Amsterdam School of Economics
Faculty of Economics and Business

Master’s Thesis

Estimating Price Elasticity of Demand using the Double Machine Learning Method

Florence Piet

Student number:	11891920
Date of final version:	March 31, 2023
MSc programme:	Econometrics
Specialization:	Data Analytics

Supervisor:	Dr. A. Juodis
External Supervisor:	Dhr. S. Hennekes, MSc.
Second reader:	Dr. N. P. A. van Giersbergen

Faculty of Economics and Business

Statement of Originality

This document is written by Florence Piet who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document is original and that no sources other than those mentioned in the text and its references have been used in creating it. The Faculty of Economics and Business is responsible solely for the supervision of completion of the work, not for the contents.

Abstract

This Master's thesis examines the appropriate model for estimating price elasticity of demand in the context of an online grocery retailer, with a focus on drinks. Two partially linear model set-ups are estimated using the Double Machine Learning (DML) method, which accounts for confounding effects in the framework. One of the set-ups includes product purchase price an instrumental variable. In the first stage of the DML method, nuisance parameters are estimated using three different models. The study reveals that the choice of the first stage model significantly impacts the final elasticity estimate. Therefore, the first stage models are evaluated based on feature importance and residual correlation. A Random Forest first stage model is identified as the most appropriate for nuisance function estimation in this context, and it is observed that the model framework without an instrumental variable is the most suitable for this dataset. This research provides valuable insights into price elasticity estimation in online grocery retail, with the potential to enhance pricing strategy optimization and revenue maximization.

Keywords: *Online Retailing, Grocery Demand, Causal Inference, Price Elasticity of Demand, Demand Modeling, Double Machine Learning, Instrumental Variables*

Title: Estimating Price Elasticity of Demand using the Double Machine Learning Method
Author: Florence Piet, floor.piet@student.uva.nl, 11891920
Supervisor: Dr. A. (Arturas) Juodis
Second Reader: Dr. N. (Noud) P. A. van Giersbergen
External Supervisor: Dhr. S. (Stan) Hennekes
Date of final version: March 31, 2023
Examination date: July 15, 2023

Amsterdam School of Economics
University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam
https://ase.uva.nl/

Picnic Technologies B.V.
Van Marwijk Kooystraat 15, 1114 AG Amsterdam
http://www.picnic.nl/

Acknowledgements

I would like to express my gratitude to Picnic for providing me with the opportunity to apply the theoretical knowledge I gained during my studies in Econometrics and Data Analytics at the University of Amsterdam to practical research during my internship. This thesis would not have been possible without their support.

I would like to extend my sincere thanks to my supervisor, Arturas Juodis, for his guidance, expertise, and thought-provoking questions during our meetings. He constantly challenged me to question my assumptions and reminded me that logic and common sense are crucial in conducting research, especially when tackling complex and abstract concepts. I am also grateful to him and Mr. van Giersbergen for their patience and flexibility.

I would like to thank my supervisor at Picnic, Stan Hennekes, for his consistent focus on the bottom line and his guidance in structuring the thesis writing process. I am grateful to him for helping me to keep the business implications in mind throughout the research. I would also like to thank Annemarijn Haasdijk for her insightful feedback and fruitful brainstorming sessions. In addition, working together within the Pricing team has been enjoyable and relaxed.

I am grateful to my colleagues at Picnic for making the process of writing this thesis more fun and for their openness to answer all of my questions. Finally, I would like to thank my family, friends, and roommates for their unwavering patience and support during the writing process.

Terms and Abbreviations

PPL	Picnic Price Line, referring to the regional pricing lines.
PPL1; PPL2; PPL3	The three regional pricing lines, of which PPL1 generally displays the lowest prices and PPL3 the highest prices.
IV	Instrumental Variable
DML	Double Machine Learning; Debiased Machine Learning; Neyman Machine Learning
PLR	Partially Linear Regression
PLIV	Partially Linear Instrument Variable Regression
PLR 1; PLR 2; PLR 3	PLR models with respectively a Lasso, Random Forest and XGBoost regression in the first stage.
PLIV 1; PLIV 2; PLIV 3	PLIV models with respectively a Lasso, Random Forest and XGBoost regression in the first stage.

Contents

Chapter 1

Introduction

The aim of this MSc thesis is to investigate the causal relationship between selling price and demand for a dataset of drinks, with the ultimate goal of estimating the price elasticity of demand for the online grocery retailer Picnic. Price elasticity is a measure of the responsiveness of demand for a product to changes in its selling price. Understanding price elasticity is essential for online grocery retailers like Picnic to make informed decisions regarding pricing strategies, sales forecasting, and revenue optimization, and it can aid in identifying the optimal price point. In the highly competitive online grocery industry, accurately estimating price elasticity can be the key to gaining a competitive advantage, retaining customers, and increasing market share. This study will contribute to the existing literature on price elasticity and provide valuable insights for Picnic in improving their pricing strategies.

This study examines the sales history and demand patterns of Picnic, which is an app-based online supermarket that seeks to revolutionize the online grocery market by offering customers an affordable, efficient, and environmentally sustainable shopping experience. Picnic currently employs a pricing model based on three regional price lines and mostly relies on monitoring competitor prices. By developing a method to estimate price elasticity, Picnic's pricing strategy can become more proactive and focus on anticipated demand.

It is hypothesized that demand and selling price have a negative causal relation, and that there are underlying confounding explanatory variables which may lead to biased estimates in case of naive estimation. To address this issue, a partially linear model is estimated using the Double/Debiased/Neyman Machine Learning (DML) method. This estimation method is grounded upon the Neyman orthogonality assumption, and has the goal of isolating and identifying the parameter of interest in the presence of confounding variables. As it can be argued that selling price is endogenous, also a partially linear regression model including an instrumental variable is estimated using the DML estimation method. The instrumental variable employed is the article purchase price.

The potentially confounding effects of covariates on selling price, order quantities measuring demand and optionally the instrumental variable are modeled as nuisance functions. In the first stage of the Double Machine Learning (DML) method, these functions are estimated in order to orthogonalize the selling price, demand and purchase price variables. The DML method exploits the qualities of Machine Learning (ML) algorithms in order to accurately estimate this. Three different models will be applied and compared to determine the most suitable approach for estimating the nuisance functions during the first stage of estimation. This includes a Lasso regression, a Random Forest regression and a XGBoost regression.

Causal inference in the supply and demand framework poses a significant challenge and therefore the objective of this study goes beyond identifying the exact price elasticity of demand for the selected articles within the allocated time frame. Another aim of the research is to develop a generally usable approach for approximating price elasticity, which can be employed to other product groups and/or time periods to facilitate further insights and comparison. In this thesis, price elasticity estimates will also be obtained based on subsets of the data, grouped by product categories.

The structure of this thesis is as follows. Chapter 2 presents a literature review, focusing on the price elasticity metric and the development of empirical demand models. In Chapter 3, the dataset construction is discussed, followed by an illustration of the data distribution. This chapter also includes information about the data transformation process and some initial analysis. Chapter 4 outlines the method, including the two model frameworks and the three first stage models that are explored. The results are presented and analyzed in Chapter 5, while Chapter 6 concludes.

Chapter 2

Literature Review

This section offers a brief summary of related work. Section 2.1 shortly explains the formal definition of price elasticity; in Section 2.2 demand models that can be used to estimate price elasticity are discussed and Section 2.3 portrays some findings related to price elasticity in the online (grocery) shopping framework.

2.1 Price elasticity of demand

In oft-cited 'Economics' by Parkin (2012), price elasticity is defined as "a units-free measure of the responsiveness of the quantity demanded of a good to a change in its price when all other influences on buying plans remain the same". One of the significant issues in determining the price elasticity of demand is the endogeneity present in the relationship between price and demand, as they are influenced by many of the same underlying factors (Pearl, 2000). This challenge and the importance of identifying cause-and-effect in non-experimental data have been widely discussed in literature, with notable contributions from Meyer (1995) and Pearl (2000).

The formal definition of the own-price elasticity of demand can be denoted as

$$\epsilon_{own} = \frac{\text{Percentage change of quantity demanded}}{\text{Percentage change of price}} = \frac{\partial Y/Y}{\partial P/P}, \tag{2.1}$$

where Y denotes demand and P denotes price. This metric evaluates the change in demand after a price change, specifically in the case when both these quantities regard the same product. The effect of a price change on the demand of a different product is called the cross-price elasticity of demand (ϵ_{cross}). This is calculated using the same formula, however in that case the price and demand quantities in (2.1) are attributed to different products. This research focuses solely on estimating ϵ_{own}, however for contextual purposes the interpretation of ϵ_{cross} is also given.

When executing a log-transformation on a basic regression model for demand, it becomes possible to interpret the target parameter directly as the price elasticity of demand (Imbens and Rosenbaum, 2005). To illustrate this, see the following example regression equation where the product order amount Y is explained by an intercept and price P:

$$\log(Y(P)) = \alpha + \theta \log(P) + \xi \quad \left(\text{with } \xi \sim N\left(0, \sigma^2\right)\right). \tag{2.2}$$

If one takes the partial derivative of both sides of (2.2) with respect to P, this will result in the parameter of interest θ, which can directly be interpreted as the price elasticity of demand:

$$\frac{1}{Y}\frac{\partial Y}{\partial P} = \theta \frac{1}{P} \Rightarrow \theta = \frac{\partial Y/Y}{\partial P/P} = \epsilon_{own}. \tag{2.3}$$

This way, the parameter of interest θ in (**??**) equals the price elasticity of demand.

The potential values of price elasticity can be split up as follows:

- $0 < |\epsilon| < 1$: Demand is relatively inelastic. This denotes the situation where a change of price results in a relatively smaller demand change. The sign of the ϵ value implies the direction of the demand change with regard to the direction of the price change. Normally speaking, the observed sign of ϵ_{own} is negative. A positive sign of ϵ_{own} implies that a higher price causes a higher demand, which we generally don't observe real-world. A positive sign of ϵ_{cross} implies that the two products compared are substitutes: a price change of the one product causes the demand of the other product to change in the same direction. A negative sign of ϵ_{cross} implies that the two products are complements.

- $|\epsilon| > 1$: Demand is relatively elastic. In this situation, a change in price entails an even greater relative change in demand. As above, we don't observe a positive value of ϵ_{own} naturally and the sign of ϵ_{cross} indicates whether the products used are substitutes or complements.

- $\epsilon = -1$: Demand is unit-elastic. Here, the relative change in price is equal to the relative change in demand it causes. If ϵ_{own} is unit-elastic, changing the price will cause the revenue associated to this product to remain constant.

- $\epsilon = 0$: Demand is perfectly inelastic or non-elastic. In this case, for ϵ_{own}, the demand of a product does not respond to a price change. When ϵ_{cross} is inelastic, the respective demand quantities of both products regarded are independent.

2.2 Demand models

As illustrated in (2.2), a way of estimating price elasticity can be done by modeling demand Y in a partial-equilibrium setting. Demand modeling can be done in various ways, of which some will be discussed in this section. Firstly, in Section 2.2.1, single- and simultaneous equation models will be discussed. In Section 2.2.2, approaches based on choice and panel data models for estimating demand are pointed out. In Section 2.2.3 it is explained how Machine Learning techniques can be exploited in demand models.

2.2.1 Single-equation and simultaneous equation models

The history of econometric models that analyse the effect of price on demand in a partial-equilibrium setting coincides with the history of econometric models in general. The initial examination of this issue is often attributed to Wright (1928), who examined the impact of tariffs on animal products and vegetable oils. His primary argument is that by solely analyzing historical data of quantities and prices, one cannot infer the shape of the demand curve. To address this issue, he used a structural equation model in his analysis, which is well-suited for inferring causal relationships. Another type of model commonly used to analyze relationships among multiple variables is the simultaneous equation model. This model is commonly used to analyze the relationships between multiple endogenous variables, such as price and demand, and allows for the estimation of the effects of changes in one variable on the other, while taking into account their interdependence. In such a model, simultaneity and/or reverse causality could induce endogeneity in the model. Moreover, in both a single-equation model and a simultaneous equation model, there could be unobserved (confounding) variables, which could also result in endogeneity and subsequently lead to biased and inconsistent estimates.

To address the endogeneity problem in a causal inference setting, Wright (1928) proposed the use of instrumental variables (IV) as a solution. Theil further expanded upon this method in 1975, and both IV and the more general Two Stage Least Squares (TSLS) methods of estimation are now widely known and used methods, covered in econometrics textbooks such as Theil's "Principles of Econometrics" (1971). In short, IV methods are commonly presented as means to address the issue of simultaneous equations bias. However, in general, IV offers a robust and adaptable estimation approach that can also be applied to address the problem of omitted-variable bias in various single-equation regression scenarios. The concept behind this approach is to utilize instruments that affect the endogenous variable (price) directly, which in turn impacts the dependent variable (demand). The instrument should not have a direct impact on the dependent variable, but only an indirect impact through the endogenous variable. Variables that pertain to the cost-side are commonly employed as instruments in demand models, as they are expected to affect the price directly, yet affect the demand indirectly. This idea is widely applied in literature (e.g. Berry et al. (1995), Nevo (2001)).

2.2.2 Individual choice and panel data demand models

Another way to estimate demand and price elasticity is to employ choice or panel data models. A big development in the field of demand estimation is made by McFadden (1986), who focused on consumer choice making based on individual preferences by introducing a mixed logit choice model including the existence of differentiated products. Berry et al. (1995) expanded the field of consumer choice models by introducing a common demand shock and by means of this created the Berry-Levinsohn–Pakes (BLP) model to estimate demand from product market shares. A lot of work is built on this utility based model. In these models of individual choice, the price is considered as given, hence simultaneous modeling cannot be a cause of endogeneity. In this case however, endogeneity may arise from omitted variable bias: the act of choosing one option may affect the availability or perceived utility of other options, which is not directly observable.

A challenge in applying individual choice models on a large scale, such as for the large grocery scanner data sets that are currently available, is the issue of "many zero's" (Chernozhukov et al., 2019). This phenomenon refers to the fact that individuals often have zero demand for a large number of products under investigation, rather than having nonzero demand for all products. This complication arises when modeling individual preferences for an array of products. In contrast to the method of individual choice modelling, structural models in the partial equilibrium setting seek to estimate demand at an aggregated level, leading to total sales or market shares to be used as the dependent variable. This approach therefore avoids the "many zero's problem."

Another possible model to employ is a panel data model, in case this data is available. For example, the Difference-in-Difference (DD) estimation method of Angrist and Krueger (1999) is suitable to identify treatment effects. This method was in the first instance based on a unit treatment, however recently Callaway et al. (2021) showed the method can also be used for a continuous change in treatment: for example in case of a price change. Additionally, Goodman-Bacon (2021) determined how the DD method can be applied with variation in treatment timing, which means that the treatment may not have been implemented at the same time for all individuals or groups, and there may be differences in the duration or intensity of the treatment. A drawback of applying the DD method in the price and demand framework is that even despite these recent advancements, the method is specifically suitable in case of quasi- natural experiments. In addition, the identification of causal effects through the use of Difference-in-Difference estimation is based on various assumptions, including the assumption of linearity, which may oversimplify the complexity of real-world scenarios. An advantage of using the panel data framework on the other hand is the inclusion of time- and product-specific effects. However, in this research setting, as a substitute for product- and time-specific effects, it also possible to include product-specific features and lagged observations to a cross-sectional data set.

2.2.3 Machine Learning

Machine Learning (ML) is a subfield of artificial intelligence (AI), which began to take shape in the 1950s. The idea behind ML is to give computers the ability to learn from data. In recent years, the field of Machine Learning has experienced a resurgence of interest and growth, driven by the availability of large amounts of data and the development of powerful computational resources. This has led to a proliferation of new techniques and applications in, among others, the economic field (Athey, 2018).

The fundamental distinction between econometric methods and Machine Learning lies in the objectives of each approach. Econometrics places a significant emphasis on generating estimates that can be comprehended as indicators of causality, in an attempt to assess (economic) theories.

In contrast, the focus of Machine Learning is on prediction. Recent developments in the field have involved integrating traditional econometric methods with Machine Learning techniques. This combined approach has proven to be mutually beneficial, as Machine Learning methods excel in handling high dimensional data and sparsity, while econometrics are well suited for determining treatment effects and counter-factuals. Many novel methods that have been developed focus on reducing bias using Machine Learning methods to estimate average treatment effects. This involves utilizing Machine Learning models to estimate nuisance parameters, referring to variables or parameters that are not of primary interest but are necessary to control for. The variables of interest can thereafter be "orthogonalized" by subtraction of these predicted nuisance functions. These orthogonalized or residualized variables are consequently used to infer average treatment effects. This design is illustrated by Chernozhukov et al. (2016), in the Double (Debiased) Machine Learning (DML) method. This method is rooted in the research of Robinson (1988), who proposed a semi-parametric model by partitioning the regression equation into two components: one being the target parameter β and the other being a collection of stochastic functions η, referring to the nuisance parameter, of arbitrary dimension. The aim of this approach is to accurately estimate β linearly, while allowing for non-parametric and/or ML techniques to estimate optionally non-linear nuisance parameter η. In 2016, Chernozhukov et al. further developed Robinson's partially linear model by constructing a general framework to make moment conditions orthogonal to η (based on the idea of Neyman-orthogonality, (Neyman, 1959)), and thereby enabling the estimation of locally robust estimates of β. The paper focuses mostly on the estimation of treatment effects in the context of randomized experiments. Chernozhukov et al. refined and expanded upon the DML approach in 2018, increasing its generality and providing guidance on its application in a wider range of scenarios, including observational studies.

The setup of this estimation technique consists of two stages. The general idea behind this method is to make the moment conditions regarding the parameter of interest orthogonal to the nuisance function η during the first stage, such that the estimation in of η has no effect on the asymptotic distribution of the parameter of interest β. The orthogonalized moment condition is then estimated in the second stage, to obtain parameter of interest β. In this framework, it is possible to use advanced ML methods to accurately estimate η, while decreasing the bias that is commonly present in estimating causal effects using ML techniques. In the DML method, this orthogonalization procedure during the first stage is combined with so-called cross-fitting, which is an efficient way of sample splitting, in order to decrease regularization and model selection bias, which is another common issue in identifying causal effects in the presence of interdependent and confounding variables.

Considering the various models presented, it has been determined that a simultaneous equation model for demand and price would be the most appropriate model to estimate price elasticity in this research scenario. This is due to several factors, including the avoidance of the "many zero's" problem associated with individual choice models which would arise within this grocery demand framework. Furthermore, as previously noted, product-specific effects (usually taking place in panel data models) can be substituted in a structural model setting with article-related variables that are available within the dataset. Finally, the DML method can be used to estimate the target parameter measuring price elasticity of demand θ from the simultaneous equation model, while allowing for nonlinear and high-dimensional terms to be included in the nuisance parts of the simultaneous equation model. This way, by utilizing Machine Mearning methods, it is possible to take advantage of their high prediction performance while also prioritizing causal inference in the model estimation process.

2.3 e-Grocery context

Conventional grocery price elasticity estimates may be outdated in the e-commerce situation, for example considering Natarajan et al.'s finding that customers' intention to use mobile shopping application has a negative relationship to price sensitivity (2017). Similarly, Chu et al. (2008) showed that for households that buy groceries both offline and online, they are less price sensitive when shopping online. A reason to substantiate this would be that consumers who feel pressured for time have a tendency to shop online, using less time to make well-oriented price comparisons. Also, online shops can offer the possibility to easily copy an earlier bought basket, which could also decrease the tendency to conscientiously compare products. The retailer under investigation, Picnic, also employs this option. This finding is consistent with research done by Pozzi (2012), who showed that in e-grocery stores brand exploration is less prevalent than in traditional brick-and-mortar supermarkets. This could entail cross-elasticities in the online environment to be of a lesser magnitude than those in the physical environment. Another explanation for lower price sensitivities in the online environment is that consumers typically have easy access to other (non-price) product information (e.g. nutrition details), which causes them to be less price-sensitive (Lynch Jr. and Ariely, 2000). Lastly, according to Morganosky and Cude (2000), consumers that do grocery shopping online experience this as advantageous and user-friendly, and therefore may take higher prices for granted in exchange for its convenience.

Another difference between researching the traditional and online supermarket settings is an expansion of observed features with regard to consumer behaviour in the online setting. In traditional retail settings, scanner data is available and already yield considerable insights in purchase patterns (Baron and Lock, 1995). In the online setting, even more attributes are observed. For example, click-data is stored, tracking what products consumers look at, what products they add and potentially even remove from their shopping baskets.

Chapter 3

Data description and preliminary analysis

This section provides a description of the nature of the dataset and its preliminary analysis. Firstly, Picnic's regional pricing strategy using price lines is explained in Section 3.1. Subsequently, in Section 3.2, the selection process for the product category under investigation is explained, along with the identification of the variables that have been incorporated in the set. Thereafter, some preliminary analysis is done in Section 3.3, illustrating the distinction of the quality tiers, the brand tiers and the subcategories of the data are illustrated, along with the development of the price and demand variable over time. Afterwards, in Section 3.4, the necessary data transformations are explained and finally, Section 3.5 portrays the relation between the (transformed) variables of interest.

3.1 Picnic Price Lines

In the context of an e-commerce grocery store, an important distinction from traditional brick-and-mortar supermarket chains is that there is typically only one operational webstore (or app) as opposed to multiple physical locations. This can make it difficult to set different prices across different locations. However, the online grocery store being studied in this research, Picnic, does employ location-based pricing. For a portion of its assortment (around 4000 products), regional pricing is applied. This pricing strategy is divided into three distinct "Picnic Price Lines" (PPL's) that are exclusive to specific regions. These regions are composed of groupings of non-contiguous municipalities, based on the regional pricing strategy of competing supermarket "Jumbo" in those areas. Roughly speaking, PPL1 concerns the lowest price point, and PPL3 the highest. This does not hold strictly however.

