ECON 470 HW2

Ben Yang

2023-02-19

1 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:

2 Summarize the data

2.0.1 Question 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.

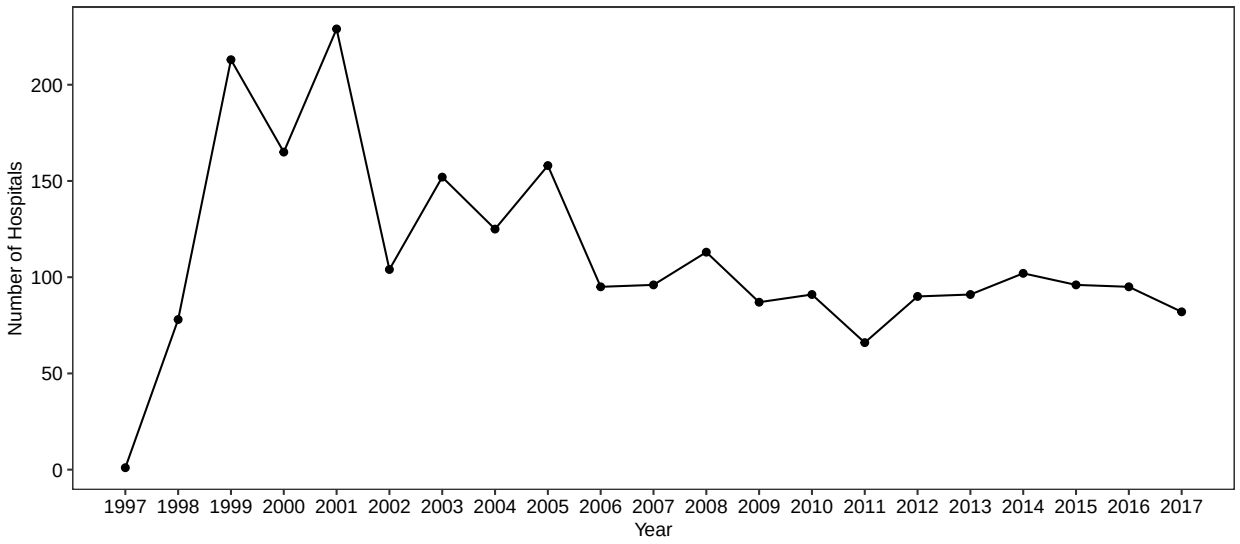

Figure 1: Number of Hospitals with More Than 1 Report in Each Year from 1997 to 2018

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

After removing/combining multiple reports, there are 9323 unique hospital IDs (Medicare provider numbers) exist in the data.

2.0.3 Question 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.

Figure 2 and Table 1 display the distribution of total charges in each year from 1997 to 2018.

