Submission2-HW2

Research Methods, Spring 2024

Caroline Hansen

https://github.com/carolinezhansen/HCRIS/tree/main

Answers for HW 2

Summarize the 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.

First I use the "deplicate hcris" data to find the count of the hospitals with more than one report in the same fiscal year. Then I used the ggplot function to create the line graph.

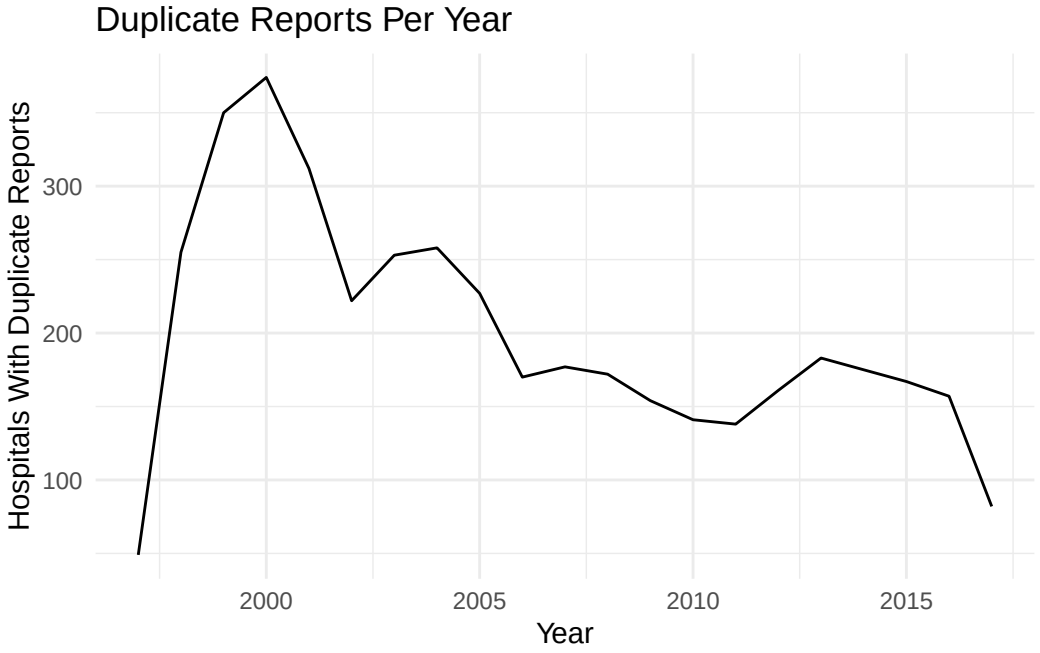

Figure 1: Duplicate Hospital Reports per Year

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

```
# A tibble: 21 x 2
   fyear num_hospitals
   <dbl>         <int>
 1  1997          1332
 2  1998          6072
 3  1999          6032
 4  2000          6018
 5  2001          6031
 6  2002          6016
 7  2003          6036
 8  2004          6125
 9  2005          6165
10  2006          6068
# i 11 more rows
```

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..

see @ totcharges

```
Warning: Removed 4748 rows containing non-finite values (`stat_ydensity()`).
```

```
Warning: Removed 4748 rows containing missing values (`geom_point()`).
```