The implementation of these distinct PPL's presents a unique opportunity to analyze the variations in prices and demand across regions by treating each PPL as a separate store observation. By doing so, each product, also referred to as a stock-keeping unit (SKU), has three distinct observations for a given period of time: one for each PPL. This approach is particularly useful if there is a significant variation in prices associated with the three PPL's. However, it is important to note that in order to do inference on the customer population as a whole, it must be assumed that among the PPL's, customers possess similar behavioral characteristics.

3.2 Composition of the data set

3.2.1 Choice of product category

In creating the comprehensive dataset, a crucial step is to define the scope of the analysis by identifying the appropriate product categories to include. This is accomplished by evaluating the "Level 1" product categories, which are the most general classifications of product types, using a set of established criteria.

Firstly, it is advantageous to choose product categories that make use of the regional pricing strategy (hence the use of separate price lines), as this allows for a larger number of observations with increased price variation to learn from. Therefore, categories such as “Meat & fish”, “Pantry”, “Meals & ease”, “Dairy”, “Vegetables, potatoes & fruits”, “Bread & pastry”, “Baby”, “Housekeeping”, and “Animals” were not included in the analysis, as they do not have the application of price lines.

Another important criterion in selecting the product category to investigate is the degree of variation among these price lines and in the prices themselves. The presence of substantial spread in prices, as well as the presence of differences among the price lines enables the examination of the varying effects of price changes on demand. Therefore, the categories of “Candy & Snacks”, “Coffee & tea”, “Vegan & vegetarian” and “Breakfast & sweet spreads” were not included in the analysis, as they were found to have possessed a limited degree of variation within and between their respective price lines. An additional factor in the selection of product categories is the frequency of promotional pricing. As the focus of this research does not include the impact of promotions, it is important to exclude categories that are frequently subject to promotional pricing. As such, the categories of “Drugstore”, “Beer & aperitifs”, and “Wine & fizzy” were not included in the analysis.

A final criterion in selecting the product category for analysis is the availability of sufficient historical data. The category of “Cold cuts, spreads & tapas” is relatively new and as such, a limited amount of data has been collected over time, making it unsuitable for inclusion in the analysis. Also, the assortment in the “Pasta, rice & international” category has undergone significant changes over time, which makes it difficult to conduct a valid analysis, and thus this category was also not included in the final dataset. Furthermore, due to a supply crisis in frozen products between August and October 2022, the “Frozen” category was excluded from the analysis.

Finally, the product categories of “Cheese” and “Drinks” have been identified as suitable for inclusion in the analysis. Among these, the drinks category seems particularly relevant, as it is a commonly studied category in research on price sensitivity. This provides additional motivation to examine this category in order to make comparisons with previous research. As a result, this category is selected for analysis.

3.2.2 Data description

The data includes articles sold from the first week of 2021 until the first week of 2023, belonging to the (level 1) drinks category. This set is focused on the 225 articles that were consistently available for purchase in the Picnic app throughout this period. This set of products contains both products that fall under the regional pricing strategy and products that are priced consistently nationally. The set is aggregated on a weekly basis and includes the following variables:

- SKU (article identifier);
- Week;
- PPL type (categorical);
- Number of regular (non-promotional) orders for the specific article in the given week and PPL;
- Average selling price of the article for the given week and PPL;
- Average purchase price of the article for the given week and PPL;
- Total number of orders for the given week and PPL;
- Level 2 Product category (categorical);
- Article quality tier (categorical);
- Article brand tier (categorical);
- Article packaging (categorical);
- Dummy variable indicating whether the product is a multi-pack;
- The content volume of the article;
- Average highest daily temperature in a given week;
- Average percentage of unavailability of the product in a given week;
- Dummy variable indicating the presence of a promotion on the specific product for the given week;
- Total number of distinct articles in the ”Drinks” category available in store that week;
- Total number of distinct articles in the given Level 2 product category available in store that week;
- A dummy indicating whether it was a national or school holiday;
- Number of weeks counted from the first week of 2021

In Section 2.2.2 it was discussed that in order to account for time-specific effects, it is possible to include lagged order demand. In this setting however, because of the high (weekly) frequency of the data, adding lagged product order amount to the model for demand will most likely cause correlated errors and therefore result in biased estimates. Additionally, in the model for price, including these lags would result in reverse causality.

Orders made during a promotion are not considered in the analysis, hence these orders are not included in the order quantity. Nonetheless, when an article is on promotion during a week, this does not have to be a national Picnic-wide promotion. Specifically, it is possible for an article to be on promotion in some regions of The Netherlands but not in others, so regular orders may still be recorded even when the promotion dummy variable equals 1.

The total number of observations in the set is calculated by multiplying the number of articles (225), the number of weeks (52*2+1 = 105), and the number of tracked price lines (3). This results in a total of 70875 observations.

3.3 Preliminary analysis

The following Table 3.1 presents some of the dataset's descriptive statistics, organized per PPL. It is important to note that the dataset comprises 225 distinct products, and as such, the weekly product order quantity, as well as the product's selling and purchase prices, are represented as averages of the weekly averages for each individual product. The purchase and selling prices are denoted in cents.

Table 3.1: Descriptive statistics per PPL

PPL	**PPL1**	**PPL2**	**PPL3**
Average weekly total order amount	349,055	1,422,735	586,658
Average weekly avg. product order amount	1,148	4,939	1,879
Average weekly avg. product selling price	209	211	216
Average weekly avg. product purchase price	151	151	151

Table 3.1 shows that PPL2 is the largest in terms of order quantity in terms of both total order quantity as well as product order quantity. As pointed out in 3.1, generally, PPL1 has the lowest prices and PPL3 the highest, which also shows in the depicted average selling prices.

The brands of the products sold at Picnic are classified into one of three tiers: "Price entry", "Private label" (indicating Picnic's own brand), or "A-brand". Additionally, the products are categorized into one of three tiers: "Good", "Better", or "Best". Table 3.2 provides statistics that depict the distribution of these two types of tiering within the dataset. Table 3.2 shows that there are only two "Price entry" articles contained in the data set. The "Private label" products are generally cheaper than the "A-Brand" products, which is to be expected. The "A-Brand" category is by far the largest of the three, comprising 84.89% of the dataset. Over the observation period, more "Price entry" and "Private label" products were added to the store, however the data set only includes articles that were available in store over the whole observation period. Regarding the article quality tiers, it can be seen that the largest amount of products is in the "Better" tier, comprising almost half of all articles. It can also be seen that the "Good" articles are generally most popular, with the highest average product order amount. As expected, the selling prices increase with the quality tier.

In Table 3.3, some more insight is given in terms of the "Level 2" category partition of the analyzed articles. Here, level 2 refers to the sub-categories within the drinks set.

Table 3.2: Distribution brand & product quality tiering

Article brand tier	**Count of articles (%)**	**Avg. product order amount**	**Average weekly avg. product selling price**
A-brand	191 (84.89%)	2,236	226.66
Price entry	2 (0.89%)	1,292	150.68
Private label	32 (14.22%)	5,246	127.30
Article quality tier	**Count of articles (%)**	**Avg. product order amount**	**Average weekly avg. product selling price**
Good	35 (15.56%)	4,221	119.50
Better	107(47.56%)	2,242	187.10
Best	83 (36.89%)	2,529	282.80

Table 3.3: Distribution category level 2

Category level 2	**Count of articles (%)**	**Avg. product orders**	**Average weekly avg. product selling price**
Fruit drinks	45 (20%)	2,198	130
Juices & smoothies	36 (16%)	1,570	221
Special soda	30 (13.33%)	1,779	214
Cola	27 (12%)	4,319	342.60
Lemonade & Syrup	25 (11.11%)	1,741	245
Small cartons	24 (10.67%)	1,465	206
Orange, Lemon & Cassis	17 (7.56%)	2,801	210.90
Water	13 (5.78%)	9,845	93.50
Sport- & energydrink	4 (1.78%)	2,733	381.30
Ice tea	4 (1.78%)	1,718	198.20

Table 3.3 shows that the “Sport- & energydrink” and “Ice tea” categories are the smallest, both only containing 5.78% of the products. The largest categories are the “Fruit drinks”, Juices & smoothies” and “Special soda” categories, containing 20%, 16% and 13.33% of the articles respectively. The most sold articles belong to the “Water” category, with an average number of weekly product orders of 9,845. This category also shows the lowest average selling price. The most expensive average prices belong to the “Cola” and “Sport & energydrinks” categories.

Note that in all Tables 3.1 - 3.3, the data set did not undergo any transformations yet. Furthermore, as mentioned, the order amount quantities exclude promotional sales, causing the regular order amount in promotional weeks to be low or equal to zero. This may cause downward bias as these observations were not removed yet.

3.3.1 Demand development

In Figure 3.1, the total order amount per week for Picnic is depicted, regardless of the products included in these orders. The total number of weekly orders increased on average during the study period, with significant peaks occurring in the weeks leading up to the holidays of Sinterklaas and Christmas in both 2021 and 2022. This 2021 “festive peak” coincides with renewed COVID-19 restrictions, which may have contributed to the increase in orders, as ordering online may be favorable in times of stricter quarantine regulations. Many of the other peaks and dips can be attributed to school holidays, as Picnic’s target audience consists of families with young children. Thus, for example, the end of summer holidays (the “back to school” period) is often associated with increased demand. The amount of orders may reduce during school breaks because families may need less groceries because of vacations or other activities. The impact of festive and school holidays can be seen in Figure 3.1 below. Please note that the fluctuations in order quantity often occur in one or two weeks before the mentioned suspected cause (holiday, shool break), as orders must be placed some number of days ahead of the associated delivery.

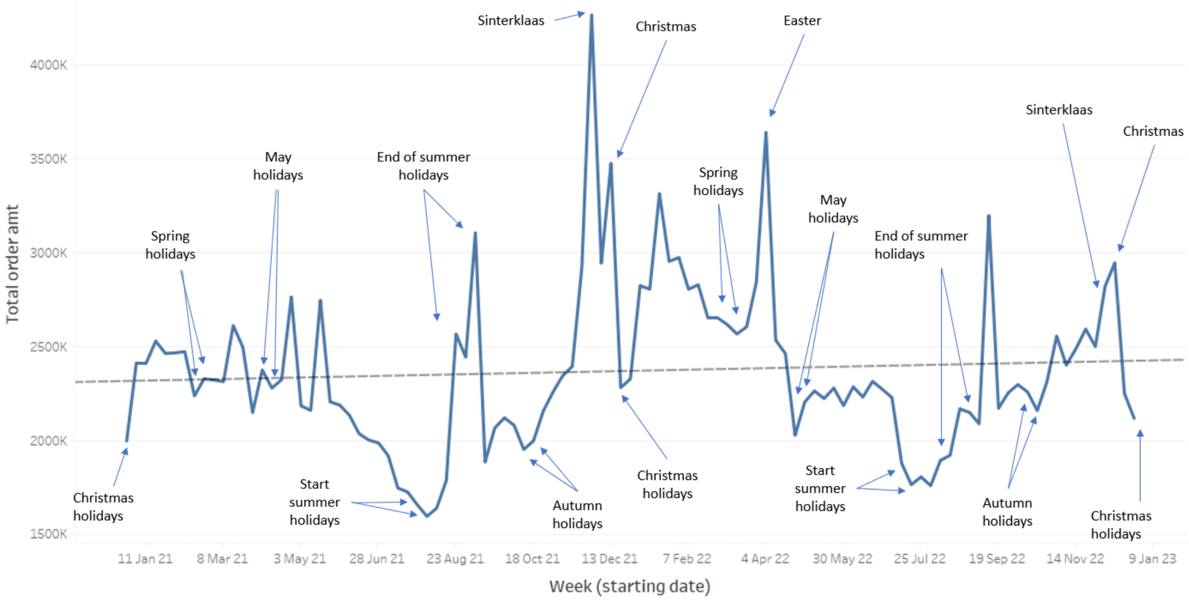


Figure 3.1: Development of total weekly average order quantity 2021-2023

Figure 3.2 presents the same trend as in Figure 3.1 (the course of total weekly order amount at Picnic), while also displaying the course of average *product*-specific order amount below it, based on the dataset under analysis. In other words, the lower graphs shows the development of weekly order amount at the SKU level, averaged over the 225 products in the dataset. It can be observed that the patterns in both measures are similar, with peaks and dips occurring at the same times. This suggests that the drinks articles under investigation are ordered substantially and consistently throughout the analyzed period. One difference between the graphs is that the dip in number of orders occurring in December 2021 is lower for the product within the drinks set than for the total orders. This point refers to orders placed with a delivery date in the week after Christmas. A hypothesis for this low number of product orders within the drinks category is that people might have had sufficient drinks left after their Christmas celebrations, resulting from the peak in article orders during the week before. The biggest outliers occur around the “back to school” period and Christmas period, both in 2021. A possible explanation for these outliers not taking place in 2022 can be that during 2021, stricter COVID-19 restrictions were in place, nudging people to stay at home and order their groceries instead of going to physical stores.

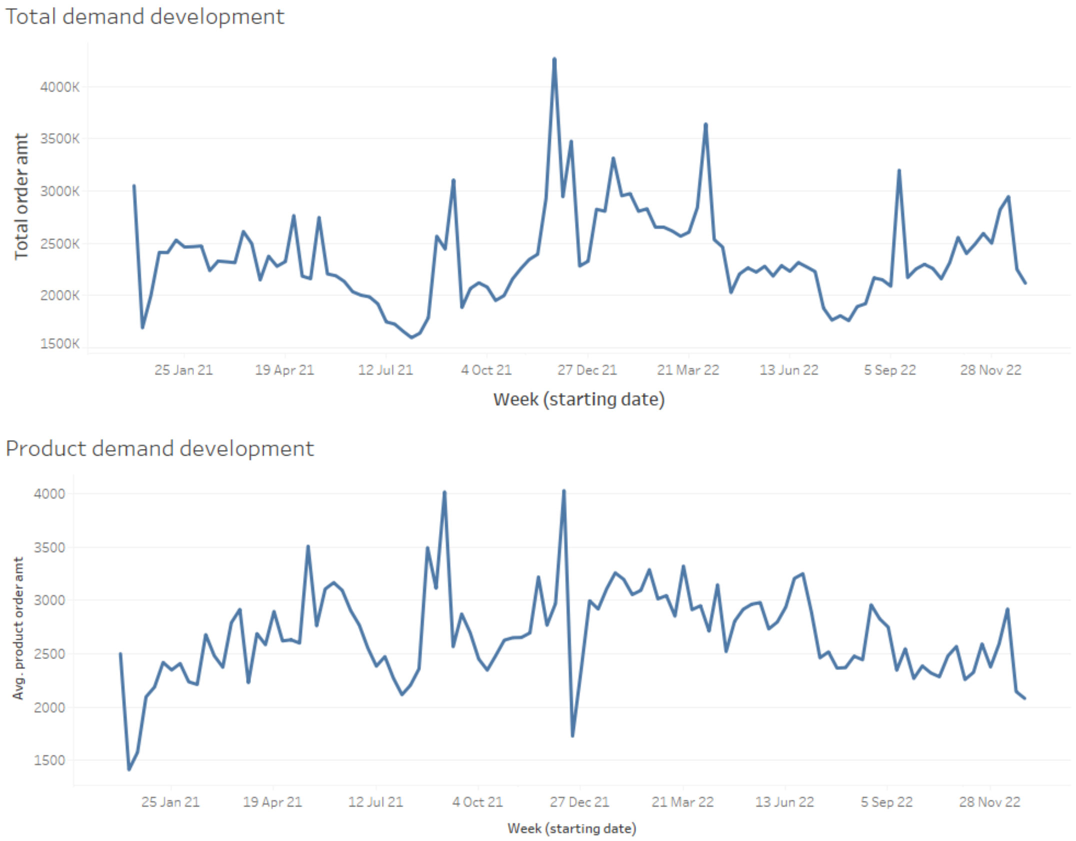


Figure 3.2: Comparing total weekly average demand with average demand on product level

As mentioned, Picnic’s customers are divided into three regions, or price line types: PPL1, PPL2, and PPL3. The regions are consistent of municipalities spread around The Netherlands. During June 2022, there was a shift in the customer composition, with many customers being moved to PPL3. This can be clearly seen in Figure 3.3, which illustrates the demand development (at the SKU level) among the three price lines. Apart from this shift, the graph shows that similar fluctuations appear for the three price lines. This could confirm the aforementioned assumption that customers assigned to the three price lines have similar characteristics and therefore show similar behaviour.

Lastly, on the left side of Figure 3.4, the distribution of the weekly product order amount is depicted, using bins of size 350. It can be seen that this distribution is very skewed.

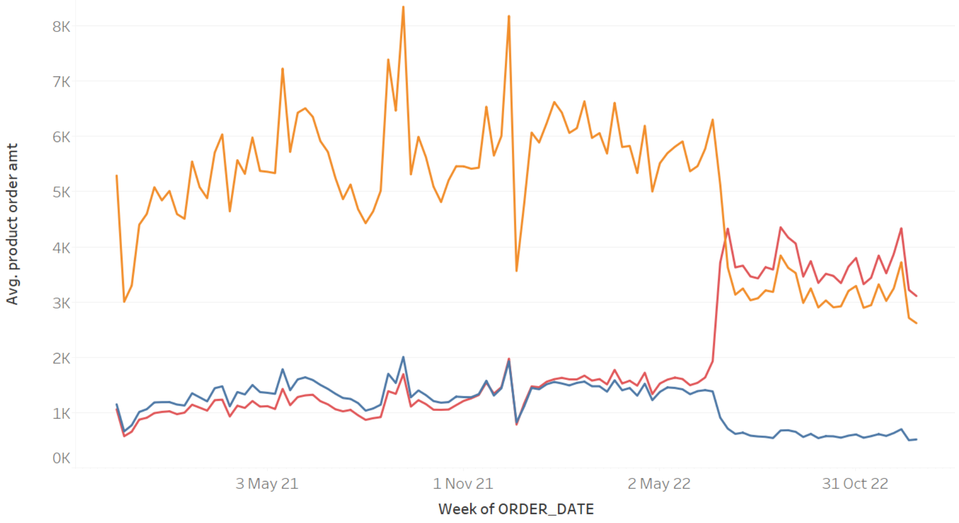


Figure 3.3: Average weekly demand on SKU level, divided in the PPL's

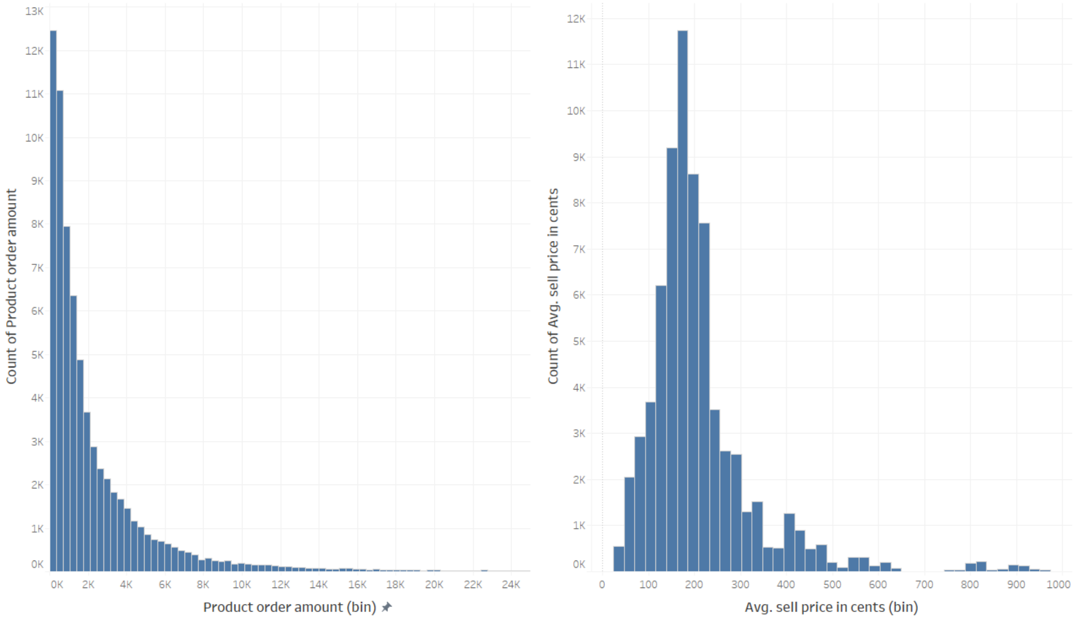


Figure 3.4: Distribution of weekly order amounts and selling prices in cents

3.3.2 Price development

For the fixed set of 225 products, the weekly average price was calculated for all of the products, per Picnic Price Line (PPL). The evolution of the weekly average article selling price, in cents, for these products under examination throughout the given period is depicted in Figure 3.5. An examination of the data reveals that there is a consistent upward trend in the weekly average prices throughout the period for all PPL's. Additionally, it can be observed that the prices per PPL exhibit a similar pattern, with fluctuations occurring at roughly the same points in time. However, it seems that the depths of these fluctuations vary somewhat among the different PPL's.

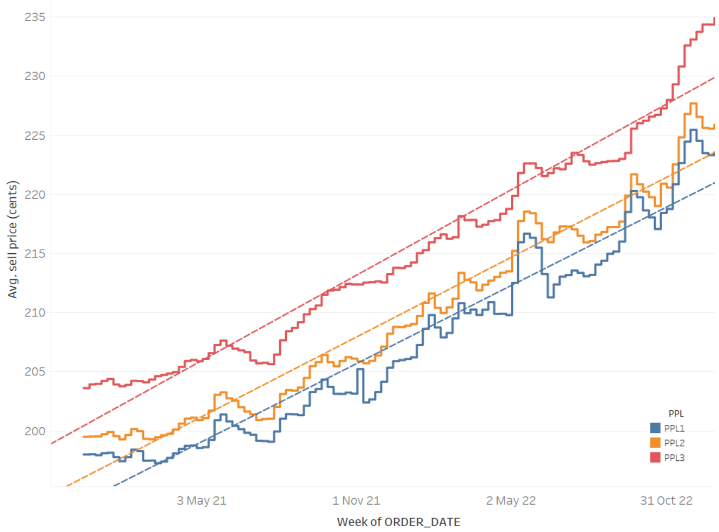


Figure 3.5: Development of the weekly average price of drinks products for the PPL's

As done with the order amount, we also look at the distribution of the average weekly prices in cents in Figure 3.4, using bin size 23. The distribution is found to be somewhat skewed.

