**Effect of Stores on Online Sales**

**Group 6**

1. **Avinash Vamshi Kanthreegala - AXK220103**
2. **Haitao Wang - HXW210009**
3. **Rakshit Mathur - RXM210132**
4. **Jaymala Chavan - JSC200005**

**Question 1 (2 points) Students should compute the ATE from equation (1) and analyze its statistical significance using a t-test. Why is this treatment effect estimate biased?**

ATE ON ONLY TREATED CUSTOMERS

ATE = YTrpost - YTrprior



T Test

A screenshot of a computer code

Description automatically generated

P value equals 2.2e-16

There is a difference in mean in the pre and post online sales revenue for the treatment group and it is significant as p <0.05.

This estimates the treatment effect as the difference in mean online revenue of treated customers before and after the treatment (store opening). However, this approach is biased because it does not account for underlying trends in online revenue over time. For instance, the increase in online revenue for treated customers from pre to post period may be partly due to overall ecommerce growth rather than the store opening.

**Question 2 (2 points) Students should compute the ATE using equation (2) and analyze its statistical significance using a t-test. Why is this treatment effect estimate biased?**

DIFFERENCE-IN-MEANS ATE

ATE = YTrpost - YCtpost



T Test

A computer code with black text

Description automatically generated

P value equals 2.2e-16

There is a difference in mean in the post treatment and post control online sales revenue and it is significant as p <0.05.

This compares mean online revenue of treated vs control groups after treatment. However, this approach is biased as well because the two groups likely differ in their underlying characteristics like, treated customers likely have higher purchase propensities than control customers. We would need to match customers properly before comparing outcomes.

**Question 3 (2 points)**

**What variables should be included in the computation of propensity scores and Why? Explain.**

Variables should be included in the computation of propensity scores: Customer-level pre-purchase behavior, their individual socioeconomic factors demographic factors.

**Customer purchase behavior variables:**

**PreTotalInt, PreTotalPur, PreTotalRev, PreOnlineInt, PreOnlinePur,PreOnlineRev,PreStoreInt,**

**PreStorePur, PreStoreRev**

We can use the observed customer-level purchase/return variables capturing the recency, frequency, and monetary values of their purchases prior to store openings. How frequently are purchases made in store/online? How many returns made? How much revenue generated by purchase in store/online. Unaffected customers, who exhibited similar purchase behavior as the affected customer prior to store opening, are likely to closely approximate the counterfactual purchase behavior of affected customers after store openings.

**Customer-level socio-demographic variables:**

**IncCat, AgeCat**

Age, Income- Similar age and income more likely to have same purchase preferences.

**Q4. Students should estimate the effect of store opening by finding a matched sample of customers and conducting a t-test to compare online purchase revenue of treated and control customers in the post-period. Why would the estimate of treatment effect be biased in this case?**

library(readxl)

DataFInal\_1\_ <- read\_excel("Downloads/DataFInal(1).xlsx", sheet = "Raw Data")

head(DataFInal\_1\_)

attach(DataFInal\_1\_)

install.packages("MatchIt")

library(MatchIt)

set.seed(42)

Match = matchit(Treat ~ PreTotalInt + PreTotalPur + PreTotalRev + PreOnlineInt + PreOnlinePur

+ PreOnlineRev + PreStoreInt + PreStorePur + PreStoreRev + IncCat + AgeCat , data = DataFInal\_1\_, method = 'nearest', caliper = 0.001)

#Checking the number of households dropped from treatment customers and control customers

dim(DataFInal\_1\_[Treat==0,])

dim(DataFInal\_1\_[Treat==1,])

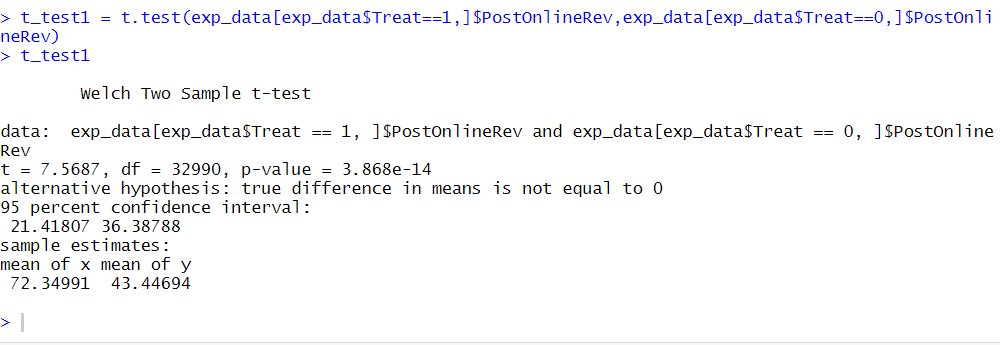
summary(Match)

#Only **450** records out of 17277 records are **not** matched in Treatment group and they are dropped so Caliper=0.001 is used.

exp\_data = match.data(Match)

t\_test1 = t.test(exp\_data[exp\_data$Treat==1,]$PostOnlineRev,exp\_data[exp\_data$Treat==0,]$PostOnlineRev)

t\_test1



#Here p value < 0.05 so we conclude that there is a significant different in means of post online purchase for treatment and control groups. So the opening a store affect the online purchases by customers.

#Although Propensity score accounts for observed characteristics of treatment and control groups customers still there are other unpredictable characterisitcs of customers which could not be accounted. To eliminate these differences between both the groups we need to consider the difference in mean of online purchases for treatment and control groups before treatment period in our analysis.

**Question 5 (2 points) We cannot use the data in the provided format to run a regression analysis to find the DiD estimator. Explain how the format of the data needs to be changed so that we can run the regression analysis using the DiD estimator**

In order to find the DiD estimator by running a regression analysis, we should have panel data which requires collection of data points over multiple time periods. In order to identify the treatment and control group in the pre-treatment and post-treatment periods. So, we could add a dummy variable for identify treatment and separate pre-treatment data and post-treatment data into two records in the raw data. Then we can identify the 4 situations (1. treatment group/pre-store opening, 2. treatment group/post-store opening, 3. control group/pre-store opening, 4. control group/post-store opening). Additionally if data is available, we could add monthly transaction data so we can add a fixed effect on time by adding month dummy variable.