Homework 2

Instructions

In this assignment, you'll recreate the HCRIS data and answer a few questions along the way. The first step is to make sure you're working with the HCRIS GitHub repository and downloaded all of the raw data sources. Once you have the data downloaded and the code running, answer the following questions. For more detailed instructions on how to submit your homework answers, please see the overview page here.

The due date for initial submission is **2/19**, the revision due date is **2/21**, and the final due date is Friday, **2/23**.

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

```
discount_factor = 1-tot_discounts/tot_charges
price_num = (ip_charges + icu_charges + ancillary_charges)*discount_factor - tot_mcare_pay
price_denom = tot_discharges - mcare_discharges
price = price_num/price_denom
```

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.
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.
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
8. With these different treatment effect estimators, are the results similar, identical, very different?
9. Do you think you've estimated a causal effect of the penalty? Why or why not? (just a couple of sentences)
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.