Hedonic Price Analysis of Online Bread Prices

Mauri Yli-Liipola
Master's thesis
Agricultural Economics
Department of Economics
and Management
University of Helsinki
August 2022

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildingsprogram – Degree Programme	
Agriculture and Forestry	Agricultural, Environmental and Resource Economics	
Tekijä – Författare – Author		
Mauri Yli-Liipola		
Työn nimi – Arbetets titel – Title		
Hedonic Price Analysis of Online Bread Prices		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track		
Agricultural Economics		
Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's thesis	August 2022	75

Tiivistelmä – Referat – Abstract

Increased awareness in health issues and nutritional elements has derived consumers to pay premium prices for functional products and producers have developed differentiated products to match these heterogenous preferences. The assessment of the versatile food commodity markets has long intrigued the attention of researchers and it is well-established that the hedonic pricing method is a prominent approach in determining an attribute's cost and perceived value to customers. With the method of hedonic pricing, this study aims to determine how those values form the online prices of bread. Specifically, it examines the marginal prices of bread sold on e-commerce platforms in Finland and whether those change over time. Moreover, this study aims to analyze the extent to which complex data extraction methods add value to prior hedonic pricing methodologies.

To test the null hypotheses that estimated marginal prices have no relationship to online prices of bread, and the marginal prices of bread attributes do not significantly change during the data collection period, a semi-logarithmic hedonic price function with HAC standard errors was specified. In the empirical part, the hedonic price function was estimated for the whole sample and each data day. Daily price quotes for the analysis were extracted from the largest Finnish grocery retailer's e-commerce platform via web scraping. The results showed that the estimated relationships between price and marginal prices were significant but that the effects on the price did not significantly change during the data collection period.

The results suggest that Finnish consumers value taste over health and place a high value on stomach-friendly and domestic bread. These findings are a valuable source for better understanding regular bread purchasing decisions and producer product differentiation strategies. Although web scraping was the only alternative to obtain detailed, up-to-date product data in the Finnish context, daily scraping seemed unnecessary as bread prices remained stable. However, daily scraping combined with the hedonic pricing method yielded valuable information regarding the holiday season's pricing strategies of Finnish retailers. On this basis, web scraping should be included in hedonic pricing applications on food products and the whole food commodity research field.

Avainsanat – Nyckelord – Keywords

Hedonic pricing, hedonic price model, web scraping, online prices, e-commerce, bread

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors

Xavier Irz

Säilytyspaikka – Förvaringställe – Where deposited

E-Thesis

Muita tietoja – Övriga uppgifter – Additional information

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildingsprogram – Degree Programme
Maatalous-metsätieteellinen	Maatalous-, ympäristö- ja luonnonvaraekonomia

Tekijä – Författare – Author
Mauri Yli-Liipola

Työn nimi – Arbetets titel – Title
Leivän verkkokauppahintojen hedoninen hinta-analyysi

Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track
Maatalousekonomia

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Maisterintutkielma	Elokuu 2022	75

Tiivistelmä – Referat – Abstract

Lisääntynyt tietoisuus elintarvikkeiden terveys- ja ravitsemustekijöistä on saanut kuluttajat maksamaan korkeampia hintoja miellyttävistä tuotteista, ja tuottajat ovat tuotekehityksellä pyrkineet vastaamaan näihin heterogeenisiin preferensseihin. Monipuoliset elintarvikemarkkinat ovat kiinnostaneet tutkijoita pitkään ja on osoitettu, että hedoninen hinnoittelumenetelmä on varteenotettava lähestymistapa määritettäessä tuotteen ominaisuuden marginaalihintaa tuottajille ja kuluttajille. Tämän tutkimuksen tavoitteena on selvittää, miten leivän eri attribuutit muodostavat lopputuotteen hinnan vähittäiskaupan verkkoalustoilla hedonisen hinnoittelumenetelmän avulla. Tämä tutkimus analysoi verkkokauppaalustoilla myytävän leivän marginaalihintoja ja niiden vaihtelua Suomessa. Lisäksi tämän tutkimuksen tavoitteena on kartoittaa, missä määrin monimutkaisemmat aineistonkeruumenetelmät, kuten web scraping, tuovat lisäarvoa aikaisempiin tutkimuksiin.

Tutkimuksen nollahypoteesit ovat muodostettu seuraavasti: estimoiduilla marginaalihinnoilla ei ole merkittävää vaikutusta leivän verkkokauppahintoihin ja leivän attribuuttien marginaalihinnat eivät merkittävästi muutu aineistonkeruujakson aikana. Nollahypoteesien testaamiseksi estimoitiin puolilogaritminen hedoninen hintafunktio HAC-keskivirheillä. Tutkimuksen empiirisessä osassa tämä funktio estimoitiin käyttäen hyväksi koko aineistoa sekä erikseen jokaiselle yksittäiselle päivälle aineistossa. Leipätuotteiden hinnat ja attribuutit kerättiin päivittäin Suomen suurimman päivittäistavarakaupan verkkokaupan alustalta web scraping -tekniikan avulla. Tulokset osoittivat, että marginaalihintojen rooli lopputuotteen hinnanmuodostuksessa on merkittävä, mutta attribuuttien vaikutukset hintaan eivät merkittävästi muuttuneet aineistonkeruujakson aikana.

Lisäksi tutkimuksen tulokset viittaavat siihen, että leivän ostopäätökseen vaikuttaa tuotteen maku enemmän kuin terveydelle hyväksi olevat ominaisuudet. Tuloksista huomattiin myös, että vatsaystävällisiä ominaisuuksia ja kotimaista leipää arvostetaan paljon. Nämä havainnot muiden tulosten ohessa luovat arvokkaan pohjan tavanomaisten leivän ostopäätösten ja tuotteiden hinnoittelustrategioiden ymmärtämiseen. Web scraping -tekniikka oli ainoa vaihtoehto yksityiskohtaisen ja ajantasaisen aineiston keräämiseksi Suomessa, mutta päivittäinen aineiston keruu osoittautui tarpeettomaksi leivän hintojen pysyessä vakaina. Päivittäinen aineiston keruu yhdistettynä hedoniseen hinnoittelumenetelmään antoi kuitenkin arvokasta tietoa suomalaisten vähittäiskauppiaiden ja tuottajien toiminnasta sekä kuluttajien käyttäytymisestä. Tulevien tutkimusten tulisi soveltaa web scraping -tekniikkaa elintarviketuotteiden hedonisiin hinnoittelututkimuksiin sekä koko elintarvikealan tutkimukseen.

Avainsanat – Nyckelord – Keywords
Hedoninen hinnoittelu, hedoninen hintamalli, web scraping, verkkokauppahinnat, elektroninen kaupankäynti, leipä

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Xavier Irz

Säilytyspaikka – Förvaringställe – Where deposited
E-Thesis

Muita tietoja – Övriga uppgifter – Additional information

Table of Contents

1 Introduction

The analysis of food markets has long occupied the attention of researchers. This is because food prices have a major effect on the standard of living of a considerable portion of society. Moreover, the movements in the food market overall reflect the global macroeconomic and consumption trends that are happening. For instance, knowledge of ethical and environmental prospects of food production have enhanced the popularity of plant-based diets. Furthermore, increased health and nutritional awareness has derived manufacturers to develop functional foods with enhanced health benefits and appealing flavors for which consumers are willing to pay premium prices. However, not all qualities are positively perceived and generate premiums.

Several methods have been used to explain the pricing of the qualities of differentiated food commodities thus far. One prominent approach, hedonic price modeling, states that a good does not possess utility. Rather, the good comprises of attributes that yield the utility (Lancaster, 1966; Rosen, 1974). By comparing the prices of commodities that possess a certain characteristic to those that do not, it becomes feasible to determine the attribute's perceived worth to customers. This feature of hedonic pricing made it an ideal method for determining the market value of qualities whose prices were not readily apparent.

In various studies, application of hedonic pricing on food commodities allowed valuation of attributes such as the presence of a health or nutritional claim, specific brand, packaging material, or method of production (Costanigro & McCluskey, 2011; Ribeiro et al., 2019). Consequently, the hedonic pricing tool has been implemented for several food products such as yogurt (Bimbo et al., 2016; Bonnano, 2016; Chen et al., 2021), olive oil (Carlucci et al., 2014), coffee (Schollenberg, 2012), and sparkling wine (Fedoseeva, 2020). These studies, among others, are aimed at examining product attributes on American and European countries as well as China and Russia.

