Problem Set - Uber Case

Shu Wang

Install the packages if you don't have them installed yet. Install them only once, and not when you run you render.

1. Read the HBS Case. What is the difference between Uber POOL and Express POOL? No more than two sentences.

Express POOL offers a cheaper ride to users in exchange for waiting and walking. Users in Express POOL need to wait for two minutes then walked to a designated pick-up point after matching.

2. How did Uber use surveys in designing Uber Express Pool?

With concerns on Express pricing, uber conducted a survey on assessing riders' willingness to wait and walk at different points. The survey measured consumers' sensitivity to different variables and Uber built a calculator to calculate the floor and ceiling for pricing based on walking and waiting parameters.

3. Suppose Uber was considering a new algorithm to recommend ride destinations in the app. Which type of research strategy should they use (A/B Test, Switchback, Synthetic Control)? No more than two sentences.

Switchbacks would be able to test a new algorithm by switches between the original algorithm and the new algorithm continuously during an observed experiment period(e.g, 2 weeks).Switchbacks would make it able to compare differences in efficiency and customer satisfaction between two groups.

4. Suppose Uber was considering a radio advertising campaign. Which type of research strategy should they use (A/B Test, Switchback, Synthetic Control)? No more than two sentences.

Synthetic(control) would be the best fit to conduct a research strategy on radio advertising because due to the nature of radio stations/radio channels, it is more controllable to conduct the advertising on a city-level rather than user level. Uber would roll out the radio advertising in a selected number of cities and create “synthetic controls” by studying the same outcome varibale in a set of another cities that is not exposed to the radio advertising campaign.

5. Create two new columns in the dataset that represent the total number of trips for both pool products and the profit from these products. (10 points)

(remember you can create a new column by: data[, new_col_name := whatever you want the new column to contain])

```
data[,total_trips := trips_pool + trips_express_pool]
data[,profit := revenue - total_driver_payout_sr]
```

6. Plot the average number of trips as a function of the time of the day. Describe a reason why this pattern exists (no more than 2 sentences). (20 points)

Hint: You can use ggplot to do this. As in assignment 1, you’ll first have to create a dataset with the average number of trips by time of the day.

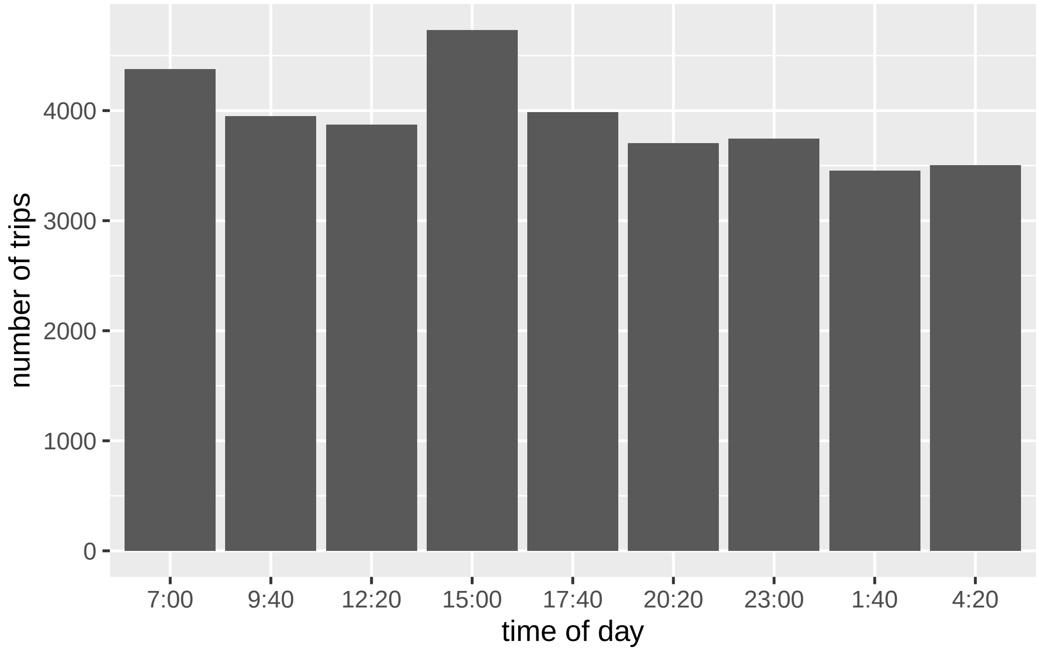
```
# calculate average number of trips based on time slot
avg_700<- mean((dplyr::filter(data, grepl('7:00', period_start)))$total_trips)
avg_940<- mean((dplyr::filter(data, grepl('9:40', period_start)))$total_trips)
avg_1220<- mean((dplyr::filter(data, grepl('12:20', period_start)))$total_trips)
avg_1500<- mean((dplyr::filter(data, grepl('15:00', period_start)))$total_trips)
avg_1740<- mean((dplyr::filter(data, grepl('17:40', period_start)))$total_trips)
avg_2020<- mean((dplyr::filter(data, grepl('20:20', period_start)))$total_trips)
avg_2300<- mean((dplyr::filter(data, grepl('23:00', period_start)))$total_trips)
avg_140<- mean((dplyr::filter(data, grepl('1:40', period_start)))$total_trips)
avg_420<- mean((dplyr::filter(data, grepl('4:20', period_start)))$total_trips)
```

```
time_of_day <- c("7:00","9:40","12:20","15:00","17:40","20:20","23:00","1:40","4:20")

avg_trips <- c(avg_700,avg_940,avg_1220,avg_1500,avg_1740,avg_2020,avg_2300,avg_140,avg_42
df <- data.frame(time_of_day, avg_trips)
```

```
#rearrange order
df$time_of_day <- as.character(df$time_of_day)
#Then turn it back into a factor with the levels in the correct order
df$time_of_day <- factor(df$time_of_day, levels=unique(df$time_of_day))

ggplot(df,aes(x = time_of_day,y = avg_trips)) +geom_bar(stat = 'identity') + xlab("time of
```



