Casual effect of stores on online sales

Introduction

A multichannel apparel and home goods retailer suffered massive losses due to closure of its stores due to COVID-19 pandemic. The CEO of the company was contemplating whether to close the stores permanently or wait till economy revives. But it could be possible that online sales are highly interdependent with the stores.

Objective: To determine the effect of stores on online purchase metrics.

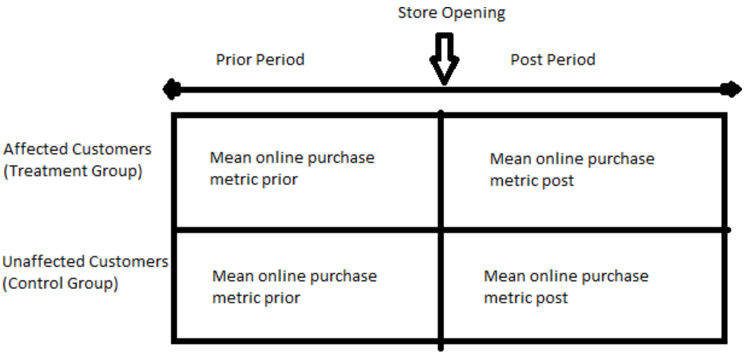
Approach
Design quasi-natural experiment using past transaction data. Quasi-natural experiments do not involve random application of a treatment. Instead, a treatment is applied due to social or political factors, such as a change in laws or natural disasters. The recipients of the treatment are thus not randomly but intentionally chosen according to some predetermined criteria.
In this case, new stores being opened is taken as the natural event. Six new stores were opened by the firm in year 200X.Data on all customer interactions for 4 years around 200x, two years prior and two years after 200X was provided. The distance of each customer from the nearest store was also provided. As a result of store openings, store distance reduced for approximately 17000 customers and it remained the same for 1.5 million remaining customers. The 17000 customers were taken as affected(treated) customers and 50,000 customers were randomly chosen from the 1.5 million customers as unaffected customers(control). Treatment being distance to nearest store affected by new store opening. The two group are then compared for treatment effect of stores opening using Difference in Difference (DID) and matching approach.

Data Dictionary:

Variable Name	**Variable Description**
CustID	Customer ID
Treat	Treatment indicator (Treat=1 for affected customer and =0 for unaffected customer
IncCat	Income category of the customer household [1: < $50K, 2: $50-$75K, 3: $75-$100K, 4: $100-$150K, 5: >$150K, and 6: unknown]
AgeCat	Age category of head of the customer household [1:< 25 Yrs, 2:25-34 Yrs, 3: 35-44 Yrs, 4: 45-54 Yrs, 5: 55-64 Yrs, 6: >65 Yrs, and 7: unknown]
Variables before store opening	
BefDist	Distance from the nearest store for the customer before store opening (miles)
PreTotalInt	Total purchase interactions by the customer before store opening
PreTotalPur	Total purchase quantity for the customer before store opening
PreTotalRev	Total purchase revenue for the customer before store opening
PreOnlineInt	Number of Online purchase interactions for the customer before store opening
PreOnlinePur	Online purchase quantity for the customer before store opening
PreOnlineRev	Online purchase revenue (USD) for the customer before store opening
PreStoreInt	Number of Store purchase interactions for the customer before store opening
PreStorePur	Store purchase quantity for the customer before store opening
PreStoreRev	Store purchase revenue (USD) for the customer before store opening
Variables after store opening	

AftDist	Distance from the nearest store for the customer after store opening (miles)
PostOnlineInt	Number of Online purchase interactions for the customer after store opening
PostOnlinePur	Online purchase quantity for the customer after store opening
PostOnlineRev	Online purchase revenue (USD) for the customer after store opening
PostStoreInt	Number of Store purchase interactions for the customer after store opening
PostStorePur	Store purchase quantity for the customer after store opening
PostStoreRev	Store purchase revenue (USD) for the customer after store opening

The metrics from data chosen to compare the two groups are online purchase interaction, online purchase quantity and online purchase revenue.



Average Treatment Effect (ATE) on only affected (treated) customers (post – pre)

We first performed a F-test to test the equality of variances and found that for all the three metrics the variances are unequal for pre- and post-period. Then, we performed the t-test for unequal variances and found all of them to be significant.
The estimates are shown below in the table for all the three metrics.