At present, there is relatively little research done on the hedonic pricing model for bread and systematic hedonic pricing research has not yet been conducted on Finnish food markets. One exception is Nganje et al. (2008), who studied price premiums of low carbohydrate bread, but the focus of their study was the US market. Bread is a major source of excessive salt intake which is a factor behind non-communicable chronic diseases such as cardiovascular diseases, diabetes, and

some types of cancers (World Health Organization, 2004). As bread has a prominent role in global diets and represents a key component of the global eating habits, it has been an important target of sodium intake reduction programs (Trieu et al., 2015; Di Vita et al., 2016). Because bread is a differentiated product with numerous attributes and nutritional ingredients that are beneficial for human health, the product has attracted much research interest (e.g., Dewettnick et al., 2008; Bruschi et al., 2015; Lobo & Ferreira, 2021).

The understanding of consumers' preferences for bread is essential given the global importance of bread consumption. Except for the hedonic pricing study of Nganje et al. (2008), previous bread studies have focused on stated preference methods with possible hypothetical biases. The absence of research on revealed regular bread consumption preferences makes the product relevant to be assessed with the approach of hedonic pricing.

Data for the hedonic pricing analyses have been obtained mainly by observing products in grocery stores (Carlucci et al., 2013; Caso et al., 2020) and purchasing scanner data (Muth et al., 2013; Edenbrandt et al., 2018; Chen et al., 2021). Additionally, in comparison to these traditional data sources used in the hedonic pricing analyses, there is also research done in the context of e-commerce, however, to a comparatively lesser extent. These few studies, which include Carlucci et al. (2014) and Fedoseeva (2020), have relied on monitoring product data from the websites in a similar manner as in grocery store studies, i.e., directly observing product information.

There is no suitable way to obtain scanner data for this analysis in Finland, and the price of consumer panel data provided by Nielsen is prohibitive. Additionally, monitoring prices from stores or websites is considered laborious and impractical. Observing product prices in stores requires the explicit agreement of the manager of the store, but additionally collecting daily price quotes manually for hundreds of products is not sustainable – even on an e-commerce platform. As such, there exists a need for developing an alternative method of collecting detailed product information in Finland, but additionally these alternative methods of data collection have been overlooked in the food commodity hedonic pricing research. There exists mounting evidence that more advanced and automated data extraction methods, like web scraping, have the potential to reduce the costs of gathering versatile data and improve the variety and precision of future studies (Cavallo & Rigobon, 2016; Hillen, 2019).

Furthermore, although economic theory provides solid justifications for hedonic pricing functions to shift over time (Pakes, 2003), this is not commonly the subject of empirical inquiry. That is, despite the abundance of research on the factors that influence the prices of various product attributes, the subject of whether marginal prices vary over time is often neglected. Time invariance approaches are related to data unavailability and price rigidity in the food industry (Fedoseeva, 2020). However, it is expected that the increasing growth of e-commerce and advanced data extraction techniques would add flexibility to these concerns. In fact, it has been shown that large internet companies, such as Amazon, utilize their ability of dynamic pricing (Hillen, 2020).

1.1 Research questions and objectives

It has been determined that there are a few gaps in the hedonic pricing literature that this study intends to fill. First and foremost, the study aims to estimate the marginal prices of bread sold on e-commerce platforms in Finland. As the objective product attributes have been argued to be more relevant in terms of price determination compared to subjective characteristics (Muñoz et al., 2015), this research aim leads to the formulation of the main research question:

What are the marginal prices of objective attributes of bread sold on an e-commerce platform in Finland?

Moreover, the study aims to examine whether the estimated marginal prices are time invariant. The analysis extends the previous research of Fedoseeva (2020) on Russian e-commerce sparkling wines. It is true that the demand for sparkling wines is intuitively much more volatile than that for bread, but it would not be justified to merely assume *à priori* that the marginal prices of bread attributes remain time invariant. As it has previously been argued that hedonic price functions may shift over time (Pakes, 2003), we formulate the following question to guide the research:

Did marginal prices of bread attributes change during the data collection period spanning several months?

Furthermore, the aim is to promote awareness of web scraping so that future studies in agricultural and food economics may adopt and develop the method. Thus, a comparison to other market data

collection methods used in hedonic pricing research is carried out. Web scraping is definitely not a new method in the area of food market research, but to the best of the author's knowledge, it has never been utilized for hedonic pricing analyses on food market studies. Extracting data via web scraping and reviewing it against existing data collection methods is a prominent contribution not only for hedonic pricing but to the whole food market research. Therefore, the last question guiding the research may be formulated in the following way:

Does web scraping provide novelty value for hedonic pricing analysis compared to existing market data collection methods?

The null hypotheses for this thesis are that the estimated marginal prices have no relationship to online prices of bread. The alternative hypothesis is that the estimated marginal prices are significantly associated with the online prices of bread. The null hypothesis for the second research question is that the marginal prices of bread attributes do not significantly change during the data collection period. The alternative hypothesis is that the marginal prices of bread attributes significantly change during the data collection period.

In order to reach the stated aims and rigorously test the hypotheses, the following issues need to be addressed. Initially, it is essential to define a theoretical framework. Secondly, relevant objective attributes need to be selected. Thirdly, it is necessary to specify an appropriate empirical model and estimation strategy. After that, a web scraping program needs to be developed to collect data. Lastly, results need to be interpreted.

This thesis is divided into six chapters: in the next chapter (#2) the existing hedonic pricing literature is comprehensively reviewed to guide the methodology of the thesis, which is presented in detail in chapter three. After that, results are presented in chapter four and analyzed in chapter five. Lastly, conclusions, limitations, and future research topics are provided in chapter six.

2 Background and literature review

To address the research questions and objectives outlined for this research, a through literature review was carried out. The following sections provide a comprehensive summary of what has previously been published in the literature on hedonic analyses of food markets. The first section provides an overview of the Finnish bread market. The theoretical framework and assumptions of hedonic price modeling are presented in the second section. Following that, the third section focuses on the empirical techniques in relevant studies, methodological issues as well as a comparison of statistical approaches typically utilized in the field of food market hedonic pricing. The fourth section discusses the commonly selected attributes in hedonic pricing food market research. Lastly, a novel data extraction approach from the perspective of hedonic price analysis is presented.

2.1 The Finnish bread market

According to Statistics Finland's (2008) Standard Industrial Classification (TOL) the primary portion of the bakery sector in Finland is the production of soft bread and fresh pastries (TOL 10710) and manufacture of crispbread, rusks, preserved pastry goods, and cakes (TOL 10720). Bread consumption in Finland in 2020 was around 41 kilos per capita, based on both domestic and imported bakery products (Ministry of Economic Affairs and Employment, 2021). This equates to around four slices of fresh bread or slightly more than ten pieces of dry bread every day. The consumption has remained relatively stable throughout the years.

The Covid crisis had a substantial influence on the bread supply chain in 2020, despite the delivery volume of domestic bread remaining unchanged from 2019 (Ministry of Economic Affairs and Employment, 2021). As dining out declined and Finns worked and spent a lot of time at home, bread consumption moved to homes. As much less bread was delivered to schools and employee eateries, cafés, and restaurants, bread purchases focused heavily on retail. Domestic bread deliveries to the hotel and catering sector fell by more than a quarter (26%) in 2020 compared to the previous year (Ministry of Economic Affairs and Employment, 2021) but still the consumption in overall remained unchanged (The Finnish Bread Information, 2020). In addition, the pandemic boosted the growth of online food and bread retailing (Sintonen et al., 2021).

The Finnish Bread Information (2020) pointed out that the trend from the early 2000s to the mid-2010s was the growing popularity of dark bread. Whole grain flour and rye bread were shown to have beneficial health advantages, which fueled the increased demand. However, consumers began to enjoy more oatmeal and oat bread in the middle of the decade, and they have since gained market share over the dark bread. Finns consider rye and oats to be the healthiest cereals, with oats being more stomach-friendly than other cereals. In recent years, the main consumer trends affecting Finnish bread consumption have been related to the increasing importance of taste, domesticity, responsibility, climate-friendliness, health, stomach-friendliness, and the growing popularity of freshly baked products (Ministry of Economic Affairs and Employment, 2021).