3.4 Data preparation

Before further analysis is done on the price-demand relationship, some data preparation is executed.

3.4.1 Outlier identification

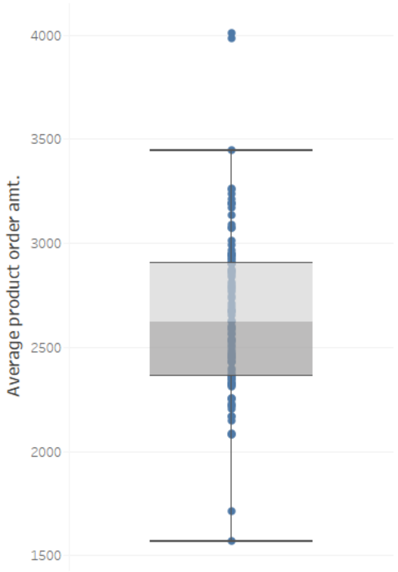


Figure 3.6: Box plot of product order amount

As illustrated in the lower part of Figure 3.2, there might be some outliers present in the data. These outliers may potentially bias the results and compromise the inferences drawn from the data. The method of Tukey et al. is utilized to identify the most prominent outliers in the data, as presented in Figure 3.6. This method, which is based on the visual analysis of a box plot, is particularly useful due to its simplicity and the fact that it does not rely on the extreme outliers when computing the spread. The two outliers identified in the figure were found to correspond to the weeks starting on September 6th and December 13th, 2021. As discussed, these data points may be attributed to the "back to school" and Christmas peaks and may indicate atypical customer behavior. To prevent drawing inaccurate conclusions, the observations referring to product orders during these weeks will be excluded from further analysis.

On the other end of the spectrum, there are observations of products with zero orders in a week. Specifically, there are 1305 instances in the data set where a product is not ordered in a week within a PPL. Note that this might be due to promotional efforts: promotional orders are excluded from the data set, and hence leave zero regular demand. Another cause of zero demand could be occurrences of unavailability caused in (part of) Picnic's supply chain. In order to safeguard the accuracy of the results in the specific context of regular, non-promotional demand in the setting without major unavailability issues, these zero demand instances are dropped. Apart from the accuracy concerns, another motivation for this is the data transformation: a log-transformation will be executed on the data for the results to be consistent with the formal elasticity definition. Transforming the zero-demand observations will output -infinity, which most statistical models cannot handle or would have to be replaced by a mean, median or similar metric. After dropping outliers and missing values, the dataset finally contains 67910 observations.

3.4.2 Log-transformation

After obtaining the final set of observations, a log-transformation is executed on the data. Log-transforming the numerical variables before modelling serves two main purposes: 1) it helps to handle the skewness of the distribution of the variables and shrinks the tails, and 2) as mentioned in Section 2.1, it allows for interpretation of the coefficient of interest as elasticity, measuring the percentage change in the response variable (product order amount) for a 1% change in the predictor variable (selling price).

After dropping outliers and zero-demand observations and executing the log-transformation, the distribution of the product order amount and selling prices in cents look as depicted in Figure 3.7. As expected, the transformation deals with the skewness of the variables initially seen in Figure 3.4.

3.4.3 PPL shift correction

In order to accurately determine the influence of each PPL, it is necessary to correct for the shift that occurred in the PPL setup in June 2022. To accomplish this, the original categorical variable, containing entries 'PPL1', 'PPL2' and 'PPL3', is reconstructed such that there are now six possible entries, including the three price line types before, as well as 'POSTPPL1', 'POSTPPL2' and 'POSTPPL3' indicating the distinction after the changes made. This allows for the effective adjustment of the intercept for each price line pre- and post-shift.

3.4.4 Inflation

Since 2021 the Dutch inflation (CPI on yearly basis) has been higher than the usual so-called "healthy" standard of 2% on a yearly basis (CBS, 2022). One of the underlying causes for this inflation has been the war between Ukraine and Russia, which started near the end of February 2022. This inflation is also reflected in food prices. As is visible in Figure 3.5, this is also the

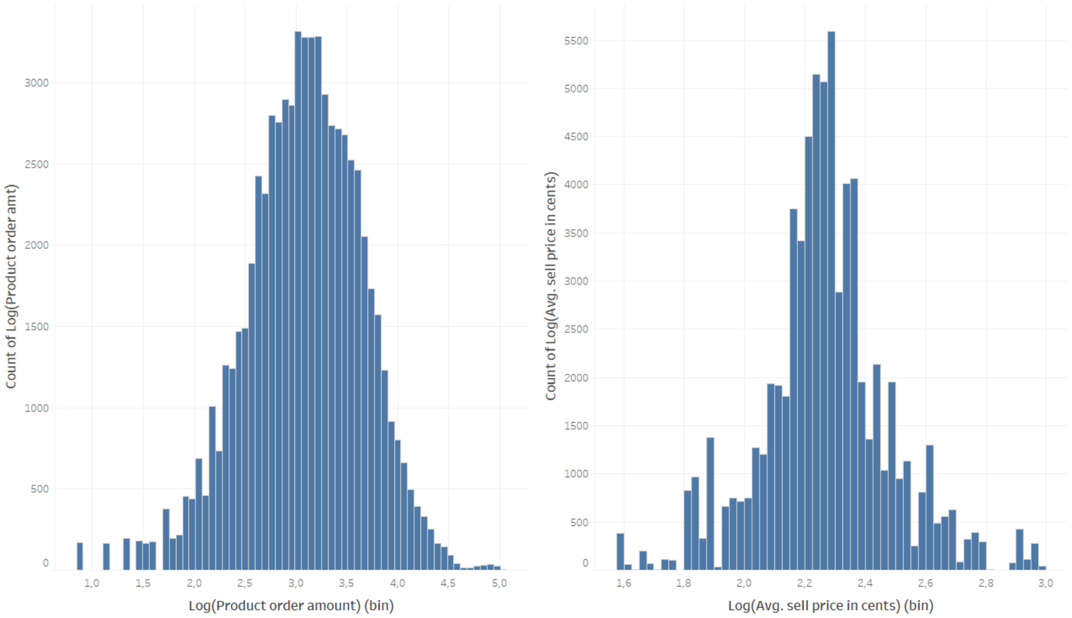


Figure 3.7: Distribution of transformed weekly order amounts and selling prices in cents

case for the products in this dataset. A numerical variable is added to the data set, reflecting the number of weeks passed since the first week in January 2021, with the objective of capturing the increasing trend in prices.

3.5 Relations of interest

In this section, the relations between the variables of interest and their log-transformed values will be analyzed further. These variables include demand (Y), selling price (P), and purchase price (Z), which will be used as an instrumental variable.

3.5.1 Demand and selling price

In most economic theory, it's posited that price and demand have a negative relationship. To see if this holds for the data under invesitation, the correlation between the product order quantity and selling price in cents is calculated, yielding in an overall correlation coefficient of -0,12. Splitting this up in the PPL's gives three similar coefficients, which are documented in Table 3.4. Here, Y denotes the weekly product order quantity, P denotes the product selling price in cents and ρ denotes Pearson's correlation coefficient. The correlations between the log-transformed variables are also included, which results in heightened correlation coefficients. Additionally, a simple regression is done where $Log(Y)$ is explained by an intercept and $Log(P)$, yielding a very rough elasticity estimate ϵ (as derived in Section 2.1). Note that when the log-transformed variables are indicated, this transformation also includes the dropping of outliers (as described in Section 3.4.1).

Table 3.4: Correlation demand (Y) and price (P)

$\rho(Y, P)$	**Overall**	**PPL1**	**PPL2**	**PPL3**
	-0,12	-0,14	-0,15	-0,11

$\rho(Log(Y), Log(P))$	**Overall**	**PPL1**	**PPL2**	**PPL3**
	-0,31	-0,37	-0,37	-0,32

$\epsilon(Log(Y), Log(P))$	**Overall**	**PPL1**	**PPL2**	**PPL3**
	-0,57	-0,61	-0,70	-0,55

3.5.2 Purchase price

In economic and econometric theory, supply-side variables such as production costs are often regarded as suitable instrumental variables in explaining demand with endogenous variable selling price (Wooldridge, 2010). Most hypotheses express that cost-related variables such as the purchase price do not directly affect the customer and therefore have no direct on the product order quantity. The purchase price does however contain information about for example market circumstances and production technology, and, in turn, directly influences selling price. Therefore, it indirectly influences demand and serves as a suitable instrument. Correlation coefficients involving the purchase price (Z), selling price (P) and demand (Y) are listed in Table 3.5.

Table 3.5: Correlation purchase price (Z), selling price (P) and demand (Y)

$\rho(P, Z)$	**Overall**	**PPL1**	**PPL2**	**PPL3**
	0,95	0,95	0,95	0,95

$\rho(Log(P), Log(Z))$	**Overall**	**PPL1**	**PPL2**	**PPL3**
	0,95	0,95	0,95	0,95

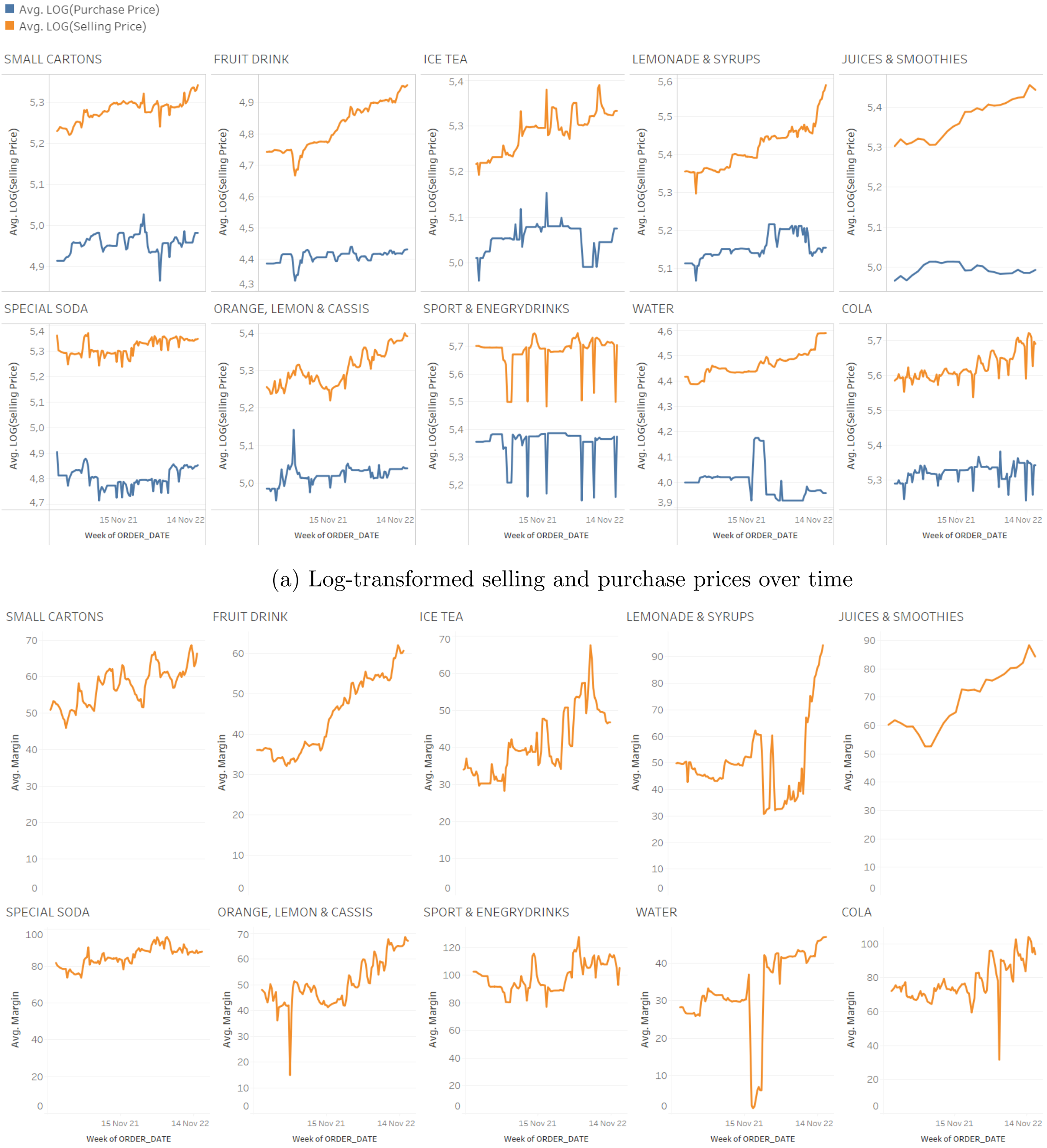
$\rho(Y, Z)$	**Overall**	**PPL1**	**PPL2**	**PPL3**
	-0,07	-0,08	-0,09	-0,09

$\rho(Log(Y), Log(Z))$	**Overall**	**PPL1**	**PPL2**	**PPL3**
	-0,26	-0,28	-0,30	-0,30

The correlation coefficients between selling and purchase price are, as could be expected, very high. The correlation between purchase price and order amount are similar to the correlation between selling price and order amount (Table 3.4), however they are slightly lower.

3.5.3 Relation selling price and purchase price over time

For all ten category subsets of the drinks dataset, the log-transformed weekly average selling and purchase prices are depicted over time in Figure 3.8a. In Figure 3.8b, the difference between these log-transformed weekly averaged variables is depicted, in order to better determine whether the relation between these variables remained constant over time or not.



(a) Log-transformed selling and purchase prices over time

(b) Difference between log-transformed selling and purchase prices over time

Figure 3.8a illustrates a noticeable increase in the weekly averaged log-transformed selling price relative to a slightly more stable purchase price for most categories, starting from 2022. This phenomenon can be attributed to two factors. Firstly, the European crisis in natural gas supply, caused by the Russian-Ukrainian conflict, resulted in above-average inflation, which could have increased costs other than the purchase price, such as transportation and electricity costs. These increased costs are then passed on to the selling price, causing them to rise, which may not be reflected in the purchase price of the items. Secondly, Picnic's pricing strategy changed in the second half of 2022, with the company no longer claiming to always offer the lowest prices for A-brand items. As the dataset mostly contains products from the A-Brand brand tier (as shown in Table 3.2), these price increases can definitely be reflected in the average selling prices. These events may have contributed to the increase in selling prices while the purchase prices remained relatively stable. Figure 3.8b shows the increasing difference between the two log-transformed variables.

Chapter 4

Method

This chapter provides an overview of the methods used to model log-transformed product order demand. First, in Section 4.1, the two possible model specifications are demonstrated, one including and one excluding an instrument, after which the naive estimation methods will be discussed shortly in Section 4.2. Thereafter, the Double Machine Learning (DML) estimation method consisting of two stages is explained in Section 4.3. During the first stage of this method, three possible methods are used to estimate the nuisance functions, which will be discussed in Section 4.3.1. The second stage of DML estimation is discussed in Section 4.3.2.

4.1 Model specification

Demand can be modeled with or without the use of instrumental variables. Both models are considered, starting with the partially linear regression model.

4.1.1 Partially Linear Regression (PLR) model

The partial linear model without instrumental variables has the following form:

Demand measured as log-transformed product order amount Y:

$$Y_j = P_j\theta + l(X_j) + U_j \quad \mathbb{E}[U|X] = 0 \tag{4.1}$$

Log-transformed product selling price in cents P:

$$P_j = r(X_j) + W_j \quad \mathbb{E}[W|X] = 0 \tag{4.2}$$

In which observation index j represents the unique combination (i, t, p), indicating an observation of product i at week t, within price line p. Y_j denotes the log-transformed number of orders, P_j the log-transformed selling price, and X_j the covariate matrix of explanatory variables including:

- a constant

- features concerning product i
- features concerning time t
- features specifying price line p

The full list of variables is described in Section 3.2.2. As can be seen in equations (4.1) - (4.2), X_j influences demand and selling price via potentially nonlinear functions $l(X)$ and $r(X)$ respectively. This set is referred to as the set of nuisance functions, denoted by $\eta = (l(X), r(X))$.

The conditions that set the conditional stochastic processes denoted by U and W to zero assume that there are no unobserved variables beyond X that influence demand and price. However, this assumption may be invalid. It is possible that there are unobserved confounding factors that impact price and demand simultaneously, causing bias. In the case of unobserved variables that influence selling prices solely, an instrumental variable can be used, which will be illustrated in the following section.

4.1.2 Partially Linear Instrumental Variable Regression (PLIV) model

The decision whether to include an instrumental variable in the model is contingent on the assumption whether all confounding variables are observed. In case the treatment variable selling price is influenced by unobserved variables, an instrumental variable (IV) can be used to isolate the causal effect of the treatment variable from the influence of unobserved variables. In case there is endogeneity and the instrument is valid, this method could provide a more accurate estimate of the causal effect of interest. In this research, it can be argued that there are unobserved variables that affect the selling prices. Examples of these variables include a part of and the aftermath of the COVID-19 pandemic and the impacts of the Russian-Ukrainian conflict on natural gas supply in Europe.

As discussed in Section 3.5.2, an article's purchase price at time t can serve as instrumental variable, because at time t, it directly affects the selling price P_j, and then indirectly affects the article's product order demand Y_j. When instrumental variable Z_j is included, the model equations expand to:

Log-transformed Demand (same as in 4.1):

$$Y_j = P_j\theta + l(X_j) + U_j \quad \mathbb{E}[U|X, Z] = 0 \tag{4.3}$$

Log-transformed Selling price:

$$P_j = r(X_j) + Z_j\gamma + W_j \quad \mathbb{E}[W|X, Z] = 0 \tag{4.4}$$

Log-transformed Purchase price:

$$Z_j = m(X_j) + V_j \quad \mathbb{E}[V|X] = 0 \tag{4.5}$$

Z_j here denotes the log-transformed purchase price, which is affected by X_j through possible nonlinear function $m(X)$. Nuisance parameter η expands and in this case refers to the collection of functions $(l(X), r(X), m(X))$.

The difference between the two partial linear models is that in the partial linear IV model, instrumental variable purchase price Z serves as a proxy for the endogenous variable P that is influenced by unobserved variables. However, it should be noted that the IV approach can only account for the influence of unobserved variables on the treatment variable P, and not directly on the dependent variable Y. In other words, Z does not account for all unobserved effects, especially when there are variables affecting Y.

4.2 Naive estimation methods

Before employing the DML estimation method to estimate equations (4.1) - (4.2) and (4.3)-(4.5), these models will also be estimated using Ordinary Least Squares (OLS) and Instrumental Variable (IV) regression respectively. These methods disregard the bias that can be induced by potential nonlinearity and/or high-dimensionality of nuisance functions η. Also, the OLS method does not correct for the potential effect of unobservable variables causing bias. These methods are used as a control measure and to compare the resulting θ estimates.

4.2.1 Ordinary Least Squares regression

Model equation (4.1) can be estimated using Ordinary Least Squares (OLS) regression, in which case $l(X_j)$ refers to the linear combination of matrix X and parameter vector β, transforming the model equation to $Y_j = P_j\theta + X_j\beta + U_j$. Parameters θ and β are estimated by minimizing the sum of squared differences between the observed values of the dependent variable Y_j and the values predicted by the linear regression model $\hat{Y_j}$.

4.2.2 Instrumental Variables regression

As a baseline estimation method for (4.3), an Instrumental Variable (IV) regression is used to approximate θ from equations (4.3) - (4.5). IV regression is a commonly used linear model to estimate the parameters of a model in case of endogeneity, i.e. models with $\mathbb{E}[P'U] \neq 0$. In this method, $l(X)$, $r(X)$ and $m(X)$ are assumed to be of linear form. The model is just-identified, as the number of endogenous variables is equal to the number of instruments ($k = r = 1$).

Generally, in linear model $Y = X\beta + U$ with endogenous covariate matrix X, the IV estimator can be obtained with two stages. During the first stage X is regressed on Z as follows:

$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

after which follows the second stage:

$$\hat{\beta}_{IV} = (\hat{X}'X)^{-1}\hat{X}'Y$$

where $\hat{X}$ is the predicted value of X from the first-stage regression of X on Z, and $\hat{\beta}_{IV}$ is the IV estimate of coefficient vector β. Mind that in this notation, endogenous variable P is included in X.

The IV model is grounded upon several assumptions, of which one states that the instrumental variable should be relevant, i.e. it should correlate with the endogenous variable. In this setting, this implies $\gamma \neq 0$ in (4.2). Table 3.5 shows that the correlation between P and Z is notably high. Additionally, for the estimation to be valid, the following condition must be satisfied: $\mathbb{E}[Z'U] = 0$. This equation can be rewritten to establish the IV estimator.

4.3 Double Machine Learning

The Double Machine Learning method is founded upon the Frisch-Waugh-Lovell (FWL) theorem, which demonstrates that inference can be conducted through using ortogonalized (residualized) variables. In this particular research setting, the effect of the confounding variables on the dependent, endogenous and possibly instrumental variables can therefore be partialled-out in order to do inference on the price elasticity of demand. This assumption intuitively means that nuisance parameter η which measures the confounding effects of X on selling price P, purchase price Z and demand Y is orthogonal to the effect that price has on demand, which is measured by the parameter of interest θ. In practice, this implies that estimation of η should have no effect on estimation of θ.

