PPHA 30545 Lab 2

Nicholus Tint Zaw

Disclaimer:

I discussed this homework problem-set with the following study group members for questions clarification and an individual approach to answering each question. However, the solution and codes were not shared among each other.

- Anna Meehan
- Aulia Larasati
- Betty Wong
- Sarah Mering

Load the dataset

```
df_beijing <- Beijing_sample
df_tianjin <- Tianjin_sample
```

4.3 Clean Data of Beijing and Tianjin Car Sales

```
# keep 2010 and 2011 only
beijing <- df_beijing %>%
  filter(year >= 2010 & year < 2012)

# collect unique MSRP values
beijing_uniqueMSRP <- data.frame(MSRP = unique(beijing$MSRP))

# keep 2010 and 2011 only
tianjin <- df_tianjin %>%
  filter(year >= 2010 & year < 2012)

# collect unique MSRP values
tianjin_uniqueMSRP <- data.frame(MSRP = unique(tianjin$MSRP))



# aggregate sales at each price for 2010 (pre-lottery)
beijing10_sales <- beijing %>%
  filter(year == 2010) %>%
```

```r
  dplyr:: group_by(MSRP) %>%
  summarize(count = sum(sales))

# merge the MSRP and sales
beijing_pre <- left_join(beijing_uniqueMSRP, beijing10_sales, by = "MSRP") %>%
  replace_na(list(count = 0)) %>%
  arrange(MSRP)

# preview data
head(beijing_pre)
##    MSRP count
## 1 20800     0
## 2 29800    47
## 3 32900  3153
## 4 33800  3678
## 5 34800   592
## 6 36800  1735
```

Exercise 4.1.

(a) Beijing car sale in 2011

```r
beijing11_sales <- beijing %>%
  filter(year == 2011) %>%
  dplyr:: group_by(MSRP) %>%
  summarize(count = sum(sales))

# merge the MSRP and sales
beijing_post <- left_join(beijing_uniqueMSRP, beijing11_sales, by = "MSRP") %>%
  replace_na(list(count = 0)) %>%
  arrange(MSRP)

# preview data
head(beijing_post)
##    MSRP count
## 1 20800    23
## 2 29800     0
## 3 32900  1393
## 4 33800     4
## 5 34800   189
## 6 36800   459
```

(b) Tianjin car sale in 2010

```r
tianjin10_sales <- tianjin %>%
  filter(year == 2010) %>%
  dplyr:: group_by(MSRP) %>%
```

```r
  summarize(count = sum(sales))

# merge the MSRP and sales
tianjin_pre <- left_join(tianjin_uniqueMSRP, tianjin10_sales, by = "MSRP") %>%
  replace_na(list(count = 0)) %>%
  arrange(MSRP)

# preview data
head(tianjin_pre)
##    MSRP count
## 1 20800     0
## 2 28800     0
## 3 29800    51
## 4 30900     0
## 5 32900   599
## 6 33300     2
```

(c) Tiajin car sale in 2011

```r
tianjin11_sales <- tianjin %>%
  filter(year == 2011) %>%
  dplyr:: group_by(MSRP) %>%
  summarize(count = sum(sales))

# merge the MSRP and sales
tianjin_post <- left_join(tianjin_uniqueMSRP, tianjin11_sales, by = "MSRP") %>%
  replace_na(list(count = 0)) %>%
  arrange(MSRP)

# preview data
head(tianjin_post)
##    MSRP count
## 1 20800    23
## 2 28800     7
## 3 29800     5
## 4 30900     1
## 5 32900   948
## 6 33300     0
```

4.4 Visualize Beijing Car Sales

```r
beijing_dist_pre <- beijing_pre %>% uncount(count)
beijing_dist_post <- beijing_post %>% uncount(count)

bdist <- ggplot() +
```

```
  geom_histogram(data = beijing_dist_pre,
                 aes(x = MSRP/1000,
                     y = ..density..),
                 binwidth = 20,
                 fill = "orange", color = "orange", alpha = 0.35) +
  geom_histogram(data = beijing_dist_post,
                 aes(x = MSRP/1000,
                     y = ..density..),
                 binwidth = 20,
                 fill = "steelblue", color = "steelblue", alpha = 0.35) +
  labs(x = "MSRP (1000 RMB)", y = "Density")

bdist
```

