UCID: 12245077
PPHA 30545
Dr. Christopher Clapp
2/17/2021

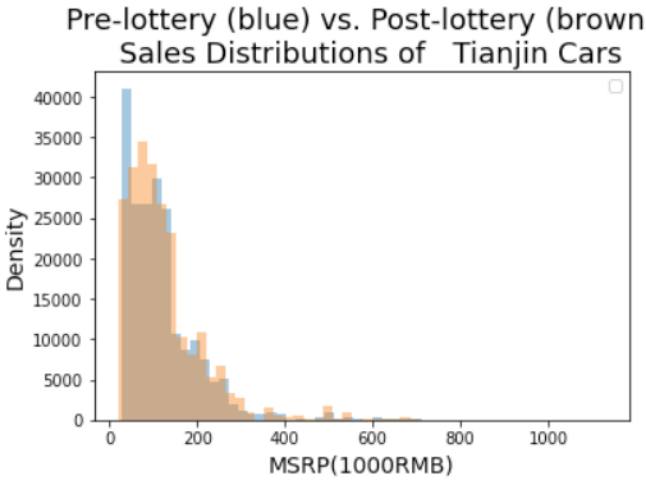
ML Mini-Lab 2

I worked with Maria Jiang (UCID: 12231837) as a partner on this Mini-Lab assignment.

All working code can be located via my GitHub.

Exercise 3.1. For each of the following, ensure that the first column is MSRP and the second column is count.
a. Clean data of Beijing car sales in 2011, and store the data frame in a variable called Beijing_post.
b. Clean data of Tianjin car sales in 2010 as a variable called Tianjin_pre.
c. Clean data of Tianjin car sales in 2011 as a variable called Tianjin_post.

Exercise 3.2. The goal of this exercise is to replicate Figure 1 for Tianjin.
a. Overlay the histograms that describe the 2010 and 2011 distribution of Tianjin car sales. Be sure to normalize the histograms so the area of the bars in each histogram sum to 1.



b. Compare and contrast the shift between the Beijing distributions with the shift between the Tianjin distributions. Based on the shift in Tianjin car sales, should we be surprised to see the shift in Beijing car sales?

Comparing the two cities, we notice the stark difference in the density in Beijing (pre-lottery) against the density in Tianjin. We should not be surprised by the shift in Beijing, for the car sales ought to be explained by the lottery system, as they reflect similar results in Tianjin.

Exercise 3.3.
a. Run the preceding code block so you have access to `placebo_1`. [check code].

b. Use `rmultinom` to sample observations from `Beijing_pre`. Store the resulting data frame in 0`placebo_2`. Be careful to draw the correct number of observations. [check code].
c. Compare `placebo_1` and `placebo_2`. Do they appear to be drawn from the same distribution?

	bandwidth	main
1	0	0.013909

No. By the literature and notes from earlier, if these placebo's were drawn from the same distribution, we might assume their optimal transport cost to be 0. Therefore, using the diftrans library, we unveil that there are differences in the distributions due to sampling variation.

Exercise 3.4.

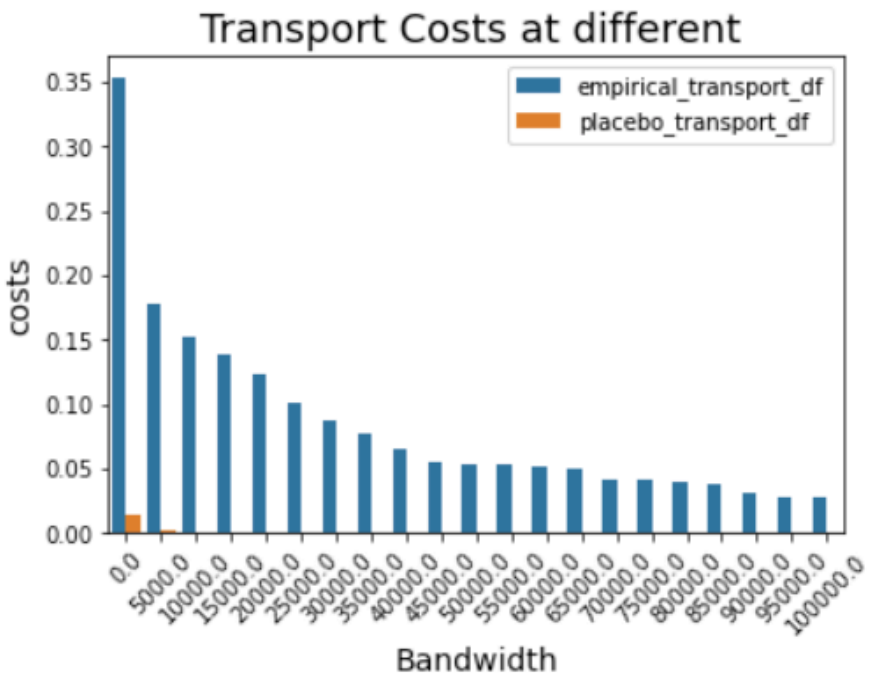
a. Compute the transport cost between the two placebo distributions for different values of d from 0 to 100,000.

