Demand estimation with double/debiased machine learning: a comparison to traditional methods

Erik Kaunismäki

Department of Finance and Economics

Hanken School of Economics

Helsinki

2021

HANKEN SCHOOL OF ECONOMICS

Department of: Department of Finance and Economics	**Type of work:** Master's thesis
Author and student number: Erik Kaunismäki, 144570	**Date:** 27.07.2021

Title of thesis:
Demand estimation with double/debiased machine learning: a comparison to traditional methods

Abstract: The thesis explores how recent advances in econometric methodology can be used for estimating demand and price elasticities of demand. Traditional econometric methods perform poorly when the number of variables is large proportional to the number of observations, usually leaving variable selection on the researcher's judgement. New advances, incorporating machine learning methods in econometric methods, provide a data driven variable selection procedure and are able to deal with sparse data sets.

Using a data set with rich product descriptions from a Finnish retail firm, the double machine learning (DML) methodology by Chernozhukov et al. (2018) is used to estimate the own price elasticity of demand. The results are furthermore compared to the estimates from traditional econometric methods. The formulation of the demand model follows as closely as possible the one laid out in Chernozhukov et al. (2017), where DML is used to estimate price elasticity of everyday retail goods.

Interestingly, the estimates of price elasticity obtained with DML differ little to those of traditional methods, which is similar to the findings in Chernozhukov et al. (2018). The results also highlight the importance of choosing the correct method as the first stage estimator and configuring it properly. Furthermore, only a few of the product characteristics are crucial for demand estimation. It is unclear whether this is due to product characteristics actually not affecting price elasticity or due to potential biases in how product characteristics are recorded.

Keywords: Demand estimation, own price elasticity of demand, machine learning, double machine learning

SVENSKA HANDELSHÖGSKOLAN

Institution: Institutionen för finansiell ekonomi och nationalekonomi	**Arbetets art:** Magisteravhandling
Författare och studerandenummer: Erik Kaunismäki, 144570	**Datum:** 27.07.2021
Avhandlingens rubrik: Demand estimation with double/debiased machine learning: a comparison to traditional methods	

Sammandrag: I min avhandling undersöker jag hur nya ekonometriska metoder kan användas i estimering av efterfråga och estimering av produkters priselasticitet. Traditionella metoder presterar dåligt då antalet variabler är relativt sett stort till antalet observationer. Det här leder vanligtvis till att forskaren på basen av sitt omdöme väljer vilka variabler hen inkluderar. Nya metoder som inkorporerar maskininlärningsmetoder med traditionella metoder erbjuder en datadriven lösning till val av variabler och ett sätt att hantera s.k. glest data.

Jag använder den nya estimeringsmetoden "double machine learning" (DML) av Chernozhukov et. al (2018) för att estimera produkters priselasticitet. Det här gör jag med hjälp av data med detaljerade produktbeskrivningar från ett finskt företag inom detaljhandeln. Utöver det jämförs resultaten från den nya metoden med resultat från traditionella ekonometriska metoder. Själva modellen för efterfråga följer så nära som möjligt den som beskrivs i Chernozhukov et al. (2017), som också använder DML för att estimera priselasticitet inom detaljhandeln.

Intressant nog, är resultaten från DML-metoden jämfört med de traditionella metoderna mycket lika, vilket stöder användningen av DML i estimering av efterfrågemodeller. Resultaten visar även vikten av att välja rätt metod i första stadiets estimering, samt att konfigurera den rätt. Få av variablerna som beskriver produkterna visade sig vara betydelsefulla. Dock är det oklart om detta beror på att produkternas egenskap de facto inte har stor betydelse i estimering av priselasticitet eller ifall variablerna är snedvridna.

Nyckelord: Estimering av efterfråga, priselasticitet, maskininlärning, double machine learning

CONTENTS

1 Introduction 1
2 Models of demand and Machine Learning 3
2.1 Price elasticity 3
2.2 Identification of demand models 4
2.3 Product space demand models 6
2.4 Characteristics space demand models 7
2.4.1 Discrete choice models 9
2.4.1.1 Logit and nested logit models 9
2.4.1.2 BLP models 11
2.5 Recent developments in characteristics space models 12
2.6 Machine Learning 13
2.6.1 Supervised Machine Learning 13
2.6.2 Machine Learning in economics 14
3 Data 16
3.1 Aggregation and structure of the data 16
3.2 Descriptive statistics & variables 18
3.3 Accounting for seasonality 21
4 Empiciral framework 24
4.1 Specification of demand models 24
4.2 Instrumental variables (IV) 25
4.3 Double/Debiased machine learning (DML) 27
4.3.1 Frisch – Waugh – Lovell theorem 27
4.3.2 First stage estimators 28
4.3.2.1 Least absolute shrinkage and selection operator (Lasso) 29
4.3.2.2 Random Forest 30
4.3.2.3 AdaBoost 32
4.3.3 Second stage estimators 33
4.4 Summary of methods 33
5 Results 35
5.1 Demand estimation results 35
5.2 DML first stage estimators and auxiliary analysis 37
6 Discussion 42
7 Conclusions 44

APPENDICES

Appendix 1 Red-bus-blue-blue example....50
Appendix 2 Descriptive statistics....52
Appendix 3 Description of Machine Learning algorithms....55
Appendix 4 Additional Results....57

TABLES

Table 1 Descriptive statistics of prices and sales....19
Table 2 Description of variables....20
Table 3 Estimates of price elasticity with the linear model....35
Table 4 Estimates of price elasticity with the partially linear model....36
Table 5 Estimates of price elasticity with the linear model without an instrument .37
Table 6 Estimates of price elasticity with the partially linear model without an instrument 37
Table 7 Descriptive statistics of all variables....52
Table 8 Second stage DML estimation with Lasso....57
Table 9 Second stage DML estimation with Random Forest....58
Table 10 Second stage DML estimation with AdaBoost....59
Table 11 Second stage DML estimation with Lasso....60
Table 12 Second stage DML estimation with Random Forest....60
Table 13 Second stage DML estimation with AdaBoost....61
Table 14 Estimation of the linear model with instrument, IV....62
Table 15 Estimation of the linear model without instrument, OLS....64

FIGURES

Figure 1 Example of price elasticity when (not) controlling for confounders....5
Figure 2 Directed acyclic graph illustrating instrumental variables....6
Figure 3 Example of the data structure of a product hierarchy....17
Figure 4 The logarithm of sold quantity and price....21
Figure 5 Power Spectral Density and modelled seasonality....22
Figure 6 Directed acyclic graph of the partially linear model....25
Figure 7 First stage of DML....29
Figure 8 Illustrative example of a simple regression tree....31
Figure 9 The second stage of DML....33
Figure 10 Correlation plot of the residualized variables and the real variables....38
Figure 11 The residualized variables for price and product cost....39
Figure 12 Price elasticity by number of folds....40
Figure 13 Variable importance when predicting sales, price and product cost....41

1 INTRODUCTION

When a consumer makes a decision to purchase a good, there are several factors affecting this decision. Previously, traditional econometric methods have not been able to model this interaction with high-dimensional, large datasets. For example, the inclusion of rich product descriptions has been problematic. It has been left to the researchers' judgement to choose the relevant variables and the left-out variables are typically modelled as unobserved. For this reason, being able to model demand with more detail, using high-dimensional data sets and machine learning (ML) methods, can yield in more precise results and a better understanding of economic decisions. This is not only an interesting development in itself, but also has implications to competition policy and to the behaviour of price optimizing firms.

The purpose of this thesis is to estimate the own price elasticities of demand using a method developed by Chernozhukov et al. (2018), called double/debiased machine learning (DML). I compare the estimates to the results from traditional econometric methods, to assess the benefits and drawbacks of DML. Using novel data from a Finnish retail firm, I restrict the estimation to a product group called "shirts" and include several variables related to product characteristics and previous realisations of the demand system. In short, the research aims to:

- estimate the own price elasticity of demand using double machine learning (DML).
- compare the estimates of DML to those obtained by a standard instrumental variables approach.

The main motivator for applying the DML framework is that it is one of the first methods to leverage the predictive power of ML methods while yielding estimates that can potentially be interpreted with causal meaning. Traditionally, ML methods are tuned to maximise out-of-sample predictive performance. This tuning entails a trade-off between what is called bias and variance and implies that the estimates cannot be interpreted as measures of causal effects. However, using Neyman-orthogonal moments and cross-fitting, Chernozhukov et al. (2018) were able to show that estimates from ML methods can essentially be debiased and therefore used to estimate causal parameters.

The results show, that estimating demand with a high dimensional data set, using the DML method, did not result in much different results than those of traditional

econometric methods. Variations of DML with different ML methods resulted in similar conclusions, which is also the case in Chernozhukov et al. (2018). However, some ML methods have a persisting high variance which lead to suboptimal results.

The thesis has the following structure: the second chapter summarizes the theoretical background of demand and demand estimation, starting from the very basics to more advanced formulations. The third chapter presents the data, after which the specification of the demand model and methods of estimation are presented in chapter four. The results of the estimations are presented in chapter five followed by a discussion and conclusions in chapter six and seven respectively.

2 MODELS OF DEMAND AND MACHINE LEARNING

In this chapter, I will give an overview of the theoretical frameworks of demand. Starting from the concept of price elasticity I continue by presenting product and characteristic space demand models. This is followed by a review of the most recent literature. Lastly, I will go through the theoretical framework for machine learning and how it has been applied in the economic literature. The area is wide so emphasis will be placed on subjects that are more relevant for this thesis.

2.1 Price elasticity

By considering a simple model of a market of one good, we can create a good understanding of price elasticity. Reservation price is referred to as a person's maximum willingness to pay for a good, and the assumption is that as we decrease the price for this good, there will be more people with the reservation price greater than or equal to the price of the good. As the price decreases more people are willing to buy the product and thus, demand increases. Plotting together all the possible combinations of price and quantity we get the demand curve which, by the reasoning just made, has a negative slope. (Varian, 2010)

The individual demand curves can be added to obtain the aggregate demand curve. Consider a market of two goods: if a consumer i has demand $x_i^1(p_1, p_2, m_i)$ for good 1 and demand $x_i^2(p_1, p_2, m_i)$ for good 2 – which both depend on prices for the respective goods (p_1, p_2) and i's income m – then the aggregate demand for e.g. good 1 is the sum of all individual demands:

$$X^1(p_1, p_2, m_1 \dots m_n) = \sum_{i=1}^{n} x_i^1(p_1, p_2, m_i).$$

The aggregate demand now depends on the prices of both goods and the distribution of income. As the individual demand curves have a negative slope, so will the aggregate. A convenient and unitless way of comparing aggregated demands for different goods is by looking at how responsive the demand is to changes in price. This is the own price elasticity of demand, usually noted as ϵ and mostly referred as just price elasticity. The price elasticity can be expressed as the slope of the demand function, in terms of derivatives or approximated with logarithms:

$$\epsilon \equiv \frac{\Delta q/q}{\Delta p/p} \quad or \ \epsilon = \frac{p}{q}\frac{dq}{dp} \quad or \ \epsilon = \frac{d\log(q)}{d\log\ (p)}$$

The price elasticity of demand has several profound properties. First of all, it gives the price setter an idea about the shape of the demand curve and how she can set a price to maximise profits. Furthermore, we can say something about the welfare allocation in markets and in monopoly contexts we can even derive precisely the optimal pricing policy for a firm. (Varian, 2010)

2.2 Identification of demand models

Price endogeneity and identification of demand systems is such a notorious challenge that it is usually covered as a separate chapter in e.g., Train (2009), Ackerberg et al. (2007). Papers such as Berry & Haile (2016), Gandhi & Houde (2019) continue to develop the field. The first study of this problem is usually attributed to Wright (1928) who studied the effect of tariffs on animal factors and vegetable oils. His main argument (laid out in appendix B of his work) is that by simply looking at historical data points of prices and quantities, one cannot infer anything about the shape nor form of the demand curve.

This is graphically demonstrated in figure 1. If we wish to estimate the price elasticity of demand, we need to be assured that the demand curve remains fixed. Using the illustrative example depicted in figure 1, we can see that drawing a best-fit line in the scatter plot of price and quantity observations in the left graph, yields a negative but small price elasticity. On the other hand, if we introduced a dummy variable in the manner that is done in the right graph, we would conclude a larger price elasticity. For purposes of illustration, the dummy variable can be thought as differentiating observations for different seasons. As noted in Train (2009), there are many factors that can lead to the endogeneity of explanatory variables, such as unobserved product attributes that correlate with observed attributes, firms' marketing efforts, dependency of other choices and so forth.

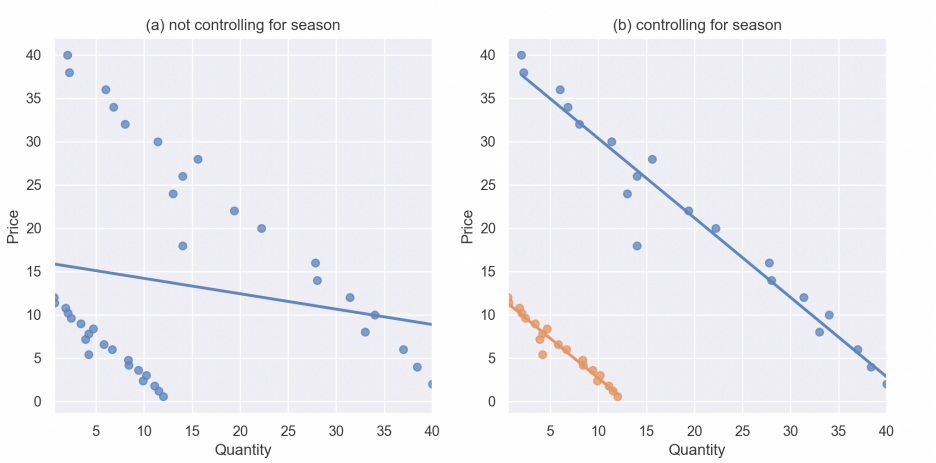

Figure 1 **Example of price elasticity when (not) controlling for confounders.** Picture inspired by Mackey, Syrgkanis & Zadik (2018)

The same problem can be spelled out in more technical terms with a log linear demand function. In this econometric model log quantity $\log(Q)$, is explained by a constant α, the price elasticity of demand δ, and the seasonal dummy variable, X. γ is the coefficient for the seasonal dummy. Lastly, u is the stochastic error:

$$\log(Q) = \alpha + \delta \log(p) + \gamma X + u$$

To consistently estimate this linear model, we need the error term to be zero in expectation conditional on all regressors, $\mathrm{E}[u|p, X] = 0$. If this requirement is met, then we could get a reliable estimate of δ which could be interpreted as the price elasticity of demand. Omitting a relevant variable, as in the left graph yields biased estimates. This requirement is nonetheless not trivial to meet. Furthermore, if there are some factors affecting demand that are unobserved, or that cannot that be quantified, our estimators remain biased. (Cunningham, 2018)

The resulting approach that rose from Wright (1928) is called instrumental variables. The idea is to find a variable z (called the instrument), that affects the endogenous variable w, which in turn affects the dependent variable y which we are interested in explaining. These relations are graphically shown in figure 2. Importantly, the instrument cannot affect the dependent variable directly, only via the variable w. As mentioned earlier cost side variables, have been very common instruments in demand equations. A cost shifter is a valid instrument if we assume that there is an exogenous factor increasing costs, which in turn increases prices and therefore affects demand indirectly. (Imbens, 2004)

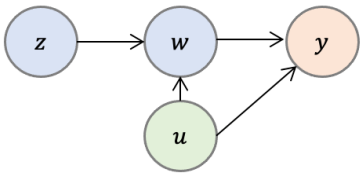

Figure 2 Directed acyclic graph illustrating instrumental variables

Further chapters introduce different models for demand, and they all deal with price endogeneity in some way as we will see. For example, Hausmann (1996) uses the price of the same product on a different geographical market. The assumption is that the price of a product is correlated across different markets. Therefore, the difference between average price and actual price in local markets should be driven by differences in local demand, making it a valid instrument. So called BLP instruments rely on the idea that products with more similar characteristics are closer substitutes and therefore should have lower markups, vice versa. However, Armstrong (2016) showed that BLP instruments are not always suitable when estimating a model with many products. Gandhi & Houde (2019) provide a new class of instruments which they call "differentiation IVs". They deal with a relatively new problem that arises from allowing demand models to exhibit non-IIA preferences and that therefore creates an endogeneity problem from simultaneously identifying the market shares and unobserved attributes.