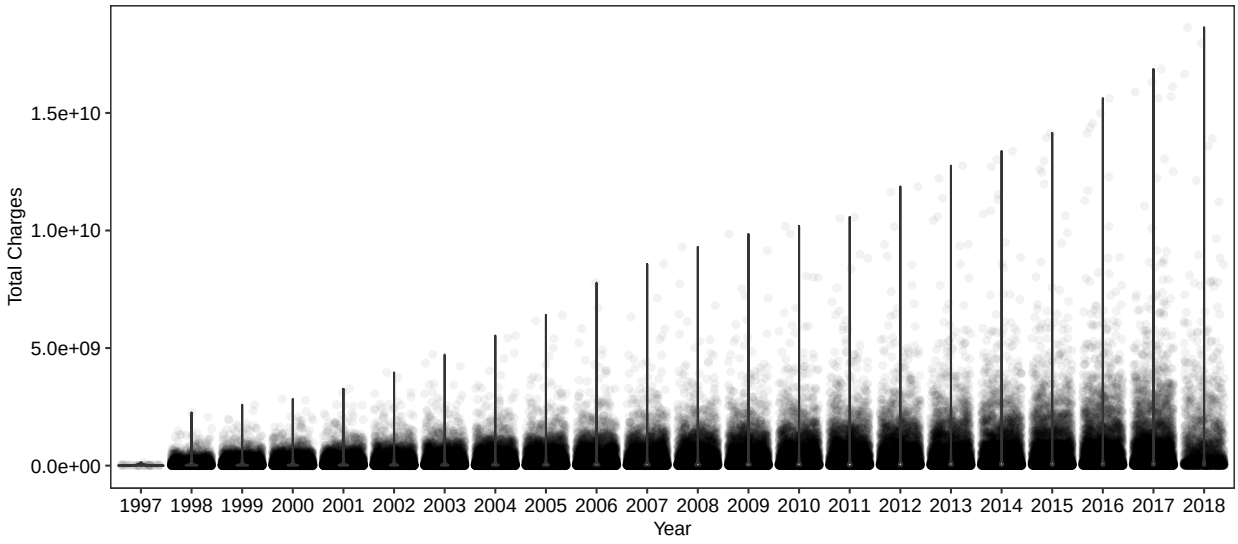

Figure 2: Distribution of Total Charges in Each Year from 1997 to 2018

Table 1: Summary Statistics of Total Charges in Each Year from 1997 to 2018

year	Mean	Min	Q1	Median	Q3	Max
1997	17,406,411	239,580	2,408,865	7,379,008	18,190,000	128,092,000
1998	106,221,397	155,387	13,999,500	40,548,466	129,857,366	2,255,621,364
1999	117,531,599	1	14,216,176	42,395,796	144,175,793	2,586,692,428
2000	131,767,289	1	14,556,663	45,006,082	157,546,224	2,823,988,041
2001	147,463,809	2,795	15,586,245	48,195,839	175,364,728	3,267,554,934
2002	170,501,538	347	17,192,218	53,816,123	202,119,743	3,957,656,325
2003	196,326,538	-1,757,898	18,857,460	59,706,752	230,519,124	4,722,758,791
2004	217,080,321	154,394	19,781,145	63,743,297	257,755,640	5,525,730,727
2005	237,504,259	1	21,781,589	67,698,741	280,002,570	6,398,553,843
2006	262,156,614	-104,189	23,630,560	75,698,500	307,081,656	7,784,094,716
2007	285,967,931	63,650	25,691,154	81,686,015	333,247,516	8,577,046,126
2008	311,419,620	4	27,352,019	87,823,449	363,096,949	9,293,788,259
2009	341,731,391	119,236	29,197,595	93,761,995	400,545,289	9,846,464,732
2010	366,800,221	306,861	31,052,856	98,454,656	426,883,234	10,185,415,748
2011	394,095,337	-27,582,223	32,756,460	104,392,453	457,465,336	10,572,291,195
2012	417,921,710	-11,799,711	33,169,626	108,541,199	481,728,422	11,865,320,139
2013	444,045,034	94,880	34,186,768	111,424,606	508,547,173	12,751,708,196
2014	477,721,391	6,624	36,027,552	118,297,638	546,186,606	13,376,352,387
2015	516,938,345	9,368	38,548,591	127,500,490	590,276,638	14,143,533,186
2016	560,688,971	84,952	40,762,872	137,437,310	635,249,151	15,618,749,067
2017	602,617,485	124,513	42,412,750	144,742,939	673,802,240	16,863,431,079
2018	680,524,568	282,914	42,185,109	144,458,496	751,891,924	18,633,710,235

2.0.4 Question 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.

Figure 3 and Table 2 display the distribution of estimated prices in each year from 1997 to 2018.

