Homework Two Submission 2

AUTHOR
Safia Read

This is my third submission of the second homework for Econ 470.

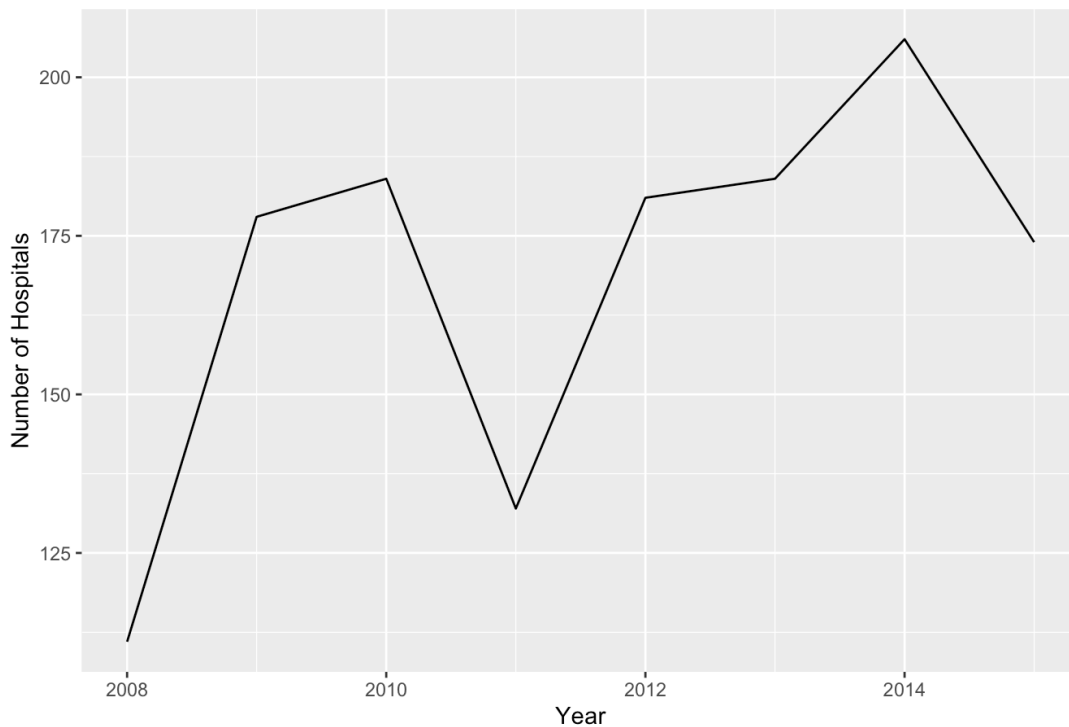Link to Github

https://github.com/safiaread/homework-2

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.

Number of Hospitals Filing Multiple Reports in a Year Over Time

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

There are 6719 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.

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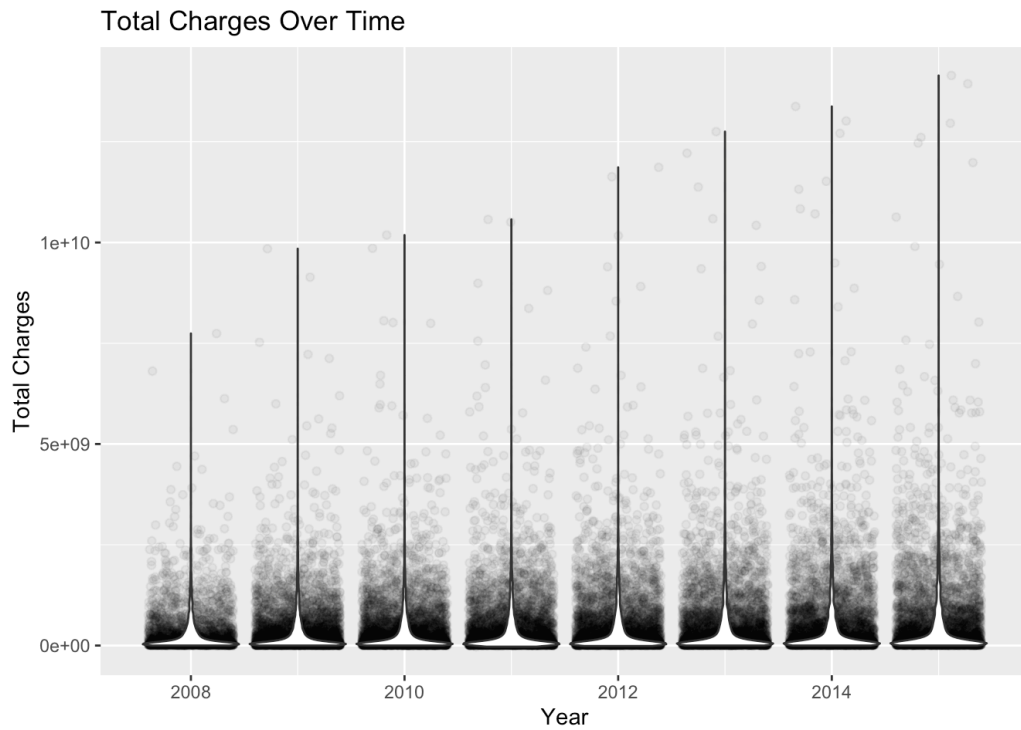
4. What is the distribution of estimated prices in each year? Again present your results with a violin plot, and recall our formula for estimating prices from class. Be sure to do something about outliers and/or negative prices in the data.