2.3 Product space demand models

Empirical applications of demand estimation for differentiated products can be divided into roughly two main segments: product space models and characteristics space models. Product space demand systems were the dominant method approximately until the 1990's after which the characteristics space systems increased in popularity due to key advantages (Aguirregabiria, 2018).

A typical product space demand system is constructed in the following way. We define a consumer's preferences over a set of products. Assume a set of varieties of a product J are indexed by $j \in \{1,2,3 \dots J\}$. We note the quantity that a consumer buys of a particular variety as q_j. From this we formulate a utility function, $U(q_1, q_2, q_3 \dots q_j)$, which depends on the quantity that the consumer buys of each variety. Thus, we can formulate the consumer's maximation problem in a very familiar fashion as:

$$\max_{\{q_1,q_2,q_3 \dots q_j\}} U(C, q_1, q_2, q_3 \dots q_j)$$

$$subject\ to: C + p_1 q_1 + p_2 q_2 + \cdots + p_j q_j \leq m$$

where C represents the consumption of an outside good and m the disposable income of the consumer. The demand function of a consumer is the solution to this maximisation problem. For each variety we can define a function that expresses the demanded quantity dependent on the prices and income, also called Marshallian demand functions: (Aguirregabiria, 2018)

$$q_1 = f_1(p_1, p_2, p_3 \ldots p_j, m).$$

The different versions and applications of this type of demand system usually differ in regard to what assumptions are made to the functional form of the utility function. The most known are the Rotterdam model (Theil, 1975) and the Almost Ideal Demand System (Deaton & Mullebauer, 1980).

One noteworthy drawback is the exclusion of the consumer heterogeneity which is a strong modelling assumption. Another more technical limitation is the number of parameters. In a basic product space model, the number of parameters increases quadratically with the number of products (or varieties of a product), which means that even with a small number of products we need to estimate a large number of parameters. This either puts a restriction on how many varieties we include in our analysis or requires a large data set. The other drawback relates to endogeneity. Since firms usually have more information at hand when pricing products than researchers have in their models, we would expect the price to be correlated with the error term and thus, making regression estimates inconsistent. The classic remedy for this problem is to make use of an instrument. That is, to find a variable, which is correlated with price and only effects the quantity via price. Good instruments are however hard to find in empirical data. (Aguirregabiria, 2018)

2.4 Characteristics space demand models

Characteristic space models differ in a fundamental way of thinking about demand and products in comparison to product space models. In product space models, demand arises from a consumer's preference for the product itself, whereas in characteristics space models, a product is seen as a bundle of different attributes/characteristics that define the product. The consumer is then assumed to have a demand for this particular bundle of attributes, not the product itself. (Aguirregabiria, 2018)

To lay out the model more formally, we begin by assuming the same indexation $j \in \{1,2,3 \dots J\}$ of product varieties. Furthermore, we distinguish product characteristics that are observable and measurable and others that are unobservable and/or not measurable but are known to affect consumer behaviour. The observable characteristics are represented in a vector $\boldsymbol{X}_j \equiv (X_{1j}, X_{2j} \dots X_{Kj})$ where X_{kj} represents the "amount" of attribute k in variety j. The unobservable characteristics are in a similar way represented in the vector $\boldsymbol{\xi}_j$. Although the difference is that we do not know the number of characteristics within $\boldsymbol{\xi}_j$, nor the magnitude of them. After this, we define a set of households $h \in \{1,2,3 \dots H\}$ and the individual household's utility $V_h(\boldsymbol{X}_j, \boldsymbol{\xi}_j)$ for consuming variety j, with respective attributes. The total utility for a household is defined as the sum of the utility of consuming product j and the utility of consuming other goods (C), thus giving us: $U_h = u_h(C) + V_h(\boldsymbol{X}_j, \boldsymbol{\xi}_j)$. To account for differences in preferences and other factors contributing to household heterogeneity we add the vector $\boldsymbol{v}_h$ to the function. This vector may or may not be completely, if at all, observable. Adding all these elements together we can fully define the households' utility as:

$$U_h = u_h(C: \boldsymbol{v}_h) + V_h(\boldsymbol{X}, \boldsymbol{\xi}: \boldsymbol{v}_h)$$

Once we have defined a utility function for each household, we can again write down the maximisation problem once we know the income y_h and prices $\boldsymbol{p} = (p_1, p_2 \dots p_j)$ of the varieties. Furthermore, we define an event $d_{hj} \in \{0,1\}$ which indicates whether the household h buys variety j. The maximisation problem thus is:

$$\max_{\{d_{h1}, d_{h2} \dots d_{hJ}\}} u_h(C: \boldsymbol{v}_h) + \sum_{j=1}^{J} V\left(\boldsymbol{X}_j, \boldsymbol{\xi}_j: \boldsymbol{v}_h\right)$$

$$subject\ to: C + \sum_{j=1}^{J} d_{hj} p_j \leq y_h$$

$$d_{hj} \in \{0,1\} \quad and \quad \sum_{j=1}^{J} d_{hj} \in \{0,1\}$$

In other words, there are $J + 1$ alternatives: $j = 0$ meaning that the household does not buy any product and J varieties of the product. Similarly, as in product space models, the solution to the maximisation problem is the demand equation. We define such a demand

function which is dependent on the observable characteristics, prices, income and household heterogeneity as: $d_j^*(\boldsymbol{X}, \boldsymbol{p}, y_h; \boldsymbol{v}_h)$. Assuming a continuum of households and that there exists a well-defined density function f_v for the household heterogeneity, we can obtain the aggregate demand functions by integrating over the mass of the households: (Aguirregabiria, 2018)

$$q_j(\boldsymbol{X}, \boldsymbol{p}, f) = \int d_j^*(\boldsymbol{p}, y_h; \boldsymbol{v}_h), \beta)\, f(\boldsymbol{v}_h, y_h)\, d\boldsymbol{v}_h\, d\boldsymbol{y_h})$$

2.4.1 Discrete choice models

The discrete choice models are an application of the characteristic space approach which have become the workhorse models in modern econometrics. The psychological foundation for discrete choice models comes from the work of Thurstone (1927) who developed the law of comparative judgement and within it, the concept of psychological stimuli (McFadden, 2001). What Marschak (1960) concluded, was that if the psychological stimuli are interpreted as utility, one can define a model for making choice. The models based on this derivation are known as random utility models (RUMs). (Train, 2009)

Looking at the RUM more formally, we denote a person n, choosing between alternatives $j \in \{1,2,3 \dots J\}$. The person receives utility U_{nj} from choosing alternative j and maximises her utility by choosing the alternative that gives the highest utility. Modelling this would be straightforward, if we could measure or observe utility. Instead, we direct our focus to what is called representative utility, noted V_{nj}. This representative utility depends on the attributes of the person making the choice and the attributes of the different alternatives. Since we know that utility is composed of what we can observe and what we cannot, we are able to deconstruct utility as $U_{nj} = V_{nj} + \varepsilon_{nj}$. The term ε_{nj} is the difference between the total utility and the observed utility. In the early applications this is simply treated as random. Combining all the random terms $\varepsilon_{n1} \dots \varepsilon_{nJ}$ gives the vector ε_n' with a joint density $f(\varepsilon_n')$. Assuming a functional from for this density allows us to integrate over the cumulative probability and gives us the probability of alternative j being chosen. The importance of the unobserved term is elaborated in appendix 1. (Train, 2009)

2.4.1.1 Logit and nested logit models

In the logit model we assume that each ε_{nj} is identically and independently distributed (i.i.d.) and that they follow a type 1 extreme value distribution. With these assumptions

the difference between any two ε_{nj} is distributed logistic. The assumption comes however with implications. The independence from irrelevant alternatives (IIA) axiom is a direct consequence of the assumptions in a logit model, and states that the ratio of probability between any alternatives j and i is independent from any other alternatives. Appendix 1 uses a common example, to illustrate the IIA in more detail. (Train, 2009)

Proposing that the unobserved utilities are independent from each other is a strong assumption and relaxing it has been a key component in developing other specifications of the RUM. While this is a strong assumption, there is an alternative way of thinking about it. If we are able to formulate a model with extreme precision and capture all the factors that affect choice in our representative utility, the importance of the unobserved term diminishes until it is essentially just random noise. This line of thought makes the independence assumption more reasonable and adopting this view makes the logit's assumptions look less restrictive in theory. However, researchers have emphasized relaxing this assumption when developing further specifications of the RUMs, such as the nested logit. (Train, 2009)

In a nested logit we construct nests of the choice sets, where the IIA holds within a nest, but not across nests. Formally, we assume that the alternatives in the choice set can be divided into K nonoverlapping nests: $B_1, B_2 \ldots B_K$. The utility from choosing nest B_k is the logit utility $U_{nj} = V_{nj} + \varepsilon_{nj}$. The key difference to the basic logit is in assuming that the cumulative distribution of the unobserved terms is:

$$\exp\left(-\sum_{k=1}^{K}\left(\sum_{j\in B_k} e^{-\frac{\varepsilon_{nj}}{\lambda_k}}\right)^{\lambda_k}\right).$$

Here it is important to note the parameter λ_k, that measures the degree of independence of choice alternatives within nest k. Specifically, if $\lambda_k = 1$ for all nests, the model boils down to the basic logit model. With this specification our model allows correlation of unobserved utility whereas the logit model assumed it to be i.i.d. There are also several ways to build upon the nested logit model. For example, by considering not just a two-level nested model but a three-level or by using a specification with overlapping nests. (Train, 2009)

2.4.1.2 BLP models

Another workhorse model is the Berry, Levinsohn & Pakes (1995) and Berry (1994) – called BLP – model. This model is a discrete choice model, that expresses consumer utility as a function of observable and unobservable product and consumer characteristics, product price and demand parameters. In Berry (1994), the following specification of consumer utility is used:

$$u_{ij} = x_j\tilde{\beta}_i - \alpha p_j + \xi_j + \varepsilon_{ij}.$$

Here a consumer i chooses over a $j \in \{1,2,3 \dots J\}$ set of alternatives where the outside good is normalized as $j = 0$. The unobserved consumer specific taste parameters are $\tilde{\beta}_i$ and ε_{ij}, whereas ξ_j can be thought of as the mean of consumers' valuations of unobserved product characteristics. Through decomposing $\tilde{\beta}_i$ into the mean taste parameter the equation is rewritten as:

$$u_{ij} = x_j\beta - \alpha p_j + \xi_j + \varepsilon_{ij}$$

and the mean utility of product j is defined as:

$$\delta_j \equiv x_j\beta - \alpha p_j + \xi_j.$$

This model is similar to the previous discrete choice models with the exception of ξ_j, unobserved product characteristics. One of the core contributions of Berry (1994), was to relate the equation of mean utility to observable market outcomes, specifically market shares of products, noted s_j. Assuming that the outside good is normalized to zero, using what is called the Berry inversion (property), the market shares of product j can be expressed linearly as:

$$\ln(s_j) - \ln(s_0) = \delta_j \equiv x_j\beta - \alpha p_j + \xi_j.$$

This linearisation also means that we can use traditional instrumental variables to estimate β and α and that we do not need to assume a functional form for ξ_j. We need an instrument to control for endogeneity. In short Berry, Levinsohn & Pakes (1995) showed that in a model like this, one can also use the observable product characteristics as a form of instrument, called "BLP-instruments". The intuitive idea behind BLP-instruments, is that in a market with more products that have similar characteristics, there should also be a fiercer competition which should put downward pressure on

margins and prices, vice versa. In more recent specifications of the BLP model, such as the one discussed in Berry & Haile (2016), there is a requirement for two types of instruments; both BLP instruments and cost side instruments. This is because the demanded quantity for a certain product depends on the endogenous price and the demand shocks of all products.

2.5 Recent developments in characteristics space models

Nevo (2001), uses empirical data to analyse the ready-to-eat cereal industry. The author starts from the claim, that the cereal industry is a classic example of almost collusive pricing behaviour, and an industry where competition is fierce in non-price areas. The characteristics that support this are high cost-price margins and relatively large advertising spend to name a few. The author is however able to empirically dispute that high markups would be due to collusive behaviour and concludes that the high margins are due to product differentiation and multi-product pricing. The demand model used in Nevo (2001) is a discrete choice model, similar to BLP models, but with some differences in how the model is identified. For example, the characteristics of cereal brands are modelled as endogenous, reflecting the assumption that firms choose brand characteristics based on expected consumer preferences.

Looking at even more recent research, the primary idea behind the approach from Gillen et al. (2019), is to automate the variable choice to a more data-driven approach rather than being selected by the researcher's intuition. Their proposal – called BLP-2LASSO – aims at using the "double-LASSO" method by Belloni, Chernozhukov & Hansen (2014) as a variable selector in BLP models. This way, heterogeneity in preferences can be modelled by many different variables available in demographic data. They showcase this approach with an empirical application explaining the effect of campaign spending on vote outcome with data from Mexican elections. The underlying model that the author use is a random-coefficients logit model (also called mixed logit), with heterogenous preferences. In other words, the model from Berry, Levinson & Pakes (1995).

Another paper that brings new advances and builds upon the mixed logit model is by Compiani (2018). The argument presented in the paper is that distributional assumptions in e.g., BLP models might affect the results of such models and therefore become an unwanted constraint. Compiani (2018) proposes as an alternative a non-parametric demand model based on constraints from consumer theory – such as monotonicity of demand. The novel method is used to assess counterfactuals regarding

taxation and multiproduct firm pricing, the latter being in the same spirit as Nevo (2001), using data from strawberry sales in California. One drawback is that in non-parametric models with many products, the number of parameters to estimates grows rapidly, which is solved by imposing additional (behavioural) constraints. On the other hand, the author also notes the future potential in using data driven approaches to reduce the dimensionality in demand models.