The DML technique follows a two stage process where firstly a set of possibly non-linear and/or high-dimensional nuisance functions is estimated in a so-called cross-fitting matter, which is thereafter used to orthogonalize the dependent, treatment and possibly instrumental variables. In the second stage, these residualized variables are used to estimate a partially linear model. As mentioned, this can either be either an OLS or IV regression. In this section, the DML method will be illustrated in context of the PLIV model, i.e. with an IV regression of the orthogonalized variables in the second stage of estimation.

4.3.1 First stage of estimation

In the first stage of the method the focus is on orthogonalizing the variables demand, price and instrumental variable purchase price. This is achieved by subtracting the estimated nuisance function values from the observational values. The process of orthogonalization is crucial for isolating the effect of interest while minimizing the regularization bias that would arise from performing naive estimation using ML techniques (Chernozhukov et al., 2018). As in most Machine Learning scenarios, a balance must be struck between bias and variance, or in other words, between regularization and overfitting. Therefore, in the first stage, the orthogonalization procedure is integrated with cross-fitting to prevent overfitting.

Cross-fitting

Cross-fitting is performed by splitting the data into $k = 1, ..., K$ different samples. Using all data except for the data in subsample k, the nuisance functions for product demand, selling price and purchase price are predicted as follows:

$$l_k(X_{j,-k}) := \mathbb{E}[Y_{j,-k}|X_{j,-k}] \tag{4.6}$$

$$r_k(X_{j,-k}) := \mathbb{E}[P_{j,-k}|X_{j,-k}] \tag{4.7}$$

$$m_k(X_{j,-k}) := \mathbb{E}[Z_{j,-k}|X_{j,-k}] \tag{4.8}$$

Subscript $-k$ denotes that subsample k is left out of the data while estimating. Subsequently, the estimated function values are subtracted from the observations originating from subsample k.

$$\tilde{Y}_{j,k} = Y_{j,k} - \hat{l}_k(X_{j,-k}) \tag{4.9}$$

$$\tilde{P}_{j,k} = P_{j,k} - \hat{r}_k(X_{j,-k}) \tag{4.10}$$

$$\tilde{Z}_{j,k} = Z_{j,k} - \hat{m}_k(X_{j,-k}) \tag{4.11}$$

This process is repeated until every fold $k \in (1, 2, ..., K)$ is left out of the estimation of the functions (4.6) - (4.8) once. The K versions of orthogonalized variables (4.9) - (4.11) will be used in the second stage of the model to estimate θ. Figure 4.1 is a visual representation of how the cross-fitting technique is used to residualize variables Y, P and Z during the first stage of the model.

The advantage of the DML method is that various models can be applied to estimate functions (4.6) - (4.8). In the context of this research, it is essential to utilize methods that are capable of effectively handling categorical variables and possess the robustness to handle high dimensionality. This second quality is particularly important in the context of online retailing, where possibly a large number of features are stored and could be integrated into the model. Moreover, cross-effects could be introduced which will therefore increase the dimensionality of the data. Candidates for this first stage estimation are Lasso Regression, Random Forest Regression and XGBoost Regression. Even though direct causal inference for the first stage models is not necessarily essential in this research, these models are still easy to interpret, which is crucial for understanding what happens during the first stage of the DML estimation. In the following sections the three mentioned methods and their expected implications will be explained shortly.

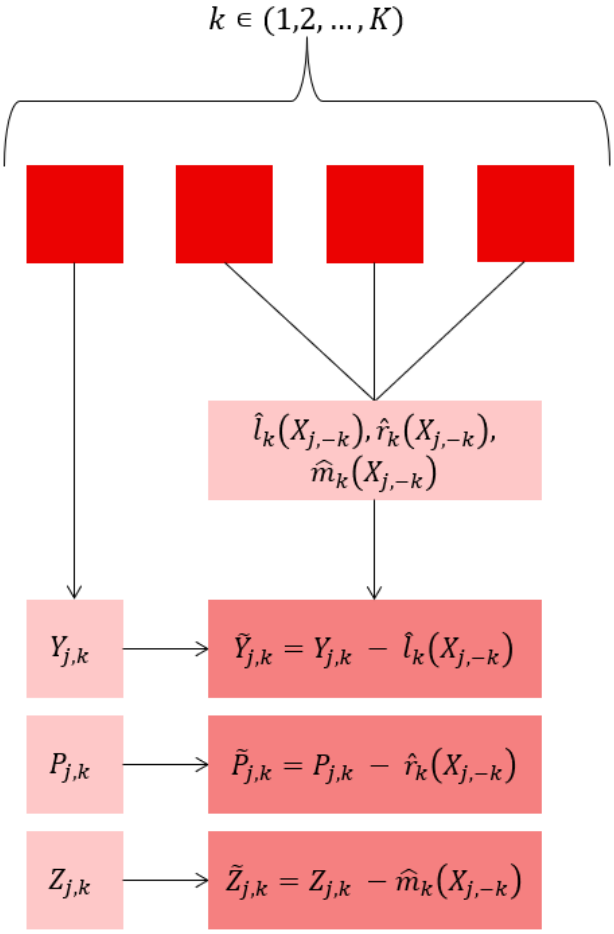

$k \in (1,2,\dots,K)$
$\hat{l}_k(X_{j,-k}), \hat{r}_k(X_{j,-k}),$
$\hat{m}_k(X_{j,-k})$
$Y_{j,k}$
$\tilde{Y}_{j,k} = Y_{j,k} - \hat{l}_k(X_{j,-k})$
$P_{j,k}$
$\tilde{P}_{j,k} = P_{j,k} - \hat{r}_k(X_{j,-k})$
$Z_{j,k}$
$\tilde{Z}_{j,k} = Z_{j,k} - \hat{m}_k(X_{j,-k})$

Figure 4.1: Visual representation of the first stage

Lasso regression

A typical method useful in high dimensional settings is Least Absolute Shrinkage and Selection Operator (Lasso) regression. Lasso regression is a linear model that incorporates L1 regularization in the objective function. The regularization term acts as a constraint on the model coefficients, shrinking their absolute values towards zero, which leads to the selection of a subset of important features. The objective function of Lasso has the general following form:

$$\beta^{Lasso} = arg\min_{\beta}\left\{(V - X\beta)^2 + \lambda\sum_{j=1}^{C}|\beta_j|\right\} \tag{4.12}$$

Where V refers to the first stage outcome variables demand, selling price or purchase price, and X includes all the C confounding variables. Parameter λ can be tuned in order to balance regularization and over-fitting.

Some advantages of using a Lasso model is that it can be used for feature selection in high-dimensional settings, it prevents over-fitting, it is easily interpretative and robust to outliers. A drawback of this method is that it is not suitable for including categorical features, and it assumes a parametric linear model. Even though the method can be performed on the log-transformed set to impose a nonlinear functional form, this latter assumption is restricting. It could be argued that the parametric as well as the linear assumption do not hold for this dataset, especially taking into account the different number of interaction effects that could influence both demand and price. The dataset is rather heterogeneous, as it is large in terms of all time period, the variety of level 2 product categories and number of unique SKU's. Interactions and nonlinear effects could capture the heterogeneity of demand and price, whereas the Lasso method is less flexible. Another drawback of the Lasso method is that because it shrinks relative insignificant coefficients to zero, the effect of these variables, albeit small, is removed, leading to biased estimates.

Random Forest regression

Random Forest is a supervised Machine Learning algorithm that can be used for classification and regression problems in both low- and high-dimensional settings. The model is tree-based, which means it uses decision trees to model decisions. In short, it works by recursively splitting the data into subsets based on the values of the input features, and assigning a decision or prediction at each leaf node of the tree. The Random Forest algorithm trains multiple decision trees based on random subsets of the data, using randomly selected features at each node at the tree. The randomization aids regularization in the model and decreases the correlation between the separate trees. The individual tree predictions are aggregated through majority voting or averaging, which is also known as bagging. To optimize the model's performance, the typically minimized loss function is the Mean Squared Error (MSE): $\sum_{i=1}^{N}(V_i - \hat{V}_i)^2$, in which V refers to the dependent variable and $\hat{V}$ its estimated value. Additionally, within each decision tree, this same loss function is applied to compare candidate splits.

When using a Random Forest model, a lot of hyperparameters can be tuned, which influence the bias-variance balance. The method presents numerous advantages. Firstly, it enables high accuracy with minimal tuning. Additionally, the method is capable of effectively handling high-dimensional data and is robust to outliers. Compared to the Lasso technique, the method has a significant advantage because due to its tree-based modeling technique, more complex nonlinear relationships can be identified. Similarly, the absence of linear or parametric assumptions regarding the data is another notable benefit of the method. The model is hypothesized to be better suited for fitting the heterogeneous data set under investigation. However, the Random Forest technique has a notable drawback in that it may be computationally demanding as it necessitates the training of several decision trees and can consume a significant amount of memory. Furthermore, the interpretability of the results may be reduced compared to linear models such as Lasso, given that Random Forest is an ensemble method that merges the predictions of multiple decision trees. Nevertheless, it is possible to infer feature importance based on the gain that variables induce while splitting.

XGBoost regression

Extreme Gradient Boosting is an optimized version of Gradient Boosting, which is another powerful ensemble method for regression and classification. Similar to Random Forest, it is a supervised algorithm that combines several base learners. Instead of bagging these results, XGBoost combines them by boosting. The objective function that is minimized is represented as:

$$\mathcal{L}(\phi)=\sum_{i=1}^{N} L\left(\hat{V}_i, V_i\right)+\sum_{m=1}^{M} \Omega\left(f_m\right)$$

where $\hat{V}_i=\sum_{m=1}^{M} f_m\left(X_i\right)$ presents the predicted value of the outcome variable, and $\Omega(f)=\gamma J+\frac{1}{2}\zeta\|w\|^2$ is a regularization term. Here, L is a differentiable loss function that quantifies the discrepancy between the predicted value $\hat{V}_i$ and the actual target value V_i. $f_m(X)$ Represents m^{th} tree $(m \in (1,2,...,M))$. The regularization terms within $\Omega(f)$ are used to discourage complexity and produce smooth final learned weights w. The objective function is used to train models to improve previous models by using a form of gradient boosting.

The XGBoost algorithm further deepens general gradient boosting algorithms by using a tree pruning technique when estimating the separate trees (expressed by γ), as well as adding the regularization term in the loss function (expressed by ζ). Comparable to Random Forest, the model hyperparameters can be tuned in order to optimize its performance. The method is found to be highly efficient and also useful for high-dimensional problems.

When comparing XGBoost to a Random Forest model, it is declared that Random Forest is a good choice when the goal is to achieve high accuracy with minimal tuning and XGBoost is particularly useful when the goal is to achieve the best possible performance with more tuning. Another advantage of the XGBoost algorithm is that it can handle missing values and is more efficient than the Random Forest algorithm. Random Forest is more resilient to noisy data however, and the XGBoost model is relatively less suitable for handling high-dimensional datasets. However, in the present study, this distinction may not be as significant since the number of covariates is limited to 31.

To summarize the discussed potential first stage models, their advantages and drawbacks are represented in Table 4.1.

Table 4.1: Advantages and drawbacks of the discussed first stage models

Method	Advantages	Drawbacks
Lasso	Handles high dimensionality	Biased estimates due to sparse coefficients
	Robust to outliers	Cannot handle categorical features and missing data
	Easily interpretable	Linear parametric assumption
	Prevents overfitting by shrinking coefficients	
Random Forest	Handles high dimensionality	Not directly interpretable
	Robust to outliers	Cannot handle missing data
	Prevents overfitting by randomization	More computationally demanding than XGBoost
	Prevents overfitting with regularization term	
	Possible to model nonlinearities	
XGBoost	Handles high dimensionality	Not directly interpretable
	Robust to outliers	Needs more tuning than Random Forest
	Prevents overfitting with regularization term	
	Prevents overfitting by pruning trees	
	Possible to model nonlinearities	
	Can handle categorical features and missing data	

4.3.2 Second stage of estimation

In the second stage of DML estimation, the K variations of residualized variables $\tilde{Y}$, $\tilde{P}$ and $\tilde{Z}$ are used in an IV regression. The final estimator $\hat{\theta}$ is calculated by equating the average of K score functions based on the orthogonalized variables to zero. The estimator can be mathematically represented as:

$$\hat{\theta} = \Big(\sum_{k=1}^{K}\sum_{i\in I_K}\tilde{Z}'_k\tilde{P}_k\Big)^{-1}\Big(\sum_{k=1}^{K}\sum_{i\in I_K}\tilde{Z}'_k\tilde{Y}_k\Big) \tag{4.13}$$

This model configuration builds on the application of an instrumental variable in the second stage, rendering a Partially Linear Instrumental Variable (PLIV) model. As previously mentioned, alternatively, a Partially Linear Regression (PLR) model can be used, obviating the need for instrument Z. Consequently, parameter vector η is reduced to the set $(l(X), r(X))$, thereby requiring its estimation in the first stage solely for the purpose of residualizing P and Y. The second stage consists of OLS of $\tilde{P}$ on $\tilde{Y}$, yielding:

$$\hat{\theta} = \Big(\sum_{k=1}^{K} \sum_{i \in I_K} \tilde{P}_k' \tilde{P}_k\Big)^{-1} \Big(\sum_{k=1}^{K} \sum_{i \in I_K} \tilde{P}_k' \tilde{Y}_k\Big) \tag{4.14}$$

Chapter 5

Results

The discussed model frameworks in combination with the different first stages yields a total of six models, which are estimated on 75% of the dataset, using 50933 observations. The resulting estimations of the price elasticity coefficient are discussed, and thereafter further analysis is conducted on the models to gain more insight into their differences. This further analysis includes investigation of the feature importances, in-sample and out-of-sample prediction residual analysis, and estimation on category subsets.

5.1 Model output

The resulting outputs of the different models to estimate price elasticity are shown below in Tables 5.1 - 5.2, including bootstrapped standard errors, confidence intervals and the p-value. These metrics are all calculated following the method by Chernozhukov et al. (2018). The number of cross-fitting folds that is used in the DML estimation is K = 5. As a baseline, naive OLS and IV estimation are also executed for both the Partial Linear Regression (PLR) model and the Partial Linear Instrumental Variable (PLIV) model. Also depicted in the tables are reference prediction errors for the first stage estimations of $\eta = (l(X), r(X), m(X))$. During estimation these functions are calculated K = 5 times for each model, so therefore the metrics in Tables 5.1 and 5.2 are approximated by estimating on the whole training sample.

5.1.1 DML-PLR Models excluding instrument

In Table 5.1 the results from estimating the Partial Linear Regression (PLR) models using Lasso, Random Forest and XGBoost regression in the first stage and OLS in the second stage are depicted, yielding models PLR 1, PLR 2 and PLR 3, respectively. The OLS model shows an elasticity of -1.0140, which is close to the value of unitary price elasticity of demand $\theta = -1$. The IV estimate is closer to zero, -0.8785, suggesting there may be endogeneity bias present in the OLS model. The largest coefficient of -1.1543 is attributed to PLR 1, characterized by a Lasso first stage. As there is no instrument used there and the model is linear, most likely there is unobserved variable bias. The small standard error of 0.0129 associated to this model is

Table 5.1: Results naive and DML-PLR estimation methods

Model	OLS	IV	PLR 1	PLR 2	PLR 3
θ Estimate	-1.0140	-0.8785	-1.1543	-1.0655	-0.8180
Std. error	0.0250	0.0347	0.0129	0.0480	0.0448
P-value	0	0	0	0	0
CI lower 2,5%	-1.0629	-0.9465	-1.1796	-1.1595	-0.9059
CI upper 97,5%	-0.9651	-0.8104	-1.1290	-0.9714	-0.7302
Instrument used?	No	Yes	No	No	No
$l(X)$ method			Lasso	RForest	XGBoost
RMSE			1.7260	1.0650	0.9355
$r(X)$ method			Lasso	RForest	XGBoost
RMSE			0.1980	0.0094	0.0094
N	50933	50933	50933	50933	50933

expected because of the fact that some coefficients are shrunk to zero, which can simultaneously bias the results. The tree-based models PLR 2 and PLR 3 produce an estimate of -1.0655 and -0.8180 with similar standard errors of 0.0480 and 0.0488 respectively. The confidence intervals of PLR 1, PLR 2 and PLR 3 show no overlap.

The estimate of PLR 2 is close to -1, suggesting that a price change entails a relatively equal change in demand. The PLR 3 model suggests, following the distinction made in Section 2.1, relative inelastic demand: suggesting a smaller proportional change in demand following a price change. The estimate of PLR 1 is relatively elastic, referring to the instance where a price change would result in a proportionally larger change in demand.

Overall, the tree based methods perform better for estimating both nuisance parameters than Lasso regression. The reference RMSE's for the first stage models $l(X)$ for log-transformed order demand Y and $r(X)$ for log-transformed selling price P show that the Random Forest and XGBoost method have a similar RMSE for their prediction of selling price, indicating that the disparities in their final θ estimate is most likely attributed to the difference in their first stage $l(X)$ demand estimation. The XGBoost model shows the lowest first stage RMSE values, implying it best predicts log-transformed demand variable Y. This does not necessarily mean it best uncovers the patterns that drive Y. Inspection of feature importances might reveal the differences in the first stage Random Forest and XGBoost models. This entails assessing the relative importance of input variables in predicting the nuisance functions $l(X)$ and $r(X)$, with the aim of identifying the factors that have a substantial impact on the demand as well as the price.

Table 5.2: Results naive and DML-PLIV estimation methods

Model	OLS	IV	PLIV 1	PLIV 2	PLIV 3
θ Estimate	-1.0140	-0.8785	-1.0591	-0.6025	-0.4408
Std. error	0.0250	0.0347	0.0119	0.0839	0.0692
P-value	0	0	0	0	0
CI lower 2,5%	-1.0629	-0.9465	-1.0824	-0.7697	-0.5765
CI upper 97,5%	-0.9651	-0.8104	-1.0359	-0.4408	-0.3052
Instrument used?	No	Yes	Yes	Yes	Yes
$l(X)$ method			Lasso	RForest	XGBoost
RMSE			1.7260	1.0650	0.9355
$r(X)$ method			Lasso	RForest	XGBoost
RMSE			0.1980	0.0094	0.0094
$m(X)$ method			Lasso	RForest	XGBoost
RMSE			0.2542	0.0073	0.0081
N	50933	50933	50933	50933	50933

5.1.2 DML-PLIV Models including instrument

Table 5.2 presents the outcomes resulting from DML estimation of the Partial Linear IV (PLIV) models. PLIV 1 refers to the model with Lasso first stage estimation, PLIV 2 refers to the first stage Random Forest regression and PLIV 3 refers to a first stage XGBoost regression. The second stage of these models is IV regression using residualized variables $\tilde{Y}$, $\tilde{P}$ and $\tilde{Z}$. The OLS and IV output remain unchanged. First stage models $l(X)$ and $r(X)$ are estimated in the same fashion as in the PLR model; additionally the RMSE values for $m(X)$ are included.

The Lasso-based method PLIV 1 estimate changes from -1.1543 to -1.0591 with respect to PLR 1. This change is not as big compared to models PLIV 2 and PLIV 3, which change from -1.0655 to -0.6025 and from -0.8180 to -0.4408 respectively compared to PLR 2 and PLR 3. The Lasso method performs better in predicting selling price P than purchase price Z, whereas for the other two methods, the prediction errors for predicting purchase price Z are smaller than the prediction errors for selling price P. The PLIV 3 estimate is just contained in the rounded confidence interval of model PLIV 2. This confidence interval is the largest of the three models, which is also consistent with it having the largest standard errors of the three models. The PLIV 1 confidence interval shows no overlap with the other PLIV confidence intervals.

It can be seen that the PLIV 1 model delivers a price elasticity estimate that is close to the PLR 2 model, being unitary elastic, i.e. the coefficient is approximately equal to -1. Models PLIV 2 and PLIV 3 suggest relative inelasticity, indicating that a price change results in a relatively smaller change in demand. In practice, this would mean that a price increase would result in a smaller loss in demand, suggesting a profitable strategy.

Generally, comparing PLR results to PLIV results, the estimations of θ get closer to zero, shrinking the effect selling price P has on demand Y. To further evaluate the differences between models PLR 1, PLR 2, PLR 3 and PLIV 1, PLIV 2 and PLIV 3, some further analysis will be conducted.

5.2 First stage analysis

In order to enhance understanding of the various models, the initial stage models $l(X)$ for outcome variable log-transformed product order demand Y, $r(X)$ for treatment variable log-transformed selling price P and $m(X)$ for the instrument log-transformed purchase price Z are examined. It should be noted that despite the inclusion of 31 covariates in confounder matrix X, dependent variables Y, P, and Z only serve as outcome variables in these models, and as such, cannot be directly interpreted as independent (explanatory) variables. For instance, the models are not structured upon the hypothesis that P has an effect on Y, and Z is related to P, as these variables are solely used as dependent variables in this stage and are not included in covariate set X. In short, the first stage models cannot be used for direct causal interpretation.

One approach to gain a better understanding of the estimated first stage model set η, the feature importances of these models can be examined. By analyzing the wieght or "importance" assigned to each variable, it is possible to determine the factors that significantly influence demand and both prices. In the following sections, the importances for $l(X)$, $r(X)$ and $m(X)$ will be discussed. In Lasso estimation, a number of features are shrunk to zero, leaving only a few impactful variables. The resulting feature importances are represented by the absolute coefficient estimates. In Random Forest regression, the feature importances are determined by the mean decrease of Gini impurity caused by splits based on the features. In XGBoost Regression, the feature importances indicate the average gain across all splits where the feature was used, measured as the contribution of the feature to the reduction in loss when creating a split in the tree. For all three models, the feature importances are normalized by dividing them by their sum. This provides valuable insights into the relative importance of each variable and makes comparison straightforward.