Finland has a rich and diversified bread culture in which small, diverse, and traditionally family-run bakery firms play a significant part according to the Ministry of Economic Affairs and Employment's industry report (2021). The bakery sector in Finland is business-intensive, unlike the rest of the food industry. For example, bakeries account for more than a third (35%) of all businesses in the food industry. In 2020, there were 620 bakery enterprises employing over 7 000 people, accounting for 20% of the food industry's employment. The bakery business generated 1.1 billion euros in revenue, accounting for 10% of the total food industry’s revenue. Fazer and Vaasan generated the largest revenue in 2020, with 251,7 and 137,4 million euros, respectively. Lantmännen Unibake Ab Finland, Lantmännen Cerelia Oy, and Myllyn Paras Oy were the next largest in terms of revenue. It is worth mentioning that Myllyn Paras Oy has been sold to Lantmännen.

2.1.1 Special product characteristics

In addition to the different types of bread and cereals as well as the market-leading brands, there are a few special product characteristics in the Finnish market. The Key flag is one of the most well-known Finnish commercial labels. For a product to have this label, it must be made in Finland and include at least 50 percent of domestic material by cost. The Good From Finland mark is the origin label for packaged Finnish products. To receive the Good From Finland label, a product must be manufactured in Finland and include at least 75 percent of domestic ingredients by cost. Meat, fish, eggs, and milk, however, are always entirely domestically if they receive this label. In addition, the Heart label signifies that the product is a healthier alternative. To get the Heart label, bread must have between 6 and 10 grams of fiber and between 0.9 and 1.2 grams of salt per 100 grams. Additionally,

as the S-Group was selected, its own private-label brands Kotimaista and Rainbow are also noteworthy.

2.2 Hedonic pricing theory

In classical consumer theory, consumer choice maximizes the utility function that depends on consumed quantities subject to a budget constraint. However, the traditional theory does not consider the introduction of new commodities or varying qualities (Lancaster, 1966). As an alternative, Lancaster's innovative consumer theory can be summarized in three main points, breaking away from the traditional approach that goods are the objects of utility. First, a good itself does not provide utility. Rather, each good contains utility-bearing characteristics instead. Second, the good has multiple characteristics of which many are shared among other goods. Third, combinations of goods may possess characteristics that differ from those of an individual good.

Lancaster's theory contributed significantly towards the hedonic pricing theory, which was, however, officially formulated a few years later, when Rosen (1974) developed a model emphasizing further the properties of the market equilibrium. Even though Lancaster's and Rosen's models both share the assumption of goods possessing a myriad of consumer-valued characteristics, they also have some fundamental differences. Lancaster's model presumes that goods belong into groups that are then consumed in combinations yielding the desired characteristics. On the contrary, Rosen's model assumes a range of goods from which preferred characteristics are not acquired by purchasing product combinations. Rather, they are acquired after analyzing the "spectrum of products" offering the good with such characteristics. Although Chin and Chau (2003) outlined that the Lancastrian model is more suited to consumer goods, like bread, Rosen (1974) argued that his model considers market equilibrium properties which is why most of the research on the hedonic pricing of food products has used this method. These studies include Carlucci et al. (2013), Szathvary and Trestini (2014), Bimbo et al. (2016), Caso et al. (2020), and Chen et al. (2021).

This paper's presentation of the theoretical framework closely follows Rosen's work on mapping multiple product attributes into price space. Consider a row vector specifying, for any number of products, k characteristics, $z_i = (z_1, z_2, \dots, z_k)$. To allow for marginal analysis, it is assumed that there are large numbers of differentiated products i in the market such that the "spectrum of products"

exists. Each product has a market price p associated with its vector of characteristics revealing the hedonic price function $p(z) = p(z_1, z_2, \dots, z_k)$. The function gives a minimum price for each embedded product characteristic. Only the cheaper one is considered if two different sellers offer the same bundle of characteristics but with different prices, i.e., the consumers are always aware of all the available products and switching between them is costless. Furthermore, many buyers and sellers without market power buy and sell products with the freedom to enter and exit the market.

Rosen's backdrop model has two individual estimation stages. The first stage estimates the marginal prices of z, given $p(z)$, by regressing product prices on corresponding product characteristics, using the best functional form and an ordinary least squares (OLS) procedure. The second stage considers regressing marginal prices against product and, for example, the consumers' sociodemographic characteristics (Costanigro & McCluskey, 2011; Ribeiro et al., 2019). The resulting functions are aggregated inverse demand and supply functions.

While Rosen (1974) did not specifically discuss the functional form for the first-stage estimation, he departed from Lancaster's assumption on the linear relationship between price and characteristics. The latter implied that the marginal prices are constant regardless of the number of characteristics unless the combination of goods consumed changes. Rosen (1974) postulated that since consumers cannot untie and repackage attributes, a non-linear relationship is more realistic. According to Rosen (1974) the implicit price of a characteristic is, thus, not constant but depends on the interaction of supply and demand. Given many heterogeneous consumers and firms, Costanigro and McCluskey (2011) and Ribeiro et al. (2019) made the same assumption. Therefore, marginal prices for each buyer and seller can be recovered from the first order condition $\partial\hat{p}(z)/\,\partial z_k = \widehat{p_{zk}}(z)$, where $\hat{p}(z)$ is the resulting estimate of the hedonic price function $p(z)$.

It is important to emphasize that consumers and producers are assumed to be price takers, thus $p(z^*)$, where asterisk indicates optimum quantities, is defined by market clearing conditions (Ribeiro et al., 2019). As $p(z)$ is assumed to be given, it represents the minimum price consumers are willing to pay. That is, only consumers who have an equal or higher willingness to pay (WTP) to the equilibrium price would consume the good (Bishop & Timmins, 2011; Ribeiro et al., 2019). Analogously, the equilibrium point represents producer willingness to accept (WTA) to sell the product subject to receiving a potential profit. Therefore, utility and profits are maximized when WTP and WTA conditions are tangent to the equilibrium point. However, it is important to note that since both cost

and WTP factors determine the marginal prices, they should not be interpreted merely as consumer WTP as is often done in the literature. Rather, marginal prices represent the price consumers must pay in the market (Costanigro & McCluskey, 2011; Schollenberg, 2012; Edenbrandt et al., 2018; Ribeiro et al., 2019).

The equilibrium condition poses an empirical simultaneity problem as the error terms are likely correlated with independent variables in either the demand or supply equations. Rosen's second stage can be used to solve this problem. In the case of a non-linear price function, however, z and $p(z)$ affect consumer choices simultaneously making z endogenous, i.e., the first-order condition $\widehat{p_{zk}}(z)$ changes with z_k (Ribeiro et al., 2019). Bartik (1987), Epple (1987), Kahn and Lang (1988) and later Bishop and Timmins (2011) among others noted that this leads to a situation where the marginal price of a characteristic varies systematically with its quantity consumed. Therefore, the OLS regression generates biased results in the second stage (Ribeiro et al., 2019). Bishop and Timmins (2011) stressed that there are very few natural exclusion restrictions and only weak instruments to solve this endogeneity problem.

As a result, with a few exceptions, the hedonics literature has ignored Rosen's second stage focusing on recovering marginal prices from the first stage – as does this study. Studies have made certain assumptions in tackling the first simultaneity problem without relying on the second stage. By assuming that the market is in the equilibrium, preferences and technology are distributed uniformly, and an individual consumer cannot affect the market price, the hedonic price function can be presented solely as a function of product attributes (Rosen, 1974; Diamond & Smith, 1985; Costanigro & McCluskey, 2011; Ribeiro et al., 2019). Thus, under those assumptions, a researcher can estimate the hedonic price function, $\hat{p}(z)$, and recover the marginal prices for each characteristic, $\widehat{p_{zk}}(z)$ (Costanigro & McCluskey, 2011; Bonnano, 2016).