Metric	**Pre-period Mean**	**Post-period Mean**	**ATE**	**F-Test p-value**	**t-test(two-tailed) p-value**
Online Purchase Interaction	0.303756439	0.532268334	0.228511894	0	2.56431E-34
Online Purchase Quantity	0.794466632	1.299357527	0.5048909	0	7.07768E-15
Online Purchase Revenue	44.79431151	83.26673612	38.4724246	4.1162E-248	7.07313E-22

This treatment effect only accounts for the treated customers and doesn't account for the other customers which are the majority. So, this treatment effect is not the right approach because it leaves out the majority of the population out of consideration. Also, there could be some other factor which could have happened in the post period which is influencing this difference. We cannot be sure whether the effect is just because of the treatment.

Difference in mean ATE of unaffected customers and affected in post period.

Metric	**Control Average**	**Treatment Average**	**ATE**	**Control Std Dev**	**Treatment Std Dev**	**t-value**
Online Purchase Interaction	0.299068385	0.532268334	0.233199949	1.583134952	1.993467081	13.94

Online Purchase Quantity	0.774905847	1.299357527	0.52445168	8.023268726	6.902439576	8.258
Online Purchase Revenue	48.89881546	83.26673612	34.36792066	437.6800663	415.8603556	9.249

We see that all of the metrics are significant with t-values much higher than critical t-value of 1.96. Online Purchase interaction increases by 0.2332, quantity by 0.5245 and revenue by 34.368 after the store opens in the vicinity. But, to compare the two groups the characteristics of the two groups of customers should be same.
Analysing the prior store opening period purchase behaviour, there was significant difference between the characteristics of the two groups as shown below:

Purchase Behaviours	**PreTotalInt**	**PreTotalPur**	**PreTotalRev**	**PreOnlineInt**	**PreOnlinePur**	**PreOnlineRev**	**PreStoreInt**	**PreStorePur**	**PreStoreRev**
Control Average	11.7536	36.5790	1759.6132	0.2076	0.5333	28.1014	4.1872	14.2140	649.3555
Treatment Average	7.5693	23.6202	1342.5828	0.3038	0.7945	44.7943	2.3679	8.4642	453.2587
Difference	-4.1844	-12.9588	-417.0303	0.0961	0.2612	16.6929	-1.8193	-5.7498	-196.0968
Control Std Dev	17.9255	69.6304	3756.4791	1.0807	3.5666	192.6768	8.4971	35.0385	1907.6877
Treatment Std Dev	12.2428	51.5109	5208.7190	1.4342	4.9985	321.5504	6.0091	26.1521	2479.2727
t-Value	-34.1145	-25.9342	-9.6956	8.0623	6.3379	6.4392	-30.6601	-22.7421	-9.4796

The purchase behaviour of the two groups before the store opening period is totally different as we can see that all the t values are greater than critical value of 1.96.
Also, for the online channel the treatment group already has a higher purchase value than the control group. That means the two groups are not statistically similar.

ATE calculated before doesn't account for the difference between the two groups before the store opening period. So, we can't comment the ATE post store opening is only due to treatment effect or not. To tackle this issue, we conduct propensity score matching on multiple variables like age, income and other purchase variables to find similar customers in treatment and control groups.

<u>Propensity score = Logit probability (Treatment of easier store access | Pre store treatment variables)</u>

We will do the analysis in the following steps:

- From here on, we transformed the dataset into a panel dataset with the unit of analysis as CustID and the time unit as Post, where Post=0 from all the pre observations and Post=1 for all the post observations. Also, we removed the PreTotal variables from the dataset as they were not available for the post period.
- CustID, Treat, Post, BefDist, IncCat, AgeCat, OnlineInt, OnlinePur, OnlineRev, StoreInt, StorePur, StoreRev & AftDist are the variables in the final dataset for regression.
- The variables that can be used for propensity score matching will be done on the demographic variables and the variables that can affect online purchase interaction, quantity and revenue.
- Here, we will use IncCat, AgeCat, BefDist, OnlineInt, OnlinePur, OnlineRev, StoreInt, StorePur and StoreRev (for Post=0) to match the customers. This way we can match each customer in the treatment group to another customer in the control group based on their demographics

