huang-r-hwk2-1

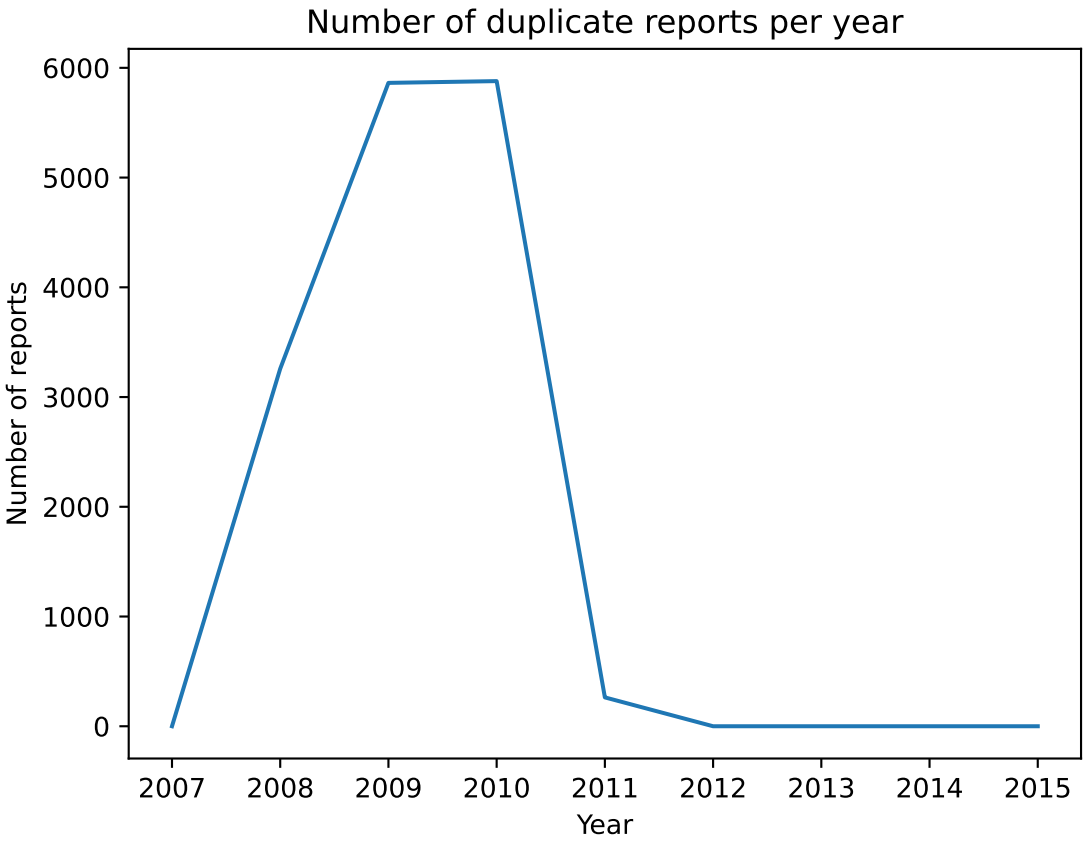
Ron Huang

Table of contents

git@github.com:Rorn001/eocn470-hwk2.git

1 Summarize the data

1.1 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.