2.3 Issues in the formulation of an empirical model

2.3.1 Common functional forms

There is not much guidance on the choice of the proper functional form for the first stage estimation in the hedonic price literature (Butler, 1982; Chin & Chau, 2003; Costanigro & McCluskey, 2011;

Ribeiro et al, 2019). Given the uncertainty surrounding the precise specification and strictly positive price data, several studies have used a strategy that considers transformations of the variables (Edenbrandt et al., 2018). However, to avoid losing the essence of the hedonic price modeling, the value of a characteristic to achieve the best possible prediction accuracy, earlier studies have focused on the simpler transformations, often in the logarithmic form. This simpler specification approach is also supported by the fact that the marginal prices are measured better if the product attribute information is incomplete (Cropper, Deck & McConnell, 1988). Therefore, linear, semi-logarithmic (log-lin), and double logarithmic functional forms are the most often applied functional forms in the hedonic pricing studies (Chin & Chau, 2003; Carlucci et al., 2013; Edenbrandt et al., 2018).

Although the linear specification is the only form for which estimated coefficients are interpreted directly as marginal prices, it has not gained much acceptance in the literature. Bonnano (2016) and Fernández et al. (2019) argued that the linear functional form is not favored because it keeps the marginal effects of the attributes on the price constant. Additionally, hedonic price functions are generally expected to be non-linear (Rosen, 1974; Rasmussen & Zuehlke, 1990). Furthermore, Costanigro and McCluskey (2011) and Fernández et al. (2019) claim that logarithmic transformations have a better control over heteroscedasticity and multicollinearity which, according to Costanigro and McCluskey (2011) and Edenbrandt et al. (2018), are commonly present in these models. Regardless, Nganje et al. (2008) and Caso et al. (2020) estimated the linear hedonic price function, justifying their choice by the ease of interpretation. The linear model has also been sometimes estimated as a comparison or reference model (Schollenberg, 2012; Fedoseeva, 2020; Yang & Dharmasena, 2020) but other specifications have been, nevertheless, chosen as the main model.

Models that employ the natural logarithmic transformation of the dependent variable appear the most frequently in hedonic price studies of food products. Many such papers have at least partly based their choice of this specification on its popularity in previous studies (Schollenberg, 2012; Muth et al., 2013; Muñoz et al., 2015; Bimbo et al., 2016; Bonnano, 2016; Giombi et al., 2018; Edenbrandt et al., 2018; Fedoseeva, 2020; Chen et al., 2021). Although the semi-logarithmic specification is seen as a canonical model, Costanigro and McCluskey (2011) outlined that the adoption of the specification on such grounds is not justified as it is not uncommon for empirical tests to reject that functional form.

The popularity of the semi-logarithmic specification is based on overcoming the mentioned flaws of the linear specification, but at the same time maintaining a convenient result interpretation (Muth et al., 2013; Giombi et al., 2018). Additionally, according to Schollenberg (2012), the semi-logarithmic specification has the potential to significantly increase the model's fit and as the semi-logarithmic functional form has been extensively used, it allows for wider comparison of findings within the same product categories or markets.

In hedonic price research, logarithmic transformations are often applied only on the price variable given that in the model there are often just a few continuous variables (Fedoseeva, 2020). If possible, some have transformed independent variables such as package and container sizes (Carlucci et al., 2013; Carlucci et al., 2014). The choice of the double logarithmic over other specifications has been based on similar reasoning as when choosing the semi-logarithmic form over the linear form. For instance, Kim and Chung (2009), Martínez-Garmendia (2010), and Carlucci et al. (2013) justified the double logarithmic form by the straightforward calculation of elasticities that it allows and its ability to accommodate non-linear patterns. Carlucci et al. (2014) decided to use the double-logarithmic equation for its better capability to fit the data over the semi-logarithmic form.

Other specifications have been seen in the hedonic pricing literature as well. However, the specifications have been rarely estimated as the main model of the study. For example, Fernández et al., (2019), in addition to the linear and the logarithmic specifications, considered a multiplicative inverse transformation of the price variable. Schollenberg (2012) estimated ¼ power and inverse square root specifications. Costanigro & McCluskey (2011) also tested various specifications but like the other mentioned studies, rejected the alternative specifications and estimated the semi-logarithmic equation.

2.3.2 Functional form selection

In addition to the previous criteria to select a functional form, the proper specification of the hedonic pricing model depends also on the nature of the data (Ribeiro et al., 2019). Therefore, data manipulation techniques of which the most popular is a Box-Cox transformation, have been frequently applied in hedonic pricing research to assist in the selection of the functional form (Costanigro & McCluskey, 2011; Szathvary & Trestini, 2014; Bimbo et al., 2016; Ballco and de-Magistris 2018; Giombi et al., 2018; Fernández et al., 2019; Fedoseeva, 2020). The basic principle of

the Box-Cox technique is to nest alternative functional forms by adding non-linear estimates of the lambda parameter (λ) on the dependent variable (Box & Cox, 1964). The parameter ranges from -5 to 5, and its optimal value results in the highest likelihood for the dependent variable and error term to be normally distributed (Box & Cox, 1964; Costanigro & McCluskey, 2011). Depending on the lambda's value, the dependent variable undergoes the transformation in the following way:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \ln y, & \text{if } \lambda = 0. \end{cases} \quad (1)$$

However, it has been shown that by applying such transformations, it is uncertain that the model will fit the data even though it is possible to use more flexible functional forms (Linneman, 1980; Cassel & Mendelsohn, 1985; So et al. 1997; Costanigro & McCluskey, 2011). Furthermore, the Box-Cox transformation has been argued to decrease the accuracy of the coefficients and make the interpretation of the results more complicated than in simpler transformation cases (Cassel & Mendelsohn, 1985). Therefore, in hedonic pricing the technique has been used merely as an indicator towards the most adequate functional form rather than applying the exact transformation as in Equation 1 (Szathvary & Trestini, 2014; Bimbo et al., 2016; Ballco & de-Magistris, 2018; Giombi et al., 2018; Fernández et al., 2019; Fedoseeva, 2020). In this sense, the resulting λ is rounded to the nearest five-tenths value and is then used as the price variable's exponent. If the value is rounded to zero, the natural logarithmic transformation is employed.

With one exception (Bimbo et al., 2016), all the reviewed papers that used the Box-Cox approach were guided to estimate the hedonic pricing function through the logarithmic transformation of the price variable (Szathvary & Trestini, 2014; Ballco & de-Magistris, 2018; Giombi et al., 2018; Fernández et al., 2019; Fedoseeva, 2020). Support provided by this type of computational approach appears to be an additional factor behind the popularity of the logarithmic transformations, given the absence of theoretical guidance regarding the suitable functional form.

Furthermore, the comparison of alternative specifications has also been based on goodness of fit and regression significance metrics. The coefficient of determination, or R^2, has been widely used to compare how well the model explains the variation in the price (e.g., Schollenberg, 2012; Carlucci et al., 2014; Szathvary & Trestini, 2014; Chen et al., 2021). More specifically, as it is common for

hedonic price analyses to use many regressors, a modified version of the R^2, adjusted R^2, has been used. Adjusted R^2 offers a more precise view as it also considers the number of independent variables, unlike the regular R^2.

The F-statistic and its p-value have been used in multiple studies to select a functional form based on the significance of the estimated regressions (Carlucci et al., 2013; Muñoz et al., 2015). Additionally, Ballco and De-Magistris (2018) employed the Akaike (AIC) and Schwarz (BIC) information criterion to choose among specifications. The AIC is an estimator of prediction error and hence it is a quality measure of a statistical model. The BIC is based partly on the likelihood function and is related closely to the AIC.

2.3.3 Time dimension

Pakes (2003) argued that the producer's cost expectations and consumer preferences change over time therefore shifting the hedonic price functions. Still, many hedonic price analyses have not incorporated the time dimension in their models (Nganje et al., 2008; Carlucci et al., 2013; Muth et al., 2013; Carlucci et al., 2014; Muñoz et al., 2014; Ballco & de-Magistris, 2018; Fernández et al., 2019; Caso et al., 2020). Such a focus on prices measured at a certain point of time does not consider Pakes' (2003) argument (Fedoseeva, 2020). This sort of static modeling may be driven by the fact that the prices of products do not change due to the price rigidity of the grocery sector as well as limitations to repeatedly collect data (Fedoseeva, 2020).