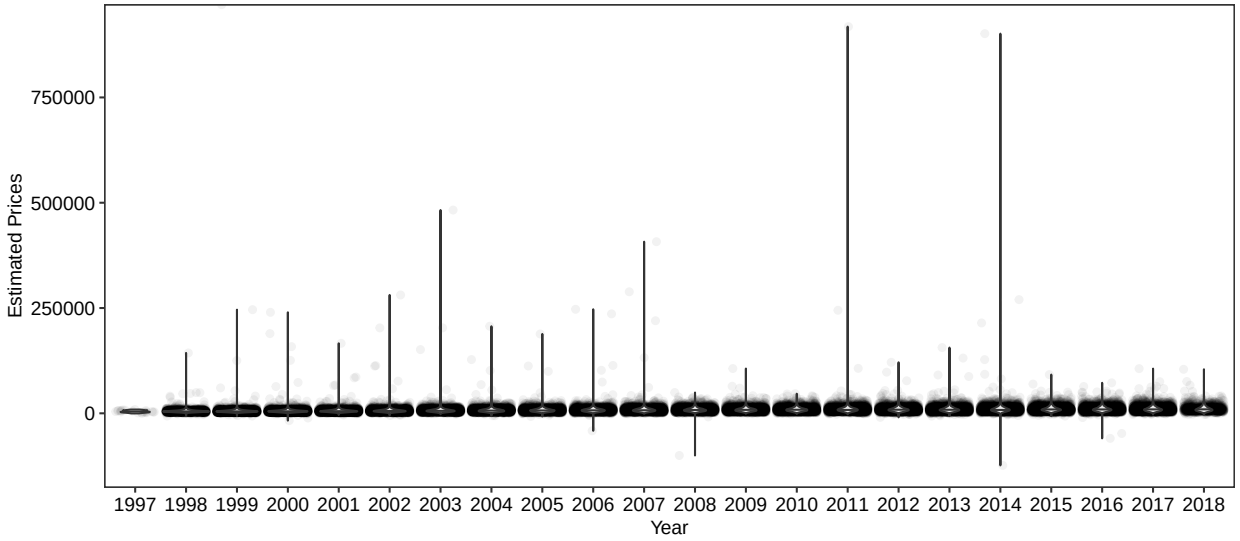

Figure 3: Distribution of Estimated Prices in Each Year from 1997 to 2018

Table 2: Summary Statistics of Price in Each Year from 1997 to 2018

year	Mean	Min	Q1	Median	Q3	Max
1997	4,410.53	-235.44	2,665.52	4,163.18	6,497.91	8,970.90
1998	5,183.28	-6,388.03	3,469.31	4,610.48	6,037.46	143,282.73
1999	Inf	-6,497.26	3,502.93	4,796.21	6,321.03	Inf
2000	5,716.97	-17,277.43	3,674.99	5,006.30	6,594.93	239,768.21
2001	5,891.01	-3,935.23	3,823.17	5,221.48	7,012.21	166,195.67
2002	6,504.55	-5,170.77	4,063.65	5,691.18	7,721.93	280,981.99
2003	7,102.01	-2,125.82	4,393.44	6,197.95	8,503.28	482,642.37
2004	7,352.90	-5,658.70	4,544.19	6,590.22	9,014.78	206,657.30
2005	7,660.43	-7,409.81	4,727.25	6,930.57	9,547.47	188,397.74
2006	8,142.47	-41,517.89	5,040.47	7,309.52	9,919.95	247,003.82
2007	8,678.09	-1,000.65	5,356.15	7,628.00	10,397.69	407,460.04
2008	8,689.38	-100,124.58	5,613.18	7,956.23	10,817.00	49,316.28
2009	9,100.19	-999.68	5,818.84	8,258.63	11,137.59	106,286.42
2010	9,320.58	-1,522.19	6,052.21	8,539.18	11,461.35	46,558.55
2011	10,087.45	-2,347.68	6,256.42	8,851.77	11,890.63	918,157.14
2012	9,635.76	-9,579.78	6,100.86	8,677.64	11,904.53	121,121.52
2013	9,631.07	-6,573.86	5,857.43	8,544.05	11,928.15	156,071.38
2014	10,420.87	-123,696.86	6,179.13	8,894.06	12,484.67	901,486.91
2015	10,385.18	-6,625.30	6,608.97	9,247.47	12,891.47	91,929.45
2016	10,362.30	-59,461.94	6,439.61	9,281.36	12,837.91	72,412.91
2017	10,603.28	-2,798.07	6,499.56	9,360.83	13,097.68	106,066.13
2018	11,140.59	-3,144.59	6,400.41	9,458.47	13,564.53	104,492.31

3 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 Section 2 slides.

