Machine Learning: Lab 2

Jingpeng Hong

February 14, 2022

1 Visualize Beijing and Tianjin Car Sales (Exercise 4.2)

Figure 1 and Figure 2 show the 2010 and 2011 distribution of car sales in Beijing and Tianjin, respectively. We can find that the 2010 and 2011 distribution of car sales in Tianjin have no significant shift. This is because Tianjin introduces the random allocation of license plates in 2014, and both 2010 and 2011 are "pre-lottery" in Tianjin. In other words, by the comparison between Beijing and Tianjin we can conclude that the observed shift in Beijing may not be driven by some common time trends (e.g. Households accumulate wealth to buy more expensive cars in 2011.). We can hypothesize that the lottery policy and accompanied illicit transactions in black markets drive this shift in Beijing.

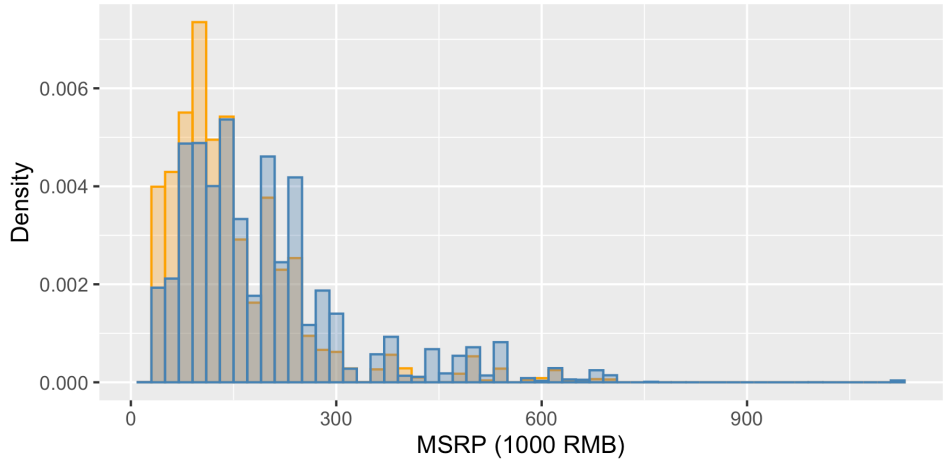

Figure 1: Pre-lottery (orange) v.s. Post-lottery (blue) Sales Distributions of Beijing Cars

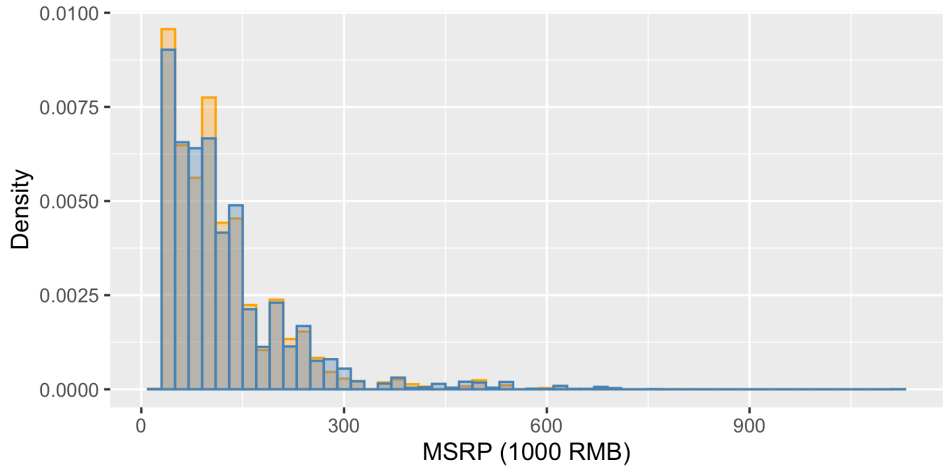

Figure 2: Pre-lottery (orange) v.s. Post-lottery (blue) Sales Distributions of Tianjin Cars

2 Compute Before-and-After Estimator

2.1 Exercise 4.3

From Figure 3, we can see the histograms of our two samples are quite similar to one another. We should notice that $\tilde{\mathcal{P}}^{(1)}$ and $\tilde{\mathcal{P}}^{(2)}$ are drawn from the distribution of $\hat{\mathcal{P}}_{\text{before, Beijing}}$, while $\tilde{\mathcal{P}}^{(1)}$ has the number of observations as $\hat{\mathcal{P}}_{\text{before, Beijing}}$ and $\tilde{\mathcal{P}}^{(2)}$ has the number of observations as $\hat{\mathcal{P}}_{\text{after, Beijing}}$. Despite of the sample size, they appear to be drawn from the same distribution. Moreover, according to diftrans, the optimal transport cost between two placebo distributions is 1.27%.