2.6 Machine Learning

In this chapter I will quickly overview the central concepts, terminology and methods used in modern machine learning. The focus will not yet be on any specific machine learning method as such, but rather more on the principles and traits that are included in all of machine learning. I will focus only on supervised machine learning. That is, models where data is labelled with clear inputs and outputs and the models aim is to learn a rule or function that maps the inputs to the outputs.

2.6.1 Supervised Machine Learning

Any supervised machine learning algorithm can be generally explained by a few key elements. We are trying to predict the value of a variable y, usually called the dependent variable, with some variables x, called independent variables, that we think are relevant for predicting y. Based on a data set containing observations of y and x, we create a function $\hat{f}$ that maps these two. Importantly, we define some loss function $L\left(y, \hat{f}(x)\right)$, that quantifies how well we were able to predict y. The purpose of the machine learning algorithm is to find such a function $\hat{f}$ that minimizes the expected loss on a new data point from the same distribution. (Mullainathan & Spiess, 2017)

The key formulation in the description above is that we want to minimize loss on a new datapoint. This means that we cannot create an algorithm that describes a dataset in detail, the algorithm has to generalize as well. Therefore, we start by dividing our data into three parts: a training set (50%), validation set (25%) and a test set (25%). We use the training data set to fit the function $\hat{f}$. This is called training the model. Once we obtain the model, we use the validation set to calculate the prediction error which will measure the performance of the predictor. Typically, any machine learning model will have some form of hyperparameters which describe how complex the model is. Therefore, we use the validation set to give feedback to the model to choose values for the hyperparameters, such that we minimize the prediction error. In essence, a model that is not complex

enough will not predict well since it is too simple, and a too complex model will not predict well on unseen data, however it will predict well on the training set. The test set is used lastly, as the best hyperparameters are found, to test the model's ability to generalize on unseen data. The method of restricting the complexity of a model is usually called regularization and a model with too high complexity is called overfitted and vice versa a model with too low complexity, underfitted. (Friedman et al., 2001)

The phenomena described above is what is referred to as the bias-variance trade-off. Variance meaning how much the prediction varies around the true target. The variance is never zero, however, a large variance in prediction, is suboptimal for minimizing the loss function. Bias on the other hand, describes how much the average of our prediction differs from the true average. Bias and variance are usually seen as mutually exclusive. A model with low complexity has a high bias but low variance, vice versa. Too much of either will lead to a lower predictive performance or increase loss. The method above of splitting the data into three sets is a very common method to find the optimal balance between bias and variance. (Friedman et al., 2001)

2.6.2 Machine Learning in economics

The main difference between econometric methods and machine learning is the objective of the methods. In econometrics, the emphasis is on producing estimates that can, at their best, be interpreted as indicators of causality and its magnitude. This is done to estimate and prove an economic theory. In machine learning on the other hand, interest lies in prediction; how well can we predict housing prices in a particular area or the future value of a particular stock? In prediction we do not necessarily need to focus on a specific model that lays out what factors affect housing price, as long as the prediction is good. This is the reason why machine learning has taken its time to become more used in economic research. Nonetheless, due to recent advances and developments its implementation in economics has become more common. (Mullainathan & Spiess, 2017)

The Causal Forest (Wager & Athey, 2018), the double machine learning (Chernozhukov et al., 2018) and Athey & Imbens (2019) are some of the notable recent advances in the cross section of machine learning and economics. I will only go through some basic examples and scratch the surface of the development based on Athey & Imbens (2019) which provides a very good overview. The recent advances have come from combining machine learning methods and traditional econometric methods. The common win-win combination has typically consisted of exploiting the fact that machine learning methods

work well with high dimensional data and sparsity, whereas traditional econometrics are suitable for estimating treatment effects and counterfactuals. Many new methods have focused on debiasing machine learning methods to estimate average treatment effects. The idea is to estimate nuisance parameters with machine learning models and thereafter to estimate the average treatment effect with the orthogonalized predictions of machine learning models. Chernozhukov et al. (2018) and Wager & Athey (2018) for example both use the principles of this approach and chapter four explores the former more in depth. Athey & Imbens (2019)

Another interesting development is the usage of reinforcement learning agents and multi-armed bandits. Some of the most known algorithms, such as described in Sutton & Barto (1998) and Lai & Robbins (1985), are applied in various economic settings, for example by Trovò et al. (2015) who modify and improve Upper Confidence Bound (UCB) algorithms in the context of price optimisation. Furthermore, on a tangent to the development in economics, there is also a field within computer sciences that explores causality. Works such as Bareinboim, Forney & Pearl (2015) explore how multi-armed bandits can be improved in a situation where there are many potential unobserved confounders. Their algorithm, multi-armed bandit problem with unobserved confounders (MABUC) does not maximise its expected reward solely based on experimentation but also on observational data from which counterfactual estimates can be made.

Machine learning is also used in research related to competition policy. For example, in research about how information about consumers' preferences can be used to extract consumer surplus and whether pricing algorithms effect competition. The latter question has been studied in Calvano et al. (2019), where the authors simulate an oligopoly model, in which the agents' pricing decision is based on Q-learning, a reinforcement learning model. The question is whether such algorithms can sustain price levels above the competitive level in an autonomous fashion. In essence, algorithms colluding without the explicit intention of doing so. Their finding indicates that these algorithms can indeed learn to systematically learn to play collusive strategies, which lead to price levels above static Bertrand equilibrium. Furthermore, Brown & MacKay (2021) also show that equilibrium profits can be higher when firms use so called high-frequency pricing algorithms. Especially if the pricing technology is asymmetric amongst firms.

3 DATA

In this chapter, I will start by exploring some of the challenges regarding data in demand estimation and the proposed solution. I will consider what the appropriate level of data is and how to structure it to be able to answer the research questions. Thereafter, I will explore the data itself, showing descriptive statistics and going through the variables in the dataset. Lastly, I consider a Fourier transform as a solution to account for seasonal variation in the data.

The data for this thesis is provided by a Finnish retail firm within the clothing sector. It is gathered from the firm's point of sales system as a dataset of all the transactions that occurred in the firm's flagship store from January first, 2018 to the 24th of February 2020. Each transaction is recorded with a date, time, price and model/stock keeping unit (SKU).

3.1 Aggregation and structure of the data

As discussed in the previous chapter, the characteristic space models interpret the notion of a good in a different way than what was previously done in product space models. Additionally, many of the related literature mentioned in chapter 2 deal with defining the product hierarchy – explicitly or inexplicitly – such as when constructing a nested logit. I will use the example hierarchy in figure 3 as a running example to illustrate the task at hand. The transactional data is recorded at the lowest level, meaning that we get a datapoint per SKU. As can be seen in the figure, this is at the most detailed level. Going up one level, we can consider the different size variations of a t-shirt as variations of what is essentially the same product. Especially if they are priced identically, which indeed is the case in the data in this thesis. This same logic can be applied to go one level up, arguing that the white and blue t-shirt are basically the same product, but with a different colour. The same iteration could be continued further up concluding that a t-shirt and long-sleeved shirt are essentially the same, only with a variation in sleeve length. Importantly – whichever level is defined as the appropriate one – there are always certain assumptions that come with it. For example, how likely does the IIA hold for a particular combination of products?

Choosing which level is the most suitable is not very different from what is done in Hausman (1996) when choosing how to define the market segment equation. He groups of ready-to-eat cereals based on industry practice, which is to divide them into segments of adult, child and family cereals. Arguably, the factors that constitute what is classified

as a cereal for children and for adults may not differ very much in terms of the cereal itself. There might be factors such as the levels of sugar or other taste related variables that are different, but it is unlikely that one could be able to distinguish what is categorised as child or adult cereal by only considering the cereal itself. Therefore, the segmentation of products done by the seller is a relevant starting point in how the consumer potentially perceives the product. This is also what is done by Chernozhukov et al. (2017), who draw a product tree hierarchy like the one in figure 3 which is based on how the firm has segmented their products. Creating a demand model with several levels was originally my aim as well but unfortunately the low number of transactions became a limiting factor.

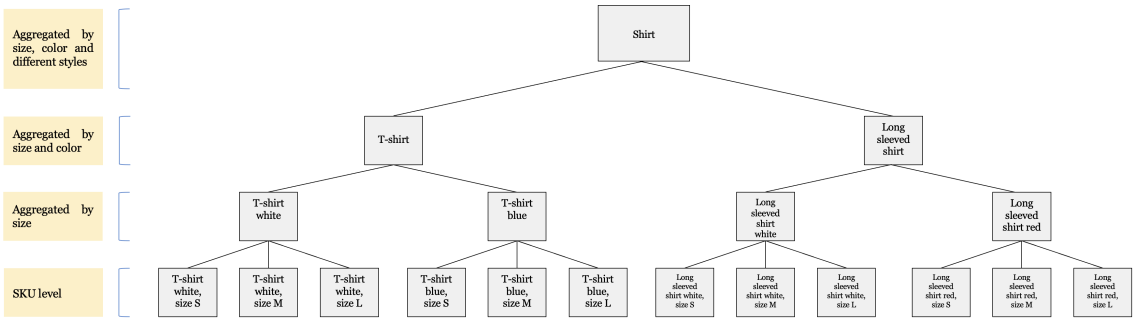

Figure 3 **Example of the data structure of a product hierarchy.** Picture inspired by Chernozhukov et al. (2017)

Switching the view from the seller to the buyer, we also need to think about how the customer behaves when making a purchase decision. As mentioned in chapter 2, the underlying assumption in a logit model is that introducing a new alternative does not alter the ratio of choice probabilities. This means that we have to make some assumption about the cross effect the products have when choosing which level to use. Using the same product hierarchy as before, it is easier to argue that the IIA holds at the lower levels than at the higher levels. For example, it is conceivable that at the SKU level the introduction of a new size does not change the ratio of choice probabilities. The same goes for the introduction of a new colour. Although, at higher levels in the hierarchy, this assumption becomes less plausible. For example, grouping long-sleeved shirts and t-shirts into the same nest might violate the IIA, if consumers do not regard them as substitutes. One can also think of more complex scenarios of substitution patterns. In some cases, a consumer needs a shirt for a formal occasion, implying that a white shirt cannot be substituted with a colourful shirt, due to etiquette requirements. However, if the etiquette is more relaxed, these two products could easily be substitutes for one

another if the etiquette is more relaxed. This is where I have to compromise and make some form of generalisation. These types of complex interactions are naturally not found in point of sales data and in general the intentions of the buyer are quite hard to quantify. Moreover, a wide range of statistical tests have been developed e.g., McFadden, Train & Tye (1977), McFadden (1987) and as well one by Hausman & McFadden (1984), to test whether a particular grouping is appropriate. Gandhi & Houde (2019) also provide a way of testing the IIA in their specification.

Even though the entire product catalogue was at my disposal, few products are consistently sold over the year and therefore suitable for this analysis. For example, winter jackets naturally do not have observations for the summer period. To make things simpler, I only consider products that are sold regardless of the season. Furthermore, there are rather few products that have enough transactions to be considered. To satisfy this requirement I chose only to focus on products that are sold in higher volumes, in this case shirts. As a technical note, the data is also aggregated to weekly sales. Meaning that one observation corresponds to the sold quantity during a week. This comes with the downside of averaging observations within weeks which might lead to a loss of information. On the other hand, due to the frequency of transactions, it comes with the conveniency of not having to deal with periods of zero observations, which means that transforming the data into logarithm form is very straightforward.

Defining the appropriate level to use in this thesis is done by following the examples of Hausman (1996) and Chernozhukov et al. (2017). The product tree hierarchy follows the industry standard, or in this case, the internal grouping hierarchy given by the firm. Thereafter, I consider a level referred to as “shirts”, which contains formal and semi-formal shirts designed for men. Conveniently, all products at this level are priced more or less similarly. This level is also chosen with the argument, that the IIA should hold within this it. Meaning, that the shirts are of similar nature and follow a similar pricing policy. For example, t-shirts and very casual shirts are not included in the analysis.

3.2 Descriptive statistics & variables

As noted earlier, the data consists of transactional (sales) data of a Finnish retail firm. The chosen segmentation is a product group named “shirts” containing formal and semi-formal shirts for men. Furthermore, one observation corresponds to one week, meaning that the total dataset – ranging from 1.1.2018 to 24.2.2020 – contains 108 observations. Due to the cross-validation procedure of double machine learning (DML), it is possible

to train machine learning methods with a relatively small number of observations. Chernozhukov et al. (2018) empirically demonstrate DML with a data set with as few as 64 observations.

Table 1 presents descriptive statistics of the two most important variables: sales quantities and price, both in logarithmic form. As we can see, the average weekly log sales are at 3.64 with a rather moderate standard deviation of 0.54. Price, does not fluctuate as much and is steady around a mean of 4.19 with a standard deviation of 0.13, reflecting the fact that the shirts in the group follow similar pricing. This is what we would expect from a typical retail firm. Price variation is required to be able to estimate price elasticities.

Table 1 Descriptive statistics of prices and sales

	Mean	StDev	Min	Max
Log of Sales	3.64	0.54	1.09	4.57
Log of Price	4.19	0.13	3.84	4.34

The main variable of interest – which is the dependent variable - for this thesis is the log sold weekly quantity of products in the product group “shirts”. The second variable of importance is log price, which is expressed as average weekly price of the occurred sales of this particular product group. The third important variable to be mentioned separately is product cost or purchase cost, which will be used as an instrument. The other variables can be roughly divided into two categories: product characteristics and time-related variables.

Variables included in the product characteristics are the hard-coded attributes of the products, retrieved from the firm’s product information management system. These include for example colour, pattern, material, collar, button colour, button thread colour and so on. The time-related variables, such as the seasonal variable are included to account for seasonality. Furthermore, lagged variables of sales quantities and prices are included to condition as much as possible on the information set available to the price setter, in the same spirit as in Chernozhukov et al. (2017). All of the variables are listed in table 2, with a brief description and an indication whether the variable is a categorical or numerical variable and if it has been transformed.

Table 2 Description of variables

Variable	Description	Form
Dependent variable		
sales	Weekly sold quantity	logarithm
Explanatory variables (product characteristics)		
Price	Weekly average of product price	logarithm
Cost	Weekly average cost of production related to products sold	logarithm
Colour	Main colour of the garment	categorical
Size	Size of the garment	numerical
Material	Main material of the garment	categorical
Fit	Different fits of the garment	categorical
Collar	Collar type	categorical
Button	Main colour of button	categorical
Button thread	Main colour of button thread	categorical
Buttonhole thread	Main colour of buttonhole thread	categorical
Cuff	Design of the cuff	categorical
Other covariates		
Linear trend	Linear trend of sales during the time period	numerical
Fourier transform	Seasonal decomposition	numerical
Lagged variables of sales and price	Lagged realisations of the demand system	logarithm

Figure 4 shows a timeline of the log sold weekly quantities and log prices. There are no observation periods with zero sales, which means that no additional transformations need to be done. Sales are also somewhat volatile and have strong seasonality, as is quite natural in the retail industry. Price follows the same cyclical pattern in an inverse manner with a pairwise correlation of -0.61. There appears to be two sales peaks during a

calendar year. One peak occurs during the end of the summer, approximately in July and August and another peak occurs before Christmas and continues approximately to January. A visual inspection of figure 4 also suggests that there does not appear to be a strong linear trend or increase/decrease over the seasonal fluctuation.