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

The average price among penalized hospitals is 9,896.31. The average price among non-penalized hospitals is 9,560.41. The mean difference is 335.9.

3.0.2 Question 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.

Table 3 displays the average price among penalized hospitals (treated group) and non-penalized hospitals (control group) for each quartile of bed size. For the first three quartiles, the average price among penalized hospitals is higher than that of non-penalized hospitals. For the fourth quartile, the average price among penalized hospitals is lower than that of non-penalized hospitals. The results seem to suggest that hospitals of small and medium bed sizes tend to have higher price on average if penalized, while hospitals of large bed sizes tend to have lower price on average if penalized.

Table 3: Average Price among Treated/Control Groups for Each Quartile

Quartile Based on Bed Size	Non-Penalized (Control)	Penalized (Treated)	Mean Difference
1st Quartile	7,684.24	8,318.71	634.47
2nd Quartile	8,510.96	8,690.89	179.93
3rd Quartile	9,856.93	10,127.13	270.20
4th Quartile	12,355.61	12,068.48	-287.13

3.0.3 Question 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

Table 4 displays the average treatment effect obtained based on quartiles of bed size, using nearest neighbor matching with inverse variance distance, nearest neighbor matching with Mahalanobis distance, nearest neighbor matching with propensity score distance, inverse propensity weighted regression, and simple linear regression.

Table 4: Average Treatment Effects by Different Estimators for Each Quartile

Estimator	Average Treatment Effect
Nearest Neighbor Matching with Inverse Variance Distance	199.528083911484
Nearest Neighbor Matching with Mahalanobis Distance	199.528083911484
Nearest Neighbor Matching with Propensity Score Distance	199.528083911484
Inverse Propensity Weighted Regression	199.528083911487
Simple Linear Regression	199.528083911475

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

The results of these different treatment effect estimators are identical in this case because the estimations are based on the dummy variables representing quartiles of bed size, which are discrete variables. The results would be different if the estimations are based on some continuous variables. The idea of nearest neighbor matching is to estimate an unobserved data point by identifying the observed data points that are most similar as measured by some other features. The nearest neighbor method minimizes the chosen measure of distance between the data points. Meanwhile, the simple linear regression minimizes the least square error and puts more weight on observations that are penalized and unexplained the bed size variables.

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

I do not think I have estimated a causal effect of the penalty because the analysis above only considers the effect of bed size and omits many other factors that could be relevant, and thus the results may be subjected to selection bias and omitted variables bias.

The aforementioned estimators of the average treatment effects provide estimates of the average difference in price between penalized and non-penalized hospitals while controlling only for the bed size of the hospitals. The nearest neighbor method estimates the potential outcomes with different measures of distance from the observed outcomes based only on quartiles of bed size.

The prices of hospitals may be affected by various factors such as the physical location, the quality of services, the type of services offered, the applicable regulations, the ownership of the hospital, etc. Whether a hospital is penalized or not is certainly not random assigned. It is likely that the penalized hospitals (treated group) was going to have different prices on average even without being penalized (treatment). The regression analysis has omitted many other factors as the models examine the impact of penalty on price while controlling only for bed size.

3.0.6 Question 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.

Initially, I got stuck at question 6 as I was not able to create the appropriate quartiles of bed size for further analysis. Later I spent much time familiarizing with the regression codes, which helps me understand the concepts discussed in class such as the estimation of average treatment effects and nearest neighbor matching.

One thing I learned is that penalized hospitals of small and medium bed sizes may have higher prices on average. If penalty has justified a causal effect on price, the policy may have made health care services less affordable to patients. What aggravated me is that the penalty may not lead to better services as the policy makers intend, but only higher prices as the healthcare providers are able to pass on the cost to patients.