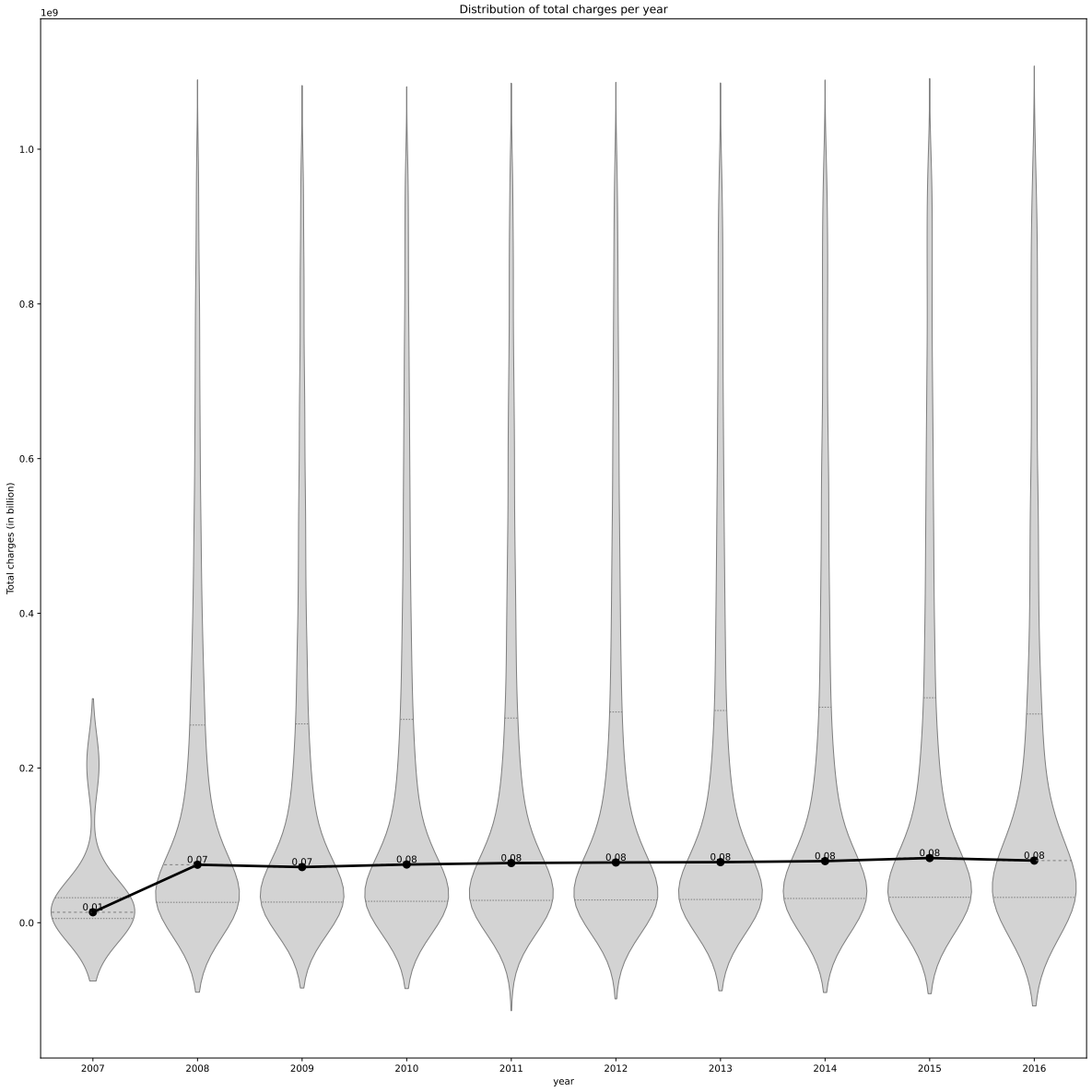
1.2 After removing/combining multiple reports, how many unique hospital IDs (Medicare provider numbers) exist in the data?

```
'Unqiue number of hospitals across all years is 6747'
```