However, there are also many applications that have analyzed price variation in daily (Fedoseeva, 2020), weekly (Schollenberg, 2012; Bonnano, 2016; Chen et al., 2021), monthly (Bimbo et al., 2016), quarterly (Kim & Chung, 2011), and annual (Muth et al., 2013; Yang & Dharmasena, 2020) data. Despite this vast research exploiting both cross-sectional data and price variations through seasons, holidays, and other time periods, the question of whether and how the WTP for an attribute changes over time is frequently overlooked.

One known exception is Fedoseeva's (2020) paper on sparkling wine bought through Russian e-commerce, where the hedonic price function was estimated for each day of the data in order to study whether marginal prices changed over time. The author pointed out high volatility in the marginal price estimates and proved that more research was needed to cover the complex dynamics of the

hedonic prices. Especially nowadays as new data sources that overcome the data availability challenges are gaining consumer attention. Although the demand for wines and champagnes is much more volatile than that for bread, the question of whether marginal prices are time invariant forms a gap in the literature. It seems well-suited as a subject of inquiry in this study.

2.3.4 Hedonic OLS regression issues

As discussed in the theoretical section, one can estimate the hedonic price function using the single equation approach. This estimation utilizing the OLS is the subject of a vast literature (e.g., Schollenberg, 2012; Carlucci et al., 2014; Bimbo et al., 2016; Bonnano, 2016; Giombi et al., 2018; Fedoseeva, 2020; Chen et al., 2021). While the OLS provides for a more straightforward understanding of the relationships between price and characteristics, it prevents non-linear and complex patterns in the data from being captured. Therefore, other methods, such as a two-stage least squares (Martínez-Garmendia, 2010), generalized method of moments (Kim & Chung, 2011; Ribeiro et al., 2019), and a maximum likelihood estimation (Nganje et al., 2008), have seen applications in hedonic price papers. However, as choosing between these alternative approaches was not considered as a key focus of the study, this part of the literature review considers only the OLS regression analysis.

To rely on the OLS regression results, five assumptions must be met: linearity of the data, multivariate normality, no multicollinearity, homoscedasticity, and no autocorrelation (Hayashi, 2000; Verbeek, 2004). These assumptions are often investigated through series of tests after performing the regression. These tests are also used to help in choosing the adequate functional form among the alternative specifications. Unfortunately, these assumptions can be quite restrictive and bending them is sometimes unavoidable leading to a situation where the confidence intervals and insights provided by the model can be misleading. However, several ways have been found to use the OLS setting despite the possible violations.

The most frequently occurring assumption violation is the varying variance of the error term across observations, or heteroscedasticity (Costanigro & McCluskey, 2011; Bimbo et al., 2016; Edenbrandt et al., 2018; Fernández et al., 2019). Heteroscedastic errors are widespread in hedonic price analyses (Edenbrandt et al., 2018) as specific patterns of heteroscedasticity are typical. For instance, the variance of the error term often increases with the magnitude of the predicted prices (Costanigro &

McCluskey, 2011). Heteroscedasticity has been detected by employing the Breusch-Pagan (Breusch & Pagan, 1979) and White's (White, 1980) tests (e.g., Nganje et al., 2008; Schollenberg, 2012; Muñoz et al., 2015; Giombi et al., 2018).

The second most frequently discussed assumption is little to no multicollinearity. As it is common for hedonic regressions to use observational data and certain characteristics tend to be correlated, like organic and healthy, some degree of collinearity between the independent variables is unavoidable. The presence of multicollinearity has been tested by simply evaluating the correlation matrices (Fedoseeva, 2020) and utilizing the variance inflation factor (VIF) as in Equation 2 below (Nganje et al., 2008; Schollenberg, 2012; Muñoz et al., 2015; Edenbrandt et al., 2018). The VIF for a variable is equal to the ratio of the total model variance to the variance of a model that only includes a single independent variable. For each independent variable, the ratio is determined accordingly:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{2}$$

where R_i^2 is the unadjusted R-squared metric for regressing the i_{th} independent variable on the others. As a rule of thumb, VIF values of ten or higher imply problems of multicollinearity (Nganje et al., 2008; Edenbrandt et al., 2018). Sometimes a VIF value of 5 is considered as a threshold value that should not be exceeded. Researchers have either excluded variables with high VIF values from the model (Nganje et al., 2008) or kept them in the analysis anyway (Edenbrandt et al., 2018).

Other assumptions, linearity, normality of residuals, and no autocorrelation, have received less attention in the literature. In fact, these assumptions were not even mentioned as prevalent issues in Costanigro & McCluskey's (2011) review article on the hedonic pricing approach in food markets. Nonetheless, in the research by Nganje et al. (2008) and Schollenberg (2012), the normality of the residuals was tested using the Komogorov-Smirnov test. However, Galarraga (2001) pointed out that numerical tests for residual normality are ineffective, and that the Shapiro-Wilk test, for example, cannot be applied to sample sizes larger than 5000 observations. As a result, Schollenberg (2012) chose to rely more on residual histograms and normal density plots to detect any issues. Furthermore, only Kim and Chung (2011) and Schollenberg (2012) examined whether the no autocorrelation assumption was violated.

Often, different data transformation techniques help to overcome violations towards the heteroscedasticity assumption (Fernández et al., 2019). For example, logarithmic transformations stabilize the variance (Costanigro & McCluskey, 2011). To overcome heteroscedasticity, one can also adopt a weighted least squares (WLS) method. However, most hedonic pricing research seems to rely on the consistency of the OLS estimator and correct the standard errors of the coefficients with heteroscedasticity consistent (HC) (White, 1980) or heteroscedasticity and autocorrelation consistent (HAC) standard errors (Newey & West, 1987,1994). Several studies, such as Muñoz et al. (2015), Bimbo et al. (2016), Edenbrandt et al. (2018), Ballco and De-Magistris (2018), Fedoseeva (2020), and Chen et al. (2021), have used HC standard errors in their estimations. Martínez-Garmendia (2010) and Schollenberg (2012) on the other hand obtained the Newey-West, or HAC standard errors.

2.3.5 Sources of data

The focus of this review is strictly on the revealed preference method, and hence sensory perceptions, and survey methods with hypothetical biases are excluded. Additionally, guides and catalogs that contain market prices are not considered because they are only available for some products, such as wine, often consist of price ranges, and are updated unfrequently (Fedoseeva, 2020). Therefore, only three common sources have been reviewed: grocery store shelves, consumer scanner datasets, and e-commerce platforms.

The method of collecting data from grocery store shelves has in each reviewed case been a direct observation of the product. Observing products in stores provides a consistent way of obtaining reliable product information as it mirrors real-life purchasing situations well (Carlucci et al., 2013; Caso et al., 2020) and it allows the recording of special, for example, highlighted attributes on the packaging, important for the analysis (Fernández et al., 2019). Additionally, it is possible to include store characteristic such as shelf space, store location as well as type, and size (Nganje et al., 2008; Carlucci et al., 2013; Muñoz et al., 2015; Fernández et al., 2019). Furthermore, while Cavallo and Rigobon (2016) thought that this approach was expensive in relation to the amount of data acquired, Carlucci et al. (2013) considered that it was fast and inexpensive. However, the data collection process has not been repeated which, in addition to the limited spatial scope, may have had an effect on the reliability of the results (Ward, Lusk, & Dutton, 2008; Carlucci et al., 2013).

Scanner data obtained from a provider e.g., Nielsen or Symphony Group are seen as an alternative for the one-time grocery store snapshot type of data collecting. Reasons behind this are that the datasets can include a wide variety of characteristics together with a long time range of transaction data across different regions (Kim & Chung, 2011; Schollenberg, 2012; Muth et al., 2013; Bimbo et al., 2016; Bonnano, 2016; Chen et al., 2021). For instance, one of the datasets acquired by Giombi et al. (2018) contained over 150 product attributes.

The approach, however, has a few flaws. Cavallo (2018) and Hillen (2019) pointed out that as the scanner data are seldom raw, errors, manipulation, and missing value imputations are impossible to discover. Hillen (2019) added that the purchased data can be costly especially when long time series, many characteristics, different retailers, and regions are needed. Additionally, scanner data suffer from publication delay (Hillen, 2019). Furthermore, result comparability between studies utilizing scanner data is constrained to some extent due to different UPC systems between data providers (e.g., Yang & Dharmasena, 2020) and datasets may not be available with the same structure by different providers for locations of interest (Hillen, 2019). Moreover, it seems that a lot of information is missing from scanner datasets as validating and collecting additional information is commonly done in such studies.