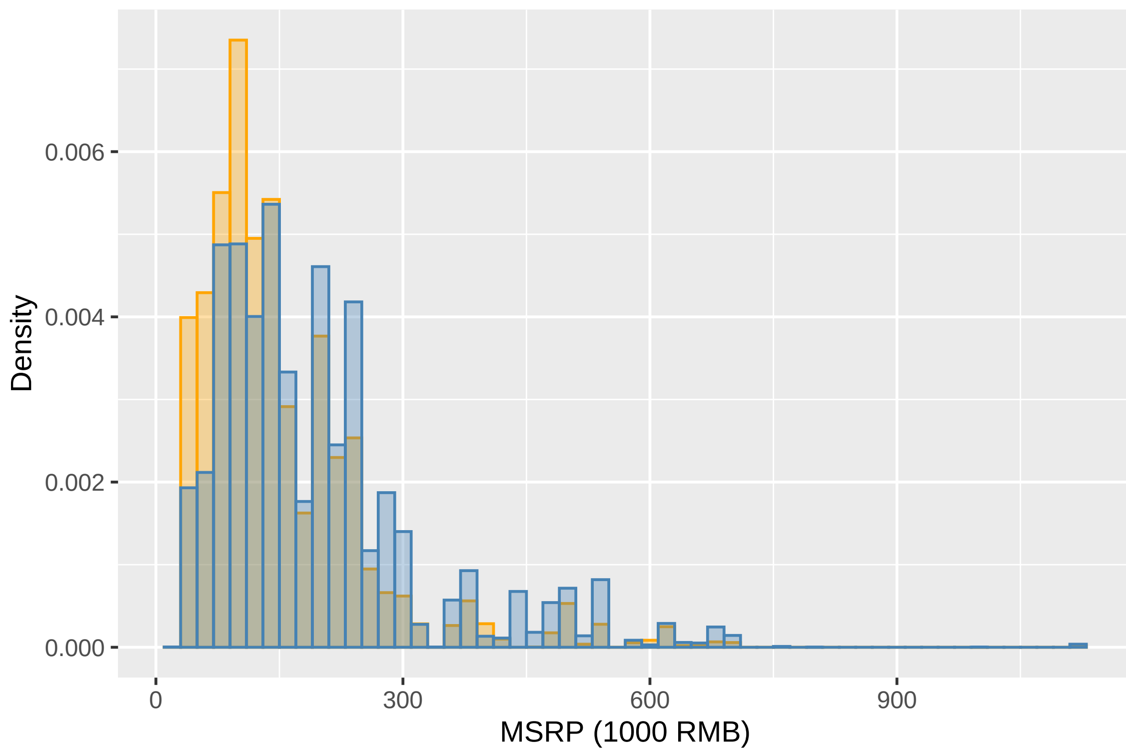

Figure 1: Beijing Car Sales Distribution 2010 vs 2011

Exercise 4.2.

(a) Tianjin car sales 2010 and 2011 distribution histograms

```
tianjin_dist_pre <- tianjin_pre %>% uncount(count)
tianjin_dist_post <- tianjin_post %>% uncount(count)

tdist <- ggplot() +
  geom_histogram(data = tianjin_dist_pre,
                 aes(x = MSRP/1000,
```

```
                    y = ..density..),
                binwidth = 20,
                fill = "orange", color = "orange", alpha = 0.35) +
  geom_histogram(data = tianjin_dist_post,
                aes(x = MSRP/1000,
                    y = ..density..),
                binwidth = 20,
                fill = "steelblue", color = "steelblue", alpha = 0.35) +
  labs(x = "MSRP (1000 RMB)", y = "Density")

tdist
```