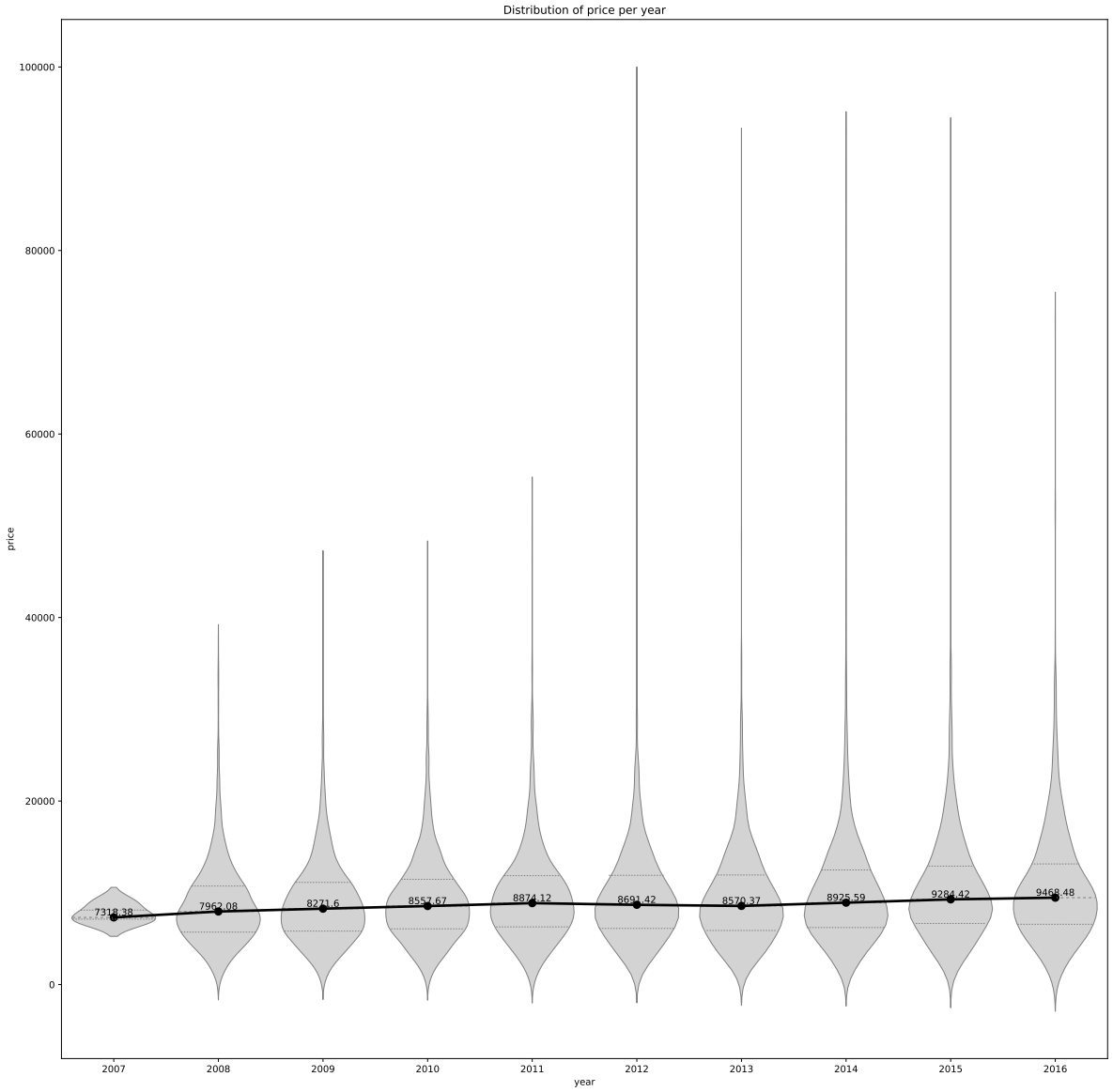
```
'Unqiue number of hospitals per year is:'
```

year	provider_number
2007	16
2008	3525
2009	6100
2010	6103
2011	6097
2012	6140
2013	6066
2014	6064
2015	6042
2016	2650

1.3 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.



1.4 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.



2 Estimate ATEs

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

	penalized	non_penalized	difference
Average price	9896.308498	9560.413227	335.895271

2.2 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.