For example, Yang and Dharmasena (2020) had to manually collect a lot of additional data because their dataset did not comprise enough goods or enough nutritional information. Consequently, the authors were limited to select only three types of dairy alternative beverages for their study. Furthermore, they were unable to merge their initial data with other data that provided sufficient nutrition information because both used different unit recording. Giombi et al. (2018) examined two separate scanner datasets provided by IRI and Gladson, of which the Gladson's did not contain any price data. As a result, prices from the IRI's dataset were employed in Gladson's. Bimbo et al. (2016) and Bonnano (2016) searched manufacturers' websites as they wanted to include more characteristics and fill missing values. Edenbrandt et al. (2018) studied how adding Dutch Choices and Danish Keyhole labels influenced product prices. To assess this, Edenbrandt et al. (2018) needed to gather product prices before and after both labels were introduced. However, as only the labeling status of the Choices was centrally registered, Keyhole labeling status information needed to be gathered manually from producers, manufacturers' websites, and stores which led to significant data losses.

Products in previously described cases were identified with a Universal Product Code (UPC) or European Article Number (EAN) which made information gathering rather convenient. The dataset of Chen et al. (2021), however, did not contain such identification numbers, forcing the authors to do a lot of additional work to even identify the products to include more characteristics.

As an alternative to scanner data and due to the rapid development of online retailing, food product studies have started using more e-commerce data (Fedoseeva, 2020). Consumers can effortlessly search for various products across different websites with a couple of clicks rather than moving from one shelf or store to another. Additionally, consumers can easily notice the wide variety of product characteristics and compare them products comfortably at home or anywhere else, not to mention deliveries to home (Carlucci et al., 2014). Furthermore, e-commerce is a new way to reach out to consumers as retailers tend to post online prices to advertise products for offline shoppers (Cavallo, 2018). Carlucci et al. (2014) noticed that e-shoppers may be willing to pay more for some characteristics due to the previous reasons.

Additionally, as the price adjustment costs are smaller on e-commerce platforms than in grocery stores, the online environment offers more flexibility to change prices (Fedoseeva, 2020). Electronic price labels, on the other hand, have been utilized in K-supermarkets in Finland for some time, and their price adjustment costs are comparable to those of online retailers. Although the evidence of more frequent price adjustments on e-commerce compared to offline stores is still scarce (Cavallo, 2018; Hillen, 2019), it has been shown that, for example, Amazon utilize their ability of dynamic pricing (Fedoseeva, 2020; Hillen, 2020). As online retailing continues to grow (Hillen, 2020), it provides new opportunities to obtain higher-frequency data fort for hedonic price studies.

Although some hedonic price studies have identified how convenient it is to obtain data from e-commerce platforms, the potential of this data source has not yet been exploited fully. This is evidenced by the fact that research (Carlucci et al., 2014; Ballco & De-Magistris, 2018; Fedoseeva, 2020) gathered prices and attributes in the same way as grocery store studies did – directly observing goods on websites. Furthermore, static modeling practices remain the norm. That is, Carlucci et al. (2014) did not repeat the collection process and Ballco and De-Magistirs (2018) monitored websites over a certain period of time but collected only one price for each product. Fedoseeva (2020), as an exception, observed product prices daily for 77 day period.

2.4 The hedonic price analysis of foods

Searching from Google Scholar with a keyword *hedonic price analysis* yielded 179 000 results, and 41 600 results published since 2010. By adding *food* to the keyword, reduced the number of results to 21 700 since 2010. Those numbers demonstrate that the hedonic price method has been extensively applied to food commodity markets which, according to Costanigro and McCluskey (2011), is due to their high heterogeneity and differentiation. The hedonic approach has the advantage of explaining how product prices change depending on a specific attribute. This is particularly beneficial in studying, for example, how the presence of a health claim is valued (Muth et al., 2013). The features of the hedonic price method are not limited to studying how a single attribute influences the price, but it also takes into account many characteristics in a broad range of products as well as both supply- and demand-side factors.

Typically, the analyzed characteristics of food products over which consumers form expectations and perceptions are divided into intrinsic and extrinsic cues (Symmank, 2019). Intrinsic, or subjective characteristics, comprise appearance of the product such as smell, taste, and texture. Extrinsic, or objective characteristics on the other hand are product-related but not part of the product, such as information on the label, price, and packaging, and are therefore visible before purchasing takes place. Dewettnick et al. (2008) described that the distinction is also made between experience and credence attributes. The former relates to intrinsic cues as they can be observed at the time of consumption. The latter, such as healthiness of an ingredient and naturalness, are perceivable only on the high level of abstraction, if at all.

In the literature on hedonic pricing of food products, the discussion has focused around the objective and subjective characteristics. It has been argued that the objective characteristics best explain the final product price because they are easier for consumers to identify (Muñoz et al., 2015; Fedoseeva, 2020) although some subjective attributes appear to be significant in markets with experienced consumers (Carlucci et al., 2013; Ballco & De-Magistris, 2018).

The objective attributes tend to be further divided into smaller categories. In this section of the literature review, the focus is on general product characteristics that have been determined to be significant, rather than product-specific attributes, e.g., whether yogurt is drinkable. Furthermore, there does not exist similar grouping for these general characteristic for food products as there exists

for real estate and housing markets made by Chin and Chau (2003). Therefore, the distinction of different characteristics in this section has been based on those that can be generalized to every food.

There has also been some debate of which sort of general attributes should be included in the first stage of Rosen's (1974) hedonic price regression. As discussed in the theoretical section of this study, consumers are expected to be price-takers, hence their characteristics should not be able to change prices available to them (Costanigro & McCluskey, 2011; Ribeiro et al., 2019). As a result, Schollenberg (2012) and Edenbrandt et al. (2018) argued that only characteristics that are related to the product and its costs should be included and any consumers characteristics, such as income, should not be included.

The majority of the research that was evaluated has followed this line of reasoning, although some exceptions exist - for example, Fernández et al. (2019) included county-level income data in their model. However, higher prices in high-income counties may be related to a variety of factors other than income, such as business pricing strategies (Chen et al., 2021), as county-level data does not necessarily accurately reflect wealth distribution. Nevertheless, consumer characteristics and their influence on the price have been left out from this review.

2.4.1 General attributes

The first category, consist of labels that relate to certain nutrition contents of the product, e.g., good source of fiber, and make health claims e.g., fiber contributes to an acceleration of intestinal transit. Whether a product label contains a nutritional claim has been found to have a significant influence on the price in multiple studies (Muth et al., 2013; Szathvary & Trestini, 2014; Bimbo et al., 2016; Bonnano, 2016; Fernández et al., 2019). The effect, however, has also been shown to be insignificant by Giombi et al. (2018). Similarly, the presence of a health claim has been found to have a significant role in price determination (Bonnano, 2016; Szathvary & Trestini, 2014). However, not only have nutritional and health claims been found to have an impact on the price when taken individually, but the presence of both simultaneously plays a crucial role (Szathvary & Trestini, 2014; Ballco & de-Magistris, 2018).

It is worth mentioning that nutritional values have also been included into the model in some studies (Nganje et al., 2008; Caso et al., 2020; Yang & Dharmasena, 2020), but many consumers may not

consider such information when making purchasing decision (Giombi et al., 2018). Therefore, they have not been reviewed separately here.

The second category, namely packaging attributes, consist mainly of packaging size and material. As expected, packaging size and weight have been found to have a negative impact on the unit price when the size or weight increases (Nganje et al., 2008; Carlucci et al., 2013; Carlucci et al., 2014; Bimbo et al., 2016; Caso et al., 2020). In terms of packaging material, glass and cardboard have seem found to have a positive influence on the price due to quality signaling and higher costs, while plastic material has been found to have a negative impact on the price (Carlucci et al., 2013; Szathvary & Trestini, 2014; Muñoz et al., 2015; Bimbo et al., 2016). Furthermore, environmentally-friendly packaging has been found to have a positive influence on the price (Kim & Chung, 2011).