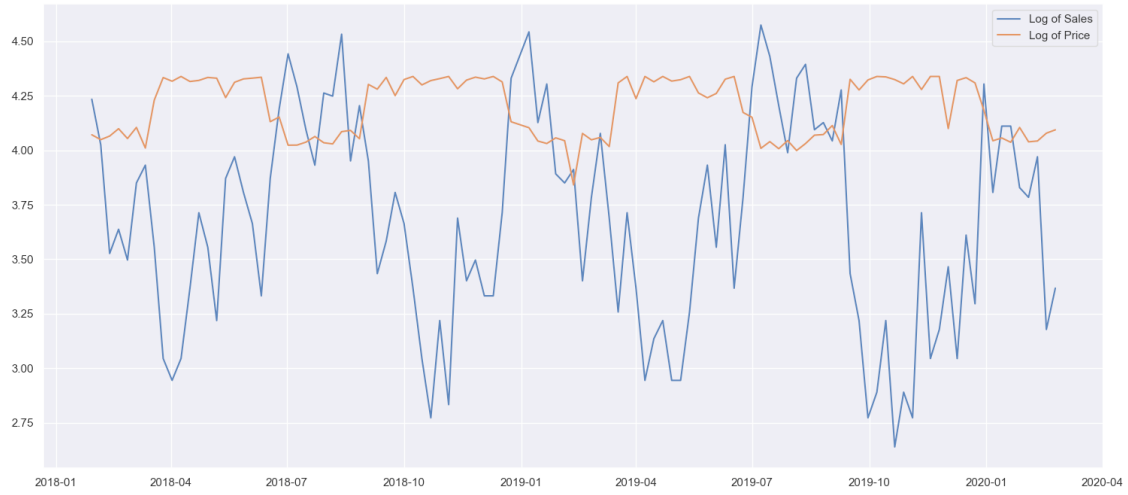

Figure 4 The logarithm of sold quantity and price

3.3 Accounting for seasonality

Looking at the data in figure 4, we can see that sales and prices follow a cyclic pattern. One of the important tasks in this thesis is to predict log sales, log price and costs as well as possible. A common approach to include seasonality in a model is by using dummy variables. For example, by having each month as a variable whose value equals 1 if the observation falls within that particular month and otherwise being 0. The drawback of this approach is that after dropping one of the months to avoid perfect collinearity, there are 11 additional variables in our model.

Another approach is wavelet analysis and frequency domain analysis. Papers such as Crowley (2007) and Schleicher (2002) give an introduction to how these methods can be used in economics. Although more importantly, since we are dealing with a purely prediction problem, we can make use of approaches presented in Lange, Brunton & Kutz (2021) and Fumi et al. (2013), of which the latter make use of a Fourier transform to predict demand in the fashion industry. There are other wavelet methods for capturing seasonality in the same manner but due to the wide availability of ready-made functions, I chose to opt for a Fourier transform to model seasonality.

The idea of a Fourier transform is to decompose a time series or wave function into the underlying components that additively create the observed time series. Given a time series $g(t)$, we can get the strength of the frequencies of $g(t)$ from the output of the function below:

$$\hat{g}(f) = \int_{t_1}^{t_2} g(t) e^{-2\pi i f t} dt$$

which is the Fourier transform. Lange, Brunton & Kutz (2021)

This way we can model seasonality in a smoother way and using only one variable. In practical terms, I am using the NumPy module in Python that features a set of functions that make the calculation of Fourier transforms very easy to implement, using the Fast Fourier Transform (FFT) algorithm. The upper graph in figure 5 shows visualized the output of $\hat{g}(f)$, with the frequencies (expressed in hertz) on the x-axis and the power spectral density (PSD) on the y-axis, which can be thought of as how impactful each frequency is. Visually, it is evident that there is one frequency, that is far more dominant than the other ones, approximately 0.03 Hz.

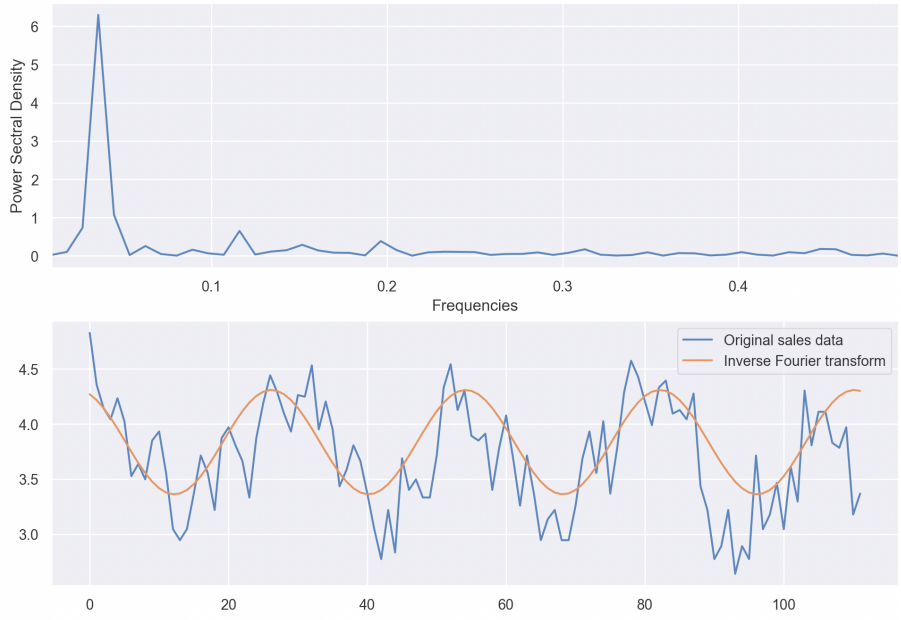

Figure 5 Power Spectral Density and modelled seasonality

From this we can create a wave function, which is based only on the frequencies that are above a cut-off value of the PSD. I define the cut-off value to be one, to capture only the most dominant frequency at 0.03Hz. This whole process can again be finished

conveniently by calling an inverse FFT function in NumPy that reverses the procedure and creates a wave function based on the 0.03Hz frequency. This is visualized in the lower graph of figure 5, alongside with the original sales data. The orange line represents the inverse Fourier transform. This newly created wave function models the seasonality of the sales and can be used as a control variable in a prediction model, in the same way as quarterly or monthly dummy variables would be used.

4 EMPICIRAL FRAMEWORK

In this chapter I will start by presenting the specification of demand models. I consider two models: one linear model and one partially linear model. Thereafter I review the instrumental variables method, which will be used for the estimation of the linear model. This is followed by reviewing double machine learning, which is the method for estimating the partially linear model. Furthermore, I will explore different machine learning methods as first stage estimators of DML. These methods are also presented in this chapter.

4.1 Specification of demand models

The choice of demand models is motivated by following the model in Chernozhukov et al. (2017) as closely as possible although, due to differences in data and modelling choices some deviations are made. The partially linear model will be estimated with DML and is similar to the model in Chernozhukov et al. (2017) whereas the linear model that is estimated with traditional econometric methods serves as a comparison. The linear model has the following form:

$$Y_{jt} = \beta_0 + P_{jt}\theta + \beta_1 X_{jt} + U, \quad \mathrm{E}[U|X, Z] = 0 \tag{4.1}$$

where Y_{jt} is the log sales of product j at time t and P_{jt} is the log price of product j at time t. X_{jt} is a matrix containing observable product characteristics and information of previous realisations of the demand system. The demand model is identified by assuming that product cost (Z) is a viable instrument, meaning that product costs affect demand via price. The partially linear model on the other hand is:

$$Y_{jt} = \beta_0 + P_{jt}\theta + g_0(X_{jt}) + U, \quad \mathrm{E}[U|X, Z] = 0 \tag{4.2}$$

$$Z_{jt} = \beta_0 + m_0(X_{jt}) + V, \quad \mathrm{E}[V|X] = 0 \tag{4.3}$$

where, similarly, Y_{jt} is the log sales, P_{jt} is the log price and X_{jt} is contains characteristics and previous realisations of the demand system. In contrast to the linear model, X_{jt} affects log sales via function $g_0(X_{jt})$. Z_{jt}, the product cost, is also dependent on X_{jt}, via function $m_0(X_{jt})$. Collectively the functions are noted as the nuisance parameters $\eta_0 = (m_0, g_0)$. It is important to note that the dimension of X_{jt} is large relative to the sample size n, which can be a problem for traditional econometric methods. This is what motivates the usage of double machine learning in the first place. The first equation is

the main equation and the parameter θ is our target parameter of interest. If the mean zero condition for both equations hold, we can interpret θ as the causal effect of price on demand, the price elasticity of demand. Note also that the behavioural implication of a log-demand model is that price elasticity of demand is constant. The second equation keeps track of confounding, and is from a technical perspective important, but interpreting it as such is not of interest to us. This model closely resembles that of Chernozhukov et al. (2017) and the lead example in Chernozhukov et al. (2018).

The identification assumption for both models is that Z_{jt} has to be a valid instrument and for the partially linear model, the mean zero after condition for Z_{jt} also has to hold. If the conditions are violated, the problem of price endogeneity arises as discussed in chapter 2.2 making our estimates of price elasticity are potentially biased. If this is the case, we cannot interpret them as estimates of causal parameters. The assumed causal relationships of the variables in the partially linear model are visualized in figure 6 with a directed acyclic graph. To summarise, after conditioning on X_{jt}, product cost affects demand via price and this way, the partially linear model is identified.

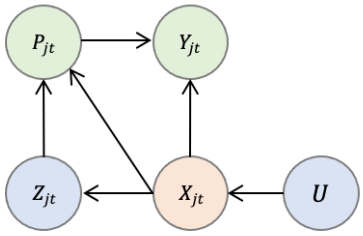

Figure 6 **Directed acyclic graph of the partially linear model.** Bach et al. (2020)

4.2 Instrumental variables (IV)

Instrumental variables methods are a well-known approach in the economics literature, dating back to Wright (1928) and Haavelmo (1943). There are many ways of implementing instrumental variables, but I consider the two-stages least squares method (TSLS or 2SLS). The assumption that needs to be made to be able to identify the demand equation in this case, is that the instrument – product cost – affects sales via price, but not sales directly.

The application of instrumental variables is what Imbens (2014) refers to as the textbook example of instrumental variables and TSLS. Recall that the linear model specified in equation 4.1 is:

$$Y_{jt} = \beta_0 + P_{jt}\theta + \beta_1 X_{jt} + U, \quad \mathrm{E}[U|X, Z] = 0$$

Y_{jt} denotes the dependent variable, P_{jt} and X_{jt} are the dependent variables and U the stochastic error. The condition that must be met is that the error is mean independent of X_{jt} and the instrument Z. In the TSLS method, we begin by regressing price on the instrument and X_{jt}, obtaining a predicted value for price:

$$\hat{P}_{jt} = \beta_0 + \beta_1 X_{jt} + Z_{jt}\beta_2 + \varepsilon_j$$

This is done by estimating the equation with standard OLS. Thereafter we use the predicted value of price to estimate θ:

$$Y_{jt} = \beta_0 + \hat{P}_{jt}\theta + \beta_1 X_{jt} + U$$

which will be our estimate of the own price elasticity. It is good to note the importance of a valid instrument. If the instrument does not affect sales via price, or if the relation is weak, we have a weak instrument problem. This can be highlighted if we rewrite our estimate of price elasticity as:

$$\theta = \frac{Cov(Y_{jt}, Z_{jt})}{Cov(P_{jt}, Z_{jt})}.$$

If the covariance of price and product cost is close to zero, our estimate of the price elasticity will be biased. Stock, Wright & Yogo (2002) provide a rule of thumb for determining if the instrument is weak by calculating the F-statistic of the regression on price. If the F-statistic if above 10, we may conclude that the instrument is not weak. However, more recent research such as Lee, McCrary, Moreira and Porter (2020) suggest that this cut-off value should actually be 104.7. Furthermore, Keane & Neal (2021) show several shortcomings in only considering an F-value when diagnosing whether an instrument is weak. Rather, they suggest using the Anderson-Rubin test, which is more robust. Their main conclusion is that estimating an OLS simultaneously is the best research strategy. The authors' argument is that for a 2SLS to confidently give better results than OLS, the first stage F-statistic should be above 50. This is also the main motivating reason for estimating the linear and partially linear demand models both with and without instruments.

4.3 Double/Debiased machine learning (DML)

The double machine learning method of Chernozhukov et al. (2018) is a useful method when one wants to estimate a low dimensional causal parameter in the presence of several potentially high dimensional confounders. Double machine learning removes the bias of ML methods using Neyman-orthogonal moments and by cross-splitting the data. This means that one can leverage the benefits of modern ML methods which are suited for high dimensional data.

Before exploring the DML method itself, I will do a quick overview of a central theorem of DML and one that helps with understanding the procedure at large. After that I will review DML in two parts: a *first stage* and *second stage* estimation. The first stage is estimated with three different machine learning methods: Lasso, Random Forest and AdaBoost. The second stage equation is estimated with a TSLS.

4.3.1 Frisch – Waugh – Lovell theorem

The Frisch – Waugh – Lovell theorem is an integral part of DML, and an intuitive understanding thereof makes DML also a lot clearer. This subchapter can be seen as a short prelude before we dive into the actual DML. The theorem is based on Frisch & Waugh (1933) and Lovell (1963). The idea will be illustrated by the example of the following regression:

$$Y = \beta_1 D + \beta_2 X + \varepsilon.$$

Assume that we are interested in understanding how D affects Y while controlling for X. The standard way of doing this is by estimating a regression and obtaining an estimate $\widehat{\beta_1}$. What the FML theorem specifically states, is that we can estimate β_1 in another way:

1. Regress Y on X and recover the fitted values, $\hat{Y}$. Calculate the residuals by subtracting the observed value from the fitted value: $\tilde{Y} = Y - \hat{Y}$. Here, I find it useful to remind that the result of regressing Y on X can be written as: $\widehat{\mathrm{E}[Y|X]}$.

2. Regress D on X and recover the fitted values, $\widehat{D}$ or $\widehat{\mathrm{E}[D|X]}$. Calculate the residuals by subtracting the observed value from the fitted value: $\widetilde{D} = Y - \widehat{D}$.

3. Estimate the regression: $\tilde{Y} = \beta_1 \widetilde{D} + \varepsilon$.

This way, we can estimate β_1 while controlling for X. The FML theorem is a key component upon which the DML method builds. When using this idea in the DML method we will substitute into step 1 and 2 the estimation of the nuisance parameters (which are the ones we want to control for) and then proceed into step 3 where we get the relation between price and sales.