Estimated Prices Over Time
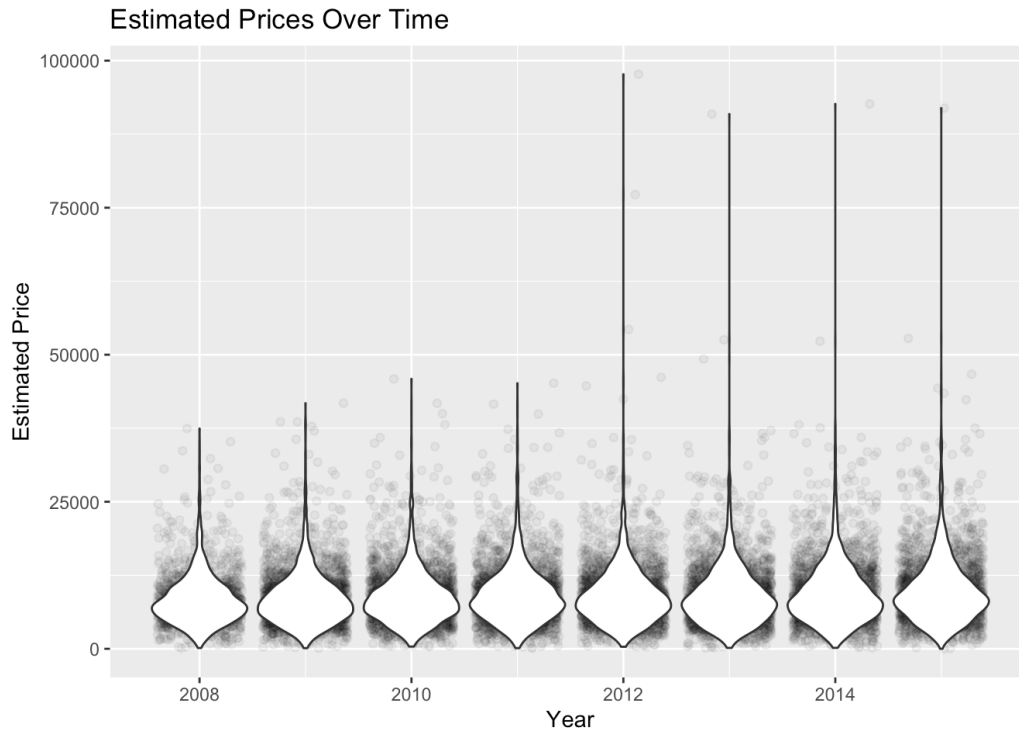
Estimate ATEs

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

Penalty	Average Price
0	9,220.558
1	9,463.758

Average Price among Penalized versus Non-Penalized Hospitals

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.

Penalty	Average Price of First Quartile Beds	Average Price of Second Quartile Beds	Average Price of Third Quartile Beds	Average Price of Fourth Quartile Beds
0	7,692.414	8,729.654	9,512.582	11,191.36
1	7,640.886	8,812.637	9,863.944	10,900.33

Average Price among Penalized versus Non-Penalized Hospitals By Bed Size

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

a. Nearest neighbor matching (1-to-1) with inverse variance distance #based on quartiles of bed size

```
Increasing memory because of ties: allocating a matrix of size 3 times :
I would be faster with the ties=FALSE option.



Estimate...  21.661
AI SE......  241.08
T-stat.....  0.089847
p.val......  0.92841

Original number of observations..............  1880
Original number of treated obs..............  483
Matched number of observations..............  1880
Matched number of observations  (unweighted).  335084
```

b. Nearest neighbor matching (1-to-1) with Mahalanobis distance based on quartiles of bed size

```
Increasing memory because of ties: allocating a matrix of size 3 times :
I would be faster with the ties=FALSE option.



Estimate...  21.661
AI SE......  241.08
T-stat.....  0.089847
p.val......  0.92841

Original number of observations..............  1880
Original number of treated obs..............  483
Matched number of observations..............  1880
Matched number of observations  (unweighted).  335084
```

c. Inverse propensity weighting, where the propensity scores are #based on quartiles of bed size

The IPW estimate is 21.6606323.

d. Simple linear regression, adjusting for quartiles of bed size using dummy variables and appropriate interactions as discussed in class

```
Call:
lm(formula = price ~ penalty + first_quartile + second_quartile +
    third_quartile + fourth_quartile + fq_diff + sq_diff + tq_diff +
    foq_diff, data = reg.dat)

Residuals:
   Min     1Q Median     3Q    Max
 -9770  -3015   -777   2023  35986

Coefficients: (2 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      11191.36     253.14  44.210  < 2e-16 ***
penalty             21.66     245.09   0.088    0.930
first_quartile   -3498.95     347.33 -10.074  < 2e-16 ***
second_quartile  -2461.71     350.94  -7.015 3.21e-12 ***
third_quartile   -1678.78     357.72  -4.693 2.89e-06 ***
fourth_quartile        NA         NA      NA       NA
fq_diff            239.51     704.63   0.340    0.734
sq_diff            374.02     672.68   0.556    0.578
tq_diff            642.40     667.07   0.963    0.336
foq_diff               NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4605 on 1872 degrees of freedom
Multiple R-squared:  0.06963,   Adjusted R-squared:  0.06615
F-statistic: 20.01 on 7 and 1872 DF,  p-value: < 2.2e-16
```

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

All four estimates are very similar, around 21.66.

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

I don't think so because only matching on bed size does not make the treatment and control groups functionally similar. There are still many other confounders we are not controlling for.

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 dropped the NAs so everything displays better and figured out how to get all my matching code to work. I was able to troubleshoot a lot of those issues I had with my code succesfully. The hardest thing this time was rendering my qmd files to pdf format. My Latex says it is 2022, not 2023 so won't render. However, Latex says it has no updates and I tried uninstalling and reinstalling but it still didn't change.