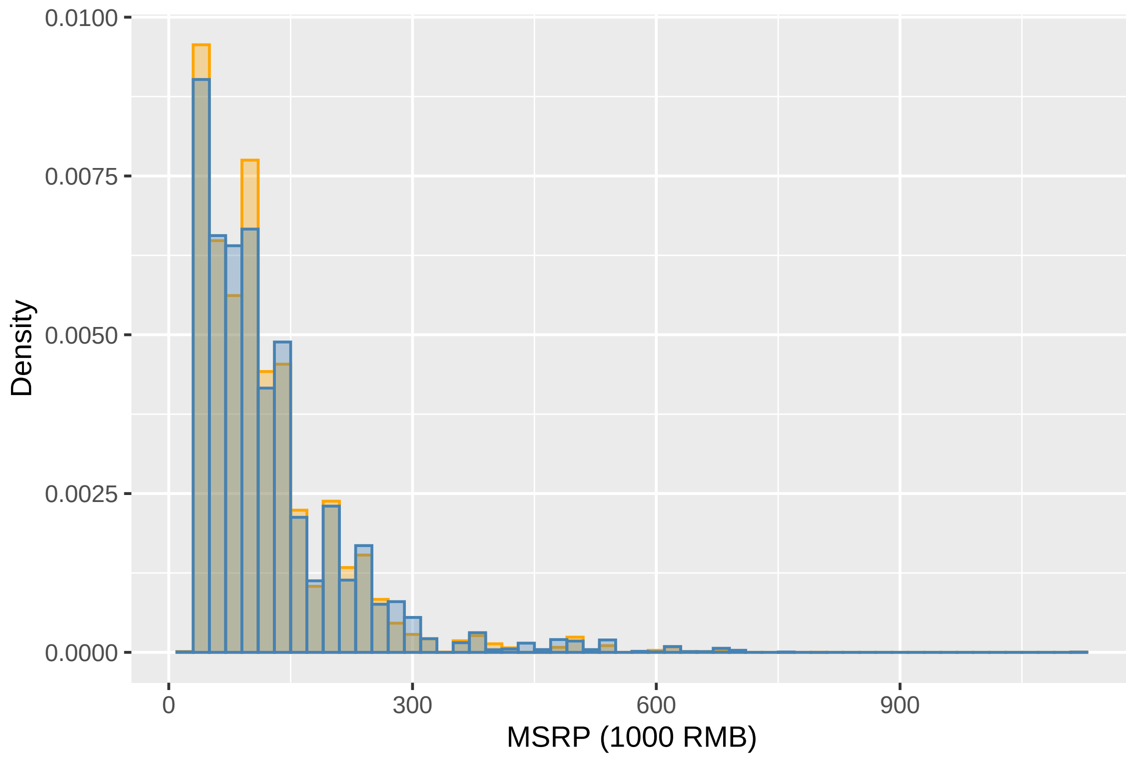

Figure 2: Tianjin Car Sales Distribution 2010 vs 2011

(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?

Both cities' care sale distributions (for 2010 and 2011) had the right-skewed distribution, which is more obvious for Beijing. The majority of car sales in both cities in both years were MSRP price less than 30,000 RMB. But, in 2011, more car with MSRP prices higher than 30,000 RMB were sold in both cities compared to 2010. From the visual inspection on both plots, Beijing had the more obvious trend changes than Tianjin. Based on the Tianjin car sale distribution shift in 2011, we can say that there were general changes across different cities in buying higher MSRP price. It is hard to say that the Beijing changes affected the new policy on the license plate lottery. However, we can not say whether this is statistically significant in these changes as no statistical analysis was performed yet to detect the significant differences.

4.5 Compute Before-and-After Estimator

```
set.seed(3453245)
n_obs <- 100000

placebo_demo <- data.frame(sample1 = rnorm(n_obs),
                           sample2 = rnorm(n_obs))

ggplot(placebo_demo) +
  geom_histogram(aes(x = sample1,
                     y = ..density..),
                 fill = "orange", color = "orange", alpha = 0.35) +
  geom_histogram(aes(x = sample2,
                     y = ..density..),
                 fill = "steelblue", color = "steelblue", alpha = 0.35) +
  labs(x = "Support", y = "Density") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 14))
```

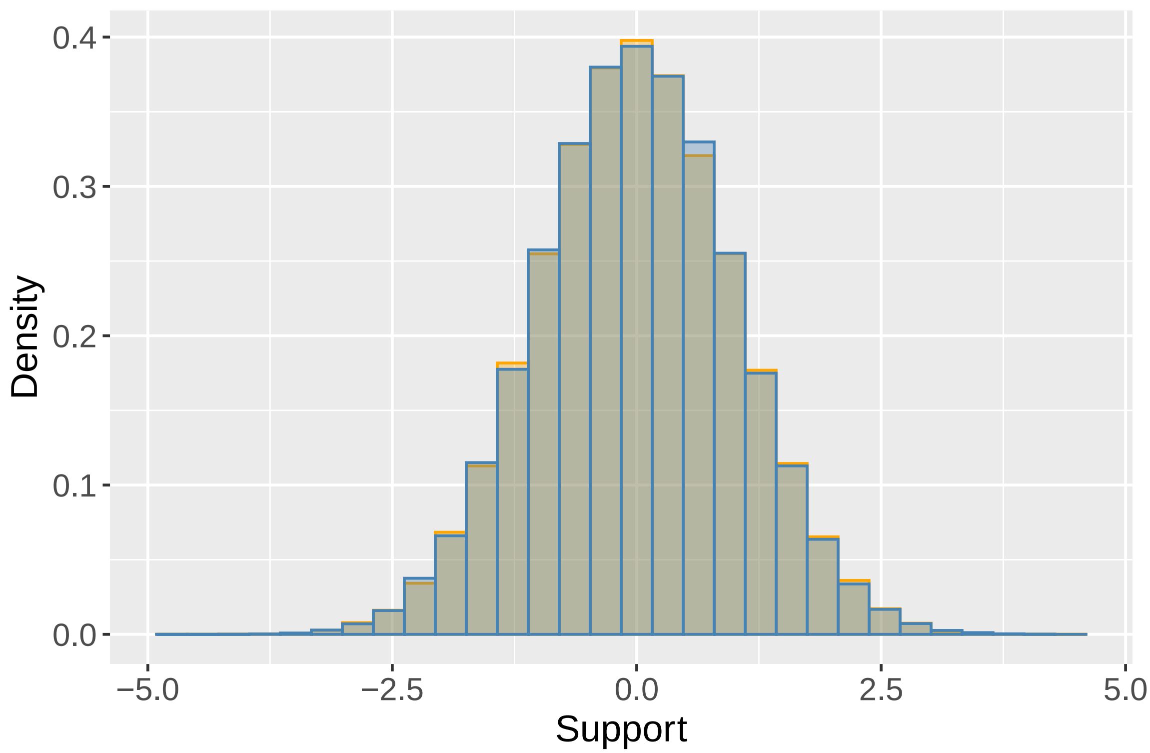

Figure 3: Two Samples from Standard Normal Distribution

Exercise 4.3.