4.3.2 First stage estimators

The DML estimation is divided into two parts, a first stage estimation and a second stage estimation. The first stage of DML corresponds roughly to the first and second stages of the Frisch – Waugh – Lovell theorem. In other words, we are estimating the tilde variables that will be needed for the second stage. Recall that we are only going to be only estimating the partially linear model – equation 4.2. and 4.3 – with DML.

In concrete terms, the first stage task is to predict price, sales and product cost using the product characteristics and cofounders in X_{jt}. As stated in Chernozhukov et al. (2018), for this prediction one can use any method that is "able to deliver sufficiently high-quality approximations to the underlying nuisance functions". Furthermore, DML entails a specific sample splitting procedure as illustrated in figure 7. We begin by dividing our sample into k different batches. Using data from all batches except k, we predict price, sales and cost with a machine learning method of our choice – the first stage estimator. The reduced form of the first stage is thus defined as:

$$l_{i0}(x) \equiv \mathrm{E}\left[Y_{jt}|X_{jt}\right] \tag{4.4}$$

$$d_{i0}(x) \equiv \mathrm{E}\left[P_{jt}|X_{jt}\right] \tag{4.5}$$

$$h_{i0}(x) \equiv \mathrm{E}\left[Z_{jt}|X_{jt}\right] \tag{4.6}$$

And the respective residuals are:

$$\tilde{Y}_{jt} = Y_{jt} - \hat{l}_{i0}(x) \tag{4.7}$$

$$\tilde{P}_{jt} = P_{jt} - \hat{d}_{i0}(x) \tag{4.8}$$

$$\tilde{Z}_{jt} = Z_{jt} - \hat{h}_{i0}(x) \tag{4.9}$$

The important note is that when calculating the residuals: Y_{jt}, P_{jt} and Z_{jt} are drawn from batch k, which was not used in estimating $\hat{l}_{i0}(x)$, $\hat{d}_{i0}(x)$ nor $\hat{h}_{i0}(x)$. We repeat this

procedure so that all batches have been left out from the first stage estimation once. Finally, we average the results to obtain the final residualized (tilde) variables. This way, we can get the benefits of sample splitting while at the same time using the entire sample size.

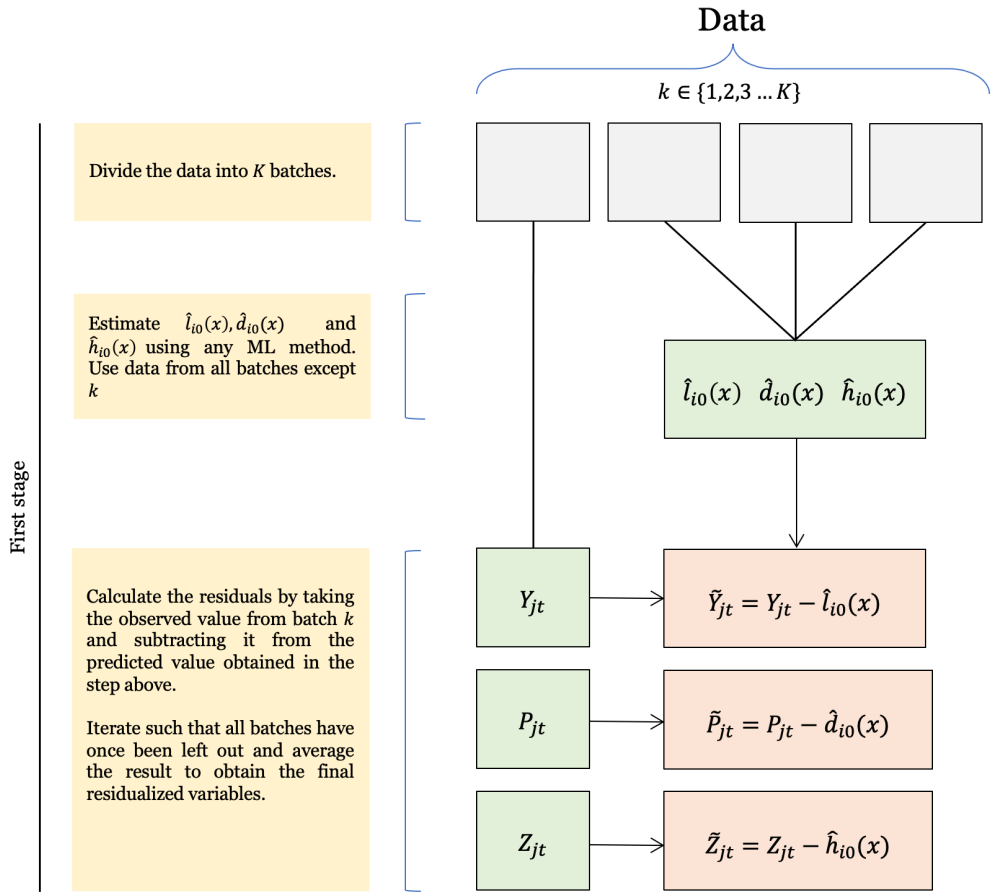

Figure 7 First stage of DML

To find the most suitable machine learning method to estimate the first stage, Chernozhukov et al. (2018) deploy several potential machine learning methods and measure out-of-sample performance. They list methods ranging from random forest to lasso to neural networks. I will use three methods: Lasso, Random Forest and AdaBoost. The motivation for the choice of these method will be explained more in detail in the following sections.

4.3.2.1 Least absolute shrinkage and selection operator (Lasso)

Lasso is a penalized regression or also called shrinkage method. It minimizes a penalized residual sum of squares, also known as L_1 regularisation. More specifically the L_1

regularisation is a penalisation in absolute value. The magnitude of the penalisation or regularisation is done by changing a hyperparameter, λ, as shown in the equation below:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmax}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

When estimating a regression with many different parameters, it is likely that some of the regressors are correlated and thus, cause the regression to exhibit high variance and suboptimal performance. By introducing the penalization parameter, we are addressing this issue. This also makes Lasso a natural choice for high dimensional estimations, such as the one in this thesis. It is worth noting, that the only difference between the classical OLS and Lasso is the last term – the penalization parameter. A higher value for λ means that we are penalizing more and if $\lambda = 0$, the equation is identical to OLS. Thus, when choosing the proper value for λ, we are balancing between improved out-of-sample predictive performance and omitting truly important variables. In reality, a cross validation procedure as described in 2.6.1 is the most common approach of tuning the hyperparameter to suit the data one is dealing with. (Friedman et al, 2001)

4.3.2.2 Random Forest

Random Forest is a popular ensemble method by Breiman (2001) used both for classification and regression. It is suitable for high dimensional data, especially with a high number of categorical variables. While this is true for regression trees in general, the main benefit of Random Forest – in contrast to other tree-based methods – is that the correlation between trees is reduced by randomly selecting the input variables when growing a specific tree, which in turn should keep the predictors' variance at modest levels.

The idea of all regression trees, be it CART or C4.5, is to make a recursive binary split based upon the independent variables of a regression. The choice of variables and splitting points is chosen such that it minimizes the sum of squared residuals (SSR). For example, in the example tree in figure 8, we have three independent variables X_1, X_2 and X_3 to predict the outcome Y. For each variable, we partition the training data in two, such that it minimizes the SSR. The variable with the lowest SSR will be chosen as the root, in this case X_1. This process is continued and if not stopped, we would describe the entire dataset, each observation being an end node – also called terminal node or leaf.

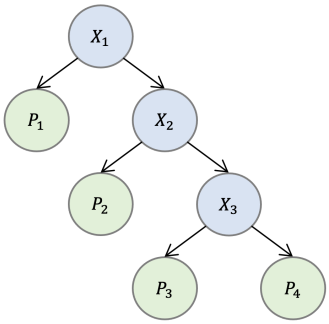

Figure 8 Illustrative example of a simple regression tree

To avoid overfitting a stopping rule needs to be defined, for example the minimum number of observations per split. In a regression tree the terminal node P_1, P_2, P_3 and P_4 are the predictions, given that an observation falls into the particular split. The prediction is calculated as the average of the observations in that split. The benefits of regression trees are a high degree of interpretability and the ability to capture non-linearities in data. On the other hand, they are known for high variance and high sensitivity. A small change in the data might cause a different choice for the root which can yield a completely different result. (Friedman et al, 2001)

The Random Forest algorithm was designed to improve the abovementioned problem of high variance in regression trees. Before Breiman (2001) introduced the Random Forest, there were other ways of dealing with this problem, such as bootstrap aggregation (bagging). The concept of bagging is straightforward: bootstrap n samples from your data set and fit a regression tree on each of the subsets after which you average the prediction which reduces variance. To be precise the Random Forest algorithm is a specification of bagging, where at each node only a subsample, m of the independent variables are considered. This subsample is different at each node and is usually one third of the total number of independent variables, as recommended in Breiman (2001). The reason for choosing a subsample for each node is that as m decreases, we are essentially de-correlating the trees. If each tree, T created by bagging is identically and independently distributed with variance σ^2, the averaged tree has a variance of $\frac{1}{T}\sigma^2$. Whereas, if the trees are identically but not independently distributed the variance of the averaged tree is:

$$\rho\sigma^2 + \frac{1-\rho}{T}\sigma^2,$$

where ρ is the positive pairwise correlation. With regular bagging, the variance is decreased by increasing T. But the first term is not affected by this. On the other hand, by using a Random Forest algorithm, we are able to affect ρ, by de-correlating the trees by choosing a smaller subset m of independent variables at each node. Note that Random Forest, and bagging in general, aims only at lowering variance of the predictors. Therefore, the main assumption is that the underlying trees in a Random Forest already have low bias. (Friedman et al, 2001)

4.3.2.3 AdaBoost

In the previous chapter, we saw how regression trees could be improved with bagging and Random Forest. Bagging is usually compared to a somewhat similar method called boosting. One of the most popular boosting algorithms is AdaBoost by Freund & Schapire (1997), which will be the third algorithm used in DML. AdaBoost is a meta-algorithm, where one estimates many so called "weak predictors" (a predictor that is slightly better than random guessing) and by aggregating them creating a powerful predictor.

More concretely, boosting is a method in which the samples in a training data set are weighted and the weak predictors, in this case shallow regression trees, also called stumps, with just two terminal nodes, are combined sequentially and not independently as in the Random Forest. One starts by giving each observation in a data set equal weights, $w_i = \frac{1}{N}, i = 1,2, \dots N$, after which the weak predictors, are fitted on the data. For each regressor, we calculate the weighted errors which become the new weights for each observation, and pass it to a new regressor. This way, the next regressor uses data with more emphasis on the datapoints which the previous regressor could not predict relatively as well. This process is continued iteratively until a stopping rule is applied. (Friedman et al, 2001)

As with Random Forest, there are many hyperparameters and model choices to be made when using these algorithms for prediction. How large to grow the trees, when to apply a stopping rule, which loss function to use, etc. Going through all the details is nonetheless slightly out-of-scope for this thesis and will therefore not be discussed in full detail. However, the exact description of the algorithms can be found in the appendix. The hyperparameters and tuning of the algorithms are chosen with a cross-validation procedure like the one described in chapter 2.6.1.

4.3.3 Second stage estimators

Once we have obtained the residualized variables from the first stage estimation, we can move on to the second stage. The second stage in DML corresponds to the third step described in the Frisch – Waugh – Lovell theorem. We use the residualized variables as defined in equations 4.7-4.9. These are sometimes collectively called tilde variables from the DML first stage regression. Now when estimating the partially linear model, we simply regress $\tilde{Y}_{jt}$ on $\tilde{P}_{jt}$, in the same spirit as in with Frisch – Waugh – Lovell. We can write the second stage residualized equation as:

$$\tilde{Y}_{jt} = \tilde{P}_{jt}\theta + U, \quad \mathrm{E}[U|\tilde{Z}] = 0 \qquad 4.10$$

Estimating this equation is what is referred to as the second stage and will be estimated by TSLS. We regress $\tilde{Y}_{jt}$ on $\tilde{P}_{jt}$ using $\tilde{Z}_{jt}$ as an instrument to obtain an estimate of $\hat{\theta}$, as shown in figure 9. Note that the second stage in DML is the same method that I use in estimating the linear model. The difference is that I use the residualized variables obtained in the first stage instead of the original data. Given that the model's assumptions hold, we can interpret the estimate $\hat{\theta}$ as the own price elasticity of demand for product j. However, we are not able to interpret any of the other parameters causally. For example, it might be tempting to find out what the effect colour has on a shirt's sales. But since we are using a naïve machine learning method to predict the sales in the first stage, $\mathrm{E}[Y_{jt}|X_{jt}]$, we obtain a measure of association not causation.

Figure 9 The second stage of DML

4.4 Summary of methods

In summation, the linear demand model is estimated with TSLS, and the partially linear model is estimated with DML where the first stage estimators are Lasso, Random Forest and AdaBoost and the second stage estimator is a TSLS regression.

I use a Python module by Bach et al. (2020) that implements the DML methodology. It is somewhat restrictive when it comes to checking different statistics for instrumental variables, which means that some parts are calculated manually. All of the first stage predictors are implemented with the Python module SciKit-Learn by Pedregosa et al. (2011) with the exception of the TSLS, which is implemented with the Linear Models module by Sheppard et al. (2021).

Finally, I will estimate both the linear and partially linear model without an instrument. In other words, regressing price and X_{jt} on sales. This means that in the partially linear model, the DML second stage is changed from TSLS to OLS. The main motivator for this is to use the regression as a robustness check as suggested by Keane & Neal (2021).

5 RESULTS

In this chapter I will review the results from the estimations discussed in the previous chapter. The linear model and partially linear model are presented separately whereafter I do a more in depth look at the diagnostics from the DML first stage estimators, such as robustness checks and variable importance.

5.1 Demand estimation results

From table 3 we can see the estimates of the linear model. A few iterations of the linear model were tested – using e.g., colour and pattern as control variables – but this version resulted in the best overall outcome. Due to the modelling assumptions regarding constant elasticity, both sales and prices can be expressed in logarithmic form which means that the parameter estimate, $\hat{\theta}$ can be directly interpreted as price elasticity. Price elasticity is the relative change in demand arising from a relative change in price.

Table 3 Estimates of price elasticity with the linear model[1]

	$\hat{\theta}$	First stage F-statistic	Std. Err.	p-value
IV	-3.40*	4.65**	1.73	0.05

The price elasticity estimate is -3.40* with a p-value of 0.05. The first stage F-statistic of the IV is 4.65** which is under the Stock, Wright & Yogo (2002) rule of thumb 10 and clearly under the suggested value of 50 as in recommended in Keane & Neal (2021). In other words, we should interpret the estimates of price elasticity with scepticism since they might be biased due to a weak instrument.

