

Homework Two Submission 2

Safia Read

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

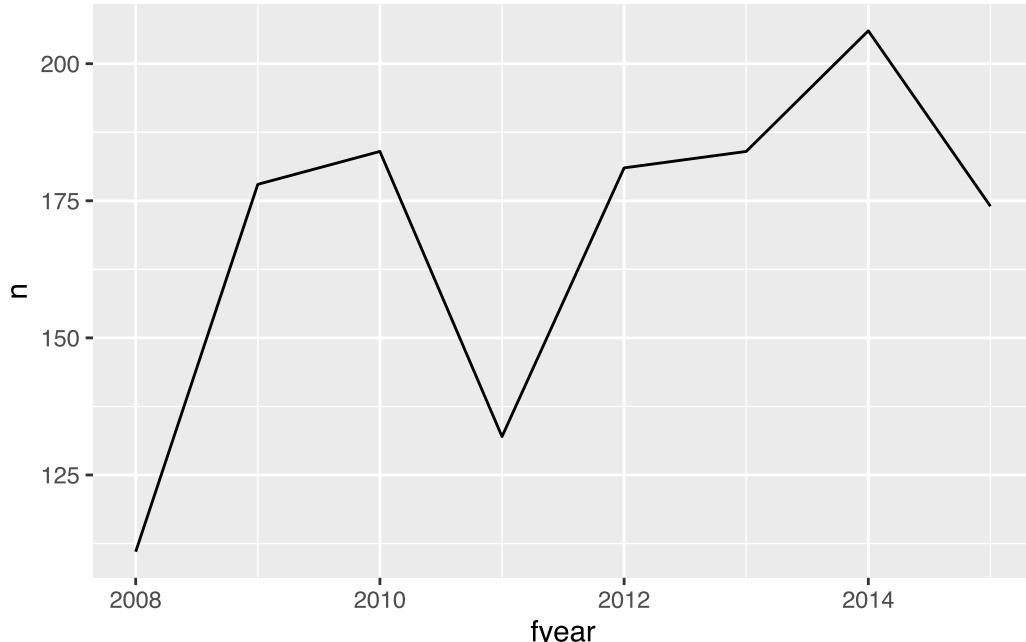
Link to Github

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

```
Loading required package: pacman
```

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.



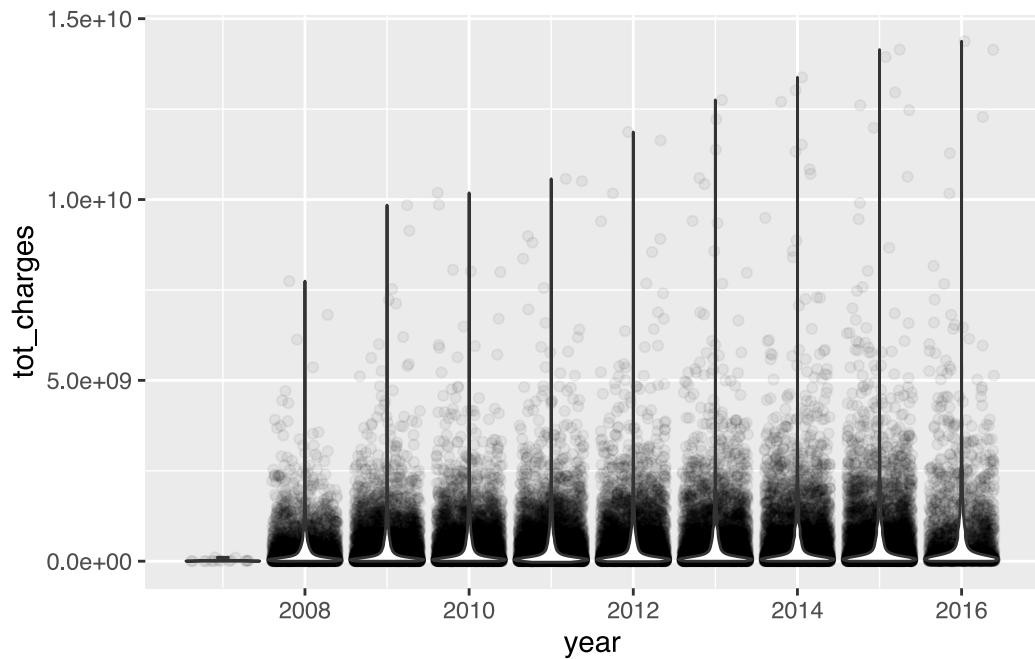
2. After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

```
[1] 6747
```

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.

```
Warning: Removed 1832 rows containing non-finite values (`stat_ydensity()`).
```