(a) placebo_1

```r
set.seed(4487989)

placebo_1 <- data.frame(MSRP = beijing_pre$MSRP,
                        count = rmultinom(n = 1,
                                          size = sum(beijing_pre$count),
                                          prob = beijing_pre$count))

head(placebo_1)
##    MSRP count
## 1 20800     0
## 2 29800    53
## 3 32900  3260
## 4 33800  3713
## 5 34800   557
## 6 36800  1695
```

(b) placebo_2

```r
set.seed(384620)

placebo_2 <- data.frame(MSRP = beijing_pre$MSRP,
                        count = rmultinom(n = 1,
                                          size = sum(beijing_post$count),
                                          prob = beijing_pre$count))

head(placebo_2)
##    MSRP count
## 1 20800     0
## 2 29800    21
## 3 32900  1293
## 4 33800  1575
## 5 34800   253
## 6 36800   739
```

(c) Compare placebo_1 and placebo_2

MSRP prices were observed at per 1000 RMB price to detect the changes in the less than 30,000 RMB car sales as more observations were accumulated in that category. Please note that each bind had a width of 2000 RMB.

```r
placebo_1_dist <- placebo_1 %>% uncount(count)
placebo_2_dist <- placebo_2 %>% uncount(count)

ggplot() +
  geom_histogram(data = placebo_1_dist,
                 aes(x = MSRP/1000,
                     y = ..density..),
```

```
                  binwidth = 20,
                  fill = "orange", color = "orange", alpha = 0.35) +
  geom_histogram(data = placebo_2_dist,
                  aes(x = MSRP/1000,
                      y = ..density..),
                  binwidth = 20,
                  fill = "steelblue", color = "steelblue", alpha = 0.35) +
  labs(x = "MSRP (1000 RMB)", y = "Density")
```

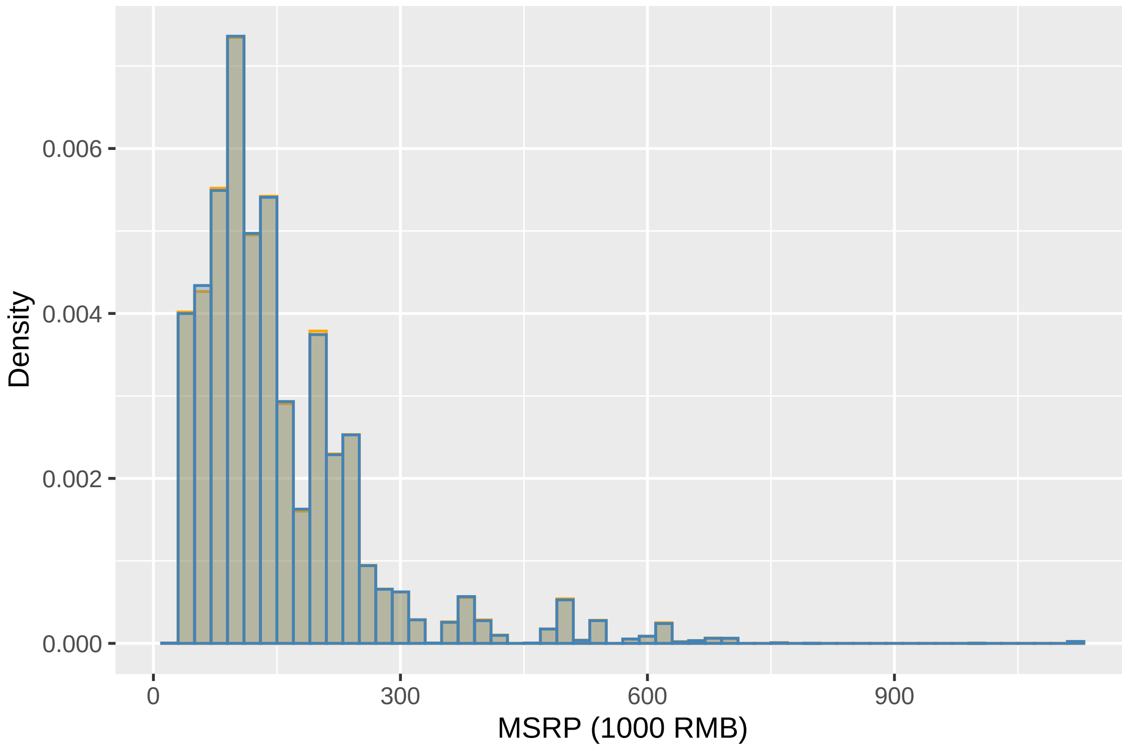

Figure 4: Comparision between placebo 1 vs 2

There were slight changes between two years in some MSRP price categories, and we can say that the optimal transport cost will be nonzero. But, it will still be very close to zero. From this visual inspection of the comparison plot, we can say that both distributions appeared to be drawn from the same distribution.