5.2.1 Feature importances for $Y = l(X)$

The Lasso $l(X)$ model, which explains log-transformed demand variable Y, sets all covariate coefficients to zero, except for three. These include the weekly average article unavailability percentage, with a normalized importance of 0.9381; the article content volume with a normalized importance of 0.0213; and the week number shows a normalized importance of 0.0867. All observations with zero demand were dropped from the dataset, and other than that there were no observations of exterme low demand present in the set. Therefore, it was not anticipated that unavailability would still have a big impact on the demand values. On the other hand, it is not surprising that the unavailability percentage is still correlated with demand, as it limits the supply and thereby restricts demand even when it is not extremely low. As expected, with only three variables included, the Lasso model is very simple. It is likely that the model is not well-equipped to account for the confounding effects adequately.

The feature importances of the tree-based models for $l(X)$ are illustrated in Figure 5.1. As mentioned, the normalized Lasso variable importances are distributed over three variables only, and therefore these importances are significantly bigger. To be able to still closely assess the differences of the tree-based feature importances, the Lasso importances are excluded from the figure. It can be seen that the two models show diverging importances. The Random Forest model indicates that, similar to Lasso, the weekly average article unavailability, article content volume, and week number are the most important variables, with their respective normalized feature importances being 0.2015, 0.1686 and 0.0866. It is surprising to note that the dummy indicating holidays shows little importance, despite its expected effect, as mentioned in Section 3.3.1. The week number variable is the third most important feature in the Random Forest

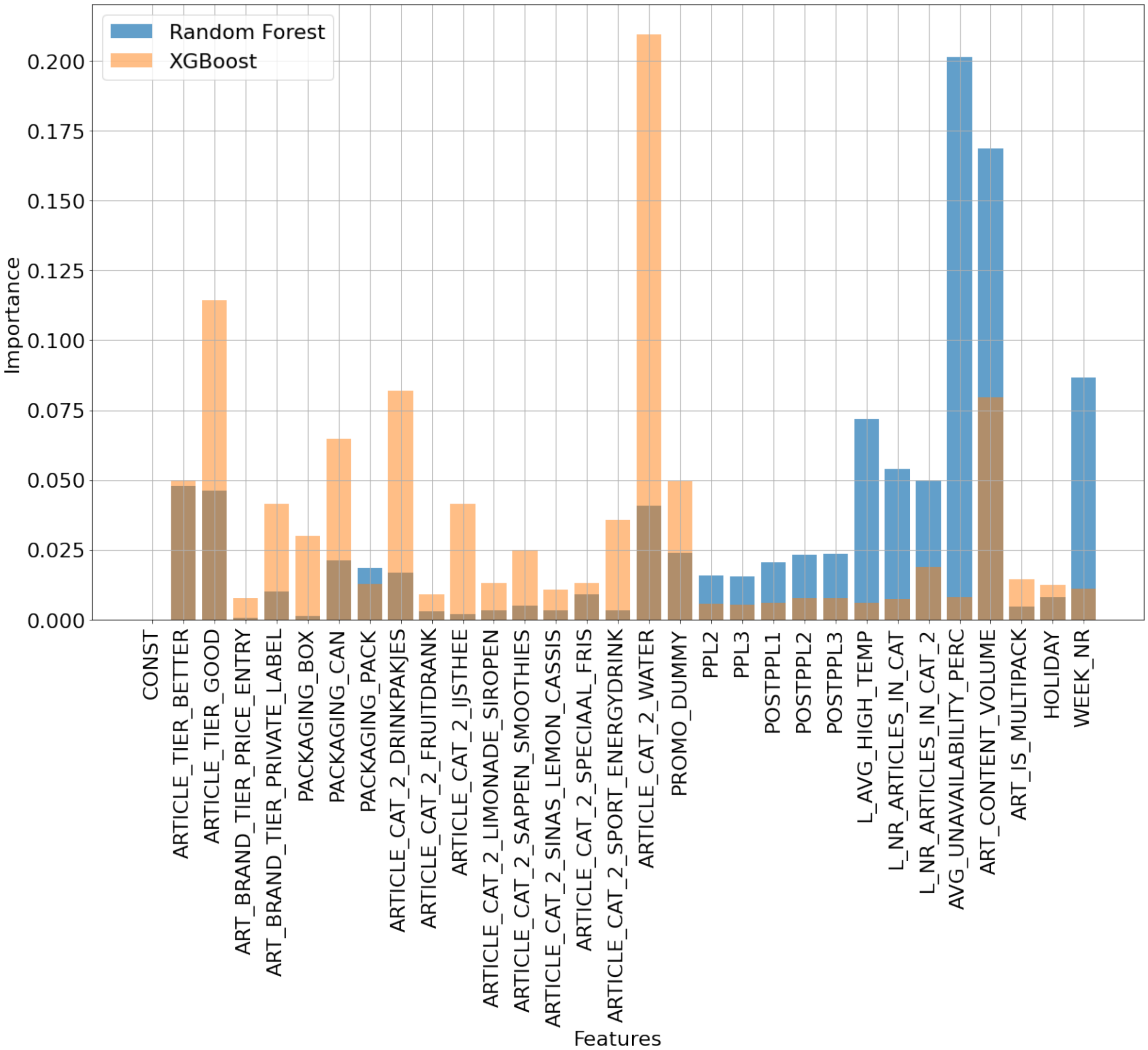


Figure 5.1: Feature importances Random Forest and XGBoost for Y

model (normalized importance of 0.0406), which is used to portray inflation. Furthermore, the Random Forest model adds average temperature to this list with an importance of 0.0720, together with the number of SKU's in the article's category, which in a sense measures cannibalization and identifies the article's category beyond the category dummies. The number of articles in category level 1 has an importance of 0.0537; the number of articles in category level 2 has an importance of 0.0505. The level 2 product sub-category reducing most impurity is "Water", attributing it an importance of 0.0407.

The XGBoost model suggests this "Water" category dummy variable has the most variable importance overall, with a normalized importance of 0.2096. Table 3.3 showed that this category has by far the highest product order amount of all ten categories, making it an evident feature to split on. Another important category feature in the XGBoost model is the "Small cartons" ("Drinkpakjes") category dummy variable, with an importance of 0.0820. XGBoost's second most important feature in predicting demand is the "Good" article quality tier, with a normalized importance of 0.1143. Similar to the Random Forest and Lasso models, the XGBoost model considers the article content volume to be of significant (0.0794) importance. Additionally, the XGBoost model considers article packaging and promotion dummies to be influential, with respective normalized importances of 0.0649 for the "Can" packaging dummy and 0.0496 for the promotional dummy variable. Even though the promotional sales are not taken into account and observations with zero regular orders are dropped, there can be instances where there is a promotion for only part of the customers, leading to a smaller number of regular orders. Therefore, the promotional dummy can serve as a demand restriction, and is therefore of importance. It is noteworthy that the XGBoost model assigns importance to all features, whereas the Random Forest model feature importances are a little less spread.

Excluding the "Water" category

Given the importance of the "Water" category in both the Random Forest and XGBoost models, it is worth exploring the potential impact of excluding this subcategory from models PLIV 2 and PLIV 3 to assess whether the category's price elasticity may be of big influence. This yields the following results:

In Figure 5.3 it can be seen that in comparison to Table 5.2 the estimates all shrink closer to zero, indicating that the "Water" subcategory might have a higher elasticity (and hence a larger coefficient in absolute terms). The Lasso first stage RMSE's decrease, although we've seen that these models are very simple hence maybe not realistic. The PLIV θ estimates change only a little: from -0.6025 to -0.5702 for PLIV 2 and from -0.4408 to 0.4098 for PLIV 3. The RMSE values for $l(X)$ by Random Forest and XGBoost increase a little bit, and the RMSE's for $r(X)$ and $m(X)$ remain constant. In section 5.5 the model will be executed on sub-categories to further analyse the differences of the price elasticity estimates for the data subsets.

Table 5.3: Results naive and DML-PLIV estimation methods without "Water" subcategory

Model	OLS	IV	PLIV 1	PLIV 2	PLIV 3
θ Estimate	-0.9868	-0.7894	-0.8749	-0.5702	-0.4098
Std. error	0.0254	0.0348	0.0134	0.0838	0.0703
P-value	0	0	0	0	0
CI lower 2,5%	-1.0365	-0.8576	-0.9012	-0.7344	-0.5476
CI upper 97,5%	-0.937	-0.7212	-0.8485	-0.4060	-0.2719
Instrument used?	No	Yes	Yes	Yes	Yes
$l(X)$ method			Lasso	RForest	XGBoost
RMSE			1.5832	1.0577	0.9099
$r(X)$ method			Lasso	RForest	XGBoost
RMSE			0.1623	0.0096	0.0094
$m(X)$ method			Lasso	RForest	XGBoost
RMSE			0.2012	0.0076	0.0081
N	47920	47920	47920	47920	47920

5.2.2 Feature importances for $P = r(X)$

The first stage models $r(X)$ for P are examined next. The Lasso model again shows three important features: the "Private label" brand tier shows a normalized absolute coefficient of 0.5164; the "Fruit drinks" category has an absolute coefficient of 0.4023 and the "Good" article tier an absolute coefficient of 0.0716. Table 3.3 showed that the "Fruit drinks" category has the second lowest average product price, which explains its importance. The lowest prices are associated with the "Water" category, however this coefficient is set to zero in the Lasso model. The average unavailability percentage and the week number both have a relative importance smaller than 0.01; the normalized feature importance of article content volume is smaller than 0.001. The other coefficients are shrunk to zero.

Figure 5.2 shows the tree-based model first stage feature importances. Again, because the (standardized) Lasso feature importances are of a larger scale, they are excluded from the graph. The Random Forest model assigns most importance to the article content volume (normalized importance of 0.2180), secondly the "Private label" brand tier (0.1623), and thirdly the dummy indicating whether the article is a multipack (0.1108). The most important feature in the XGBoost model is this same multipack dummy variable (normalized importance of 0.3373). The multipack dummy and article content volume determine the size of the SKU's, making these importances meaningful in predicting selling price. Second and third most important in the XGBoost model for $r(X)$ are the "Private label" brand tier (0.2301) and the "Sport- and energydrink" category (0.1242). Table 3.3 shows that the prices associated with the "Sport- and energydrinks" are the highest of all ten categories, and Table 3.2 showed the "Private label" tier

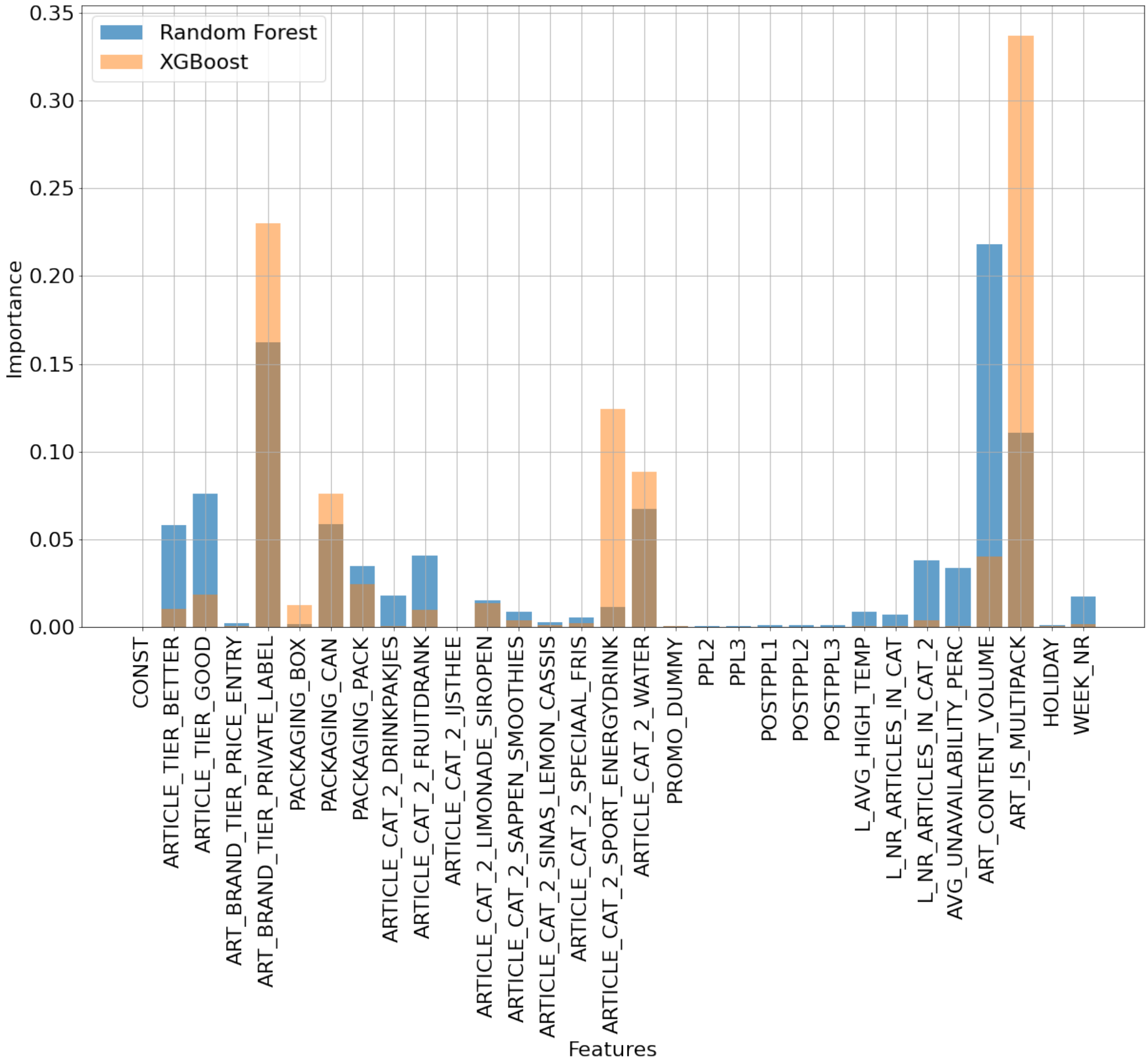


Figure 5.2: Feature importances Random Forest and XGBoost for P

has the lowest prices of the three brand tiers. Hence, it is not surprising that these variables show importance in the tree-based selling price models.

Unlike the first stage model $l(X)$ for demand Y, there are now some features that show little to zero importance. It is unexpected that the dummy variables representing different price lines ("PPL2", "PPL3", "POSTPPL1‘", "POSTPPL2" and "POSTPPL3") do not exhibit any importance in the model, despite the fact that they vary in price, with PPL3 and POSTPPL3 generally showing highest prices. It is possible that the changes made to the distinction of price lines in June 2022, doubling the number of PPL variables, resulted in a less clear distinction of prices. Additionally, there are also a number of observations where one default price was asked for all price line types. This introduces further noise in the difference between regional price lines.

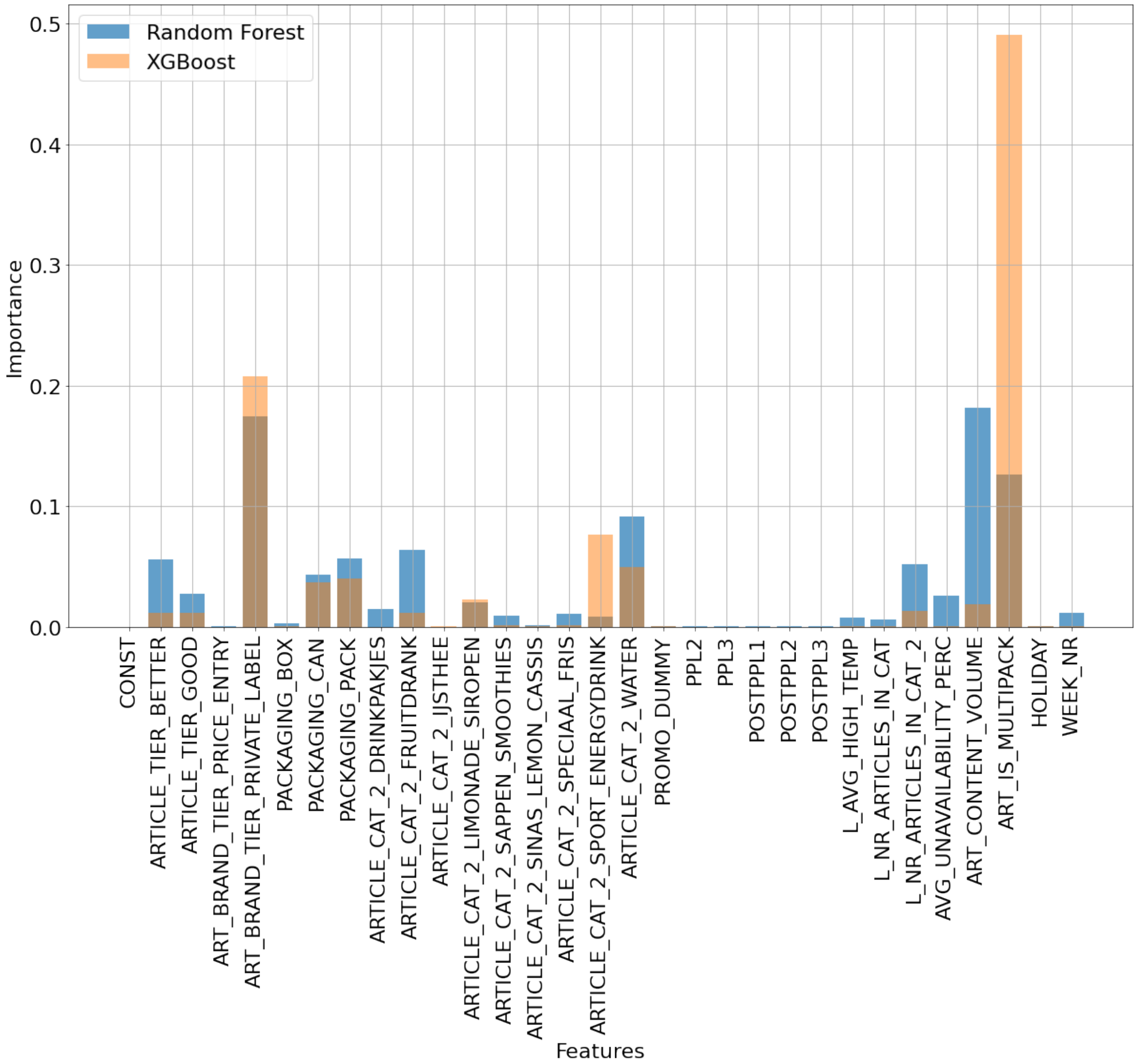


Figure 5.3: Feature importances Random Forest and XGBoost for Z

5.2.3 Feature importances for $Z = m(X)$

The feature importances for first stage model $m(X)$ for log-transformed purchase price Z are similar to those for first stage model $r(X)$ for P. The Lasso model attributes 0.5155 to the “Fruit drinks” category and 0.4700 to the “Private label” brand tier. The feature importances for the Random Forest and XGBoost models are depicted in Figure 5.3, and show a similar picture as in Figure 5.2 where the importances for $r(X)$ are depicted. The biggest difference for the Random Forest model is that the normalized feature importance for the article content volume decreases, yielding more spread in the importances. The XGBoost importances decrease in spread: all importances are smaller, except for the multipack dummy variable, which grows from 0.3373 to 0.4911.

5.2.4 Comparison of first stage models

The difference between the first stage model predictions performances depicted in the RMSE values in Tables 5.1 and 5.2 suggest that there is nonlinearity or another form of complexity in demand Y as well as selling price P and purchase price Z, which the Lasso models are less equipped to fit than the tree-based models Random Forest and XGBoost regression. Examination of the feature importances confirm Lasso's comparative deficiency by revealing that the Lasso first stage models are fairly simplistic. Figures 5.1 - 5.3 display the varying feature importances associated with the Random Forest and XGBoost first stage models, indicating similar features of significance but differing in their relative importances. Generally, the XGBoost models attribute more relevance to the product category dummies than the Random Forest models. In models $r(X)$ and $m(X)$, XGBoost shows a little less spread by attributing comparatively more importance to less features.

5.3 Residual analysis

Even though the DML approach is not designed in order to do predictions, residual analysis based on predictions can be helpful in order to judge the estimated coefficients more thoroughly, especially regarding the validity of using the estimated coefficients in practice.

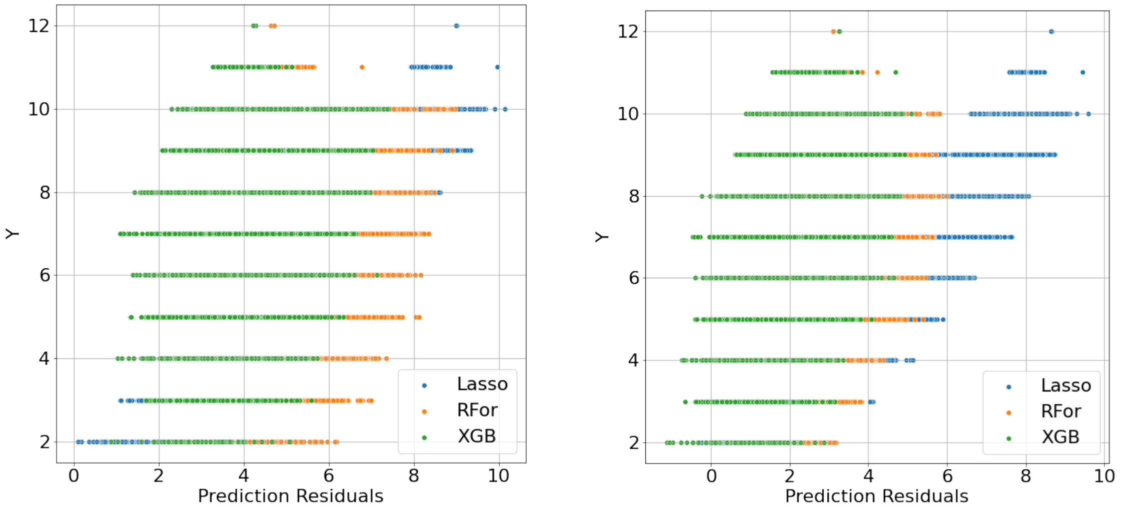
The advantage of looking at the correlation between observations and prediction residuals instead of the correlation between observations and predictions, is that it helps distinguishing whether the model's prediction errors are randomly distributed, which would yield low correlations, or whether the prediction errors might be systematic: indicating a structural bias in the model's predictions and a poor fit of the model. In the end, the objective is not to find the best prediction $\hat{Y}$ of Y, but to find the model that best fits and explains the data at hand without any structural deficits. Residual analysis could help give an insight in this model fit.