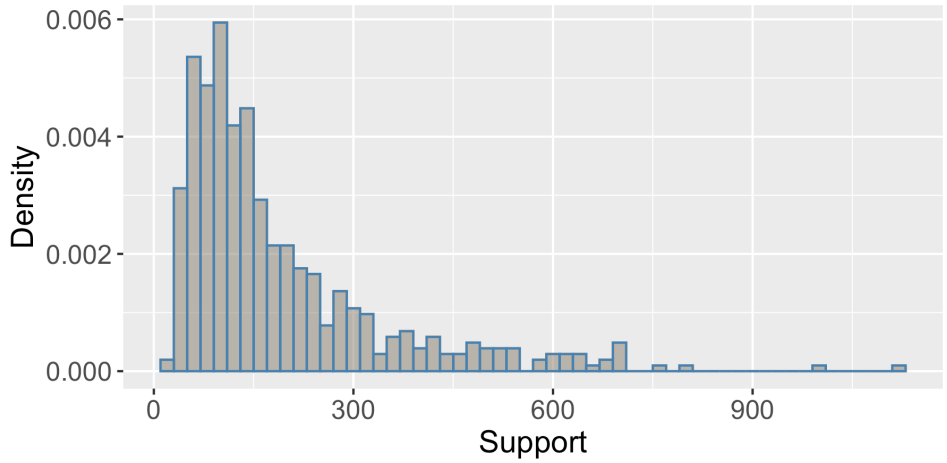

Figure 3: Placebo distribution $\tilde{\mathcal{P}}^{(1)}$ and $\tilde{\mathcal{P}}^{(2)}$ sampled from $\hat{\mathcal{P}}_{\text{before, Beijing}}$

2.2 Exercise 4.4

Figure 4 shows the placebo costs and the empirical costs with the bandwidth as the x-axis. Here, we choose the values of bandwidth from 0 to 100,000 with step size 1,000.

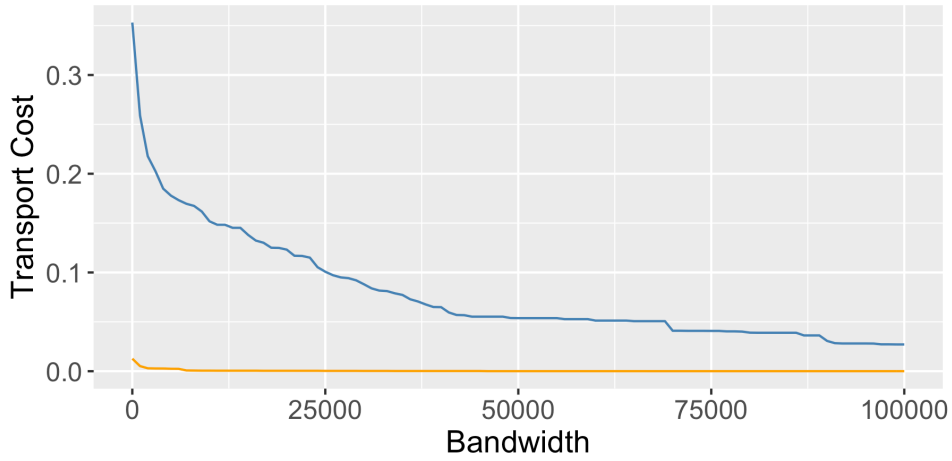

Figure 4: Placebo costs (orange) and Empirical costs(blue) with the bandwidth

From the placebo distributions, we find that when bandwidth $d \geq 17,000$, the placebo costs are less than 0.05%. Hence, we choose $d = 17,000$ to estimate our empirical transport cost. This value of bandwidth p is appropriate, because it's large enough to filter out shifts in our empirical distributions that arise from sampling bias, but not too large to mask shifts that arise due to the presence of a black market.

For $d = 17,000$, the empirical transport cost is 13.01%. This estimate for the lower bound on the volume of black market transactions is what we call the before-and-after estimate.

3 Compute Differences-in-Transports Estimator (Exercise 4.5)

Figure 5 is the plot of the absolute value of the placebo differences-in-transports estimator on the y-axis and the bandwidth on the x-axis.

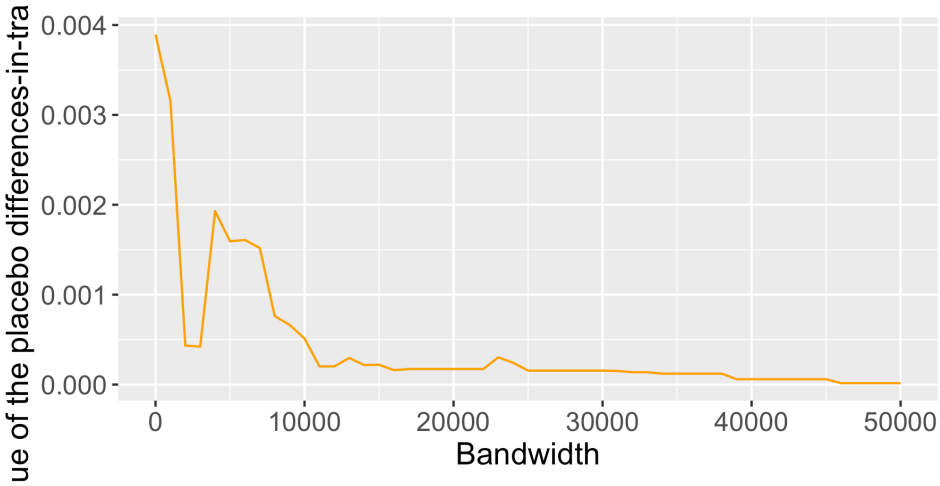

Figure 5: Placebo differences-in-transports estimator with the bandwidth

Since the absolute difference is not monotonically decreasing and we can ignore temporary increases, the absolute value of the placebo differences-in-transports estimator stay below 0.05% when $d \geq 11,000$.

When bandwidth $d = 11,000$, we can find the largest value diff2d from the empirical distribution of Beijing and Tianjin Cars. This largest value is 9.71 %, which is the difference-in-transports estimator.