	bandwidth	main
0	[0]	[0.013909162594171466]
1	[5000]	[0.0019489641981209477]
2	[10000]	[0.0007573809696677039]
3	[15000]	[0.0007316603132025103]
4	[20000]	[0.0006484126024571933]
5	[25000]	[0.0006093066340512542]
6	[30000]	[0.0005464930020972581]
7	[35000]	[0.0005421327904816576]
8	[40000]	[0.0005421327904816581]
9	[45000]	[0.0005421327904816614]
10	[50000]	[1.5670451007602535e-05]
11	[55000]	[1.5670451007602498e-05]
12	[60000]	[1.567045100760258e-05]
13	[65000]	[1.5670451007602525e-05]
14	[70000]	[1.0668323828738157e-05]
15	[75000]	[1.0668323828738157e-05]
16	[80000]	[1.0668323828738257e-05]
17	[85000]	[1.0668323828738157e-05]
18	[90000]	[1.0668323828738157e-05]
19	[95000]	[1.0668323828738157e-05]
20	[100000]	[1.0668323828738157e-05]

b. For the same values of d, compute the transport cost between the observed distributions for 2010 and 2011 Beijing car sales.

	bandwidth	main
0	[0]	[0.35312341198976444]
1	[5000]	[0.17785942748226927]
2	[10000]	[0.15183080916094385]
3	[15000]	[0.13805857605749003]
4	[20000]	[0.12315956106076102]
5	[25000]	[0.10076861095451133]
6	[30000]	[0.08807505784276998]
7	[35000]	[0.07725061893640303]
8	[40000]	[0.06488865219442809]
9	[45000]	[0.05525307829038914]
10	[50000]	[0.053762156781739324]
11	[55000]	[0.05375779736212339]
12	[60000]	[0.05126130306206454]
13	[65000]	[0.05068876595250505]
14	[70000]	[0.04098760416717745]
15	[75000]	[0.04081468052241197]
16	[80000]	[0.039060507085339355]
17	[85000]	[0.03895175520653571]
18	[90000]	[0.030705186433058418]
19	[95000]	[0.028104066062217124]
20	[100000]	[0.027102290368887762]

c. Plot the placebo costs and the empirical costs obtained in the previous two steps with the bandwidth as the x-axis.



d. For which values of d is the placebo cost less than 0.05%?

bandwidth	main
0	0.013909
25	0.013909
100	0.012443
500	0.011387
1000	0.006507
1500	0.005513
1750	0.005513
1850	0.005435
1900	0.004946
2000	0.003268
5000	0.001949
25000	0.000609
30000	0.000546
32500	0.000542
40000	0.000542
45000	0.000542
50000	0.000016
100000	0.000011

By the time we view $d = 25000$, we begin to notice placebo costs near 0.05%; however, in between the 40,000 and 50,000 range for d, does the cost actually dip below 0.05%.

e. For the smallest value of d found in the previous step, what is the empirical transport cost? This estimate for the lower bound on the volume of black market transactions is what we call the *before-and-after estimate*.

bandwidth	main
0	0.353049
25	0.353049
1000	0.258904
5000	0.177841
10000	0.151903
25000	0.100852
30000	0.088156
40000	0.064967
45000	0.055306
50000	0.053819
60000	0.051318
80000	0.039114
100000	0.027194

Proceeding with d= 50,000 from the previous question, I notice my transport cost is roughly 5.4%.

Exercise 3.5.

a. Compute the (3) for different values of d from 0 to 50,000. Unlike before, we go up to 50,000 because we are using the conservative bandwidth of $2d$ for the Beijing transport cost.

	bandwidth	main	main2d	control	diff	diff2d
0	0.0	0.353123	0.353123	0.298681	0.054443	0.054443
1	2500.0	0.215985	0.177859	0.103619	0.112366	0.074240
2	5000.0	0.177859	0.151831	0.045617	0.132243	0.106214
3	7500.0	0.169615	0.138059	0.025194	0.144421	0.112864
4	10000.0	0.151831	0.123160	0.020093	0.131737	0.103066
5	12500.0	0.148234	0.100769	0.018438	0.129796	0.082331
6	15000.0	0.138059	0.088075	0.017838	0.120220	0.070237
7	17500.0	0.130099	0.077251	0.017821	0.112278	0.059430
8	20000.0	0.123160	0.064889	0.013056	0.110104	0.051833
9	22500.0	0.115663	0.055253	0.007405	0.108258	0.047848
10	25000.0	0.100769	0.053762	0.007158	0.093610	0.046604
11	27500.0	0.095018	0.053758	0.007158	0.087859	0.046599
12	30000.0	0.088075	0.051261	0.006841	0.081234	0.044420
13	32500.0	0.081166	0.050689	0.006663	0.074504	0.044026
14	35000.0	0.077251	0.040988	0.006663	0.070588	0.034325
15	37500.0	0.070634	0.040815	0.006663	0.063972	0.034152
16	40000.0	0.064889	0.039061	0.006377	0.058512	0.032684
17	42500.0	0.056961	0.038952	0.004576	0.052385	0.034376
18	45000.0	0.055253	0.030705	0.004319	0.050934	0.026386
19	47500.0	0.055253	0.028104	0.004310	0.050943	0.023794
20	50000.0	0.053762	0.027120	0.004310	0.049453	0.022811

b. Using what you learned from Exercise 3.3, construct a placebo distribution that is sampled from Beijing_pre whose size is the number of Beijing cars in 2010. Call this distribution placebo_Beijing_1. [check code].