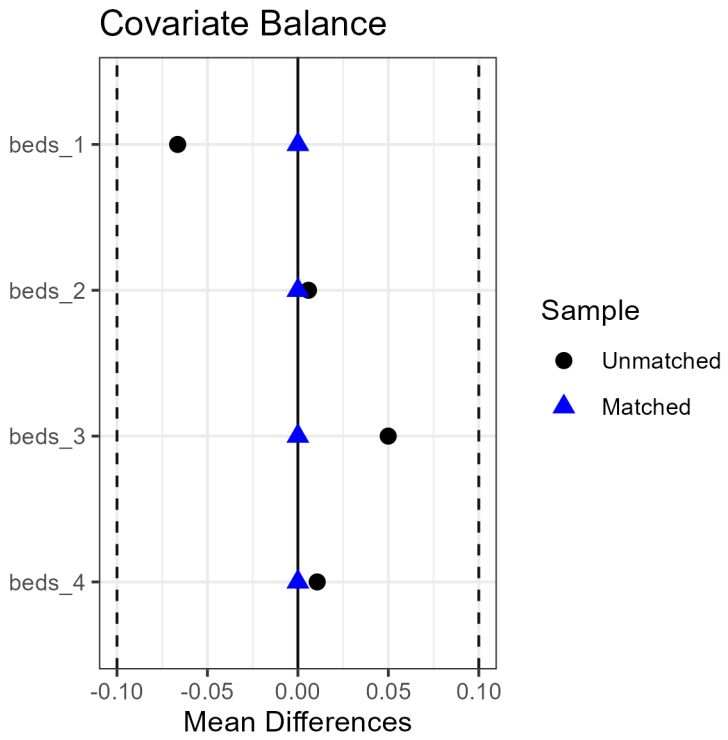
	beds_quartile	penalized_average_price	non_penalized_average_price
0	1st	8286.337994	7696.470378
1	2nd	8721.033188	8525.607482
2	3rd	10132.314630	9848.403610
3	4th	12068.478828	12367.332086

2.3 Find the average treatment effect using each of the following estimators, and present your results in a single table

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

Estimate... 193.83
AI SE..... 236.08
T-stat..... 0.82103
p.val..... 0.41163

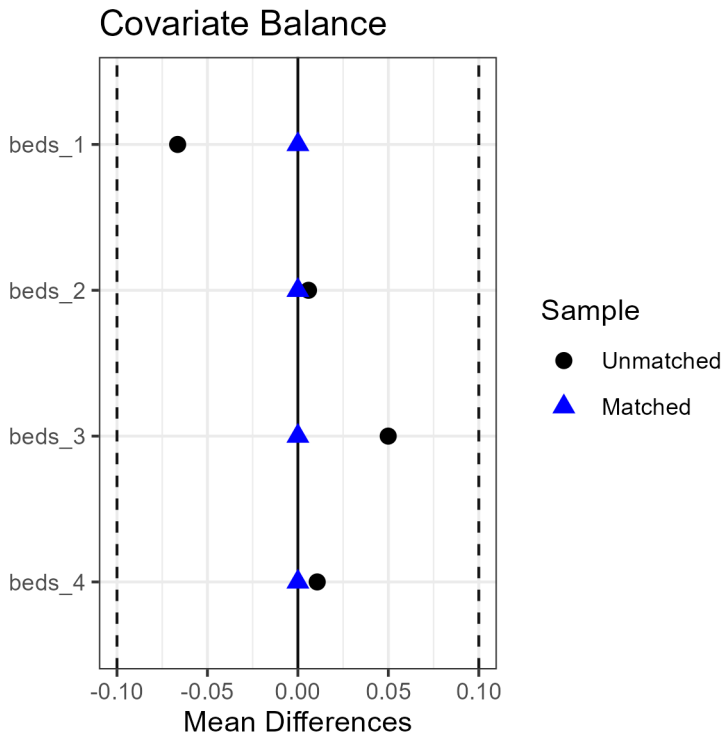
Original number of observations..... 2733
Original number of treated obs..... 704
Matched number of observations..... 2733
Matched number of observations (unweighted). 710030