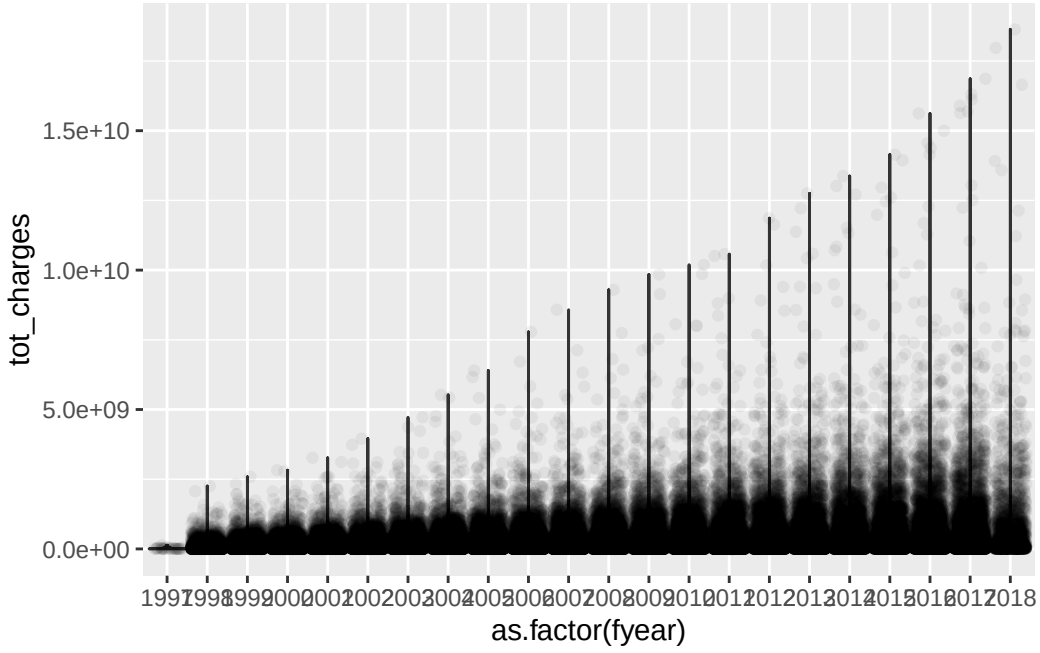

Figure 2: Unique Hospital Reports per Year

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.

```
Warning: Removed 1004 rows containing non-finite values (`stat_ydensity()`).
```

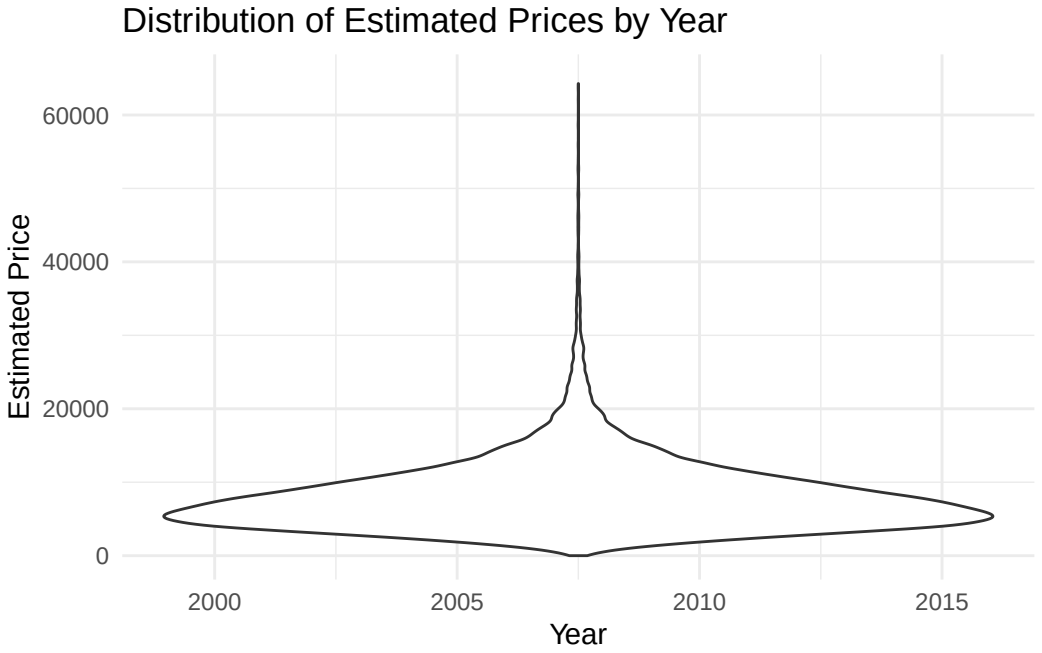

Figure 3: Estimate Prices Violin Plot

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 class slides.

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

```
# A tibble: 57,739 x 5
# Groups:   provider_number [4,033]
   provider_number fyear mean.pen na.rm mean.nopen
   <chr>           <dbl>    <dbl> <lgl>      <dbl>
 1 010001           1998    9863. TRUE       9397.
 2 010001           1999    9863. TRUE       9397.
 3 010001           2000    9863. TRUE       9397.
 4 010001           2001    9863. TRUE       9397.
 5 010001           2002    9863. TRUE       9397.
 6 010001           2003    9863. TRUE       9397.
 7 010001           2004    9863. TRUE       9397.
 8 010001           2005    9863. TRUE       9397.
 9 010001           2006    9863. TRUE       9397.
10 010001           2007    9863. TRUE       9397.
# i 57,729 more rows
```

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.

```
# A tibble: 2 x 3
  quartile avg_price_treated avg_price_control
     <int>             <dbl>             <dbl>
1        1             9863.             9397.
2       NA             9863.             9397.
```

summary table

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.

I created the matching and weighting for all 4 estimators, but still have to put it in a table. There is an attempt at this problem, but I am unable to put it in a quarto document.

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

The results are every different for each estimate.

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

I think that it is hard to see the causal effect due to the large difference in all of the estimators. While we were able to see the impact of the penalty, I am still confused about which estimator I should rely on to look at causal effect.

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 you.

I am getting better and creating graphs and sending it to quarto and github. I was really aggrivated when I was trying to push things to quarto and one small error would make the entire document not run.