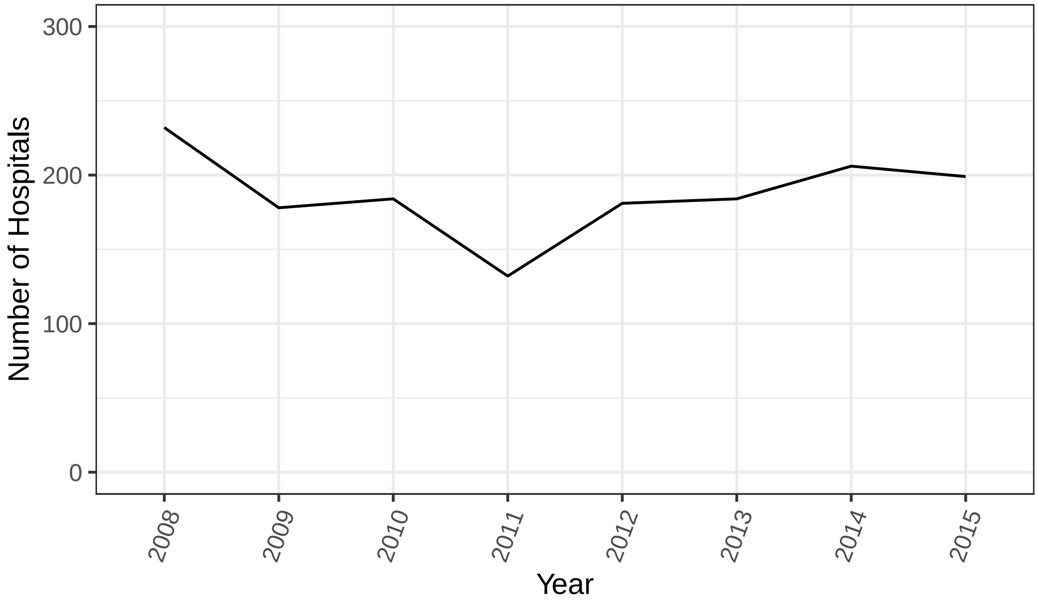

Figure 1: Duplicate Reports

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

The count of unique hospitals by year is presented in Figure 2.

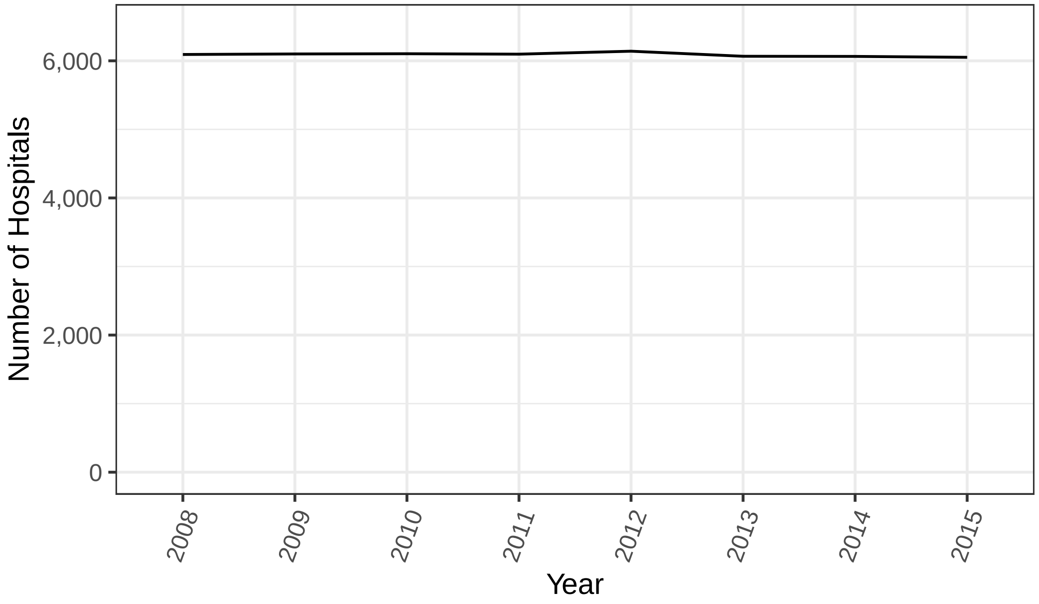

Figure 2: Unique Hospitals

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-asix. For a nice tutorial on violin plots, look at Violin Plots with ggplot2.

A violin plot of total charges is presented in Figure 3. The figure presents charges in logs after dropping the highest and lowest 1 percent in each year.