The third category of attributes corresponds to brand and store types. Brand represents an important factor used by businesses to signal value and quality to consumers and has indeed been found to have a large influence on the price (Carlucci et al., 2013; Szathvary & Trestini, 2014; Bonnano, 2016). In terms of price and shelf space competition, manufacturers may also agree to produce retailer and store brands, or private label brands (Bimbo et al., 2016), which generally possess lower prices (Schollenberg, 2012; Muth et al., 2013; Szathvary & Trestini, 2014; Edenbrandt et al., 2018). Discount stores (e.g., Lidl) have had a negative impact on the price compared to traditional supermarkets (Carlucci et al., 2013; Szathvary & Trestini, 2014). Specialized and niche shops have had higher marginal prices compared to supermarkets (Nganje et al., 2008; Fernández et al., 2019). In terms of e-commerce, Carlucci et al. (2014) inferred that structure, organization, and functionality of the website may affect the performance of the e-commerce sales.

The last category consists of other attributes. The organic mode of production has been found to have a positive impact on the price in multiple studies (Kim & Chung, 2011; Muth et al., 2013; Bimbo et al., 2016; Giombi et al., 2018; Caso et al., 2020). Additionally, domestic origin has generally been found to have a positive impact on the price (Schollenberg, 2012; Muñoz et al., 2015; Fedoseeva, 2020), although in some cases the opposite result has been found (Fernández et al. 2019).

2.5 Web scraping

Various data collectors, such as AC Nielsen, Information Resources, Inc (IRI), and Growth from Knowledge (GfK) Consumer Scan, gather store-level point-of-sale data from retailers as well as data on purchasing behavior through consumer panels in Finnish retail market. The Finnish Competition and Consumer Authority (FCCA) initiated an inquiry into potential antitrust issues connected to Nielsen's ScanTrack system's use in the Finnish food retail industry in February 2007 (Koski, 2018). The information purchased from the ScanTrack service was found to be anti-competitive, resulting in the collection and exchange of ScanTrack data being terminated.

The Finnish Food and Drinks Industries' Federation developed a replacement tool in 2011 that provided data on the overall food market, development, and direction. However, compared to ScanTrack, data from this service were less precise and had a limited scope (Koski, 2018). As a result, the only way to obtain detailed product information for our study would be to use Nielsen's consumer panel data, which price was prohibitive.

Therefore, the only practical option was to collect data from grocery store shelves or e-commerce platforms. The latter was chosen as, in 2020, the value of Finnish online shopping grew approximately a quarter, and 28 percent of Finnish consumers increased online shopping (Sintonen et al., 2021) yielding a rather interesting pricing and purchasing environment. However, it was deemed impracticable to monitor product prices on e-commerce platforms “manually” on a daily basis. Hence, this study utilizes a novel data extraction method in the context of hedonic pricing, namely web scraping. The method itself is not entirely new for food sector research (Hillen, 2019). For example, The Billion Prices Project, a huge web scraping project at MIT, was initiated in 2008 to collect massive amounts of prices (Cavallo & Rigobon, 2016). By 2010, they were scraping five million prices every day from more than 300 retailers in 50 countries (Cavallo & Rigobon, 2016). Still, no hedonic food price analysis applications utilizing web scraping seems to be available.

Technical details and directions on how to build a web scraper are omitted from the following sections. Instead, the goal is to broaden the understanding of web scraping's potential by examining its benefits and drawbacks and compare them to those of other reviewed data collection methods. Although the focus of the discussion is on hedonic pricing applications, the implications are relevant to other food and agricultural studies as well.

2.5.1 Advantages

New data collection methods have the potential to reduce the costs of collecting extremely comprehensive and complex data, allowing future research to be more versatile and accurate (Cavallo & Rigobon, 2016). Developing an automated scraping software allows anyone to acquire precise information despite the fact that products sold online are dispersed over thousands of websites.

The frequency of the scraped data is one of their most appealing features. A written script extracts data as frequently as needed, avoiding quality adjustments and errors such as time averages, which are commonly used with scanner data sets and can result in erroneous pricing changes (Cavallo & Rigobon, 2016). Furthermore, such high sampling frequencies enable data to be obtained and analyzed without any publication delay, which is especially important for extensive and up-to-date pricing dynamics analysis and forecasting (Hillen, 2019). Moreover, combining previous aspects with automation allows results to be updated constantly. Of course, one may gather data in the conventional way by monitoring the relevant information but doing so on a daily basis for hundreds or thousands of goods is not practical or even feasible in the long term.

Although developing a web scraping script requires coding effort and time, it is far cheaper than hiring people to visit stores or purchase ready datasets from commercial providers (Cavallo & Rigobon, 2016) especially when time series of current, detailed and unaggregated data are needed (Hillen, 2019). Therefore, relating to the previous point, scraped data have low costs per observation.

Furthermore, instead of selecting a few retailers for sampling, as with scanner data (Cavallo, 2018) or direct product observation methods, it is easy to choose many retailers to be scraped. Additionally, scraped datasets are already comparable and combinable across markets because data are collected with identical methods (Cavallo & Rigobon, 2016; Cavallo, 2018). In the reviewed hedonic pricing studies result comparisons relied on different data sources, collection processes, time periods, and data treatments. Collecting data remotely and scaling the collection process to cover products across retailers, markets, and time periods decrease the marginal cost per observation close to zero (Hillen, 2019).

Moreover, although scanner data are reliable and well-structured, they do not always contain all the needed information as turned out to be the case in studies of, for example, Bimbo et al. (2016),

Bonnano (2016), and Chen et al. (2021). Furthermore, Hillen (2019) noted that scanner datasets may be aggregated or be available in the same structure by different providers for locations of interest – which was also noted repeatedly in the data source section. Apart from customizing datasets to meet the demands of researchers, scraped data are always transparent as they are raw, hence do not contain any sort of missing variable imputation, censored price spells, or other data manipulation (Cavallo, 2018; Hillen, 2019).

2.5.2 Disadvantages

Despite many of its advantages, web scraping has some shortcomings. To begin with, because the scraping script collects real-time data, one must begin gathering data from day zero to acquire long time series data. The lack of historical data is indeed a drawback and studying how certain economic actions years ago have influenced prices seems like an impossible task. This issue does not concern scanner data, as evidenced by the pre-post label analysis conducted by Edenbrandt et al. (2018). To minimize this shortcoming careful planning ahead and anticipation is required. Proper planning is also essential to avoid gathering excessive amounts of data to prevent wasting time in cleaning and processing those.

Furthermore, scraped data may not always include information on how frequently products are viewed or purchased (Cavallo, 2018). Because transaction data are lacking, it is unclear whether some goods are irrelevant to customers (Hillen, 2019). Gorodnichenko et al. (2018), however, examined the quantity weighted and unweighted prices and found that they were similar. Categorizing according to *best seller* or *most popular* can give indication on frequently purchased and relevant products (Hillen, 2019).

Online price availability especially in emerging markets and developing countries limits the applicability of web scraping (Cavallo & Rigobon, 2016; Hillen, 2019). Obtaining precise product information, however, could be a problem with other data collection methods as well in such cases. Nonetheless, as the internet coverage widens globally and availability of online prices increases, the severity of this problem will decrease. As a result, the increasing usage of this effective method, ethical and legal viewpoints are given more weight as there probably will be more abusers. Although website scraping is neither illegal nor legal outright, it is necessary to assess the situation for each application (Hillen, 2019). Only approved public content should be scraped for legitimate purposes

according to Hillen (2019), and copyright restrictions, as well as Terms of Use or Terms of Service, must be obeyed. In addition to legal considerations, ethical issues such as whether it is permissible to access public information by circumventing regulations such as the robots.txt protocol and security measures set by the website administrator should be examined.

2.6 Summary

Table 2 summarizes the reviewed hedonic price studies by showing their authors, publication years, products, countries, the chosen functional forms, and data sources. All the reviewed studies have used market data and have in their theoretical framework assumed perfectly competitive markets that are in equilibrium. Therefore, only a single hedonic pricing equation was estimated in each application. The studies used different methods to determine this equation's functional form: the Box-Cox transformation technique, model performance analysis with diagnostics or other tests to achieve the best fit and using the proven practices from previous studies. Among the main functional forms, the semi-logarithmic specification is the most popular and best performing.