Predictions of $Y = \log(\text{product order amount})$ are made according to (4.1), the first stage estimations of $l(X)$ and the estimated $\hat{\theta}$ values depicted in Tables 5.1 - 5.2.

5.3.1 Using training data

Firstly, the product order demand is predicted using the same data set on which the coefficient is established (the training dataset). In Figure 5.4 the relation between (log-transformed) demand observations Y and the associated prediction residuals is visualized. In Figure 5.4a this is done for the PLR prediction residuals and in Figure 5.4b this is done for the PLIV prediction residuals.

It seems that for the PLIV and PLR models using Lasso orthogonalization there is comparatively the most correlation between Y and its prediction errors. To further clarify this, in Table 5.4 Pearson's correlation coefficients measuring the correlation between observations Y and prediction residuals $Y - \hat{Y}$ for both the PLR and PLIV models are depicted.



(a) DML-PLR train set prediction residuals (b) DML-PLIV train set prediction residuals

Figure 5.4: Scatter plots of Y and their prediction errors on the train set

Table 5.4: Correlation between Y and associated prediction errors $Y - \hat{Y}$ based on train set

	PLR 1	PLR 2	PLR 3	PLIV 1	PLIV 2	PLIV 3
Corr(Y,$Y - \hat{Y}$)	0.8911	0.1899	0.4541	0.9045	0.3540	0.5629
N	50933	50933	50933	50933	50933	50933

It can be seen from Figure 5.4 as well as from Table 5.4 that the Lasso-based models have prediction errors that correlate substantially with the observed Y values: the correlation coefficients are 0.8911 and 0.9045 for PLR 1 and PLIV 1 respectively. This correlation indicates that the Lasso models are not adequately capturing some of the underlying patterns in the data. More specifically, the graph shows that these models overestimate low demand and underestimate high demand. This indicates the presence of nonlinear and/or interaction patterns in the data, which, as discussed in Section 4.3.1, the Lasso method is not able to capture.

The tree-based ensemble models seem to fit the demand distribution more appropriately, resulting in lower correlation coefficients of the observations and prediction residuals. The Random Forest regression-based models show the lowest correlation between the log-transformed product order amount observations and their prediction residuals: 0.1899 for PLR 2 and 0.3540 for PLIV 2. This shows that within the training sample, the Random Forest based models show the best fit and prediction to the data of log-transformed demand. Models using the XGBoost method show a correlation coefficient of 0.4541 for the PLR model and 0.5629 for the PLIV model. In conclusion, it seems that the Random Forest first stage models are most suitable to fit and predict the demand variable within-sample.

5.3.2 Using test data

The analysis is repeated on data that was not included in the original training sample. Initially, a dataset from the same time period as the training data is used. Subsequently, a dataset from a time period that is contiguous to the training data is used.

Test data originating from the same time period

In order to accurately assess the demand predictions based on estimated elasticities, it must be assumed that price elasticity remained consistent between the estimation (training) period and the prediction (test) period. Therefore, accurate and valid predictions can only be made if the test data is from the same time period as the training data. For this study, a test set was created by randomly withholding 25% of the data before estimation, which satisfies the assumption of constant elasticity. Demand predictions are made according to (4.1), using the prediction output of first stage model $l(X)$ and the estimated price elasticities $\hat{\theta}$ on test data. Table 5.5 shows the correlation between the observations from this test set and the associated prediction errors.

Table 5.5: Correlation between Y and associated prediction errors $Y - \hat{Y}$ based on test set

	PLR 1	PLR 2	PLR 3	PLIV 1	PLIV 2	PLIV 3
Corr(Y,$Y-\hat{Y}$)	0.8890	0.4187	0.4933	0.9026	0.5177	0.5841
N	16978	16978	16978	16978	16978	16978

Table 5.5 shows big increases in correlation for the Random Forest-based model correlation coefficients when compared to Table 5.4: the coefficient for PLR 2 increases from 0.1899 to 0.4187 and the coefficient for PLIV 2 from 0.3540 to 0.5177. This indicates that the Random Forest model might be a little overfitted on the training data. Meanwhile, the correlation stays more or less constant for the Lasso- and XGBoost-based methods: 0.889 for PLR 1; 0.4933 for PLR 3; 0.9026 for PLIV 1 and 0.5841 for PLIV 3. Despite the increase in correlation, the Random Forest model maintains the lowest coefficients. Furthermore, for the Random Forest and XGBoost first stage models, the differences between the correlation coefficients of the PLR and PLIV residuals become slightly smaller. The difference between models PLR 2 and PLIV 2 changes from 0.1641 (using Table 5.4) to 0.099 (using Table 5.5). The difference between PLR 3 and PLIV 3 decreases from 0.1088 (using Table 5.4) to 0.0908 (using Table 5.5).

Test data originating from a different time period

In order to test the assumption whether price elasticity is indeed constant over time, additional residual analysis is done using a test sample based on observations for a time period successive to the training sample period. To be precise, a data set composed of observations during the first two months of 2023 is created, using the same articles that were used in the training sample.

Again, the test set is used to predict demand using the estimated price elasticity coefficients. The resulting residual correlation coefficients are depicted in Table 5.6.

Table 5.6: Correlation between Y and associated prediction errors $Y - \hat{Y}$ based on 2023 test set

	PLR 1	PLR 2	PLR 3	PLIV 1	PLIV 2	PLIV 3
Corr(Y,$Y - \hat{Y}$)	0.9274	0.6236	0.6411	0.9354	0.8146	0.7998
N	22077	22077	22077	22077	22077	22077

Table 5.6 shows the correlation coefficients between the log-transformed product order demand observations and the model prediction residuals, based on the 2023 test set. All correlation coefficients increase w.r.t. both the training sample and initial test sample (see Tables 5.4 and 5.5). The correlation coefficients are all larger than 0.6 and this indicates that either the models are not suited to do out-of-sample prediction, or the price elasticity changed in 2023 compared to the 2021-2022 sample, or both. Either way, the results show that the model is not relevant for predicting demand during a time period outside the training period.

5.3.3 Comparison

Tables 5.4 - 5.6 show that Lasso-based models PLR 1 and PLIV 1 are the least suitable method to use for prediction. In the training sample and the initial test sample, which consists of observations from the same time period, the Random Forest regression-based model demonstrated the lowest correlation coefficients. The differences between the correlation coefficients obtained from the Random Forest-based models of these two samples suggest that the Random Forest model is more susceptible to overfitting. This finding, however, can be interpreted positively since the objective of this research is not to forecast future demand accurately and to avoid overfitting, but rather to identify the underlying patterns in the given (training) sample to estimate price elasticity in a valid manner. The low correlation coefficients in Table 5.4 associated with the Random Forest models imply that when these models are used, prediction errors seem roughly randomly distributed. Results obtained from the 2023 test sample show no big differences in correlation coefficients for the Random Forest-based and XGBoost-based prediction methods. The correlations are high, confirming this model is not suited for prediction on a new time period.

Finally, in all samples, the PLR models exhibit lower correlation between product demand observations Y and the associated prediction residuals than the PLIV models. A possible explanation for this finding is that the PLIV models predict an additional function $m(X)$, which models the purchase price Z based on confounding variables X. This additional function increases the degrees of freedom and introduces additional noise. However, the assumptions of including the IV also play a role. To determine which model provides the most realistic θ estimates, it is essential to evaluate the model assumptions, rather than its predictive performance. Although the PLR model may appear to have less correlation between the predicted residuals and associated observations, the assumption that no unobserved variables affect the selling price P may be implausible.

5.4 PLIV model variation

Table 3.5 showed that P and Z are highly correlated, suggesting that a high-dimensional model for P might not be necessary when Z is included. Therefore, a model variation is executed where the first stage model for price $r(X) = \mathbb{E}[P|X]$ is a simple OLS model, while the first stage for demand $l(X) = \mathbb{E}[Y|X]$ is higher in complexity and the tree-based models are employed

Table 5.7: Results DML-PLIV estimation methods with simple model for $m(X)$ and $r(X)$.

Model	**PLIV 4**	**PLIV 5**
θ Estimate	-0.1281	-0.1209
Std. error	0.0309	0.0294
P-value	0	0
CI lower 2,5%	-0.1887	-0.1784
CI upper 97,5%	-0.0674	-0.0633
Instrument used?	Yes	Yes
$l(X)$ method	RForest	XGBoost
RMSE	1.0650	0.9355
$r(X)$ method	OLS	OLS
RMSE	0.0376	0.0376
$m(X)$ method	OLS	OLS
RMSE	0.0407	0.0407
Corr(Y,$Y - \hat{Y}$) train data	0.5463	0.6401
Corr(Y,$Y - \hat{Y}$) test data	0.6068	0.6474
N	50933	50933

The estimation results for models PLIV 4 and PLIV 5 are presented in Table 5.7. These models use an OLS first stage for nuisance functions $r(X)$ and $m(X)$ to estimate selling price P and purchase price Z.

Both models show a similar θ estimate, -0.1281 for PLIV 4 and -0.1209 for PLIV 5, which are closer to zero than all other estimates presented before. The RMSE's of the first stage models of $r(X)$ and $m(X)$ are 0.0376 and 0.0407, which means they predict better than the Lasso model, which is to be expected due to Lasso's sparsity. Furthermore, due to the inclusion of many dummy-variables in X, the OLS models are somewhat suitable to estimate non-linear dependencies. On the other hand however, the Random Forest and XGBoost model still perform better in predicting selling price and purchase price, as can be seen by the RMSE values from Table 5.2. The correlation of the models' prediction residuals with observations based on the train data set is 0.5463 for PLIV 4 and 0.6401 for PLIV 5, whereas these were 0.3540 for model PLIV 2 and 0.5629 for PLIV 3. The analysis demonstrates that the PLIV 2 model, which employs a Random Forest algorithm to estimate prices, produces the most random prediction errors. This finding suggests that among all the PLIV models tested, the PLIV 2 model provides the best fit to the available data, and that the first stage tree-based models for P and Z do improve the model fit within-sample.

Models PLIV 4 and PLIV 5 employ a straightforward estimation approach for determining prices. In this scenario, while Z is strongly correlated with P, it may not possess sufficient information to account for all the variability in P. This means that residualizing P and Z with a basic linear model will leave other confounding variables present in the second stage that influence prices, thereby introducing noise when estimating the impact of residualized prices $\tilde{P}$ on the residualized demand $\tilde{Y}$. Put differently, a linear first stage model for P and Z is hypothesized to be too simplistic to account for confounding variables.

5.5 Results based on category level 2 subsets

It could be argued that the data set is too large in terms of time period and/or the amount of different articles to support one unique value for price elasticity. To take the latter argument into account, the estimation is also done on subsets of the data set, based on the level 2 partition of product categories. Another argument to employ the model on subsets is to assess and compare the price elasticity estimates of various article groups. In case the elasticity estimates show differences, these characteristics can be used in steering the pricing strategies of the product subcategories. For example, subcategories with true elasticities close to zero are befitted for less competitive pricing than articles with larger elasticities (in absolute terms).

5.5.1 OLS and IV model output

Firstly, simple OLS and IV are executed, yielding the results shown in Table 5.8.

In each model estimation, inclusion of all covariates within X lead to rank deficiency due to multicollinearity. To be precise, linear combinations of article attributes cause this. The columns that mostly introduce mutlicollinearity were the columns representing the article brand tier, the article quality tier and the article packaging type. For each subset, the columns causing multicollinearity were removed before estimation, which means that for the ten subsets, the composition of X is not consistent. Table 5.8 shows that for six out of ten categories, the estimated elasticity is positive: which is an unexpected result. In general, as discussed, it is expected that a price elasticity estimate is negative: a price change is usually reciprocated with a response of demand in the opposite direction of the associated price change. For most categories, the IV and OLS results have (almost) overlapping confidence intervals, yielding similar output.

Table 5.8: OLS and IV θ results on data subsets

Category	N	OLS	IV
Fruit drinks	10955	0.0199	0.5227***
CI		(-0.08, 0.12)	(0.36 ,0.68)
Juices & smoothies	8923	-1.7096***	-2.0331***
CI		(-1.82, -1.60)	(-2.15, -1.91)
Special soda	7083	-2.0352***	-2.1663***
CI		(-2.15, -1.92)	(-2.31, -2.03)
Cola	6285	0.0603*	0.3764
CI		(-0.01, 0.13)	(0.29, 0.46)
Lemonade & syrups	6229	-1.1944***	-1.8903***
CI		(-1.37, 1.02)	(-2.21, -1.41)
Small cartons	5873	0.1753*	2.3730***
CI		(-0.01, 0.36)	(2.09, 2.65)
Orange, lemon & cassis	3915	0.1637	0.9134***
CI		(-0.04, 0.37)	(0.65, 1.18)
Water	3168	-0.3124**	-0.8782***
CI		(-0.62, -0.01)	(-1.40, -0.35)
Sport- & energydrinks	959	0.5677***	0.4040***
CI		(0.32, 0.82)	(0.16, 0.65)
Ice tea	938	0.9548***	1.0224***
CI		(0.69, 1.22)	(0.75, 1.29)

* Significant at 10% level

** Significant at 5% level

*** Significant at 1% level

This is not the case for the “Small cartons” category however, where the OLS estimate is 0.1753 and the IV estimate 2.3730, which is a big difference. Figure 3.8a shows that for this category the margin between log-transformed selling prices P and purchase prices Z is fluctuating a lot, which might explain why the model estimates including and excluding Z diverge. This margin remained relatively most consistent in the “Special soda” subcategory, for which Table 5.8 shows only a small difference between the OLS and IV θ estimates of -2.0352 and -2.166 respectively. However, Figure 3.8b shows that for almost all subcategories the difference (margin) between log-transformed selling and purchase price is fluctuating and increasing over time, while this does not always show in the differences between the OLS and IV estimates illustrated in Table 5.8. For example, the “Ice Tea” category shows a lot of fluctuation in prices and the difference between transformed selling an purchase prices, which would suggest that IV regression would show different output than OLS. However, the OLS and IV estimates are very similar: 0.9548 and 1.0024.

The table shows OLS and IV θ estimates between $(-1.7, -2.2)$ for the "Special soda", "Juices & smoothies" and "Lemonade & syrups" categories. The "Water" category shows estimates closer to zero, but still negative. The "Fruit drinks"; "Cola"; "Orange, lemon & cassis"; "Sport & energydrink" and "Ice tea" coefficients all have a positive estimated θ coefficient between $(0, 1.1)$. Lastly, as mentioned, the estimates for "Small cartons" are 0.1753 for OLS and 2.3730 for IV.

5.5.2 DML model output

Next, DML is used to estimate the price elasticity of the subsets, to compare with the results from Table 5.8. The DML method has the benefit of avoiding the problem of rank-deficiency by reason of the tree-based first stage Machine Learning methods employed. The results above show that due to the linear parametric assumption the Lasso first stage models are not appropriate to model the complex and potentially nonlinear variables demand, selling price and purchase price. Moreover, the sparsity of the Lasso-based models' feature importances and the high correlation between its prediction residuals and observations confirm that these models are not suitable as a first stage model in this research setting. For this reason, the subset analysis is done by using models PLR 2, PLR 3, PLIV 2 and PLIV 3 only.

Table 5.9 shows the results of using these four models to estimate θ. The elasticity estimates within each subcategory are diverging and seem to show no consensus. For example, the PLR models give a negative coefficient estimate for "Fruit drinks", whereas the PLIV model yields positive θ estimates for this category. The "Cola" θ estimates are positive for all four models. The OLS and IV estimates (shown in Table 5.8) are also positive for this category. In this dataset, a lot of the articles contained in the "Cola" set are of the 'Coca-Cola" brand, which is a very popular, strongly established brand. Of these products, it is not unsurprising that the demand is relatively insensitive to price changes. However, a true positive price elasticity of demand is not very plausible.

When assessing the PLR 2 and PLR 3 elasticities comparatively, the estimates for the "Cola" subcategory are both positive and larger than 0.7; the "Water" and "Orange, lemon & cassis" categories show one negative and one positive coefficient within $(-1, 1)$; the "Fruit drinks", "Small cartons" and "Sport & energydrink" estimates are both contained in $(0, -1)$; estimates for the "Lemonade & syrups", "Special soda" and "Juices & smooties" categories are both contained in $(-1, -3)$ and finally the "Ice tea" estimates are both smaller than -3.

Comparing the PLIV 2 and PLIV 3 elasticity estimates in the same manner, it is found that the estimates for "Sport & energydrinks" are both very large: 58.23 for PLIV 2 and 24.90 for PLIV 3. The "Ice tea" estimate is 2.24 using PLIV 2 and 11.31 using PLIV 3. Moreover, it can be seen that again both "Cola" estimates are positive and in this case both within $(4.5, 5)$. Also, both estimates for "Small cartons" are postive, within $(3.5, 4.5)$, and the "Orange, lemon & cassis" estimates are within $(1, 3)$. Both "Fruit drinks" estimates are positive and contained in $(0.35, 0.45)$. Negative estimates are found for the “Water”; “Lemonade & syrups”, ‘Special soda” and “Juices & smooties” categories, where both estimates are contained in $(-1, -3)$. The latter three categories show similar coefficients in Table 5.8.

The two categories that contain observations of four products only, “Ice Tea” and “Sport & energydrink”, show some very large PLIV elasticity estimates. These estimates are not significant, and differ a lot from the PLR 2 and PLR 3 estimates for those categories, which are all negative: within $(0, -1)$ for “Sport & energydrinks” and both around -3.8 for “Ice tea”. These unexpected and noisy results are presumably due to the fact that these categories possess the least SKU’s: only four in this set. This means that a large part of the explanatory confounding variables, that are article-related, show little to no variation and are multicollinear. Therefore, the first stage models won’t perform well, which causes the residualized variables to still be correlated with the outcome. In other words, the residualized variables will most likely still posses substantial underlying, unobserved patterns that were not partialled-out during the first stage, causing the final θ estimates to be biased.

The PLR models yield estimates that are closer to zero when compared to PLIV. This is contradictory to the results for the whole set, where PLR models showed larger estimates than PLIV in absolute terms (displayed in Tables 5.1 and 5.2). Both PLR 2 and PLR 3 estimates are positive for only one out of ten categories, whereas for in six out of ten subcategories, models PLIV 2 and PLIV 3 yield positive coefficient estimates. The linear models (estimated in section 5.5.1) show equivalent signs as PLIV 2 and PLIV 3.

All DML models agree on some ordering: the “Cola” articles show the largest positive elasticity estimates, which are also positive in the linear models. In addition, all DML models show negative elasticities for “Special soda” articles, even lower estimates for the “Juices & smoothies” articles and finally the lowest, most negative estimates for “Lemonade & syrups”. These estimates are also negative and show similar coefficients in the estimated linear models. The estimates for “Fruit drinks” are also similar for all linear and DML models, contained in $(-0.5, 0.5)$. The estimates for “Ice Tea” are the lowest in the PLR models, however they are positive in the PLIV and linear models, indicating some noise, which might be caused by the low number of articles in the set. Besides, Figures 3.8b shows that the difference between selling price P and purchase price Z fluctuated and increased over time for this category, which may have also introduced additional noise.

5.6 Difference between PLR and PLIV

The differences between the PLR and PLIV findings can be explained by the use of purchase price as instrument Z: when this instrument is included, essentially the residualized selling price $\tilde{P}$ is proxied by residualized purchase price $\tilde{Z}$, resulting in a variable roughly measuring the margin or surplus between purchase and selling price. In this setting, the DML method then evaluates the response in residualized demand $\tilde{Y}$ after a change in this “margin” between $\tilde{P}$ and $\tilde{Z}$. For simplicity, θ could in this PLIV case be named as the “margin elasticity of demand”.

A positive estimate for θ would then imply that an increase in margin would result in an increase in demand. Even though this exact causal connection can be doubted, positive "margin elasticities" could occur in situations where a margin decrease is observed simultaneously with a decrease in demand. A situation like this is likely to occur, for example considering rising purchase prices that stem from increasing production costs due to the natural gas crisis in 2022 following the Russian-Ukrainian war. This same natural gas crisis could have ensued irregular high inflation, which potentially lead to changing consumer behavior and therefore lower demand. As a result, both the margin and demand decrease due to an external circumstance, leading to positive "margin" elasticity estimates. In short, a positive correlation between margin and demand is not unlikely, whereas a positive causal relation between the two is less expected.

To generalize this aforementioned example, there could still be unobserved external variables influencing both P, Z and Y. Obviously, these effects influence the accuracy of all models, also the models where the instrument is excluded. In the setting where instrument Z is included however, the assumption that Z captures all unobservables affecting P is then specifically deceptive. Another reason to argue leaving out instrumental variable Z is that in this research, it is attempted to approximate the price elasticity of demand in order to quantify how customers respond to changes in selling prices. In this framework, it is proclaimed that the consumer takes prices as given when shopping, and subsequently decides based on these prices whether to buy the associated articles (or not). Using the instrument will change the interpretation of the estimated θ coefficients to the effect of "margin" changes on demand, however these margin changes are generally not within sight of the consumer. Customers only fully observe margin changes in case the selling price changes and the purchase price and other costs stay constant. Yet, customers don't directly observe changes in costs, which obviously influence the margin. To ensure that the interpretation of θ refers to the change in consumer demand after a change in selling price, it therefore seems sensible to use the PLR estimates over the PLIV estimates. A last argument to substantiate this is to look at Figure 3.8b, in which it can be seen that the relation between P and Z is very changeable during the observed time period.