Table 4 shows the estimation results from the partially linear model, each first stage estimator shown separately. All of the DML results are estimated with 30 folds, meaning that the data is split into 30 batches. The first column shows the estimate of price elasticity and the third and fourth column show the out-of-sample root-mean-square-error (RMSE) of the first stage predictors. In other words, how well the algorithms predict sales, price and product cost, given X_{jt}. There are many other error measures besides RMSE that are widely used for measuring performance of machine learning

[1] ***=the estimate is significant at 1% level and **=the estimate is significant at the 5% level and *=the estimate is significant at the 10% level

models but doing a comprehensive survey of them all is somewhat out-of-scope for this thesis. Interestingly, there is little difference in the ML methods predictive power. The RMSE for predicting sales is around 0.3 for Random Forest and AdaBoost and only slightly worse: 0.484, for the Lasso. The RMSE for price follows a similar pattern and is around 0.1 for Random Forest and AdaBoost and a bit higher for the Lasso. In other words, we can conclude that the methods perform rather well. The lowest and best RMSE is marked with bold in each column in table 4. One of the benefits of DML, is that one is able to horserace the ML methods and construct the best performers into a combined method. Ergo, one does not have to predict price and sales using the same method. Since AdaBoost was the best at predicting sales and price and Random Forest was best at predicting sales, we could create a combined DML of them. Although the RMSEs were so similar that creating a combined estimation did not give any added value.

Table 4 Estimates of price elasticity with the partially linear model

	$\hat{\theta}$	First stage F-statistic	RMSE for $l_{i0}(x)$	RMSE for $d_{i0}(x)$	RMSE for $h_{i0}(x)$
Lasso	-2.80***	9.04***	0.401	0.135	0.094
Random Forest	-2.91	0.18	0.337	0.078	**0.054**
AdaBoost	-1.25	1.93	**0.331**	**0.074**	0.059

As in the linear model, all of the F-values are low. Lasso is the highest with an F-statistic of 9.04* and furthermore the only significant F-statistic. Nonetheless, the F-statistic is still below the Stock, Wright & Yogo (2002) rule of thumb 10. Furthermore, only the price elasticity estimate of Lasso is significant below the 1%-level. Regardless of the first stage estimator all estimates are negative and closer to zero than the estimate of the linear model. As in the empirical estimations in Chernozhukov et al. (2018), DML did not produce results that would be contradictory to that of the traditional econometric method. Although, the results differ somewhat – especially when comparing the results from the AdaBoost – they all indicate a that an increase in price leads to a decrease in sales.

5.2 DML first stage estimators and auxiliary analysis

To get a better understanding of what the ML methods actually do in the first stage of DML, it is useful to look at metrics beyond RMSE. In this subchapter I am going to present more in-depth analysis of the first stage estimators and robustness checks.

Keane & Neal (2021) suggest that estimating a linear regression with controls for endogeneity can in many cases be a good empirical strategy when there is concern for a weak instrument. Table 5 presents the results of an OLS estimation of the linear model when regressing price and the control variables on sales. The price elasticity estimate is -1.68***, which is closer to zero than the estimate obtained with an instrument.

Table 5 Estimates of price elasticity with the linear model without an instrument

	$\hat{\theta}$	Std. Err.	p-value
OLS	-1.68***	0.31	0.00

Table 6 presents the results of the partially linear model without an instrument. The results were obtained by using OLS as the second stage estimator instead of TSLS in the DML method. In comparison to the estimation with an instrument, all of the estimates are significant at the 1%-level and similarly as in the linear model, all the price elasticity estimates are closer to zero. As argued in Keane & Neal (2021), an estimation without an instrument seems to outperform when controlling for endogeneity.

Table 6 Estimates of price elasticity with the partially linear model without an instrument

	$\hat{\theta}$	RMSE for $l_{i0}(x)$	RMSE for $d_{i0}(x)$
Lasso	-1.89***	0.401	0.135
Random Forest	-1.60***	0.337	0.078
AdaBoost	-1.76***	**0.331**	**0.074**

Both the linear and partially linear model had low F-statistics which indicates that the estimates might be biased due to a weak instrument. To understand this better, figure 10 illustrates correlation plots of the relevant variables. Recall, that the underlying assumption for an instrumental variable regression is that product cost affects sales via price. If the $Cov(P_{jt}, Z_{jt})$ is close to zero, the estimate will become biased. Our expectation is that sales and price have a negative correlation, product cost and price should have a positive correlation and product cost and sales to have no correlation. Log sales is noted

as Y, log price as P and product cost as Z. The residualized variables $\tilde{Y}$, $\tilde{P}$ and $\tilde{Z}$ follow the definition in chapter 4. The upper left graph is the observed data – which is used directly in the linear IV – and the other graphs show the residualized (tilde) variables from the first stage ML estimators of the partially linear model, one correlation plot per ML method.

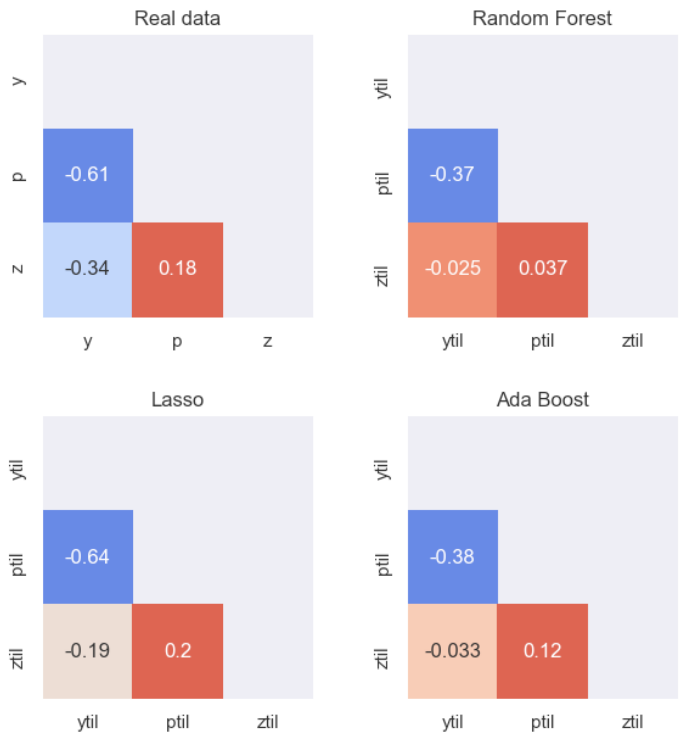

Figure 10 Correlation plot of the residualized variables and the real variables

From all subplots, we can see that price and sales indeed have a negative correlation, the correlation is somewhat stronger in the observed data than what it is for most of the residualized variables. When it comes to the Random Forest and AdaBoost we can without doubt see that the correlation between product cost and sales is minimal. The same correlation is slightly higher in the observed data and Lasso. But the crucial part is the correlation between the instrument: product cost, and price. As we can see the correlation is positive in the observed data, something we would anticipate although it is somewhat weak. Interestingly, this correlation is a lot weaker for the residualized variables from the Random Forest and AdaBoost. This could explain why the F-statistics of Random Forest and AdaBoost are clearly lower than that of the Lasso and IV.

Focusing on the residualized variables of the Random Forest and AdaBoost, we need to analyse what could be the cause of diminishing the correlation between variables in

comparison to the underlying data and the residualized variables of Lasso. For example, if there is something inherent in the tree-based methods that would cause this. One hypothesis could be that the higher variance in the Random Forest and AdaBoost predictions might cause this outcome. From a visual analysis, as shown in figure 11, we can compare the residualized variables of price (blue) and product cost (orange). A higher positive correlation would be visually seen as the lines moving in similar manners. In the case of the Random Forest and AdaBoost we can see a general tendency of covariance but at the same time rather high deviations. The occurrence of spikes in the prediction is more common for Random Forest and AdaBoost, meaning that at times, the prediction is further away from the observed value, causing the residual to increase. This hypothesis is further supported by the results in the next section that explore the variance in estimation due to the number of splits.

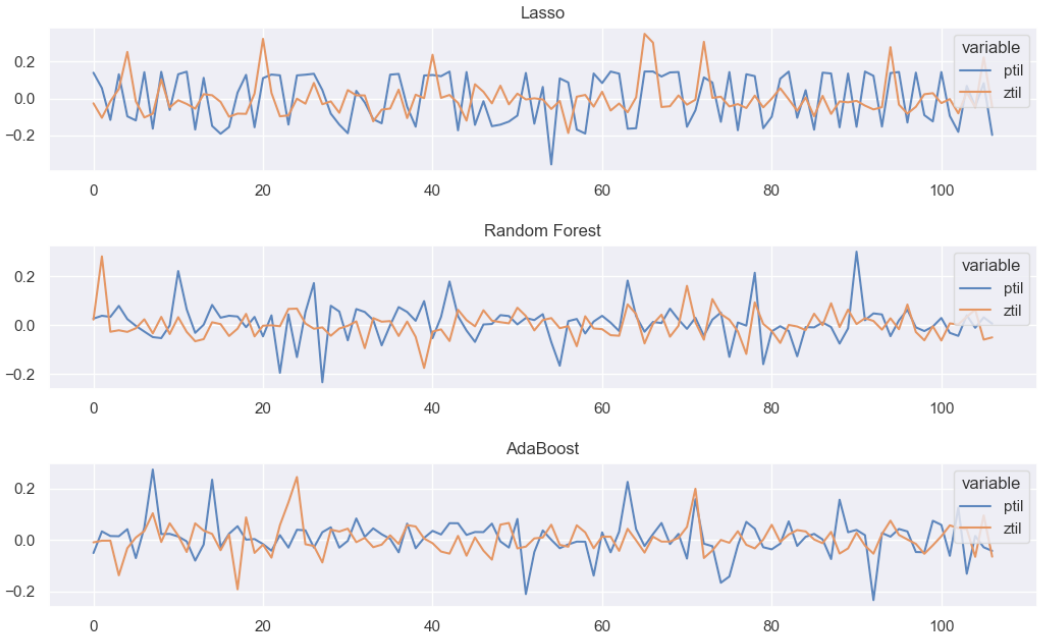

Figure 11 The residualized variables for price and product cost

Figure 12 illustrates how sensitive the estimate of price elasticity is with regards to the number of partitions of the data, that is, the number of folds. Each ML method is displayed separately both with and without an instrument, where the y-axis is the price elasticity estimate and the x-axis shows number of folds. Recall from the DML sample splitting procedure illustrated in figure 7, that if the data is split into K folds, $\frac{(K-1)}{K} * number\ of\ observations$ datapoints are used in the ML prediction and $\frac{1}{K} * number\ of\ observations$ datapoints are used to calculate the residuals. For each method

and each fold, 10 iterations were made with different random partitions, which are represented by the shaded area. As is pointed out in Chernozhukov et al. (2018), the number of folds has no asymptotic impact in their tests, but they also note that with smaller samples the choice of folds can matter. This is seen in figure 12 where the estimate of price elasticity stabilizes for most of the methods at approximately 5 folds. The AdaBoost with an instrument stabilizes at around 27 folds and the Random Forest with an instrument does not seem to stabilize at any point within the 40 rounds tested. This means that the methods exhibit high variance. Interestingly, AdaBoost and Random Forest do not exhibit such variance when estimated without an instrument.

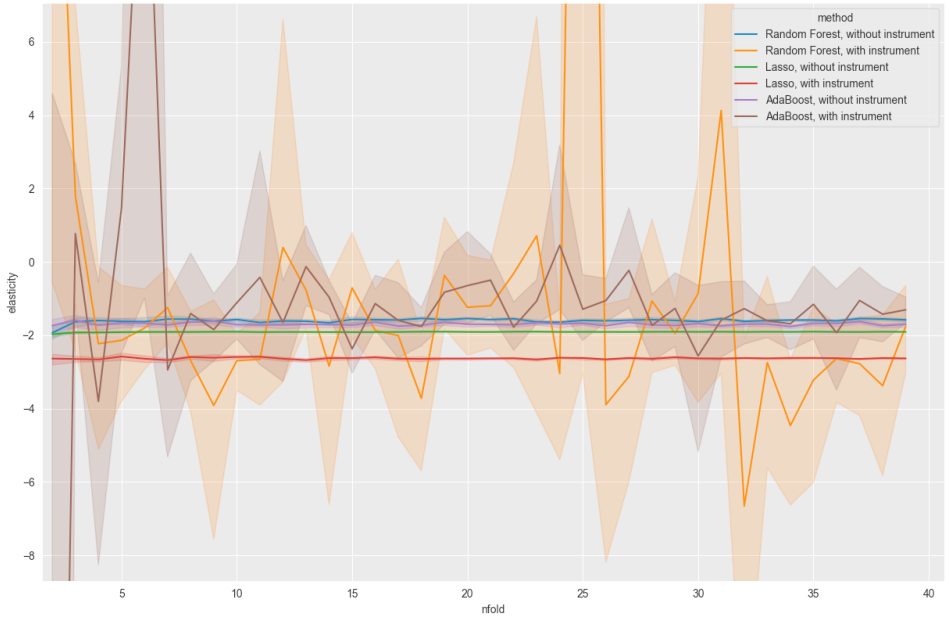

Figure 12 Price elasticity by number of folds

This supports the hypothesis that the high variance of the estimates with Random Forest and AdaBoost have to do with the same phenomena that was explored earlier when analysing at the correlation of the residualized variables. Since the predictions are more sensitive to partitioning, we see lower correlations in the residualized variables than in the underlying data. This would also cause the variation in the estimate of price elasticity that is observed in figure 12.

The next analysis concerns variable importance. In other words, which variables in X_{jt} are important for the prediction of sales, price and product cost. Figure 13 shows the variable importance for the 10 most important variables for $\hat{l}_{i0}(x)$, $\hat{d}_{i0}(x)$ and $\hat{h}_{i0}(x)$, calculated based on decrease in node impurity (MDI) – also called Gini importance. The

MDI is calculated for the Random Forest and AdaBoost, since it is a metric that is only relevant for tree-based methods. MDI measures the number of times a variable is used to split a node, relative to the number of sample splits (Friedman et al., 2001). Looking at the variables that are important for predicting sales – that is $\hat{l}_{i0}(x)$ – we can see that seasonality and lagged variables of the demand system are most important. This means that capturing these variables is important for estimating price elasticity well. As for the actual product characteristics, pattern, colour, collar and button colour were also in the top ten.