Optimal transport cost calculation

```
bandwidths <- c(0)

placebo_at_0 <- diftrans(pre_main = placebo_1,
                         post_main = placebo_2,
                         var = MSRP,
                         bandwidth_seq = bandwidths)
## ===========================================================================
```

```
placebo_at_0
##   bandwidth    main
## 1         0 0.01433
```

Exercise 4.4.

(a) Compute the transport cost between the two placebo distributions

```
bandwidths <- c(0, 500, 10000, 30000, 35000, 40000, 45000, 45150, 45198, 45200, 45500, 46000,

placebo_at_bw <- diftrans(pre_main = placebo_1,
                         post_main = placebo_2,
                         var = MSRP,
                         bandwidth_seq = bandwidths) %>%
  mutate(cat = "placebo")
## ========================================================================


placebo_at_bw
##    bandwidth        main     cat
## 1          0 0.014330328 placebo
## 2        500 0.011423255 placebo
## 3      10000 0.000929756 placebo
## 4      30000 0.000515733 placebo
## 5      35000 0.000512244 placebo
## 6      40000 0.000512244 placebo
## 7      45000 0.000512244 placebo
## 8      45150 0.000512244 placebo
## 9      45198 0.000512244 placebo
## 10     45200 0.000431990 placebo
## 11     45500 0.000418032 placebo
## 12     46000 0.000027825 placebo
## 13     50000 0.000027825 placebo
## 14     80000 0.000024361 placebo
## 15     90000 0.000024361 placebo
## 16    100000 0.000024361 placebo
```

(b) compute the transport cost between the observed distributions for 2010 and 2011 Beijing car sales

```
emprical_at_bw <- diftrans(pre_main = beijing_pre,
                         post_main = beijing_post,
                         var = MSRP,
                         bandwidth_seq = bandwidths) %>%
  mutate(cat = "emprical")
## ========================================================================
```

```
emprical_at_bw
##    bandwidth     main      cat
## 1          0 0.353123 emprical
## 2        500 0.326794 emprical
## 3      10000 0.151831 emprical
## 4      30000 0.088075 emprical
## 5      35000 0.077251 emprical
## 6      40000 0.064889 emprical
## 7      45000 0.055253 emprical
## 8      45150 0.055253 emprical
## 9      45198 0.055253 emprical
## 10     45200 0.055253 emprical
## 11     45500 0.055253 emprical
## 12     46000 0.055253 emprical
## 13     50000 0.053762 emprical
## 14     80000 0.039061 emprical
## 15     90000 0.030705 emprical
## 16    100000 0.027120 emprical
```

(c)

```
df_merged <- rbind(placebo_at_bw, emprical_at_bw)

ggplot(df_merged, aes(x = bandwidth, y = main, color = cat)) +
  geom_line() +
  labs(x = "Bandwidths",
       y = "the fraction of \n optimal transport cost",
       color = "between two distributions of") +
  theme(legend.position = "bottom")
```

(d) values of d, the placebo cost less than 0.05%

```
placebo_at_bw %>%
  arrange(-main) %>%
  filter(main < 0.0005)
##   bandwidth        main     cat
## 1     45200 0.000431990 placebo
## 2     45500 0.000418032 placebo
## 3     46000 0.000027825 placebo
## 4     50000 0.000027825 placebo
## 5     80000 0.000024361 placebo
## 6     90000 0.000024361 placebo
## 7    100000 0.000024361 placebo
```

From the bandwidth unit 45200, the optimal transfer cost become less tan 0.05%.

