The Impact of Online Grocery Shopping (OGS) on the Healthiness of Households' Food Purchases

Project Datapreparation & Workflow Management 2024 - Team 14

Introduction

The demand for healthier food consumption has increased significantly in recent years. For many households, supermarket retailers are still the primary source for their food purchases. By shaping the environment in which purchasing decisions are made, retailers play a crucial role in the forming of consumers' food purchasing behavior and habits. Recent literature has investigated how the healthiness of online grocery purchases differs from offline (in-store) purchases (Chintala et al, 2024; Harris-Lagoudakis, 2022; Huyghe et al, 2017). While online grocerybaskets tend to be healthier compared to offline baskets, it is unclear if the adoption of OGS contributes to healthier consumption or if it results in a redistribution across channels, where consumers simply shift the more healthy purchases online and purchase unhealthy products mainly offline. More research is needed that examines how households alternate between online and offline shopping trips and how they allocate their purchases across both channels. As such, the central question in this project is:

How does the transition to hybrid grocery shopping affect the healthiness of food purchases across both online and offline grocery channels?

Methodology

Data

This study uses real world panel purchase data of supermarkets in the Netherlands in 2019. The dataset contains real-world purchases by consumers in two grocery retailers and is collected by an international market research company. To ensure confidentiality, the market research company, retailers, and brands are anonymized. Additionally, the data has been modified to prevent any potential link back to the original entities. Given that this data is not publically accessible, the dataset has been stored in a separate branch in this GitHub directory. This was done to satisfy the automation requirements of the project: the makefile will download the data from this branch and store it into the main/master branch. Listed below are the final variables used in this analysis.

The dataset contains purchase data of 150 unique households in the Netherlands in 2019. These households made over 180.000 individual product purchases across 26 different retailers. Listed below are the relevant variables after cleaning the data.

Variable	Description
Panelist	Unique identifier for each household
Date	Date on which the purchase was made
Barcode	Barcode of the product purchased
Retailer	Identifier for the supermarket where the purchase was made
Brand	The brand of the product
Units	Total purchased units of the product
Value	Total value of the sales for the product (in Euro cents)
Volume	Total volume of the product sold
Channel	Method of purchase (i.e. offline or online)
Measurement unit	The unit at which the volume of a product is measured
Segment	A category indicator to identify product groups

Table 1: Description of variables in dataset

Model specification

The formal analysis examines how the healthiness of households' grocery baskets changes once they start shopping in online channels in addition to their in-store purchases. I employ a Difference-in-Difference (DiD) approach to compare the changes in healthiness of households weekly grocery purchases for households that transition to hybrid shopping against those that continue shopping exclusively offline. Given that households get treated (i.e. adopt OGS) at different times, I use the staggered DiD approach proposed by Sant'Anna & Callaway (2021) to analyse the data and examine the research question. To measure the healthiness of grocery baskets, I use the proportion of total expenditure that is spent on the product category fruits and vegetables.

The staggered DiD method from Sant'Anna & Callaway (2021) works by creating cohorts for each distinct week in which households begin shopping online. For each cohort, a separate DiD regression is conducted, comparing the change in the healthiness of grocery purchases before and after the adoption of online shopping for that specific group, while using households that do not adopt OGS ('never-treated') as the control group. The final estimate is an aggregated average, where the treatment effects from each cohort-specific DiD analysis are weighted to provide an overall measure of how online grocery shopping influences household food healthiness. This approach ensures that the variation in adoption times is accurately reflected in the analysis.

Results

The overall effect of the DiD regression can be found in the table below. Given that the Sant'Anna & Callaway method runs a DiD regression for every cohort, we can not judge significance based on a p-value for the overal effects. Rather, by constructing a 95% confidence interval, I find that the proportion of expenditure towards fruits and vegetables does not significantly change after adoption of OGS. In the gen/paper directory more plots can be found that analyse cohorts individually and that look at the change in healthiness over a 8-week time window around adoption. This latter plot thus looks at the DiD results through the lens of an event study to look at changes in trends around adoption. However, this plot does not show significant results either.

ATT	**Std_error**	**CI_lower**	**CI_upper**
-2.44	1.574	-5.526	0.646