Table 5.9: DML θ results on data subsets

Category	N	PLR 2	PLR 3	PLIV 2	PLIV 3
Fruit drinks	10955	-0.3300***	-0.4214***	0.4145***	0.3740**
CI		(-0.47, 0.19)	(-0.58, -0.26)	(0.14, 0.69)	(0.025, 0.72)
Juices & smoothies	8923	-1.8658***	-1.4713***	-2.2377***	-1.9233***
CI		(-2.08, -1.66)	(-1.70, -1.24)	(-2.51, -1.97)	(-2.23, -1.62)
Special soda	7083	-2.3164***	-1.5157***	-1.7500***	-1.3296***
CI		(-2.62, -2.01)	(-1.87, -1.16)	(-2.20, -1.30)	(-1.95, -0.70)
Cola	6285	0.7424***	1.0153***	5.3464***	4.8083***
CI		(0.31, 1.18)	(0.52, 1.51)	(4.63, 6.07)	3.87, 5.75)
Lemonade & syrups	6229	-2.0132***	-1.7921***	-2.3889***	-2.1369***
CI		(-2.25, -1.78)	(-2.07, -1.52)	(-2.86, -1.92)	(-2.66, -1.61)
Small cartons	5873	-0.6391***	-0.1421	3.8443***	4.4455***
CI		(-1.02, -0.26)	(-0.60, 0.32)	(2.355, 5.33)	(2.43,6.45)
Orange, lemon & cassis	3915	-0.2874	0.0582	2.6745***	1.2808*
CI		(-0.86, 0.28)	(-0.50, 0.62)	(1.47, 3.88)	(-0.10, 2.66)
Water	3168	-0.6162**	0.5301	-2.5452***	-1.1870*
CI		(-1.19, -0.05)	(-0.18, 1.24)	(-3.48, -1.61)	(-2.40, 0.03)
Sport- & energy-drinks	959	-0.145869	-0.2115	58.2324	24.90058
CI		(-1.73, 1.44)	(-2.16, 1.74)	(-2155.32, 2271,78)	(-1556, 1606.20)
Ice tea	938	-3.8418***	-3.7835***	2.2353	11.3140
CI		(-5.78, -1.91)	(-6.56, -1.01)	(-51.50, 55.97)	(-37.03,59.66)

* Significant at 10% level

** Significant at 5% level

*** Significant at 1% level

Chapter 6

Conclusion

To conclude the research conducted in this thesis, in this chapter the main findings are repeated in Section 6.1. Lastly, in Section 6.2, some recommendations for further analysis are presented.

6.1 Findings

In this research it is attempted to model price elasticity of product demand using two model specifications. The first model specification is a Partially Linear Regression (PLR) model and the second model a Partially Linear Instrumental Variable (PLIV) model. Both models are estimated using naive estimation methods and using the Double Machine Learning (DML) method. This is done on the whole dataset regarding drinks in the time period 2021 - 2023; and for subsets of the data grouped by the product subcategories for the same time period.

The distinction between the PLR and PLIV model specifications plays a central role in the analysis. The PLIV model is argued to be appropriate in case it is assumed that there are unobserved variables directly affecting selling price P, for which instrument purchase price Z can proxy. This assumption posits that the unobservable variables do not directly affect demand Y and instrument Z. It could be reasoned however that there exist unobservable variables influencing not only selling price P, but also purchase price Z and demand Y. In other words, it is questionable whether the assumption on which the IV estimator is grounded upon is valid. Nonetheless, as this assumption is difficult to assess, both models are estimated.

Results on the subsets show some questionable positive PLIV elasticity estimates, and the PLIV model's prediction residuals are more correlated with observations than those of the PLR models. This suggests that the PLIV models may either overestimate or underestimate demand structurally. Additionally, the data indicates that the relationship between the endogenous variable selling price and the instrumental variable purchase price is unstable, which may lead to biased final estimates. Another argument against including the instrumental variable is that the ultimate goal of estimating price elasticity in this study is to inform pricing strategies for retailer Picnic. To achieve this goal, it is necessary to assume that the elasticity estimate reflects consumer behavior based on the selling prices they observe. Using the purchase price as the instrumental variable to proxy for the selling price involves the supply-side of the framework, rather than the consumer's perception and response. Thus, including an instrument may alter the interpretation of the final elasticity estimates.

While prediction is not the primary objective of this model, residual analysis can demonstrate how well the models used fit the available data. The analysis is performed on the training sample, a test set consisting of observations from the same time period as the training sample, and a test sample comprising observations from a time period subsequent to the training sample period. Random Forest-based models exhibit lower correlation between prediction residuals and observations than Lasso and XGBoost models. In particular, the PLR 2 model, which employs a Random Forest in the first stage and a linear regression in the second stage, exhibits the least correlated residuals across all prediction samples.

Three Machine Learning models are used in the first stage: Lasso, Random Forest and XGBoost regression. It seems the Lasso first stage model is too simplistic, which was confirmed by assessing the feature importances and the residual analysis. Feature importances of the tree based models showed that the two algorithms split on different features. In the model for demand, XGBoost and Random Forest attributed the same features importance, however they differed in their amounts. Specifically, the Random Forest model assigned most importance in its demand model to the average unavailability percentage, the article's content volume and the week number, which is the same group of variables that was important in the Lasso model. The XGBoost model assigned most importance to the "Water" subcategory, the "Good" article quality tier and the article's content volume. In the first stage models for purchase and selling price, article content was most important for the Random Forest model, and the multipack dummy variable was most important for the XGBoost model. These variables determine article size and this therefore makes sense. Both models also assigned importance to the "Private label" brand tier, which is low in price. As mentioned, the analysis of the correlations between prediction residuals and observations shows that the Random Forest model seems to fit the training data best, as it shows the most random prediction residuals. This indicates it is the best suitable first stage model to use in this setting.

The model is applied on subsets in order to compare relative elasticities. The various models did not align completely, but there was some unanimity in relative coefficients, electing "Lemonade & syrups", "Special soda" and "Juices & smoothies" as the three most price elastic drinks categories. The subcategory "Cola" got positive elasticity estimates in each model setting, which may not represent the true values of elasticity, but this can imply that the products within this category are not that price sensitive. This can be supported by the fact that this category contains a lot of "Coca-Cola" -branded products, which are very popular and well-established, causing the demand of these products to become less price sensitive.

6.2 Recommendations for future research

To further investigate the elasticity estimation, it may be valuable to examine alternative combinations of product categories and time periods. A longer time frame may provide a better estimate by capturing greater variations in prices and demand. However, as observed with selling and purchase prices, relationships between variables may not be constant over time, and therefore it is essential to avoid excessively long time frames. Shortening the time frame could result in fewer observations, which can be compensated by increasing the number of observed articles. However, it should be noted that the estimated elasticity is presumed to be applicable to all included articles. In other words, striking a balance between an adequate number of observations with enough variation and a homogeneous set for consistent evaluation is crucial.

Another recommendation could be to investigate how to measure and account for macroeconomic variables that possibly play a role and are omitted now, such as the effects of the natural gas crisis and inflation on customer frugality or other variables that portray additional context and influence demand and supply. Moreover, additional instrumental variables can be searched and considered, which would result in an overidentified PLIV model. This can help to assess the robustness and sensitivity of the results by further analysis. Lastly, it could be explored how to inspect the effects of cross-elasticities and cannibalization, which are also presumed to influence demand but were mostly disregarded in this research.

Bibliography

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In *Handbook of labor economics*, volume 3, pages 1277–1366. Elsevier.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pages 507–547. University of Chicago Press.

Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2022). DoubleML – An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research*, 23(53):1–6.

Baron, S. and Lock, A. (1995). The challenges of scanner data. *Journal of the Operational Research Society*, 46(1):50–61.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2021). Difference-in-differences with a continuous treatment. *arXiv preprint arXiv:2107.02637*.

CBS (2022). Inflatie stijgt naar 14,5 procent in september.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation.

Chernozhukov, V., Hausman, J. A., and Newey, W. K. (2019). Demand analysis with many prices. Technical report, National Bureau of Economic Research.

Chu, J., Chintagunta, P., and Cebollada, J. (2008). Research note—a comparison of within-household price sensitivity across online and offline channels. *Marketing science*, 27(2):283–299.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Imbens, G. W. and Rosenbaum, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):109–126.

Lynch Jr., J. G. and Ariely, D. (2000). Wine online: Search costs affect competition on price, quality, and distribution. *Marketing science*, 19(1):83–103.

McFadden, D. (1986). The choice theory approach to market research. *Marketing science*, 5(4):275–297.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of business & economic statistics*, 13(2):151–161.

Morganosky, M. A. and Cude, B. J. (2000). Consumer response to online grocery shopping. *International Journal of Retail & Distribution Management.*

Natarajan, T., Balasubramanian, S. A., and Kasilingam, D. L. (2017). Understanding the intention to use mobile shopping applications and its influence on price sensitivity. *Journal of Retailing and Consumer Services*, 37:8–22.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pages 213–234.

Parkin, M. (2012). *Economics (10th Edition).* Addison-Wesley.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* Cambridge university press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pozzi, A. (2012). Shopping cost and brand exploration in online grocery. *American Economic Journal: Microeconomics*, 4(3):96–120.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Theil, H. (1971). Principles of econometrics john wiley and sons inc. *New York.*

Theil, H. (1975). *Economic forecasts and policy.* Contributions to economic analysis 15. North-Holland Publ. Co., Amsterdam, 2nd, rev. ed., 4th printing edition.

Tukey, J. W. et al. (1977). *Exploratory data analysis*, volume 2. Reading, MA.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.

Wright, P. G. (1928). *Tariff on animal and vegetable oils.* Macmillan Company, New York.

Appendix

IV/TSLS derivation of estimator

In order to proof how the validity condition $\mathbb{E}[Z'U] = 0$ leads to the IV estimator, firstly $l(X) = \mathbb{E}[Y|X]$ should rewritten as linear combination $l(X) = X\beta$. The $C \times 1$ parameter vector β can consequently be united with θ into parameter vector $\dot{\beta} = [\theta, \beta']'$. Analogously, P $(N \times 1)$ and X $(N \times C)$ are rewritten as $\dot{X} = [P, X]$. Using this notation, the structural model can be rewritten compactly as $Y = \dot{X}\dot{\beta} + U$. Employing the condition established above, the TSLS estimator for $\dot{\beta}$ can be derived as follows:

$$\begin{aligned} &\mathbb{E}[Z'U] = 0 \iff \mathbb{E}[Z'(Y - \theta P - l(X))] = \mathbb{E}[Z'(Y - \dot{X}\dot{\beta})] = 0 \\ &\iff \mathbb{E}[Z'Y] = \mathbb{E}[Z'\dot{X}\dot{\beta}] = \mathbb{E}[Z'\dot{X}]\dot{\beta} \iff \dot{\beta}_{IV} = (\mathbb{E}[Z'\dot{X}])^{-1}\mathbb{E}[Z'Y] \end{aligned}$$

provided $\mathbb{E}[Z'\dot{X}]$ is of full rank. Using the observational data for Y, Z and $\dot{X}$, $\hat{\beta}_{IV}$ can be obtained as:

$$\hat{\dot{\beta}}_{IV} = \mathbb{E}_N[Z'\dot{X}]^{-1}\mathbb{E}_N[Z'Y]$$

This result can also be obtained by first regressing the endogenous variable P on the instrument Z using OLS, yielding $\hat{P}$. In the second stage, structural equation (4.1) is estimated after replacing P with the estimated $\hat{P}$ to obtain estimate $\hat{\theta}_{IV}$.

As already discussed, the validity and relevance conditions are both important in estimating $\dot{\beta}$ and therefore θ. The relevance condition can be tested, for example by computing the F-statistic in the first stage regression. It is not possible to empirically test the validity condition due to the presence of the unobservable residual term U. As a result, acceptance of this condition must be based on established economic theory.

XGBoost regression algorithm

The XGBoosting algorithm consists of the following steps:

- Initialize the ensemble with a simple model $f_0(x)$, such as a decision tree with a single leaf node (a stump).;
- For each iteration $m = 1, 2, \ldots, M$:
  - Calculate $g = \left[\frac{\partial L(y,f)}{\partial f}\right]_{f=f_{m-1}}$

- Calculate $h = \left[\frac{\partial^2 L(y,f)}{\partial f^2}\right]_{f=f_{m-1}}$
- To build tree f_m, suppose that I_L and I_R are the sets of instances that belong to the left and right nodes, respectively, following a split in the data. The tree structure is determined by selecting splits that result in the greatest decrease in loss:

$$\mathcal{L}_{\text{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \zeta} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \zeta} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \zeta}\right] - \gamma$$

  In which ζ, γ are predetermined hyperparameters of the model.
- Determine the optimal weights w_j^* of leaf node $j \in (1, 2, ..., J))$ as $w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \zeta}$, where I_j denotes the set of tree predictions in leaf node j
- Complete the base learner $\hat{b}(x) = \sum_{j=1}^{J} w_j^* I_j$;
- Add trees $f_m(x) = \hat{b}(x) + \nu f_{m-1}(x)$

Steps 2-5 are repeated until convergence or a stopping criterion is met.

- Result: $f(x) = \sum_{m=0}^{M} \nu^{(M-m)} f_m(x)$

Moment conditions DML estimation

PLIV models

The first stage estimators are used to compose $\tilde{Y}, \tilde{P}$ and $\tilde{Z}$. In the second stage of the DML method, the following condition is exploited in order to estimate target parameter θ_0:

$$\mathbb{E}\left[\psi\left(W; \theta_0, \eta_0\right)\right] = 0, \tag{1}$$

with $W = (Y, P, X, Z)$ the data in the whole sample. As discussed in section 4.3.1, cross-fitting is used to avoid overfitting bias, hence within each fold the second stage is continued with $W_k = (Y_k, P_k, X_k, Z_k)$ denoting the data in the k^{th} subsample and $\eta_{0,k}$ the nuisance functions $(l_{0,k}, m_{0,k}, r_{0,k})$ trained on W_{-k}: the data *excluding* subsample k.

ψ Represents the score function associated with the partial linear model as follows:

$$\begin{aligned} \psi(W_k; \theta, \eta_k) :=&\ [Y_k - l_k(X) - \theta(P_k - r_k(X))][Z_k - m_k(X)] \\ =&\ -(P_k - r_k(X))(Z_k - m_k(X))\theta + (Y_k - l_k(X))(Z_k - m_k(X)) \\ =&\ \psi_a(W_k; \eta_k)\theta + \psi_b(W_k; \eta_k) \end{aligned} \tag{2}$$

with

$$\begin{aligned} \psi_a(W_k; \eta_k) &= -(P_k - r_k(X))(Z_k - m_k(X)) = -\tilde{P}_k \tilde{Z}_k \\ \psi_b(W_k; \eta_k) &= (Y_k - l_k(X))(Z_k - m_k(X)) = \tilde{Y}_k \tilde{Z}_k. \end{aligned} \tag{3}$$

The value of θ_0 can then be estimated using all K folds by solving for $\hat{\theta}$ in the empirical analog of equation (1):

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k} \psi\left(W_k; \hat{\theta}, \hat{\eta}_k\right) = 0, \tag{4}$$

Which yields the following estimator:

$$\hat{\theta} = -\frac{\sum_{k=1}^{K} \sum_{i \in I_K} \psi_b(W_k; \eta_k)}{\sum_{k=1}^{K} \sum_{i \in I_K} \psi_a(W_k; \eta_k)} = -\frac{\sum_{k=1}^{K} \sum_{i \in I_K} \sum_{i \in I_K} \tilde{Y}_k \tilde{Z}_k}{\sum_{k=1}^{K} \sum_{i \in I_K} \tilde{P}_k \tilde{Z}_k} \tag{5}$$

PLR models

The function ψ changes to:

$$\psi(W; \theta, \eta) = \ \psi_a(W; \eta)\theta + \psi_b(W; \eta) \tag{6}$$

with

$$\begin{aligned} \psi_a(W; \eta) &= -(P - r(X))(P - r(X)) \\ \psi_b(W; \eta) &= (Y - l(X))(P - r(X)). \end{aligned} \tag{7}$$

With $W = (Y, P, X)$.

Results gridsearch optimization of first stage models

Gridsearch Lasso models

Table 1: Lasso gridsearch optimization results

Model: l(X)	**Parameter**	**Grid searched:**	**Optimal value**
	Regularization term λ	[0,1]	0.2347
Model: r(X)	**Parameter**	**Grid searched:**	**Optimal value**
	Regularization term λ	[0,1]	0.0633
Model: m(X)	**Parameter**	**Grid searched:**	**Optimal value**
	Regularization term λ	[0,1]	0.0813

Gridsearch Random Forest models

Table 2: Random Forest gridsearch optimization results

Model: l(X)	**Parameter**	**Grid searched:**	**Optimal value**
	n_est	500	500
	max_depth	4, 6, 10, 13, 14, 15, 20, None	None
	min_samples_leaf	1, 2, 3	1
	min_samples_split	2, 4, 6	2
	max_features	0.5, 0.65, 0.7, 0.9, None	0.65

Model: r(X)	**Parameter**	**Grid searched:**	**Optimal value**
	n_est	500	500
	max_depth	15, 16, 17, None	None
	min_samples_leaf	1, 3	1
	min_samples_split	2, 4	2
	max_features	0.65, 0.8, 0.9, None	0.65

Model: m(X)	**Parameter**	**Grid searched:**	**Optimal value**
	n_est	500	500
	max_depth	14, 15, 16, 17, None	None
	min_samples_leaf	1, 3	1
	min_samples_split	2, 4	2
	max_features	0.65, 0.8, None	0.65

Gridsearch XGBoost models

Table 3: XGBoost gridsearch optimization results

Model: l(X)	**Parameter**	**Grid searched:**	**Optimal value**
	n_est	500	500
	learn_rate	0.25, 0.2, 0.15, 0.1, 0.05, 0.01	0.2
	reg_alpha	0,1	0
	max_depth	6, 8, 11, 15, 17, None	None
	col_sample_by_tree	0.55, 0.6, 0.7, 0.8, 1	1
	min_split_loss_gamma	0, 0.1	0
Model: r(X)	**Parameter**	**Grid searched:**	**Optimal value**
	n_est	500	500
	learn_rate	0.25, 0.2, 0.15, 0.1, 0.05, 0.01	0.2
	reg_lambda	0,1	1
	reg_alpha	0,1	0
	max_depth	6, 8, 11, 15, 17, None	None
	col_sample_by_tree	0.55, 0.6, 0.7, 0.8, 1	1
	min_split_loss_gamma	0, 0.1	0
Model: m(X)	**Parameter**	**Grid searched:**	**Optimal value**
	n_est	500	500
	learn_rate	0.25, 0.2, 0.15, 0.1, 0.05, 0.01	0.2
	reg_lambda	0,1	1
	reg_alpha	0,1	0
	max_depth	6, 8, 11, 15, 17, None	None
	col_sample_by_tree	0.55, 0.6, 0.7, 0.8, 1	1
	min_split_loss_gamma	0, 0.1	0

XGBoost model output without gridsearch optimization (default parameters)

Table 4: XGBoost with default parameters

Model	**PLR 3**	**PLIV 3**
θ Estimate	-0.7964	-0.3343
Std. error	0.03934	0.0571
P-value	0	0
CI lower 2,5%	-0.8735	-0.4463
CI upper 97,5%	-0.7193	-0.2224
Instrument used?	No	Yes
$l(X)$ method	XGBoost	XGBoost
RMSE	0.9117	0.9117
$r(X)$ method	XGBoost	XGBoost
RMSE	0.0124	0.0124
$m(X)$ method	XGBoost	XGBoost
RMSE		0.0115
N	50933	50933

Feature importances

Feature importances Lasso models

Table 5: Lasso feature importances (normalized)

Feature	$l(X) = Y$	$r(X) = P$	$m(X) = Z$
CONST	0.0	0.0	0.0
ARTICLE_TIER_BETTER	0.0	0.0	0.0
ARTICLE_TIER_GOOD	0.0	0.0716	0.0
ART_BRAND_TIER_PRICE_ENTRY	0.0	0.0	0.0
ART_BRAND_TIER_PRIVATE_LABEL	0.0	0.5164	0.4700
PACKAGING_BOX	0.0	0.0	0.0
PACKAGING_CAN	0.0	0.0	0.0
PACKAGING_PACK	0.0	0.0	0.0
ARTICLE_CAT_2$_DRINKPAKJES$	0.0	0.0	0.0
ARTICLE_CAT_2_FRUITDRANK	0.0	0.4023	0.5155
ARTICLE_CAT_2_IJSTHEE	0.0	0.0	0.0
ARTICLE_CAT_2_LIMONADE_SIROPEN	0.0	0.0	0.0
ARTICLE_CAT_2_SAPPEN_SMOOTHIES	0.0	0.0	0.0
ARTICLE_CAT_2_SINAS_LEMON_CASSIS	0.0	0.0	0.0
ARTICLE_CAT_2_SPECIAAL_FRIS	0.0	0.0	0.0
ARTICLE_CAT_2_SPORT_ENERGYDRINK	0.0	0.0	0.0
ARTICLE_CAT_2_WATER	0.0	0.0	0.0
PROMO_DUMMY	0.0	0.0	0.0
PPL2	0.0	0.0	0.0
PPL3	0.0	0.0	0.0
POSTPPL1	0.0	0.0	0.0
POSTPPL2	0.0	0.0	0.0
POSTPPL3	0.0	0.0	0.0
L_AVG_HIGH_TEMP	0.0	0.0	0.0
L_NR_ARTICLES_IN_CAT	0.0	0.0	0.0
L_NR_ARTICLES_IN_CAT_2	0.0	0.0	0.0
AVG_UNAVAILABILITY_PERC	0.9381	0.00427	0.0132351
ART_CONTENT_VOLUME	0.0213	0.000537	0.00110
ART_IS_MULTIPACK	0.0	0.0	0.0
HOLIDAY	0.0	0.0	0.0
WEEK_NR	0.0867	0.004951	0.000533