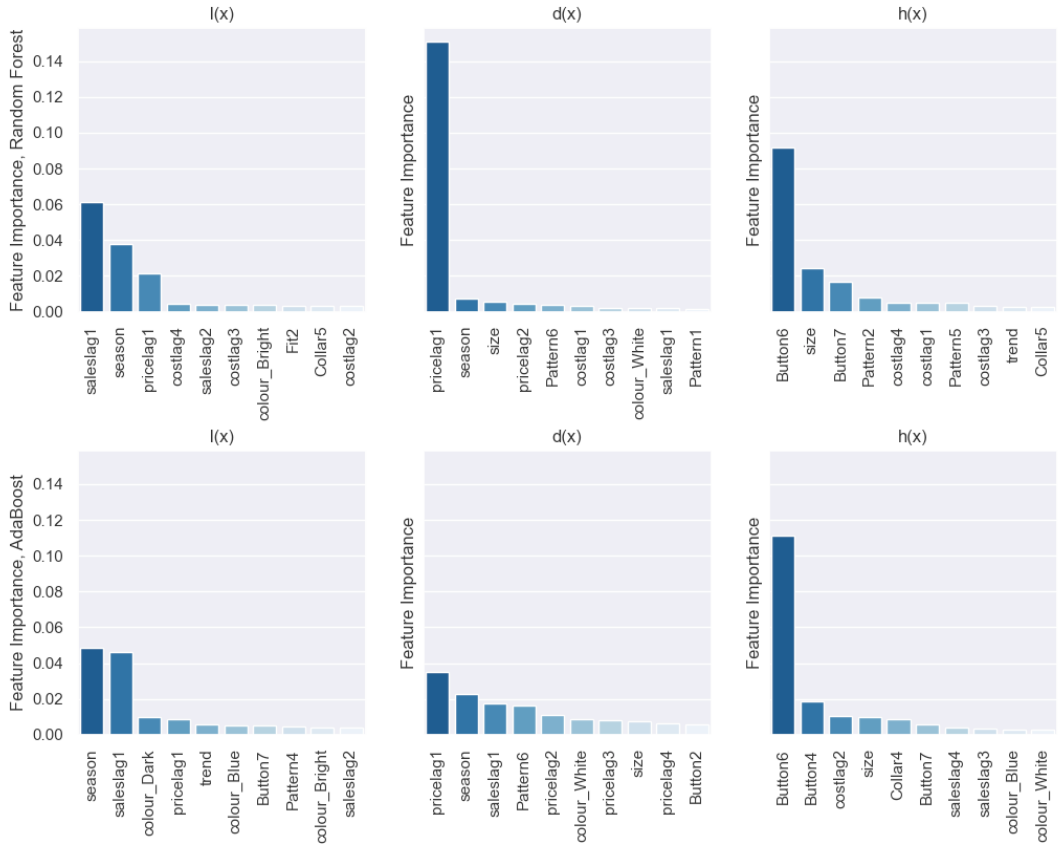

Figure 13 Variable importance when predicting sales, price and product cost

In the middle charts of figure 13 we can see the important variables for predicting price – $\hat{d}_{i0}(x)$. The result is very similar as before: seasonality and the one period lag are the most important followed by the product characteristics: colour, pattern, collar and button colour. Product cost does not exhibit the same type of seasonal behaviour as sales or prices do and the actual product characteristics play a more important role in prediction. We can see that both for Random Forest and AdaBoost, button is in the top two, following size and the lagged variable of cost.

6 DISCUSSION

As stated in chapter one, the goal of the thesis was to estimate the own price elasticity of demand using double machine learning (DML) and to compare the results to traditional econometric methods. In this chapter I will discuss the results, their implications, and limitations of the thesis. The results show that using DML to estimate price elasticity is suitable and does not lead to differing conclusions regarding the price elasticity of the product group “shirts” when compared to traditional methods. This is in line with the empirical estimations in Chernozhukov et al. (2018).

However, some of the ML methods exhibited high variance in the DML first stage prediction, which seems to have been the primary cause of diverging results. Since the estimators predicted poorly at times, the correlation between the residualized instrument and residualized sales was low, biasing the estimate of price elasticity. Although all estimations indicated a weak instrument, this was most notable when using Random Forest and AdaBoost as first stage estimators. Furthermore, the results from estimations without an instrument showed lower divergence. This highlights the importance of the first stage estimation. If the prediction is a truly hard problem, and one is unable to reach satisfactory performance, DML might not be a suitable alternative. Furthermore, this thesis assumes a simple demand model and it is not clear whether more complex models would have yielded in greater difference between the different methods of estimation.

The results from the auxiliary analysis also provide insightful information on variable importance, and on how the variable selection can be automatised with ML methods. This is one of the central reasons to use ML methods in economics in the first place (Varian, 2014). The results show that the product characteristics are not of most importance when estimating the first stage equations, indicating that they are not crucial for demand estimation either. This is most likely the reason why the results of the linear and partially linear models where similar. On the other hand, the unimportance of product characteristics might also be due to the fact that hard-coded characteristics are susceptible to human biases. For example, coding a shirt as “blue” does not fully capture the visual component of the shirts colour. DML opens the possibility of including product characteristics in demand estimation via a convolutional neural network (CNN), where the algorithm learns the characteristics from images as explored in Quan (2019). Additionally, other forms of data such as text become possible to implement in demand estimations.

Even though this thesis does not analyse the cross-price effects of products, Chernozhukov et al. (2017) show how DML can be used to estimate cross-price elasticities as well. Using CNNs to include product characteristics and estimating the substitution patterns of products with an ML method could provide insight into automated ways of defining the relevant market for a product; a field of future research that Compiani (2018) also acknowledges as important.

The results also give feedback to the target firm regarding its pricing policy. The estimates of price elasticity range from -1.25 to -3.4, indicating that a 1% increase in price would lead to a 1.25 to 3.4 per cent decrease in demand. Using the Lerner equation:

$$\frac{P - MC}{P} = \frac{1}{\epsilon}$$

one can calculate the optimal pricing policy for a monopolist (Varian, 2010). The target firm is not a monopoly but via the Lerner equation one can at least assert if the current pricing policy is too high, meaning above monopoly pricing. This is not the case the for the target firm. In general, the results demonstrate how DML can be a useful way for firms to estimate price elasticities to improve their pricing. Especially if a firm already has a pipeline consisting of prediction methods, one can build a DML based method on top of existing infrastructure. My analysis does not include rich data of demographic variables, but firms with access to such variables can easily include them in a method such as DML. This potentially opens the door to more efficient personalised pricing and price discrimination. As shown in Shiller (2013), the implication on welfare allocations that rise from personalised pricing are not obvious, but nonetheless the use of more sophisticated pricing technologies can become concern for competition policy.

In summation, the results from this thesis are in line with to those of Chernozhukov et al. (2018) and support the use of DML in empirical applications in economics. However, whether DML should be considered as the method for estimating price elasticity depends on the problem at hand, e.g., on what assumptions are made and how well do they fit DML. Furthermore, there is another pragmatic concern regarding what data is available and how it looks like. Lastly, of course, the question about what first stage equations are predictable with modern ML methods persists.

7 CONCLUSIONS

Machine learning (ML) methods have become more common in the economics literature and new methods, combining traditional econometric approaches with ML, have become available (Athey & Imbens, 2019). This thesis explores this trend by looking at how a new econometric method called double/debiased machine learning (DML) by Chernozhukov et al. (2018) performs in demand estimation. The results of DML are compared to those of a standard linear instrumental variables regression. The DML approach allows the use of a high dimensional dataset, which includes variables of product characteristics, seasonal and temporal variables.

The results from this thesis show little difference in the price elasticity estimates obtained by DML to those obtained by traditional methods. Similarly, as in Chernozhukov et al. (2018), the estimates from DML are sensitive to the performance of the first stage estimators, which highlights the importance of choosing the appropriate method and configure it properly. However, as done in my analysis, one can use several different ML methods as first stage estimators and essentially horserace them to find the best performer. Furthermore, the results indicate that many product characteristics are not important for the first stage estimators, meaning that these characteristics are not crucial for estimating price elasticity either. Although, it is not clear whether this is because the characteristics are truly unimportant for demand estimation or if the hard-coded variables do not properly capture the visual appearance of the products.

In conclusion, this thesis supports the general argument in Athey & Imbens (2019), Chernozhukov et al. (2018) and Varian (2014), that ML methods can be used in applied economics and even combined with traditional econometric methods. By doing this, one can use high-dimensional datasets, images, or text as data and explore other than linear relationships between variables. This can all be done while still obtaining estimates that can be interpreted causally. These new algorithms open many doors and depending on one's goal, they can provide a very useful tool.

REFERENCES

Ackerberg, D., Benkard, C. L., Berry, S., & Pakes, A. (2007). Econometric tools for analyzing market outcomes. *Handbook of econometrics*, 6, 4171-4276.

Aguirregabiria, V. (2018). Empirical industrial organization: Models, methods, and applications. *University of Toronto*.

Armstrong, T. B. (2016). Large market asymptotics for differentiated product demand estimators with economic models of supply. Econometrica, 84(5), 1961-1980.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685-725.

Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2020), Doub*leML - Double Machine Learning in Python*. URL: https://github.com/DoubleML/doubleml-for-py, Python-Package version 0.1.2.

Bareinboim, E., Forney, A., & Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 1342-1350.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608-650.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 242-262.

Berry, S., & Haile, P. (2016). Identification in differentiated products markets. *Annual review of Economics*, 8, 27-52.

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

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brown, Z. Y., & MacKay, A. (2021). Competition in pricing algorithms (No. w28860). *National Bureau of Economic Research*.

Calvano, E., Calzolari, G., Denicolo, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267-97.

Chernozhukov, V., Goldman, M., Semenova, V., & Taddy, M. (2017). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. arXiv preprint arXiv:1712.09988.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.

Compiani, G. (2018), Nonparametric Demand Estimation in Differentiated Products Markets. Available at SSRN: https://ssrn.com/abstract=3134152 or http://dx.doi.org/10.2139/ssrn.3134152

Crowley, P. M. (2007). A guide to wavelets for economists. *Journal of Economic Surveys*, 21(2), 207-267.

Cunningham, S. (2018). Causal inference: The mixtape (V. 17). *Tufte-Latex. GoogleCode. com*.

Deaton, A., & Muellbauer, J. (1980). An almost ideal demand system. *The American economic review,* 70(3), 312-326.

Drucker, H. (1997). Improving regressors using boosting techniques. *ICML* (Vol. 97, pp. 107-115).

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*: Journal of the Econometric Society, 387-401.

Fumi, A., Pepe, A., Scarabotti, L., & Schiraldi, M. M. (2013). Fourier analysis for demand forecasting in a fashion company. *International Journal of Engineering Business Management*, 5(Godište 2013), 5-30.

Gandhi, A., & Houde, J. F. (2019). Measuring substitution patterns in differentiated products industries (No. w26375). *National Bureau of Economic Research*.

Gillen, B. J., Montero, S., Moon, H. R., & Shum, M. (2019). BLP-2LASSO for aggregate discrete choice models with rich covariates. *The Econometrics Journal*, 22(3), 262-281.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 1-12.

Hausman, J. A. (1996). Valuation of new goods under perfect and imperfect competition. *The economics of new goods* (pp. 207-248). University of Chicago Press.

Hausman, J., & McFadden, D. (1984). Specification tests for the multinomial logit model. Econometrica: *Journal of the Econometric Society*, 1219-1240.

Imbens, G. (2014). Instrumental variables: an econometrician's perspective (No. w19983). *National Bureau of Economic Research*.

Keane, M. P., & Neal, T. (2021) A New Perspective on Weak Instruments. UNSW Economics Working Paper No. 2021-05, Available at SSRN: https://ssrn.com/abstract=3846841 or http://dx.doi.org/10.2139/ssrn.3846841

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1), 4-22.

Lange, H., Brunton, S. L., & Kutz, J. N. (2021). From Fourier to Koopman: Spectral Methods for Long-term Time Series Prediction. *J. Mach. Learn. Res.*, 22, 41-1.

Lee, D. L., McCrary, J., Moreira, M. J., & Porter, J. (2020). Valid t-ratio Inference for IV. arXiv preprint arXiv:2010.05058.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304), 993-1010.

Mackey, L., Syrgkanis, V., & Zadik, I. (2018). Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning* (pp. 3375-3383). *PMLR*.

Marschak, J. (1960), 'Binary choice constraints on random utility indications', in K. Arrow, ed., *Stanford Symposium on Mathematical Methods in the Social Sciences*, Stanford University Press, Stanford, CA, pp. 312–329.

McFadden, D. (1987). Regression-based specification tests for the multinomial logit model. *Journal of econometrics*, 34(1-2), 63-82.

McFadden, D. (2001). Economic choices. *American economic review*, 91(3), 351-378.

McFadden, D., Tye, W. B., & Train, K. (1977). An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model (pp. 39-45). *Berkeley: Institute of Transportation Studies*, University of California.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

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

Pardoe, D., & Stone, P. (2010). Boosting for regression transfer. *ICML*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.

Quan, W. T. (2019). Extracting Characteristics from Product Images and its Application to Demand Estimation. University of Georgia, Department of Economics.

Schleicher, C. (2002). An introduction to wavelets for economists (No. 2002-3). Bank of Canada.

Sheppard, K., Lewis, B., Guangyi, Wilson, K., Thrasibule, Xavier RENE-CORAIL, & vikjam. (2021). bashtage/linearmodels: Release 4.24 (Version v4.24). Zenodo. http://doi.org/10.5281/zenodo.4671862

Shiller, B. R. (2013). First degree price discrimination using big data (p. 32). Brandeis Univ., Department of Economics.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4), 518-529.

Sutton, R. S., & Barto, A. G. (1998). Introduction to reinforcement learning (Vol. 135). Cambridge: MIT press.

Theil, H. (1975). The theory of rational random behavior and its application to demand analysis. *European Economic Review*, 6(3), 217-226.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4), 273.

Train, K. E. (2009). Discrete choice methods with simulation (chapter 4). *Cambridge university press.*

Trovò, F., Paladino, S., Restelli, M., & Gatti, N. (2015). Multi-armed bandit for pricing. In *12th European Workshop on Reinforcement Learning* (pp. 1-9).

Varian, H. R. (2010). Intermediate microecomomics: Modern approach Ed. 4.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-28.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

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

APPENDIX 1 RED-BUS-BLUE-BLUE EXAMPLE

The red-bus-blue-bus example is a common illustration in the demand estimation literature and this version follows largely Train (2009). The aim is to illustrate the importance of the unobserved term in discrete choice models. The unobserved term will in many cases imply restrictions and behavioural assumptions on our model. Furthermore, the example suits well to demonstrate the IIA in logit models.