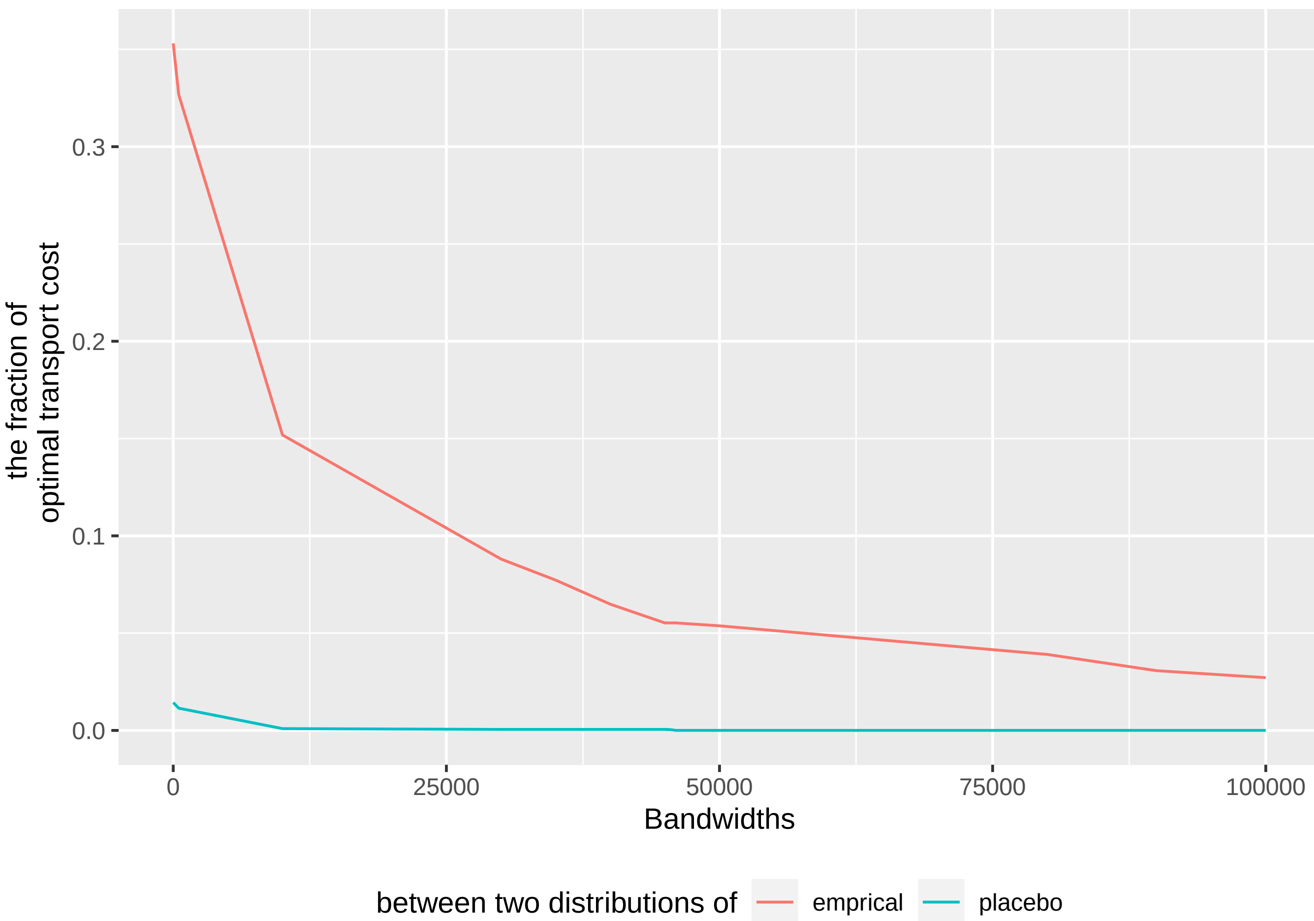

Figure 5: Comparision between Placebo costs vs Emprical costs

(e) The empirical transport cost at lowest value of d

```
emprical_at_bw %>%
  arrange(main) %>%
  filter(bandwidth == 45200)
##   bandwidth     main      cat
## 1     45200 0.055253 emprical
```

The optimal transfer cost at the smallest bandwidth we got from two placebo distributions was 5.5%.

4.6 Compute Differences-in-Transports Estimator

```
dit_at_0 <- diftrans(pre_main = beijing_pre,
                     post_main = beijing_post,
                     pre_control = tianjin_pre,
                     post_control = tianjin_post,
                     var = MSRP,
                     bandwidth_seq = c(0),
                     conservative = TRUE)
## ==========================================================================


dit_at_0
##   bandwidth    main  main2d control     diff   diff2d
## 1         0 0.35312 0.35312 0.29868 0.054443 0.054443
```

Exercise 4.5.

(a) compute diff2d for different values of d from 0 to 50,000.

```
bandwidths <- c(0, 1000, 2000, 3000, 3500, 3700, 3900, 3950, 4000, 4500, 4700, 4900, 4950, 500

dit_at_seq <- diftrans(pre_main = beijing_pre,
                     post_main = beijing_post,
                     pre_control = tianjin_pre,
                     post_control = tianjin_post,
                     var = MSRP,
                     bandwidth_seq = bandwidths,
                     conservative = TRUE)
## ==========================================================================

dit_at_seq
##    bandwidth     main   main2d   control     diff   diff2d
## 1          0 0.353123 0.353123 0.2986805 0.054443 0.054443
## 2       1000 0.258948 0.217756 0.1773211 0.081627 0.040435
```

```
## 3        2000 0.217756 0.184919 0.1136129 0.104143 0.071307
## 4        3000 0.202585 0.173243 0.0834465 0.119138 0.089797
## 5        3500 0.196131 0.169615 0.0813056 0.114825 0.088310
## 6        3700 0.196036 0.169615 0.0811506 0.114885 0.088465
## 7        3900 0.192364 0.169615 0.0732725 0.119091 0.096343
## 8        3950 0.192364 0.169615 0.0732725 0.119091 0.096343
## 9        4000 0.184919 0.167421 0.0655517 0.119368 0.101869
## 10       4500 0.182098 0.161738 0.0628204 0.119278 0.098917
## 11       4700 0.181991 0.161315 0.0627769 0.119214 0.098538
## 12       4900 0.179426 0.161315 0.0564635 0.122963 0.104851
## 13       4950 0.179426 0.161315 0.0564635 0.122963 0.104851
## 14       5000 0.177859 0.151831 0.0456166 0.132243 0.106214
## 15      10000 0.151831 0.123160 0.0200933 0.131737 0.103066
## 16      20000 0.123160 0.064889 0.0130559 0.110104 0.051833
## 17      25000 0.100769 0.053762 0.0071584 0.093610 0.046604
## 18      40000 0.064889 0.039061 0.0063769 0.058512 0.032684
## 19      50000 0.053762 0.027120 0.0043097 0.049453 0.022811
```