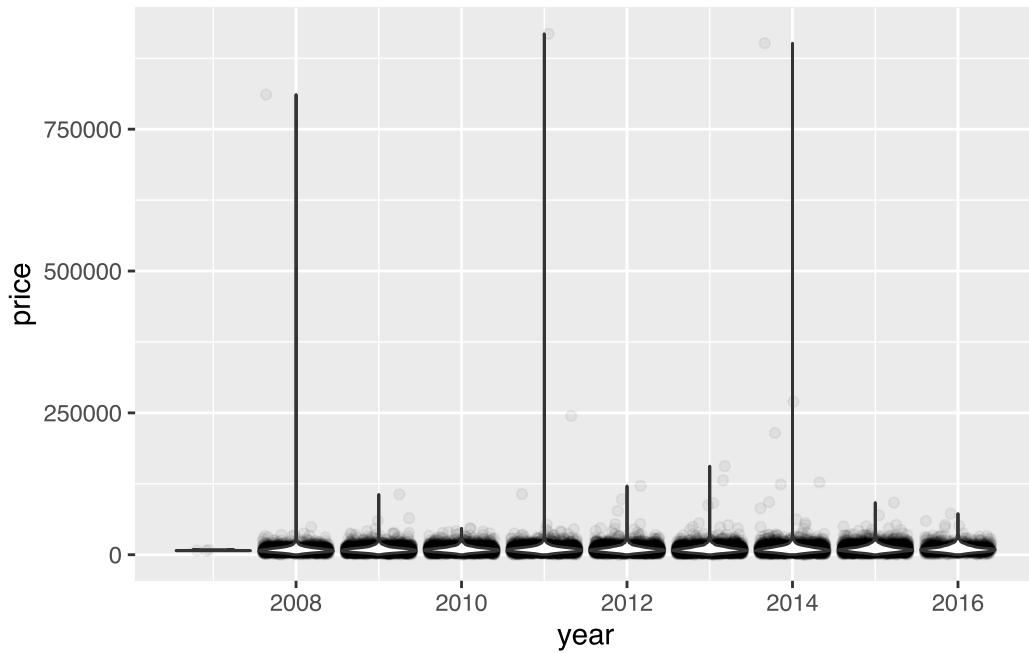
```
Warning: Removed 1832 rows containing missing values (`geom_point()`).
```



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 26829 rows containing non-finite values (`stat_ydensity()`).
```

```
Warning: Removed 26829 rows containing missing values (`geom_point()`).
```



Estimate ATEs

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

```
# A tibble: 2 × 2
  penalty avg_price
  <dbl>     <dbl>
1     0      9864.
2     1      9796.
```

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.

0	1
410	137

```
# A tibble: 2 × 5
  penalty first_mean second_mean third_mean fourth_mean
  <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1     0      7481.    9332.    10074.    12435.
2     1      7297.    9079.    10598.    12195.
```

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

```
library("dplyr")
library("tidyverse")
hcris.vars <- hcris_2012 %>%
ungroup() %>%
  select(penalty, price, first_quartile, second_quartile, third_quartile,
fourth_quartile) %>%
  na.omit()
hcris.covs <- hcris.vars %>%
select(first_quartile, second_quartile, third_quartile, fourth_quartile)

m.nn.var2 <- Matching::Match(Y=hcris.vars$price,
                               Tr=hcris.vars$penalty,
                               X=hcris.covs,
                               M=1,    #<<
                               Weight=1,
                               estimand="ATE")

summary(m.nn.var2)
```

```
Estimate... -41.133
AI SE..... 547.01
T-stat..... -0.075196
p.val..... 0.94006

Original number of observations..... 547
Original number of treated obs..... 467
Matched number of observations..... 547
Matched number of observations (unweighted). 18476
```

```
#v.name=data.frame(new=c("Beds", "Medicaid Discharges", "Inpatient Charges",
# "Medicare Discharges", "Medicare Payments"))
```

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

```
Estimate... -41.133
AI SE..... 547.01
T-stat..... -0.075196
p.val..... 0.94006
```

```

Original number of observations..... 547
Original number of treated obs..... 467
Matched number of observations..... 547
Matched number of observations (unweighted). 18476

```

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

```
[1] -41.13285
```

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 
-9309  -2828   -744   1801  26478 

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12435.12    903.74 13.760 < 2e-16 ***
penalty      -41.13     548.15 -0.075 0.940211    
first_quartile -4953.67   1325.15 -3.738 0.000205 ***  
second_quartile -3102.66   1240.67 -2.501 0.012688 *   
third_quartile  -2360.67   1588.87 -1.486 0.137929    
fourth_quartile       NA       NA       NA       NA      
fq_diff        55.28    1442.69  0.038 0.969451    
sq_diff       -13.96    1368.13 -0.010 0.991860    
tq_diff        763.16   1685.47  0.453 0.650884    
foq_diff       NA       NA       NA       NA      
...
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4334 on 539 degrees of freedom
Multiple R-squared:  0.1508,    Adjusted R-squared:  0.1398 
F-statistic: 13.67 on 7 and 539 DF,  p-value: 2.505e-16

```

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

All four estimates are very similar, around -41.13.

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 also had a little trouble figuring out how to get the code for the first to display since I wrote it in the HCRIS_Data.rds folder. I have to format everything as a workspace so the code is neater. The code I am working on is displayed so I can track my progress, but will be removed for the final submission. I also need toneaten my graphs and tables.