Feature importances Random Forest models

Table 6: Random Forest feature importances (normalized)

Feature	$l(X) = Y$	$r(X) = P$	$m(X) = Z$
CONST	0.0	0.0	0.0
ARTICLE_TIER_BETTER	0.04762	0.05519	0.0544807
ARTICLE_TIER_GOOD	0.04639	0.0733	0.02606962
ART_BRAND_TIER_PRICE_ENTRY	0.0006798	0.001913	0.0003108
ART_BRAND_TIER_PRIVATE_LABEL	0.00992	0.1682372	0.1795753
PACKAGING_BOX	0.0014775	0.001942	0.002997075
PACKAGING_CAN	0.0214900	0.0582702	0.04446342
PACKAGING_PACK	0.01850	0.0356419	0.0546903
ARTICLE_CAT_2_DRINKPAKJES	0.0174264	0.01851307	0.01630322
ARTICLE_CAT_2_FRUITDRANK	0.00297	0.0396530	0.06387224
ARTICLE_CAT_2_IJSTHEE	0.0019631	0.00010374	0.0002670
ARTICLE_CAT_2_LIMONADE_SIROPEN	0.003360933	0.0152444	0.0199395
ARTICLE_CAT_2_SAPPEN_SMOOTHIES	0.005050	0.00864164	0.00885226
ARTICLE_CAT_2_SINAS_LEMON_CASSIS	0.002985	0.0026987	0.00174594
ARTICLE_CAT_2_SPECIAAL_FRIS	0.00842	0.0058834	0.0108287279
ARTICLE_CAT_2_SPORT_ENERGYDRINK	0.0031841	0.01129414	0.008554370
ARTICLE_CAT_2_WATER	0.039812	0.065531354	0.08764689
PROMO_DUMMY	0.024025	0.0003326	0.0003072
PPL2	0.01568	0.000590	0.0002723985
PPL3	0.0152603	0.0005831	0.00027849
POSTPPL1	0.020578	0.0011149	0.000643457
POSTPPL2	0.023132	0.0011731	0.00064390
POSTPPL3	0.023628	0.0011687	0.00064802
L_AVG_HIGH_TEMP	0.0722102	0.008299140	0.0081308
L_NR_ARTICLES_IN_CAT	0.053682	.00687670	0.00629895
L_NR_ARTICLES_IN_CAT_2	0.049992	.0367651	0.05255
AVG_UNAVAILABILITY_PERC	0.2017838	0.033337	0.02592656
ART_CONTENT_VOLUME	0.16978	0.2190998	0.178020
ART_IS_MULTIPACK	0.004370	0.11050	0.133200038
HOLIDAY	0.00800	0.000800	0.00072730
WEEK_NR	0.086596	0.017260	0.0117508959

Feature importances XGBoost models

Table 7: XGBoost feature importances (normalized)

Feature	$l(X) = Y$	$r(X) = P$	$m(X) = Z$
CONST	0.0	0.0	0.0
ARTICLE_TIER_BETTER	0.0497773	0.010276	0.011670
ARTICLE_TIER_GOOD	0.11431	0.0181204	0.0114849
ART_BRAND_TIER_PRICE_ENTRY	0.0075874	0.0005502	5.78784e-05
ART_BRAND_TIER_PRIVATE_LABEL	0.0414079	0.23011264	0.2079840
PACKAGING_BOX	0.0299669	0.012536	0.0009230
PACKAGING_CAN	0.064874	0.075760	0.0371421
PACKAGING_PACK	0.0125887	0.0241634	0.03991644
ARTICLE_CAT_2_DRINKPAKJES	0.081986	0.000309	0.00021630
ARTICLE_CAT_2_FRUITDRANK	0.0090868	0.009712	0.01133683
ARTICLE_CAT_2_IJSTHEE	0.041593	0.0001465	0.00094759
ARTICLE_CAT_2_LIMONADE_SIROPEN	0.0132107	0.013391	0.02292507
ARTICLE_CAT_2_SAPPEN_SMOOTHIES	0.024878	0.0038713	0.00160642
ARTICLE_CAT_2_SINAS_LEMON_CASSIS	0.0106314	0.001146	0.000529495
ARTICLE_CAT_2_SPECIAAL_FRIS	0.0131550	0.0019449	0.001132739
ARTICLE_CAT_2_SPORT_ENERGYDRINK	0.0355541	0.124151	0.076818
ARTICLE_CAT_2_WATER	0.20963	0.08847494	0.04980224
PROMO_DUMMY	0.049643	0.000342	0.000321883
PPL2	0.0055468	7.94948e-05	2.728493e-05
PPL3	0.005341	4.707167e-05	3.823552e-05
POSTPPL1	0.0059404	6.180422e-05	2.171282e-05
POSTPPL2	0.0077782	0.00010056	5.821774e-05
POSTPPL3	0.007535	7.44785e-05	2.403623e-05
L_AVG_HIGH_TEMP	0.0061247	0.0003105	0.0002748857
L_NR_ARTICLES_IN_CAT	0.0073477	0.0004207	0.00045223
L_NR_ARTICLES_IN_CAT_2	0.0188008	0.003925	0.0129851
AVG_UNAVAILABILITY_PERC	0.008145	0.0005388	0.000460
ART_CONTENT_VOLUME	0.0794	0.04032928	0.01868
ART_IS_MULTIPACK	0.0145417	0.337280	0.4911301
HOLIDAY	0.0124044	0.00032986	0.00034686
WEEK_NR	0.011199	0.0014903161	0.000676

Subset analysis with model trained on complete dataset

The subset analysis is also executed using the first stage Random Forest an XGBoost models trained on the whole dataset instead of the subset data only. This gave the following results:

Table 8: DML θ results on data subsets (with model trained on complete dataset)

Category	**N**	**PLR 2**	**PLR 3**	**PLIV 2**	**PLIV 3**
Fruit drinks	12313	-0.3539***	-0.3744***	0.4653***	0.4831***
CI		*(-0.50, -0.21)*	*(-0.54, -0.21)*	*(0.19, 0.74)*	*(0.14, 0.82)*
Cola	7098	0.8055***	0.9966***	5.9370***	5.9908***
CI		*(0.39, 1.22)*	*(0.53, 1.46)*	*(5.14, 6.73)*	*(4.91, 7.07)*
Lemonade & syrups	6951	-2.1941	-1.7314***	-2.3051***	-2.5547***
CI		*(-2.43, -1.96)*	*(-2.02, -1.44)*	*(-2.79, -1.82)*	*(-3.10, -2.01)*
Water	3576	-0.3806	0.8795**	-2.6758***	-0.4736
CI		*(-0.96, 0.20)*	*(0.19, 1.57)*	*(-3.56, -1.79)*	*(-1.83, 0.89)*
Special soda	7979	-2.2545***	-1.4491***	-1.7115***	-1.3547***
CI		*(-2.55, -1.96)*	*(-1.81, -1.09)*	*(-2.15, -1.27)*	*(-1.98, -0.73)*
Juices & smoothies	9975	-1.7534***	-1.4602***	-2.1480***	-2.0189***
CI		*(-1.96, -1.54)*	*(-1.69, -1.23)*	*(-2.40, -1.90)*	*(-2.33, -1.71)*
Small cartons	6649	-0.4685**	0.2038	4.6372***	2.1885***
CI		*(-0.85, -0.09)*	*(-0.29, 0.70)*	*(3.13, 6.15)*	*(0.59, 3.79)*
Orange, lemon & cassis	4412	-0.8250***	0.1047	2.0694***	1.8233***
CI		*(-1.40, -0.25)*	*(-0.48, 0.69)*	*(0.92, 3.22)*	*(0.47, 3.18)*
Ice tea	1073	-5.0455***	-5.1726***	69.4982	19.5462
CI		*(-7.08, -3.01)*	*(7.75, -2.59)*	*(-150.34, 289.34)*	*(-48.25, 87.34)*
Sport- & energy-drinks	1093	-0.1180	-0.8043	-126.0950	5.9894
CI		*(-1.78, 1.54)*	*(-3.27, 1.66)*	*(-1674.79, 1422.60)*	*(-14.05, 26.03)*

* Significant at 10% level
** Significant at 5% level
*** Significant at 1% level

Table 8 shows the results of the four model θ estimates. The elasticity estimates within each subcategory are diverging and seem to show no consensus. For exampl,e the PLR models give a negative coefficient estimate for “Fruit drinks”, whereas the PLIV model yields positive θ estimates for this category. In general, it is expected that a price elasticity estimate is negative: a price change is usually reciprocated with a response of demand in the opposite direction of the associated price change. The “Cola” θ estimates are positive for all four models, which should most likely not be interpreted to be equal to the true price elasticity of demand. Yet, the estimate may be interpreted to illustrate the demand for articles belonging to this category is not very elastic.

Comparing the PLR 2 and PLR 3 elasticities comparatively, the estimates for the “Cola” subcategory are both positive and larger than 0.5; the “Water” and “Small cartons” categories show one negative and one positive coefficient within $(-1, 1)$; the “Fruit drinks”, “Orange, lemon & cassis” and “Sport & energydrink” estimates are both contained in $(0, -1)$; estimates for the “Lemonade & syrups”, “Special soda” and “Juices & smooties” categories are both contained in $(-1, -3)$ and finally the “Ice tea” estimates are both smaller than -5.

When comparing the relative PLIV 2 and PLIV 3 elasticity estimates in the same manner, it is found that the estimates for “Ice tea” are both very large: 69.5 for PLIV 2 and 19.55 for PLIV 3. Moreover, it can be seen that again both “Cola” estimates are positive and in this case both within $(5.9, 6)$. Both estimates for “Small cartons” and “Orange, lemon & cassis” are also positive, in this case within $(1.5, 5)$. Both “Fruit drinks” estimates are positive and contained in $(0.4, 0.5)$. For the “Lemonade & syrups”, “Water”; ‘Special soda” and “Juices & smooties” categories, both estimates are contained in $(-0.5, -2.6)$. Lastly, the “Sport & energydrink” elasticity estimates are conflicting: PLIV 2 yields an estimate of -126.10 and PLIV 3 yields an estimate of 5.99.

The aforementioned relative groupings of product categories in terms of estimated coefficients offer some insight in comparable price elasticity; nevertheless, the PLR and PLIV methods do not exhibit complete alignment. Despite this, all methods agree on some ordering: the “Cola” articles show the largest positive elasticity estimates for all models; all methods show negative elasticities for “Lemonade & syrups”, “Special soda” and “Juices & smoothies” that are between $(-1, -3)$.

The two categories that contain observations of four products only, “Ice Tea” and “Sport & energydrink”, show some very large PLIV elasticity estimates, in absolute terms. $\theta_{\text{Ice Tea}}$ Is estimated as 69.50 using PLIV 2 and as 19.55 in PLIV 3; $\theta_{\text{Sport- \& energydrinks}}$ is estimated as -126.10 using PLIV 2 and 5.99 using PLIV 3. These four estimates are not significant, and differ a lot from the PLR 2 and PLR 3 estimates for those categories, which are around -5 for “Ice tea” and withitn $(-1, 0)$ for “Sport- & energydrinks”. These unexpected results are presumably due to the fact that these categories possess the least SKU’s and observations.

Generally, the results show that within the subcategories, significance does not offer unanimity: for example, within the “Fruit drink”; “Water”; “Small cartons” and “Orange, lemon & cassis” categories, in terms of sign, diverging θ estimates are given significance, using any of the 10, 5 and 1% significance levels. On the other hand, the significant estimates for “Cola”, “Lemonade & syrups”, “Special soda”, “Juices & smoothies” and “Ice tea” are consistent in terms of sign.

Model output with lagged demand included in covariate matrix X

Following next are the results of the model estimated with lagged order quantities included in X. Lags of one and two weeks were used.

DML-PLIV Models including instrument and lagged demand

Table 9: Results naive and DML-PLIV estimation methods

Model	**OLS**	**IV**	**PLIV 1**	**PLIV 2**	**PLIV 3**
θ estimate	-0.1242	-0.1160	-0.1705	0.0114	0.0177
Std. error	0.0136	0.0186	0.0063	0.0302	0.0311
p-value	0	0	0	0.7055	0.5692
CI lower 2,5%	-0.1508	-0.1524	-0.1828	-0.04781	-0.0432
CI upper 97,5%	-0.0976	-0.0796	-0.1582	0.0706	0.0786
Instrument used?	No	Yes	Yes	Yes	Yes
$l()$ method			Lasso	RForest	XGBoost
RMSE			0.4139	0.1891	0.1620
$r()$ method			Lasso	RForest	XGBoost
RMSE			0.2334	0.0091	0.0095
$m()$ method			Lasso	RForest	XGBoost
RMSE			0.1736	0.0076	0.0084

Table 9 presents the outcomes of the DML estimation of the partial linear IV (PLIV) models. PLIV 1 indicates the model with Lasso first stage estimation, PLIV 2 refers to the Random Forest regression for the first stage and PLIV 3 refers to a first stage XGBoost regression. The findings indicate that the elasticity is significant across the three linear models applying OLS, TSLS and PLIV 1. Specifically, the Lasso estimate displays the highest absolute value coefficient of -0.1705 and the smallest standard error. The coefficients of OLS and IV regression are slightly smaller at -0.1242 and -0.116, respectively. In contrast, PLIV models 2 and 3, which employ more sophisticated tree-based ML methods in the first stage, exhibit conflicting estimation results. The associated coefficients, 0.0114 and 0.0177 respectively, are close to zero and positive. The insignificance of these estimates highlights the need for an examination of the reasons behind these discrepant results.

Also depicted in table 9 are prediction errors for the first stage estimations of $l(X)$, $m(X)$ and $r(X)$. The cross-fitting element of the Double ML method requires each model to implement K = 5 estimations of the first stage models. As a reference, in order to evaluate the quality of these first stage models, prediction errors are computed based on the test set left out of estimation. The resulting reference RMSE's show that the tree based methods perform better for all three nuisance parameters than Lasso regression. Overall, the XGBoost algorithm performs best in predicting all Y, P and Z. It should be noted however that the Random Forest Regression first stage prediction metrics are quite similar, which could explain the final θ coefficients of PLIV models 2 and 3 being comparable as well.

DML-PLR Models excluding instrument and lagged demand

Table 10: Results naive and DML-PLR estimation methods

Model	OLS	TSLS	PLR 1	PLR 2	PLR 3
θ estimate	-0.1242	-0.1160	-0.3011	-0.0409	-0.0407
Std. error	0.0136	0.0186	0.0086	0.0120	0.0182
p-value	0	0	0	0.0401	0.0259
CI lower 2,5%	-0.1508	-0.1524	-0.3178	-0.0801	-0.0764
CI upper 97,5%	-0.0976	-0.0796	-0.2842	-0.0018	-0.0049
Instrument used?	No	Yes	No	No	No
$l()$ method			Lasso	RForest	XGBoost
RMSE			0.4139	0.1891	0.1620
$r()$ method			Lasso	RForest	XGBoost
RMSE			0.2334	0.0091	0.0095

In table 10 the results from estimating the partial linear regression models using Lasso, Random Forest and XGBoost regression in the first stage and employing OLS in the second stage are depicted, yielding models PLR 1, 2 and 3 respectively. The OLS and TSLS remain unchanged. The PLR models yield larger elasticity coefficients (in absolute terms) than those resulting from the PLIV models. The PLR 1 elasticity estimate is -0.3011, which is substantially higher than the PLIV 1 estimate of -0.1705, while both of them display definitive significance. PLR models 2 and 3 produced significant estimates of -0.0409 and -0.0407 respectively at a 5% significance level. While these two estimates are similar to each other, they show a considerable difference when compared to the PLR 1 model output. This discrepancy is not expected while applying Double ML and will be investigated further. The models for $l(X)$ and $r(X)$ are the same w.r.t. the PLIV models.

In both the DML-PLIV and DML-PLR models, the choice of first stage is recognized to affect the final result for θ. Specifically, switching from a Lasso first stage estimation to Random Forest or XGBoost regression produces substantial differences in the final result. The latter models perform better in predicting Y, P, and possibly Z, suggesting a non-linear relationship between these variables and X. Using more sophisticated first stage models could improve the model by adequately filtering out the effect of X in the first stage. However, using potentially complex models for residualizing the variables could also result in the orthogonalized variables having little to no variation left, which might explain why the results of these models differ significantly. In other words, the effect of price on product demand might unintentionally be captured by the first stage model. For example, in PLIV models 2 and 3, the insignificance of the price elasticity estimates might be a result of this. While PLR models 2 and 3 may exhibit statistical significance, their coefficients are still notably small, which raises concerns regarding their validity. Further analysis will be done to determine which first stage model is complex enough without filtering out the relationship of interest.

Residual analysis of model including lagged demand

Using training data

Even though the Double ML approach is not designed in order to do predictions, residual analysis can be helpful in order to judge the estimated coefficients more thoroughly. Predictions of $Y = \log(\text{demand})$ are made according to equation 4.1 and the estimated θ values depicted in tables 9 and 10, using the same data set on which the coefficient is established.

Table 11: Correlation between Y and associated prediction errors based on the train set

Corr	PLIV 1	PLIV 2	PLIV 3	PLR 1	PLR 2	PLR 3
Y	0.6566	0.3427	0.2542	0.6686	0.2951	0.2450

It can be seen from Table 11 that the Lasso-based DML models have prediction errors that correlate substantially with the observed Y values: the correlation coefficients are 0.6566 and 0.6686 for PLIV 1 and PLR 1 respectively. This correlation indicates that the Lasso models are not adequately capturing some of the underlying patterns in the data. More specifically, this result suggests Y might have a nonlinear distribution, which the tree-based models seem to capture more appropriately. Models using either the Random Forest or XGBoost regression show smaller correlation coefficients for Y and their associated prediction residuals. Models using the XGBoost method shows the lowest coefficients, namely 0.2542 for the PLIV model and 0.2450 for the PLR model.

Using test data originating from the same time period

Additionally, out-of-sample predictions could offer further understanding. It is important to note that in order to accurately assess the demand predictions based on estimated elasticities, it must be assumed that the elasticity remained constant between the estimation (training) period and the prediction (test) period. Thus, reliable predictions can only be made if the out-of-sample period is from the same time period as the training sample. For this study, a test set was created by randomly withholding 25% of the sample data before estimation, which satisfies the constant elasticity assumption. The correlation between the resulting prediction errors and observations are denoted below:

Table 12: Correlation between Y and associated prediction errors based on the test set

Corr	PLIV 1	PLIV 2	PLIV 3	PLR 1	PLR 2	PLR 3
Y	0.6558	0.3488	0.3037	0.6671	0.3138	0.2844

In both PLIV and PLR cases, the Lasso-based prediction residuals correlate similarly with Y, that is, correlation coefficients of 0.6558 and 0.6671 respectively. The Random Forest- and XGBoost-based models show quite an improvement in this by approximately halving the size of these coefficients. The XGBoost-based models show the lowest correlation between Y and the prediction residuals: 0.3037 and 0.2844 for PLIV and PLR respectively. There seem to be no considerable differences between the PLIV and PLR framework when comparing them for each first-stage. Finally, the correlation coefficients in table 12 are similar to those in table 11. Especially the Lasso-based prediction residuals seem to stay constant, while the tree-based predictions show a slight increase in correlation. To conclude, these out-of-sample predictions add no further insight on model performance and/or validity.

Using test data originating from a later time period

In order to test the assumption whether price elasticity is indeed constant over time, additional residual analysis is done for a test sample based on observations successive in time to the training sample. Concretely, a data set composed of the first two months of 2023 is used to test the model performance. This yields the following results:

Table 13: Correlation between Y and associated prediction errors based on the 2023 test set

Corr	PLIV 1	PLIV 2	PLIV 3	PLR 1	PLR 2	PLR 3
Y	0.6919	0.6035	0.5225	0.7331	0.5962	0.5541

Table 13 shows that the correlation between the Lasso-based prediction residuals and observations increases slightly to 0.6919 and 0.7331 for PLIV and PLR respectively. For PLIV models 2 and 3 the correlation increases substantially: from 0.3488 to 0.6035 for PLIV 2 and from 0.3037 to 0.5225 for PLIV 3. The PLR models show similar increases: PLR 2 correlation increases from 0.3138 to 0.5962 and the correlation coefficient associated with PLR 3's prediction residual goes up from 0.2844 to 0.5541. As this increase in correlation occurs for the tree-based models, it might be suggested that these models tend to over-fit on train data. However, another explanation would be that the assumption of constant price elasticity would hold: the price elasticity estimates of the training sample would then not fit the 2023 data, leading to biased predictions. It could be hypothesised that the correlation coefficient associated with the Lasso-based models would in this case not increase because the associated price elasticity estimates were biased to begin with. This hypothesis would also explain the initial higher residual correlation coefficients of Lasso-based models depicted in tables 11 and 12. Ultimately, it is hard to state whether the increase in 2023 residual correlation for the tree-based models is due to the fact that these models are simply not suitable for prediction, or whether this is due to the fact that the assumption of constant price elasticity can be presumed, leading to incorrect predictions for the 2023 data.

Estimation details

Estimation is done in Python, using the Scikit-learn module (Pedregosa et al., 2011) for Lasso regression, Random Forest regression, and gridsearch of the first-stage model hyperparameters. A separate module is used for XGBoost (Chen and Guestrin, 2016). For the OLS and IV regressions, the linearmodels (release 4.27) package from Kevin Sheppard et al. is used. The DoubleML module by Bach et al. (2022) is used for Double ML estimation.