(b) placebo_Beijing_1

```
set.seed(4487989)

placebo_Beijing_1 <- data.frame(MSRP = beijing_pre$MSRP,
                                count = rmultinom(n = 1,
                                                  size = sum(beijing_pre$count),
                                                  prob = beijing_pre$count))

head(placebo_Beijing_1)
##    MSRP count
## 1 20800     0
## 2 29800    53
## 3 32900  3260
## 4 33800  3713
## 5 34800   557
## 6 36800  1695
```

(c) placebo_Beijing_2

```
set.seed(384620)

placebo_Beijing_2 <- data.frame(MSRP = beijing_pre$MSRP,
                                count = rmultinom(n = 1,
                                                  size = sum(beijing_post$count),
                                                  prob = beijing_pre$count))

head(placebo_Beijing_2)
##    MSRP count
## 1 20800     0
```

```
## 2 29800    21
## 3 32900  1293
## 4 33800  1575
## 5 34800   253
## 6 36800   739
```

(d) placebo_Tianjin_1

```
set.seed(4487989)

placebo_Tianjin_1 <- data.frame(MSRP = tianjin_pre$MSRP,
                        count = rmultinom(n = 1,
                                          size = sum(tianjin_pre$count),
                                          prob = tianjin_pre$count))

head(placebo_Tianjin_1)
##    MSRP count
## 1 20800     0
## 2 28800     0
## 3 29800    57
## 4 30900     0
## 5 32900   561
## 6 33300     4
```

(e) placebo_Tianjin_2

```
set.seed(384620)

placebo_Tianjin_2 <- data.frame(MSRP = tianjin_pre$MSRP,
                        count = rmultinom(n = 1,
                                          size = sum(tianjin_post$count),
                                          prob = tianjin_pre$count))

head(placebo_Tianjin_2)
##    MSRP count
## 1 20800     0
## 2 28800     0
## 3 29800    59
## 4 30900     0
## 5 32900   708
## 6 33300     5
```

(f)

```
dit_at_seq_placebo <- diftrans(pre_main = placebo_Beijing_1,
                     post_main = placebo_Beijing_2,
                     pre_control = placebo_Tianjin_1,
```

```
                      post_control = placebo_Tianjin_2,
                      var = MSRP,
                      bandwidth_seq = bandwidths,
                      conservative = TRUE)
## ===========================================================================

dit_at_seq_placebo
##    bandwidth        main      main2d     control         diff
## 1          0 0.014330328 0.014330328 0.018068490 -0.003738162
## 2       1000 0.005401229 0.003025417 0.006996325 -0.001595097
## 3       2000 0.003025417 0.001940311 0.003982996 -0.000957578
## 4       3000 0.002522998 0.001518236 0.002289604  0.000233395
## 5       3500 0.002519258 0.001251755 0.002269466  0.000249791
## 6       3700 0.002519258 0.001234603 0.002269466  0.000249791
## 7       3900 0.001971174 0.001228782 0.002269466 -0.000298292
## 8       3950 0.001971174 0.001228782 0.002269466 -0.000298292
## 9       4000 0.001940311 0.001181409 0.002212858 -0.000272546
## 10      4500 0.001896917 0.001005903 0.002204186 -0.000307268
## 11      4700 0.001878805 0.001000090 0.002203893 -0.000325087
## 12      4900 0.001706366 0.001000090 0.001952998 -0.000246632
## 13      4950 0.001706366 0.001000090 0.001952998 -0.000246632
## 14      5000 0.001541944 0.000929756 0.001870240 -0.000328295
## 15     10000 0.000929756 0.000739296 0.000687044  0.000242712
## 16     20000 0.000739296 0.000512244 0.000469484  0.000269812
## 17     25000 0.000579671 0.000027825 0.000225770  0.000353901
## 18     40000 0.000512244 0.000024361 0.000194606  0.000317638
## 19     50000 0.000027825 0.000024361 0.000013221  0.000014604
##          diff2d
## 1  -0.003738162
## 2  -0.003970908
## 3  -0.002042684
## 4  -0.000771368
## 5  -0.001017712
## 6  -0.001034863
## 7  -0.001040685
## 8  -0.001040685
## 9  -0.001031449
## 10 -0.001198283
## 11 -0.001203803
## 12 -0.000952908
## 13 -0.000952908
## 14 -0.000940483
## 15  0.000052252
## 16  0.000042760
## 17 -0.000197945
## 18 -0.000170245
## 19  0.000011139
```