Assume a person chooses between taking a red bus or the car as a form of transportation. The observable variables impacting this choice are the duration of the travel (T) and the cost of transportation (C). The utilities for the respective choices are modelled linearly:

$$U_c = \alpha T_c + \beta C_c + \varepsilon_c$$

$$U_{rb} = \alpha T_{rb} + \beta C_{rb} + \varepsilon_{rb}$$

where the representative utility for each choice is:

$$V_c = \alpha T_c + \beta C_c$$

$$V_{rb} = \alpha T_{rb} + \beta C_{rb}.$$

With data, one could estimate the equations and obtain estimates for the parameters α and β. Note, that even though we estimate the equation we are still ambiguous to the persons choice due to the unobservable part. If $V_c = 4$ and $V_{rb} = 5$, then the person chooses to take the car if the unobserved factors for the car are more than one unit than that of the bus. Therefore, to determine the probability that the person chooses the car, we need to know the probability that $\varepsilon_c - \varepsilon_{rb} > 1$. (Train, 2009)

In a logit model we assume that ε_c and ε_{rb} are distributed identically and independently and that they follow a type 1 extreme value distribution. Thus, we can formulate the choice probabilities for the respective choices as:

$$P_c = \frac{e^{\alpha T_c + \beta C_c}}{e^{\alpha T_c + \beta C_c} + e^{\alpha T_{rb} + \beta C_{rb}}}$$

$$P_{rb} = \frac{e^{\alpha T_{rb} + \beta C_{rb}}}{e^{\alpha T_c + \beta C_c} + e^{\alpha T_{rb} + \beta C_{rb}}}.$$

According to the IIA, the ratio of the choice probability,

$$\frac{P_c}{P_{rb}} = \frac{e^{\alpha T_c + \beta C_c}}{e^{\alpha T_{rb} + \beta C_{rb}}} = e^{\alpha T_c + \beta C_c - \alpha T_{rb} + \beta C_{rb}}$$

is independent of other alternatives. Assuming that $P_c = P_{br} = \frac{1}{2}$, adding a new alternative to the choice set – a blue bus – would imply that the new choice probabilities must be $P_c = P_{br} = P_{bb} = \frac{1}{3}$. This does not fall well with our intuition. We would not expect the probability of choosing a car to decrease by introducing the blue bus alternative. In reality we might expect something closer to $P_c = \frac{1}{2}$, $P_{br} = P_{bb} = \frac{1}{4}$. The key takeaway is that by introducing new alternatives that are similar to some already existing ones, we might overestimate how much the probability of other alternatives decrease. Also, in economic terms the IIA would imply that all choices can be substituted in the same way, that cross-price elasticities are the same for all choices. (Train, 2009)

APPENDIX 2 DESCRIPTIVE STATISTICS

Table 7 Descriptive statistics of all variables

Variable	n	mean	std	min	max
Log(sales)	108	3.67	0.48	2.64	4.57
Log(price)	108	4.19	0.13	3.84	4.34
Log(cost)	108	2.73	0.09	2.55	3.08
size	108	39.98	1.16	31.48	41.53
colour_Blue	108	0.57	0.1	0.31	0.81
colour_Bright	108	0.04	0.04	0.0	0.18
colour_Dark	108	0.02	0.03	0.0	0.14
colour_Red	108	0.04	0.04	0.0	0.16
colour_White	108	0.33	0.1	0.08	0.57
Fit1	108	0.16	0.1	0.0	0.43
Fit2	108	0.54	0.12	0.28	0.79
Fit3	108	0.3	0.12	0.06	0.72
Pattern1	108	0.07	0.05	0.0	0.24
Pattern2	108	0.35	0.12	0.17	0.72
Pattern3	108	0.03	0.04	0.0	0.21
Pattern4	108	0.04	0.05	0.0	0.21
Pattern5	108	0.41	0.13	0.11	0.74
Pattern6	108	0.09	0.06	0.0	0.26
Material1	108	0.95	0.08	0.71	1.0
Material2	108	0.05	0.08	0.0	0.29
Collar1	108	0.1	0.11	0.0	0.37
Collar2	108	0.01	0.03	0.0	0.19
Collar3	108	0.0	0.0	0.0	0.03

Collar4	108	0.07	0.09	0.0	0.36
Collar5	108	0.2	0.12	0.0	0.46
Collar6	108	0.62	0.14	0.31	1.0
Collar7	108	0.0	0.0	0.0	0.03
Button1	108	0.0	0.0	0.0	0.03
Button2	108	0.24	0.11	0.06	0.61
Button3	108	0.0	0.01	0.0	0.05
Button8	108	0.06	0.06	0.0	0.22
Button4	108	0.01	0.02	0.0	0.12
Button5	108	0.0	0.0	0.0	0.01
Button6	108	0.02	0.05	0.0	0.28
Button7	108	0.68	0.17	0.17	0.93
Cuff1	108	0.04	0.04	0.0	0.24
Cuff2	108	0.96	0.04	0.76	1.0
season	108	3.82	0.34	3.36	4.31
trend	108	-0.15	0.08	-0.29	-0.01
saleslag1	108	3.67	0.48	2.64	4.57
saleslag2	108	3.68	0.48	2.64	4.57
saleslag3	108	3.68	0.48	2.64	4.57
saleslag4	108	3.69	0.49	2.64	4.83
pricelag1	108	4.19	0.13	3.84	4.34
pricelag2	108	4.19	0.13	3.84	4.34
pricelag3	108	4.19	0.13	3.84	4.34
pricelag4	108	4.19	0.13	3.84	4.34
costlag1	108	2.73	0.09	2.55	3.08
costlag2	108	2.73	0.09	2.55	3.08

costlag3	108	2.73	0.09	2.55	3.08
costlag4	108	2.73	0.09	2.55	3.08

APPENDIX 3 DESCRIPTION OF MACHINE LEARNING ALGORITHMS

This appendix shows in detail the Random Forest and AdaBoost algorithms used in this thesis. Both the Random Forest and AdaBoost use a CART algorithm (simply referred as regression tree) as a base learner which is described in Friedman et al. (2001) pages 305-307. Both the Random Forest and AdaBoost have classification and regression variants and only the latter are presented here.

Random Forest

The thesis uses the specification of a Random Forest as laid out in Breiman (2001) as it is the version that is implemented in the SciKit-learn library. As recommended by the authors, I use $m = \frac{p}{3}$ that is typically seen in regression versions of the Random Forest. The notation follows Friedman et al. (2001).

Random Forest for regression:

1. For $b = 1$ to B
    a. Draw a bootstrap sample $\boldsymbol{Z}^*$ of size N from the training data.
    b. Grow a regression tree T_b to the bootstrapped data, for node of the tree, until the minimum node size n_{min} is reached.
        i. Select m variables at random from p at each split, $m = \frac{p}{3}$.
        ii. Choose the best variable split point among the m variables.
        iii. Split the node into two subnodes.
2. Output the ensemble of regression trees $\{T_b\}_1^B$.
3. Make a prediction at a new point x:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

AdaBoost

The version of AdaBoost that is implemented in the SciKit-learn package is called AdaBoost.R2 and is specified in Drucker (1997). The notation follows the one used in Pardoe & Stone (2010).

AdaBoost.R2:

1. Initialize the weiths $w_i = \frac{1}{N}, i = 1,2, \dots N$
2. For $n = 1$ to N
    a. Fit a regression tree on the training data using the weiths w_i and obtain a hypothesis $h_t: X \longrightarrow \mathbb{R}$.
    b. Calculate the adjusted error e_i^t for every instance:
        i. Let $D_t = max_{j=n}^{n} \left| y_j - h_t(x_j) \right|$ and $e_i^t = \frac{|y_i - h_t(x_i)|}{D_t}$
    c. Calculate the adjusted error for h_t:
        i. $\epsilon_t = \sum_{i=1}^{n} e_i^t \, w_i^t$
        ii. If $\epsilon_t \geq 0.5$, stop and set $N = t - 1$
    d. Let $\beta_t = \epsilon_t (1 - \epsilon_t)$
    e. Update the weight vector: $w_i^{t+1} = w_i^t {\beta_t}^{1 - e_i^t} / Z_t$
3. Obtain the hypothesis $h_f(x)$ = weighted median of $h_t(x)$, for $1 \leq t \leq N$, using $\ln\left(\frac{1}{\beta_t}\right)$ as the weights.

APPENDIX 4 ADDITIONAL RESULTS

Estimation results of the partially linear model with instrument (DML second stage results)

Table 8 Second stage DML estimation with Lasso

IV-2SLS Regression Results

Dep. Variable:	ytil	R-squared:	0.3105
Estimator	IV-2SLS	Adj. R-squared:	0.3039
No. Observations:	108	F-statistic:	6.9558
Cov. Estimator:	robust	P-value (F-stat)	0.0084

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	-0.0001	0.0319	-0.0035	0.9972	-0.0627	0.0625
dtil	-2.8013	1.0622	-2.6374	0.0084	-4.8831	-0.7195

First Stage Estimation Results

Dep. Variable	dtil
R-squared:	0.0420
Partial R-squared	0.0420
Partial F-statistic	9.0388
P-value (Partial F-stat)	0.0026
Partial F-stat Distn	chi2(1)
Intercept	0.0001 (0.0086)*
ztil	0.2931 (3.0065)*

*t-statistics reported in parenthesis

Table 9 Second stage DML estimation with Random Forest

IV-2SLS Regression Results

Dep. Variable:	ytil	R-squared:	0.0442
Estimator	IV-2SLS	Adj. R-squared:	0.0351
No. Observations:	108	F-statistic:	0.0806
Cov. Estimator:	robust	P-value (F-stat)	0.7765
		Distribution:	chi2(1)

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	0.0123	0.1376	0.0893	0.9288	-0.2574	0.2820
dtil	-2.9111	10.256	-0.2839	0.7765	-23.012	17.189

First Stage Estimation Results

Dep. Variable	dtil
R-squared:	0.0014
Partial R-squared	0.0014
Partial F-statistic	0.1760
P-value (Partial F-stat)	0.6749
Partial F-stat Distn	chi2(1)
Intercept	0.0125 (1.6912)*
ztil	0.0530 (0.4195)*

*t-statistics reported in parenthesis

Table 10 Second stage DML estimation with AdaBoost

IV-2SLS Regression Results

Dep. Variable:	ytil	R-squared:	0.1308
Estimator	IV-2SLS	Adj. R-squared:	0.1226
No. Observations:	108	F-statistic:	0.1452
Cov. Estimator:	robust	P-value (F-stat)	0.7032
		Distribution:	chi2(1)

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	-0.0229	0.0443	-0.5178	0.6046	-0.1097	0.0639
dtil	-1.2465	3.2711	-0.3811	0.7032	-7.6577	5.1647

First Stage Estimation Results

Dep. Variable	dtil
R-squared:	0.0151
Partial R-squared	0.0151
Partial F-statistic	1.1928
P-value (Partial F-stat)	0.2748
Partial F-stat Distn	chi2(1)
Intercept	0.0078 (1.1487)*
ztil	0.1519 (1.0921)*

*t-statistics reported in parenthesis

Estimation results of the partially linear model without instrument (DML second stage results)

Table 11 Second stage DML estimation with Lasso

OLS Regression Results

Dep. Variable:	ytil	R-squared:	0.404
Estimator	OLS	Adj. R-squared:	0.399
No. Observations:	108	F-statistic:	71.92
Cov. Estimator:	nonrobust	P-value (F-stat)	1.44e-13

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	-0.0002	0.030	-0.007	0.994	-0.060	0.059
dtil	-1.8906	0.223	-8.481	0.000	-2.333	-1.449

Table 12 Second stage DML estimation with Random Forest

OLS Regression Results

Dep. Variable:	ytil	R-squared:	0.135
Estimator	OLS	Adj. R-squared:	0.127
No. Observations:	108	F-statistic:	16.57
Cov. Estimator:	nonrobust	P-value (F-stat)	9.07e-05
		Log-Likelihood:	-27.722

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	-0.0042	0.031	-0.138	0.890	-0.065	0.057
dtil	-1.5990	0.393	-4.070	0.000	-2.378	-0.820

Table 13 Second stage DML estimation with AdaBoost

OLS Regression Results

Dep. Variable:	ytil	R-squared:	0.143
Estimator	OLS	Adj. R-squared:	0.135
No. Observations:	108	F-statistic:	17.67
Cov. Estimator:	nonrobust	P-value (F-stat)	5.51e-05

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	-0.0186	0.030	-0.616	0.539	-0.079	0.041
dtil	-1.7558	0.418	-4.203	0.000	-2.584	-0.928

Estimation results of linear model

Table 14 Estimation of the linear model with instrument, IV

IV-2SLS Regression Results

Dep. Variable:	Log(sales)	R-squared:	0.5468
Estimator	IV-2SLS	Adj. R-squared:	0.4848
No. Observations:	108	F-statistic:	163.55
Cov. Estimator:	robust	P-value (F-stat)	0.0000
		Distribution:	chi2(1)

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	2.7837	13.841	0.2011	0.8406	-24.344	29.911
Size	-0.0101	0.0318	-0.3183	0.7503	-0.0725	0.0523
Fit2	-1.5365	0.5326	-2.8850	0.0039	-2.5803	-0.4927
Fit3	-1.3786	0.5224	-2.6387	0.0083	-2.4025	-0.3546
Button2	17.024	9.4178	1.8076	0.0707	-1.4348	35.482
Button3	26.082	10.298	2.5328	0.0113	5.8988	46.265
Button4	13.529	9.7287	1.3907	0.1643	-5.5386	32.597
Button5	29.627	23.678	1.2513	0.2108	-16.781	76.035
Button6	17.538	9.6592	1.8157	0.0694	-1.3935	36.470
Button7	16.433	9.4397	1.7408	0.0817	-2.0688	34.934
Button8	16.269	9.3976	1.7311	0.0834	-2.1503	34.688
Cuff2	-0.1337	1.4484	-0.0923	0.9264	-2.9725	2.7050
Season	0.0860	0.3979	0.2160	0.8290	-0.6940	0.8659
Log(price)	-3.3970	1.7379	-1.9547	0.0506	-6.8032	0.0091

First Stage Estimation Results

Dep. Variable	Log(price)
R-squared:	0.6193
Partial R-squared	0.0413
Partial F-statistic	4.6511
P-value (Partial F-stat)	0.0336
Partial F-stat Distn	chi2(1)
Intercept	6.8580 (2.6560)*
Size	0.0184 (2.1259)*
Fit2	-0.2646 (-2.8607)*
Fit3	-0.2577 (-2.3686)*
Button2	-0.0343 (-0.0137)*
Button3	-1.2612 (-0.4590)*
Button4	1.0498 (0.4065)*
Button5	0.6599 (0.1042)*
Button6	-0.4699 (-0.1841)*
Button7	-0.3644 (-0.1450)*
Button8	-0.0655 (-0.0261)*
Cuff2	-0.8930 (-3.8391)*
Season	-0.2289 -8.6620)*
Log(cost)	-0.4325 (-2.1566)*

*t-statistics reported in parenthesis

Table 15 Estimation of the linear model without instrument, OLS

OLS Regression Results

Dep. Variable:	Log(sales)	R-squared:	0.6743
Estimator	OLS	Adj. R-squared:	0.6293
No. Observations:	108	F-statistic:	223.63
Cov. Estimator:	nonrobust	P-value (F-stat)	0.0000

Parameter Estimates

	Coef	Std err	t	P > |t|	[0.025	0.975]
Intercept	-7.1699	8.5854	-0.8351	0.4036	-23.997	9.6571
Size	-0.0253	0.0248	-1.0206	0.3075	-0.0739	0.0233
Fit2	-1.1334	0.3116	-3.6379	0.0003	-1.7441	-0.5228
Fit3	-1.0394	0.3536	-2.9396	0.0033	-1.7324	-0.3464
Button2	17.040	8.3216	2.0476	0.0406	0.7295	33.350
Button3	27.158	9.0502	3.0008	0.0027	9.4198	44.896
Button4	12.363	8.5355	1.4484	0.1475	-4.3665	29.092
Button5	30.319	20.913	1.4497	0.1471	-10.671	71.308
Button6	19.022	8.4356	2.2549	0.0241	2.4884	35.555
Button7	17.004	8.3260	2.0423	0.0411	0.6853	33.323
Button8	16.480	8.3017	1.9851	0.0471	0.2089	32.751
Cuff2	1.0596	0.7399	1.4322	0.1521	-0.3905	2.5098
Season	0.4672	0.1111	4.2066	0.0000	0.2495	0.6849
Log(price)	-1.6787	0.3120	-5.3802	0.0000	-2.2903	-1.0672