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

Project Data Preparation & Workflow Management

Melle Klein Goldewijk

2024-09-05

Research Motivation

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. Over the past decade, there has been a noticeable rise in Online Grocery Shopping (OGS), and the adoption of OGS has significantly accelerated since the start of the COVID-19 pandemic. 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). This research generally finds that online shopping baskets tend to be healthier compared to offline baskets. Yet, these studies often neglect the fact that households that adopt OGS may be different from households that do not, and that their shopping baskets may be healthier to begin with. Second, they also neglect the fact that most households that shop online still buy the majority of their groceries in physical stores. These households are called hybrid shoppers, which implies that they utilize both online and offline shopping channels to make their purchases. Consequently, there is a gap in our understanding regarding the impact of OGS on the overall healthiness of groceries when considering both online and offline shopping behaviors together. While online grocery baskets 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?

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. We will employ a Difference-in-Difference (DiD) approach to compare the changes in healthiness of grocery baskets for households that transition to hybrid shopping against those that continue shopping exclusively offline. This approach allows for controlling for general trends that affect all households (e.g. COVID-19 pandemic). By using a DiD model, we can isolate the effect that OGS has on the healthiness of households' grocery food purchases.

Setup

Note:

The official project for my PhD uses household purchase data provided by AiMark. A non-disclosure agreement prevents me from sharing this data publicly, creating some complications with uploading the data on Github. Given that the nature of this course lies in creating an automated project and Github Repisitory, and not on the actual outcome of analysis, I have created a sample dataset similar to the acutal data that I use for my PhD. In this sample dataset, I have annonymized the households, barcodes and retailers, making it impossible to trace back to the actual purchase data provided by AiMark. This sample dataset will be uploaded on GitHub, so I am able to fully complete the project assignment as intended.

Required packages

This markdown uses several packages to download, analyze and print data. The packages that will be used are listed below and will be automatically installed, if not already present on the system.

- dplyr
- data.table
- xtable
- tinytex

Downloading & opening data

The dataset has been uploaded on GitHub and will be automatically downloaded and opened when running the RMarkdown. The downloaded data takes form of a Excel (csv) file of around 15MB.

Data exploration

The dataset is a panel data time series that captures detailed purchasing behavior at the household and retailer level for the year 2019 in the Netherlands. Each row represents a specific product purchased by a unique household on a given date, providing insights into individual consumption patterns over time. Key variables include the date of purchase, household and retailer identifier, and product details, such as barcode, quantity purchased and price. The panel structure allows for tracking changes in purchasing habits, retailer preferences, and product choices across time.

Note that the original dataset contains 15 variable, but not every one of those is relevant for this study. In de data preparation phase I will trim and recode the dataset in such a way that best allows for analysis of the study. The table below presents a brief description of the variables relevant for this study.

Variable	Description
Panelist	Unique identifier for each household
Date of purchase	Date on which the purchase was made
Barcode	Barcode of the product purchased
Retailer	Numeric indicator of the supermarket at which the purchase was made
Brand	The brand of the product
Total unit sales	Total units of the product sold
Total value sales	Total value of the sales for the product (in Euro cents)
Total volume sales	Total volume of the product sold
Purchase method	Method of purchase (i.e. offline or online)
category	Category indicator to group products (e.g. vegetables)
Measurement unit	The unit at which the volume of a product is measured
Volume per unit	The volume per unit
segment	A generalized category indicator, based on the 'category' variable

Table 1: Description of variables in dataset

This research is focused on distribution of purchases by the online and offline channel. 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. Table 2 below summarizes some key metrics from the dataset. Out of the 150 households in the panel, 26 have purchased online at least once in 2019. These 26 households combine for just over 6000 online purchases, meaning that approximately 3% of the purchases is made online. The 'online households' are the households of interest for this study, and will serve as the treated group in the difference-in-differences analysis I will perform later on.

Metric	Count
Number of Households	150
Number of Retailers	26
Number of Purchases	182706

Table 2: Summary structure dataset