(g) absolute value of the placebo differences-in-transports estimator

```r
ggplot(dit_at_seq_placebo, aes(x = bandwidth, y = abs(diff2d))) +
  geom_line() +
  labs(x = "Bandwidths",
       y = "transport cost")
```

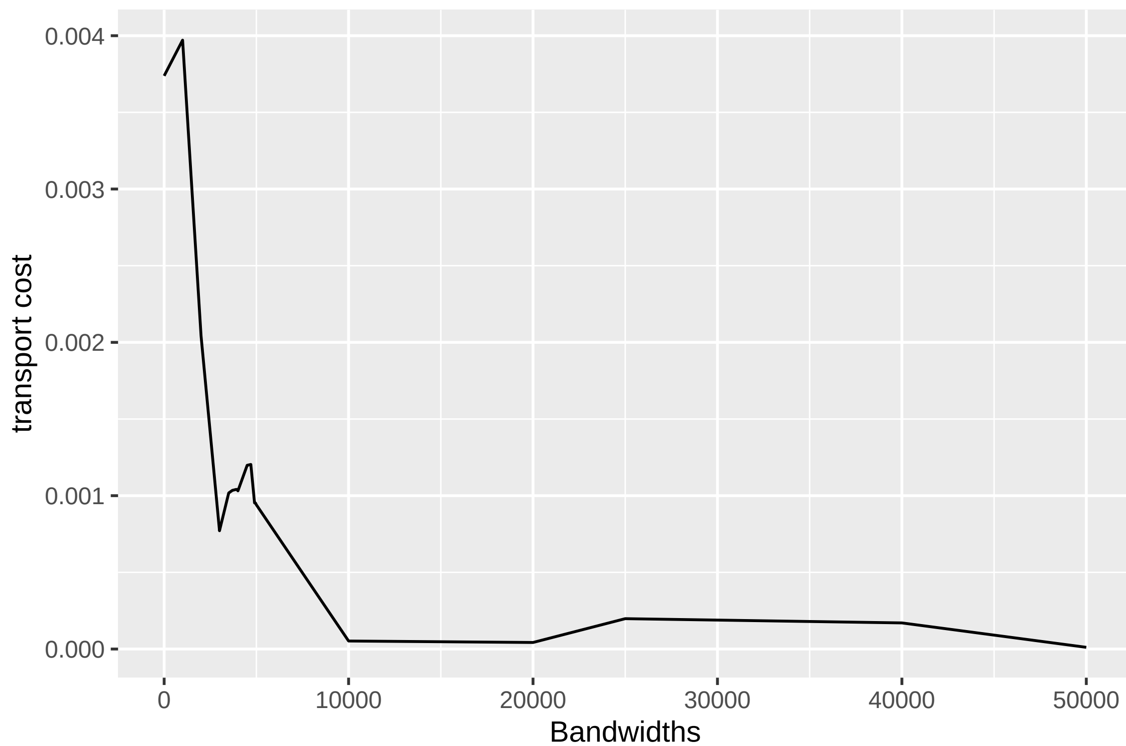

Figure 6: Placebo distribution differences in transport cost

(h) the absolute value of the placebo differences-in-transports estimator stay below 0.05%

```r
lower_bound_d <- dit_at_seq_placebo %>%
  mutate(diff2d_abs = abs(diff2d)) %>%
  arrange(bandwidth) %>%
  filter(diff2d_abs < 0.0005)

lower_bound_d
##   bandwidth        main      main2d     control        diff
## 1     10000 0.000929756 0.000739296 0.000687044 0.000242712
## 2     20000 0.000739296 0.000512244 0.000469484 0.000269812
## 3     25000 0.000579671 0.000027825 0.000225770 0.000353901
## 4     40000 0.000512244 0.000024361 0.000194606 0.000317638
## 5     50000 0.000027825 0.000024361 0.000013221 0.000014604
##          diff2d  diff2d_abs
## 1  0.000052252 0.000052252
## 2  0.000042760 0.000042760
## 3 -0.000197945 0.000197945
## 4 -0.000170245 0.000170245
## 5  0.000011139 0.000011139
```

As the unit of bandwidth increases, the placebo differences-in-transports estimators' values become smaller. At bandwidth unit 10000, the transport cost estimator became less than 0.05%. This trend can also be observed in the above plot.

(i) emprical differences-in-transports estimator

```
lower_bound_d <- lower_bound_d %>% select(bandwidth)

inner_join(lower_bound_d, dit_at_seq, by = c("bandwidth" = "bandwidth")) %>%
  arrange(-diff2d) %>%
  slice(1)
##   bandwidth    main  main2d  control    diff  diff2d
## 1     10000 0.15183 0.12316 0.020093 0.13174 0.10307
```

The largest value of the empirical differences-in-transports estimator is 10.31% with bandwidth unit value 10000.