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

Estimate... 193.83
AI SE..... 236.08
T-stat..... 0.82103
p.val..... 0.41163

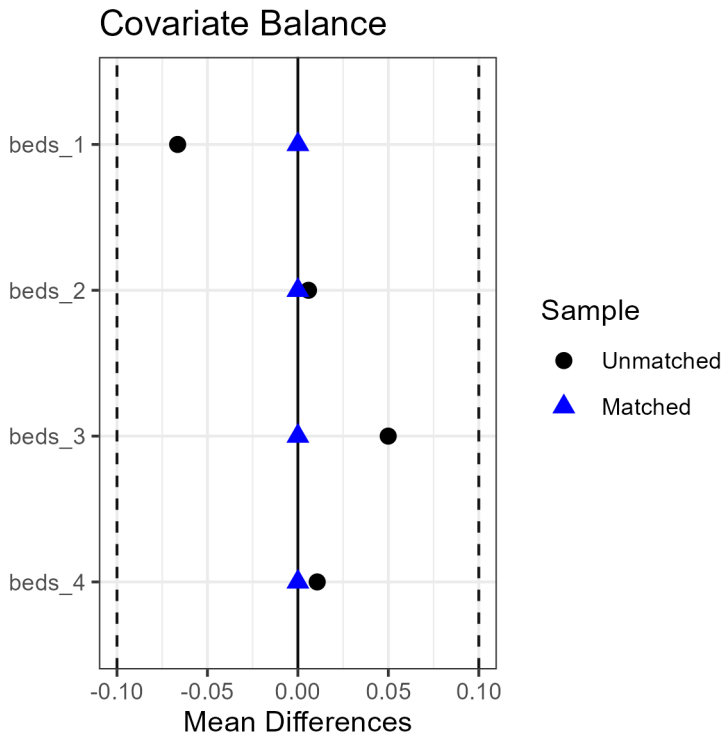
Original number of observations..... 2733
Original number of treated obs..... 704
Matched number of observations..... 2733
Matched number of observations (unweighted). 710030



(3). Inverse propensity weighting, where the propensity scores are based on quartiles of bed size

Estimate... 193.83
AI SE..... 236.08
T-stat..... 0.82103
p.val..... 0.41163

Original number of observations..... 2733
Original number of treated obs..... 704
Matched number of observations..... 2733
Matched number of observations (unweighted). 710030



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

$$Y_i = \delta D_i + \beta \Sigma_{k=1}^{k} X_{ik} + \gamma D_i \Sigma_{k=1}^{k} (X_{ik} - \bar{X}_k) + \epsilon_i$$

	term	estimate	std.error	statistic	p.value
0	(Intercept)	12367.332086	243.674026	50.753592	0.000000e+00
1	beds_1	-4670.861708	337.123415	-13.855050	3.100624e-42
2	beds_2	-3841.724604	343.922689	-11.170315	2.330312e-28
3	beds_3	-2518.928476	348.549704	-7.226885	6.387115e-13
4	beds_4	NaN	NaN	NaN	NaN
5	penalty	193.831294	239.899857	0.807968	4.191798e-01
6	penalty:var_1	888.720874	696.626975	1.275749	2.021531e-01
7	penalty:var_2	494.278964	669.451561	0.738334	4.603750e-01
8	penalty:var_3	582.764278	658.191387	0.885402	3.760177e-01
9	penalty:var_4	NaN	NaN	NaN	NaN

Summary table

	Method	ATE	Standard.Error
0	Single Inverse Variance Weighting	193.831294	236.083015
1	Single Mahalanobis Distance Matching	193.831294	236.083015
2	Propensity Score Matching	193.831294	236.083015
3	Linear Regression (Two Step)	193.831294	239.899857

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

The results are identical across all estimators. The average treatment effect is 193.83, and they are al statistically insignificant. This might be because the covariates we used are the dummies for the bed size quartiles so that all four methods create the subclasses and do the matching identically. Namely, each hospital is only compared with another hospital in the same quartile no matter what matching model we use.

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

Since we rely on the assumption of selection on observables, it is hard to determine whether we have controlled for all possible confounders that could affect the penalty and the price simultaneously. Suppose the assumption holds, we only consider four quartiles of bed size as covariates, while some other omitted variables may still exist. Therefore, the ATEs here might only partially explain the causal effect of the penalty.

2.6 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.

This is a way smaller dataset than the previous one, so it is easier to handle it. However, one thing is that the matching package is only available in R, so I have to use both R and python for different questions. It is hard to read both languages in .qmd file at the same time. For some tables, I can save them in dataframe and read them in python so that I can make sure that there aren't multiple languaes when compiling Quarto, but the regression summary for matching cannot be saved using tidy() function as lm or glm models. Currently, I haven't figure out how to do that so I can only copy and past the results, which is not very effcient.