As we observe from the bar plot, commute hours has a higher average number of total trips compared to non-commute hours.7am and 3pm are two time periods that people started to going to school/work and going back to home. This directly leads to more traffic on road hence more trips by uber.

7. Conduct a regression analysis of the experiment (considering the outcomes: revenue, total_driver_payout_sr, rider_cancellations, total_trips). Make sure to think carefully about the correct regression specification. The regression output should be easy to read, so use 'etable' or 'modelsummary'. What do you learn in words from this regression analysis (no more than 5 sentences but it can be less)?

Hint: We should control for the fact that different times of the day and different days have different demand patterns. (Please refer to p.13 of the HBS article to see why) Hint: The syntax for fixed effects is: feols(outcome ∼ treatment_name | fixed_effect_name1 + fixed_effect_name2, data = data, se = 'hetero') Hint: You can output multiple regressions in this way: etable(reg1, reg2)

```
# take weekend/weekday into consideration
time_raw <- strptime(data$period_start,"%m/%d/%y %H:%M")
day_of_week <- format(time_raw,"%u")
data[, day_of_week := day_of_week ]

reg1 = feols(total_trips ~ treat | commute + day_of_week,data = data, se = 'hetero')
reg2 = feols(revenue ~ treat | commute + day_of_week,data = data, se = 'hetero')
reg3 = feols(total_driver_payout_sr ~ treat | commute + day_of_week,data = data, se = 'het
reg4 = feols(rider_cancellations ~ treat | commute + day_of_week,data = data, se = 'hetero
etable(reg1,reg2,reg3,reg4)
```

```
                          reg1           reg2                   reg3
Dependent Var.:    total_trips        revenue total_driver_payout_sr

treatTRUE       -88.37 (70.64) -272.1 (695.1)    -2,106.8*** (621.4)
Fixed-Effects:  -------------- -------------- ----------------------
commute                    Yes            Yes                    Yes
day_of_week                Yes            Yes                    Yes
_______________ ______________ ______________ ______________________
S.E. type       Heterosk.-rob. Heterosk.-rob. Heteroskedastici.-rob.
Observations               126            126                    126
R2                     0.55141        0.71334                0.61695
Within R2              0.01320        0.00131                0.08947

                               reg4
Dependent Var.: rider_cancellations

treatTRUE          26.52*** (5.148)
Fixed-Effects:  -------------------
commute                         Yes
```

```
day_of_week                        Yes
--------------- -------------------
S.E. type       Heteroskedast.-rob.
Observations                    126
R2                          0.71332
Within R2                   0.18494
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above regression results,with commute and day_of_week as fixed effects, it is clear that treatment has a significant effect on ubers' costs(total_driver_payout) and rider_cancellations at 5% significance level.In other words, treatment increases cancellations and decreases drivers' payout. Treatment also shows negative impact on revenue generated and total trips completed but the results is not significant at 5% significance level. It might be helpful to change fixed effect variables and test the effect of treatment.

8. One of your data scientists suggests that the optimal wait time may differ by whether it's a commuting period. Test whether the effects of a 5 minute wait period on total trips and cancelations differ by whether it's a commuting period (the column 'commute'). Which policy works better during commute times? (10 points)

```
reg_k <- feols(rider_cancellations ~ commute * wait_time,
data = data, se = 'hetero')
etable(reg_k)
```

```
                                           reg_k
Dependent Var.:              rider_cancellations

Constant                         149.1*** (3.673)
commuteTRUE                      96.31*** (16.27)
wait_time5mins                   20.81*** (4.813)
commuteTRUE x wait_time5mins       35.99. (18.32)
---------------------------- -------------------
S.E. type                    Heteroskedast.-rob.
Observations                                 126
R2                                       0.72625
Adj. R2                                  0.71952
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
reg_l <- feols(total_trips ~ commute *
                 wait_time,
data = data, se = 'hetero')
etable(reg_l)
```

```
                                          reg_l
Dependent Var.:                     total_trips

Constant                     3,764.7*** (49.04)
commuteTRUE                  1,280.6*** (160.3)
wait_time5mins                   -44.28 (71.46)
commuteTRUE x wait_time5mins     -277.7 (229.5)
---------------------------- ------------------
S.E. type                    Heteroskedas.-rob.
Observations                                126
R2                                      0.54891
Adj. R2                                 0.53782
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on our above regression which contains interaction between commute and treatment,during commute times, treatment(5_min wait time) has lead to an additional cancelled trips of (20.8+36) on average,the result is significant at 10% significance level. For total number of trips,the regression shows that treatment has lead to an additional decrease of total numbers of (44.28+277.7) orders on average. However,this result is not significant.

Combined these two results together, during commuting times, control group(waiting time = 2 min) performs slightly better compared to treatment group in total trips and # of cancelled trips.