Homework Two Submission 1

Safia Read

This is my first 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.

nrow(duplicate.hcris) q1 <- duplicate.hcris%>% group_by(fyear)%>% count() ggplot(q1, aes(x = fyear, y = n))+ geom_line()+ geom_point() summary(duplicate.hcris)

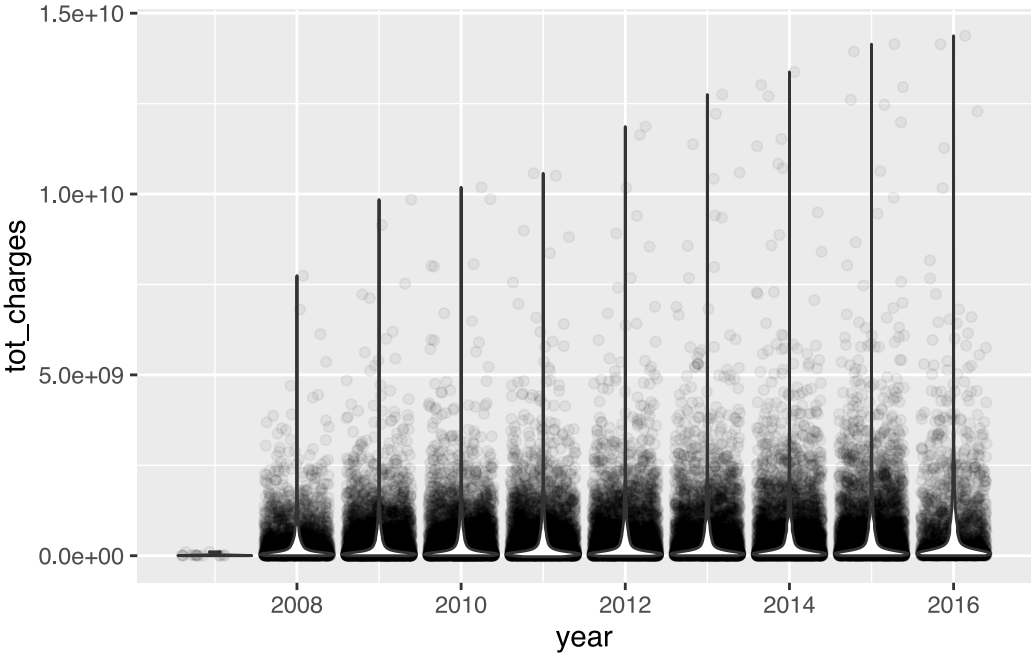
2. After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

```
[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()`).
```

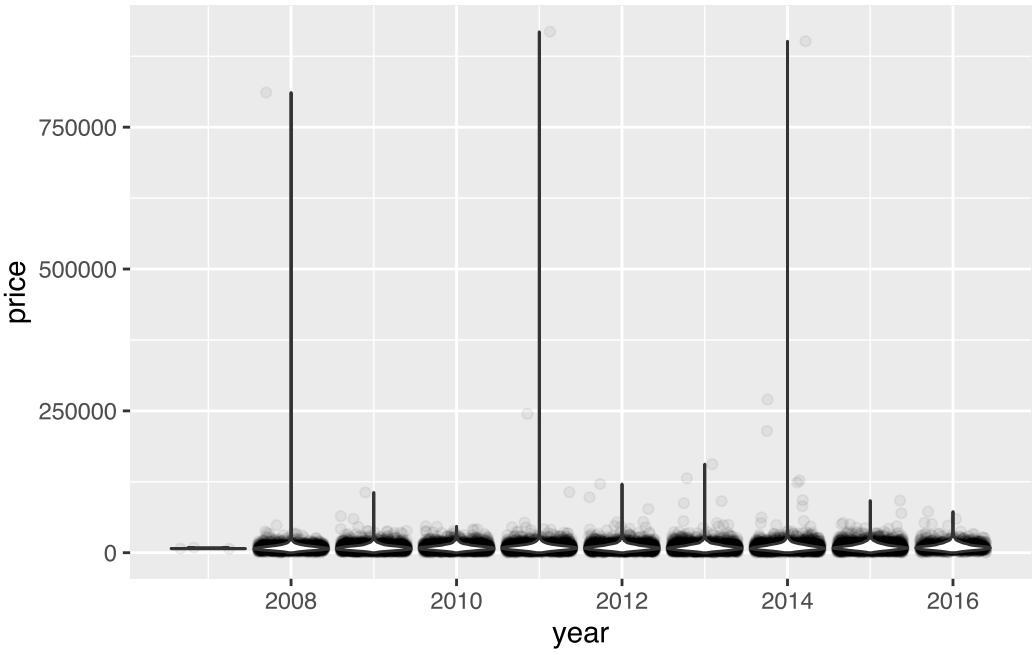
```
Warning: Removed 1832 rows containing missing values (`geom_point()`).
```



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: 3 × 2
  penalty `mean(price, na.rm = TRUE)`
    <dbl>                       <dbl>
1       0                      10731.
2       1                       9665.
3      NA                       9603.
```

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
4339 1689
```

```
# A tibble: 3 × 5
  penalty first_mean second_mean third_mean foruth_mean
    <dbl>      <dbl>       <dbl>      <dbl>       <dbl>
1       0       NaN       14462.      9557.      11032.
2       1       NaN        8456.      8066.      11106.
3      NA      7522.       7419.      8828.      11213.
```

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

install.packages("Matching") library("dplyr") library("tidyverse") hcris.vars <- hcris_2012 %>% select(penalty,price, first_quartile, second_quartile, third_quartile, fourth_quartile) %>% filter(-complete.cases(.)) hcris.covs <- hcris_2012 %>% select(first_quartile, second_quartile, third_quartile, fourth_quartile) m.nn.var <- Matching::Match(Y=hcris_2012$price, Tr = hcris_{2012}$penalty, X=hcris.covs, M=4, #<< Weight=1, estimand="ATE")

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

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

m.nn.md <- Matching::Match(Y=hcris.vars$price, Tr = hcris.vars$penalty, X=hcris.covs, M=1, Weight=2, estimand="ATE")

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

hcris.vars <- hcris.vars %>% mutate(ipw = case_when(penalty==1 ~ 1/ps, penalty==0 ~ 1/(1-ps), TRUE ~ NA_real_)) mean.t1 <- hcris.vars %>% filter(penalty==1) %>% select(price, ipw) %>% summarize(mean_p=weighted.mean(price,w=ipw)) mean.t0 <- hcris.vars %>% filter(penalty==0) %>% select(price, ipw) %>% summarize(mean_p=weighted.mean(price,w=ipw)) mean.t1$mean_p - mean.t0$mean_p

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

reg.dat <- hcris.vars %>% ungroup() %>% filter(complete.cases(.)) %>% mutate(beds_diff = penalty*(beds - mean(beds)), mcaid_diff = penalty*(mcaid_discharges - mean(mcaid_discharges)), ip_diff = penalty*(ip_charges - mean(ip_charges)), mcare_diff = penalty*(mcare_discharges - mean(mcare_discharges)), mpay_diff = penalty*(tot_mcare_payment - mean(tot_mcare_payment))) reg <- lm(price ~ penalty + beds + mcaid_discharges + ip_charges + mcare_discharges + tot_mcare_payment + beds_diff + mcaid_diff + ip_diff + mcare_diff + mpay_diff, data=reg.dat) summary(reg)

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

Still working on it!

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

Still working on it, but matching is hard to confirm causality becausr you never know if there are other variables you have not controlled 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 had some trouble with getting the matching code to work because the completecases function isn't working. 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 also 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.