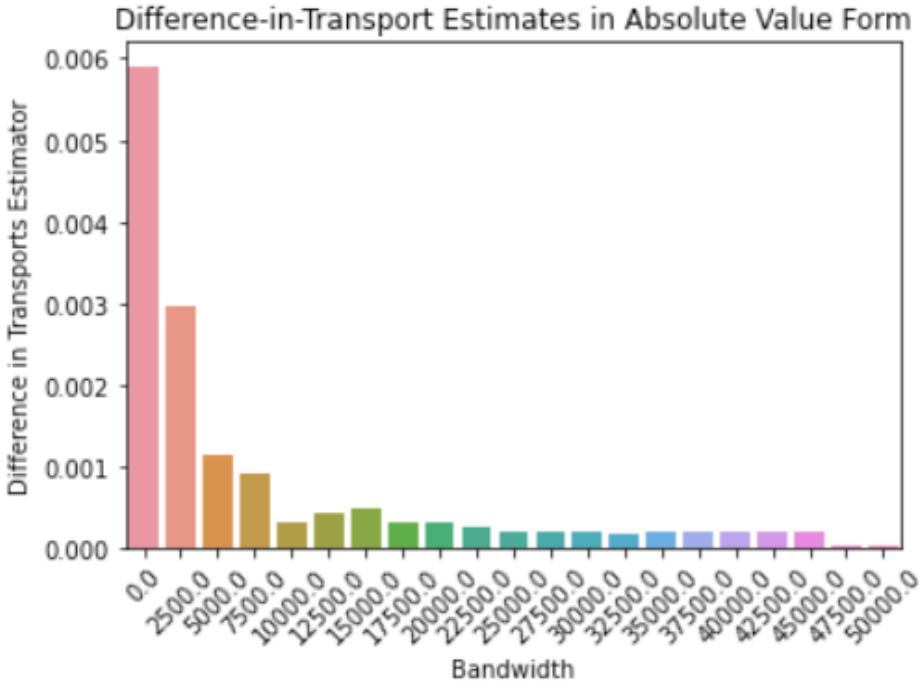
c. Construct another placebo distribution called placebo_Beijing_2 that is also sampled from Beijing_pre but is of size is the number of Beijing cars in 2011. [check code].

d. Construct a placebo distribution called placebo_Tianjin_1 that is sampled from Tianjin_pre and whose size is the number of Tianjin cars in 2010. [check code].

e. Construct a placebo distribution called placebo_Tianjin_2 that is sampled from Tianjin_pre and whose size is the number of Tianjin cars in 2011. [check code].

f. Using the four placebo distributions, compute the placebo counterpart of (3) for the same values of d that you used in part a. [check code].

g. Create a plot of the *absolute value* of the placebo differences-in-transports estimator on the y-axis and the bandwidth on the x-axis.



h. For which values of d does the absolute value of the placebo differences-in-transports estimator stay below 0.05%? Note that the absolute difference is not a monotonically decreasing object, so this difference may even increase as we increase the bandwidth. Temporary increases above the 0.05% threshold can be ignored.

	bandwidth	main	Dif-In-Transports
0	0.0	0.012020	0.005913
1	2500.0	0.003116	0.002972
2	5000.0	0.001905	0.001124
3	7500.0	0.000632	0.000899
4	10000.0	0.000611	0.000301
5	12500.0	0.000608	0.000418
6	15000.0	0.000608	0.000470
7	17500.0	0.000608	0.000301

8	20000.0	0.000608	0.000298
9	22500.0	0.000608	0.000253
10	25000.0	0.000445	0.000193
11	27500.0	0.000445	0.000193
12	30000.0	0.000393	0.000178
13	32500.0	0.000388	0.000169
14	35000.0	0.000388	0.000185
15	37500.0	0.000388	0.000185
16	40000.0	0.000388	0.000185
17	42500.0	0.000388	0.000185
18	45000.0	0.000388	0.000185
19	47500.0	0.000022	0.000011
20	50000.0	0.000022	0.000011

By the time we reach d = 10,000 , we notice the differences-in-transports estimator to stay belo w 0.05%.

i. Among all the values of d that you found in the previous step, which one yielded the largest value of (3) from part a? This is the difference-in-transports estimator.

When d = 10,000, we also notice the largest value of (3). Therefore, this is the difference-in-transports estimator.