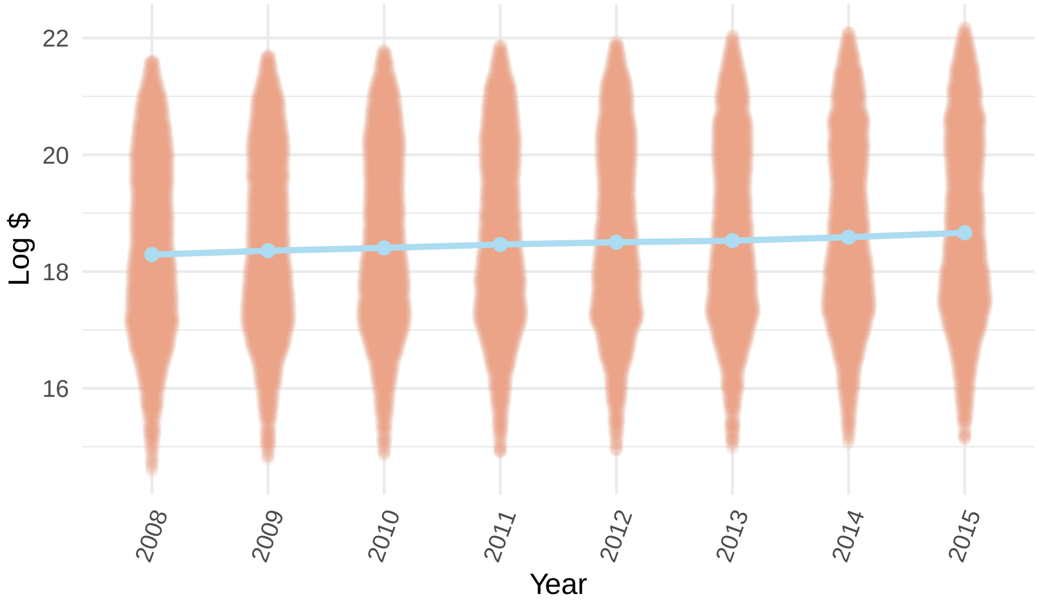

Figure 3: Distribution of Hospital Charges

4. Create the same violin plot with estimated prices on the y-axis.

A violin plot of estimated prices is presented in Figure 4. Similar to the analysis of charges, we need to do some cleanup before we plot the data. In this case, I've removed hospitals with fewer than 100 discharges, total charges of 0, bed sizes fewer than 30, and with an average price of greater than 100,000.

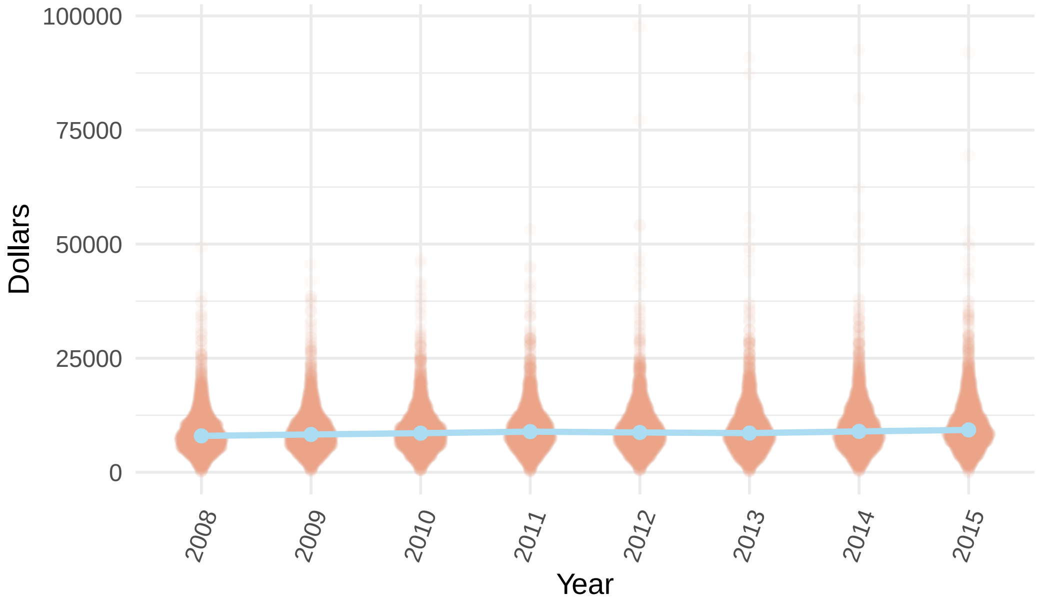

Figure 4: Distribution of Hospital Prices

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). Code to do this is in the Section 2 slides.

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

The average price among non-penalized hospitals in 2012 is 9,562.865 and the average price among penalized hospitals is 9896.308.

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.

There are lots of ways to do this, but to me, the easiest way is to create 4 new variables that represent the values of each quartile. Then, we can create indicator variables for whether the bed size of a given hospital falls into the relevant range. My code for this is in the "run" file. The resulting table of mean prices by penalty status, within each quartile, is presented below in Table 1.

Table 1: Mean prices by penalty status and bed size quartile

Bed Size	Not Penalized	Penalized
1	7,678.757	8,355.043
2	8,507.619	8,662.349
3	9,869.173	10,102.451
4	12,367.332	12,068.479

7. Find the average treatment effect using the estimators listed in the homework, and present your results in a single table.

Results are presented in Table 2. Note that all point estimates *should* be identical provided you are using the same data in each step. If your estimates differ across estimators, then it's probably because you have some missing observations somewhere that are dropped in one analysis and not in another.

Table 2: ATE Estimates

	INV	MAH	IPW	OLS
Penalty	191.230 (236.533)	191.230 (236.533)	191.230 (214.027)	191.230 (240.185)

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

We've tried lots of different estimators, and in all cases, we've gotten the exact same answer! That's pretty cool and shows us how these estimators are all trying to do the same things. Note that the equivalence between these estimators is not true in general…it's only because we're considering dummy variables as our only covariates. If we had continuous variables as covariates, then these different estimators would be similar but not identical as they are here (which is what we found in class)

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

I would **not** claim that we've estimated a casual effect of the penalty. Mainly, this is just cross-sectional data, and we know there are lots of systematic differences between hospitals that likely affect price and are also correlated with penalties. So I would not believe any claim that we meet the "selection on observables" assumption.

10. Briefly describe your experience working with these data.

No wrong answers here. I'm just looking for you to reflect a little bit on this assignment.