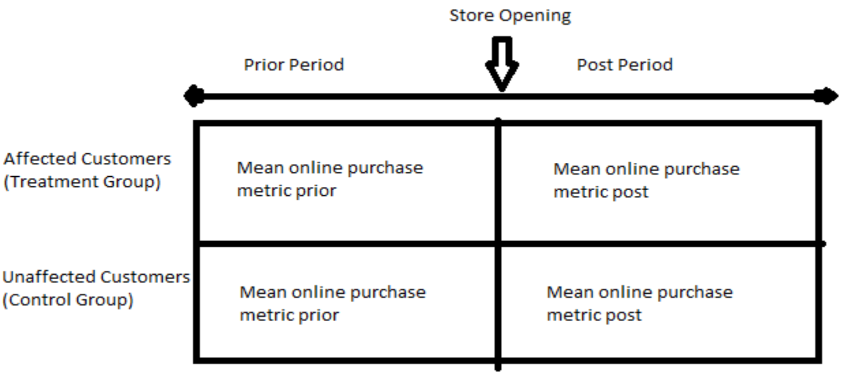
and the Distance of the store in the before period, and their purchase behaviors in the Pre-period.

- We matched the customer using matchit function in R with the following specification: - matchit(Treat~IncCat+AgeCat+BefDist+OnlineInt+OnlinePur+OnlineRev+StoreInt+StorePur+StoreRev ,data = MyData[Post == 0], method = 'nearest', caliper=0.001)

which matched 12320 treatment customers to the customers in the control group.The ATE is then computed as the difference in online purchase behaviour of treated(affected) customers from that of their matched control(unaffected) customers,which is given by coefficient of post. But, matching only accounts for observed differences between the two groups.

Difference in Difference (DID) Approach

DID approach is used to account for unobserved differences.Difference in difference approach is as shown below:



Potential effect of stores opening on online sales=[mean online purchase metric of affected customers post store opening-mean online purchase metric of unaffected customers post store opening]- [mean online purchase metric of affected customers prior store opening- mean online purchase metric of unaffected customers prior store opening].

Using the matched data, we performed a panel regression for online purchase behaviour using the following specification:

Post is 0 for prior store opening period and 1 for post. Treat is 0 for unaffected customers and 1 for affected.

- plm (OnlineInt ~ Post +Post:Treat, data= MyData[match == TRUE],index =c("CustID","Post"),model="within") (Online Interaction)
- plm (OnlinePur ~ Post +Post:Treat, data= MyData[match == TRUE],index =c("CustID","Post"),model="within") (Online Purchase)
- plm (OnlineRev ~ Post +Post:Treat, data= MyData[match == TRUE],index =c("CustID","Post"),model="within") (Online Reversal)

R Output:

Effect of Store Opening on Online Channel

	Dependent variable:		
	OnlinePurchaseInteraction	OnlinePurchaseQuantity	OnlinePurchaseRevenue
	(1)	(2)	(3)
Post	0.078***	0.212***	18.388***
	(0.012)	(0.045)	(2.913)
Post:Treat	0.114***	0.193***	13.423***
	(0.019)	(0.067)	(4.171)
Observations	49,280	49,280	49,280
R2	0.009	0.004	0.006
Adjusted R2	-0.982	-0.993	-0.988
F Statistic (df = 2; 24638)	114.878***	46.885***	77.605***
Note:	*p<0.1; **p<0.05; ***p<0.01		

Interpreting the R Output:

The coefficient estimates of Post: Treat gives us the estimate of DID with matching. We see that it is significant and positive for all the metrics, thus suggesting an increase in the purchase behaviour. The Online Purchase interaction increases by **0.114**, Online Purchase Quantity increases by **0.193** and Online Purchase Revenue increases by **13.423** for the affected customers as compared to prior store opening. This indicates stores have a complimentary effect on online sales.

Why do stores increase online sales?

- Stores help customers get assurance about a product. Let's take an example of a skirt, If I want to buy a skirt, I will see it online and select a few samples. If I have a store in the vicinity, I will go to the store to try them on and select the best one. Then I will buy that skirt online (as most of the items online have lesser price than the store).
- If a customer doesn't find his desired product in the store, the store employees help them find it online, resulting in an online purchase, which is caused due to the store.
- It's easier to return products purchased online when there is a store nearby. Hence making online purchasing less risky.
- Stores attract new customers as well and when these customers become loyal customers they start purchasing from the stores' online channel as its easier than travelling to the store.