Table 1. Hedonic pricing papers selected for the empirical analysis. An asterisk indicates that the main functional form is determined via the Box-Cox technique.

Authors	Year	Product	Country	Functional form	Data source
Nganje et al.	2008	Bread	The US	Linear	Grocery store
Schollenberg	2012	Coffee	Sweden	Semi-log	Scanner data
Kim & Chung	2011	Eggs	Korea	Double log	Scanner data
Carlucci et al.	2013	Yogurt	Italy	Double log	Grocery store
Muth et al.	2013	Breakfast products	the US	Semi-log	Scanner data
Carlucci et al.	2014	Olive oil	Italy	Double log	E-commerce
Szathvary & Trestini	2014	Fruit beverages	Italy	Semi-log	Scanner data
Muñoz et al.	2015	Olive oil	Chile	Semi-log	Grocery store
Bimbo et al.	2016	UHT milk	Italy	Semi-log*	Scanner data
Bonnano	2016	Yogurt	The US	Semi-log	Scanner data
Ballco & de-Magistris	2018	Yogurt	Spain	Semi-log*	E-Commerce

Edenbrandt et al.	2018	Several products	The Netherlands, Denmark	Semi-log	Scanner data
Giombi et al.	2018	Soup	-	Semi-log*	Scanner data
Fernández et al.	2019	Beef	Chile	Semi-log*	Grocery store
Caso et al.	2020	Yogurt	Italy	Linear	Grocery store
Fedoseeva	2020	Sparkling wines	Russia	Semi-log	E-commerce
Yang & Dharmasena	2020	Beverages	The US	Semi-log & linear	Scanner & grocery
Chen et al.	2021	Yogurt	China	Semi-log	Scanner data

Furthermore, various product characteristics are extracted from different sources of data. In the hedonics literature, the distinction of characteristics is often made between objective and subjective characteristics of which the former seemed to be more relevant in terms of price determination. Most commonly examined objective characteristics in hedonic price analyses have been labeling statements, packaging material and size, brand, store types, website functionality, and other attributes such as production methods (e.g., organic) and origin.

Data for the hedonic price analyses were gathered through observing products in stores or on e-commerce platforms and by acquiring scanner data. One promising alternative method for overcoming the flaws of the traditional data collection methods is web scraping. Its most advantageous features compared to other data collection methods were high frequency, versatility, low cost, transparency, scalability, and comparability. However, the method suffered from a few drawbacks such as the lack of historical and transaction data. Additionally, the method is limited to markets where online prices are available. The application of web scraping also raises ethical and legal questions.

Table 1 summarizes the web scraping discussion presented earlier and compares the method to three other reviewed market data collecting methods. Observing products from e-commerce platforms shared some benefits of web scraped data as both share the same source, but the actual collecting process is the main distinguishing factor here.

Table 2. Data collecting method comparison based on the reviewed literature and own elaboration. The collection method in grocery store and e-commerce columns is direct observation.

	Scraped data	Scanner data	Grocery stores	E-commerce
Cost per observation	Low	Moderate, high	High	Moderate
Data frequency	Daily	Weekly, monthly	No	No, daily
Real-time availability	Yes	No	Yes	Yes
Time consumed	Low	Low	High	Manageable
Historical price data	No	Yes	No	No
Transaction data	No	Yes	No	No
Product range	Large	Large	Low	Large
Product details	Many	Moderate, many	Limited	Many
Comparability	Yes	Limited	Limited	Yes
Scalability	Yes	No	No	Possibly

This study contributes to the existing literature in several different ways trying to fill a few identified gaps. As can be noticed hedonic pricing method has been widely applied to agricultural commodities and other differentiated products. With one exception (Nganje et al., 2008), only little research has been done to examine the link between attributes and price to explore the pricing mechanisms of bread. However, in an era of rapidly diversifying food markets, the older hedonic pricing bread study may not be adequate for understanding present-day links and pricing mechanisms of the product. Additionally, despite many markets assessed, this type of analysis for food products, and especially for bread, has not been conducted in Finland.

E-commerce platforms are a promising, but little used, source of data for hedonic price analysis. E-commerce was, however, also the only alternative for this study. Consequently, the second contribution is to take full advantage of possibilities provided by e-commerce by collecting data for the analysis via web scraping. Collecting data in such a unique and unprecedented way overcomes many problems of other data collecting methods in the context of hedonic pricing.

Scraping data from e-commerce platforms yields an excellent possibility to study whether marginal prices are time invariant as it is convenient to extract data daily. Therefore, estimating the hedonic price function for each point of time in the data is a prominent contribution to static modeling practices and a continuation to Fedoseeva's (2020) paper on Russian sparkling wines purchased by e-commerce.

3 Methodology

The chapter contains all the methods utilized in this thesis's empirical research. Firstly, the hedonic pricing theory is briefly reviewed followed by defining an empirical specification and estimation procedure. Following that, the second section consists of assessing the robustness of the estimation procedure. The third section describes additional diagnostic methods used in assistance to evaluate the robustness of the estimation process. The rest of the chapter introduces the data extraction and cleaning process as well as summarizes the data.

3.1 Model

Let us recall that every product has a market price p associated with its vector of characteristics revealing the hedonic price function $p(z) = p(z_1, z_2, \dots, z_k)$. In Rosen's (1974) hedonic price model marginal prices for each buyer and seller can be recovered from the first order condition $\partial\hat{p}(z)/\,\partial z_k = \widehat{p_{zk}}(z)$, where $\hat{p}(z)$ is the resulting estimate of regressing product prices on corresponding product characteristics, or the hedonic price function $p(z)$.

As part of the empirical research, a multiple regression model was chosen to draw results from sample data. The multiple regression model allows for the examination of the effect of a change in one independent variable on the dependent variable y_i, while holding other independent variables constant (Stock & Watson, 2019). Multiple regression is the most straightforward and simple way to interpret the results. In addition, it is permissible to examine the relationship between a particular variable and the explained factor while also considering other relevant variables. The function for a linear regression model has the following form:

$$P_i = \alpha + \beta_1 z_{1i} + \cdots + \beta_k z_{ki} + \epsilon_i \tag{3}$$

where P_i is the price for i_{th} product which is equal to the sum of the marginal monetary values of the characteristics, α is the intercept, β_k is the regression coefficient of the k_{th} characteristic z, and ε_i is the error term.

The OLS is used to estimate the coefficients of the multiple regression model, and the empirical portion of the paper is based on a model derived from the hedonic price theory. This study focuses on the OLS regression because, it is a straightforward and natural starting point for econometric analyses and it is not the worst alternative, as evidenced by its popularity in prior research. Let us recall that in the hedonic model presented by Rosen (1974), the marginal value of an attribute is presented as the partial derivative of the price function over the given attribute, which, in the case of an OLS regression, is equal to the value of the attribute's coefficient β_k.

3.2 Diagnostic tests

To rely on the OLS estimation results the five assumptions of the multiple linear regression model, that should not be violated and on which the diagnostic measures were based on, were tested:

- **Linearity:** the relationship between the dependent and the independent variables is linear. This assumption is not that restrictive as variables can themselves be transformations.
- **Multivariate normality:** the residuals of the model should be roughly normally distributed.
- **Little to no multicollinearity:** the independent variables should not be highly correlated.
- **Little to no autocorrelation:** the residuals should not be correlated.
- **Homoscedasticity:** the variance of the error term should be constant across observations.

In assessing the suitability of the models, the first two assumptions, linearity and multivariate normality, were measured by looking at specific plots. The linearity assumption was analyzed using a plot comparing model's residuals to the fitted values. Only visual interpretation approach was relied upon as the assumption originally is not that restrictive and data transformation techniques such as the Box-Cox and logarithms were believed to be enough to solve possible violations while dealing with parametric models. The normality assumption was examined by plotting the standardized residuals against the normal quantiles.

Afterward, multicollinearity was analyzed with a correlation plot and the VIF. The VIF, however, is not fully applicable to the models which include a set of categorical regressors. Consequently, Fox and Monette (1992) introduced a Generalized Variance Inflation Factor (GVIF) calculated as follows:

$$GVIF = \frac{\det(R_{11})\det(R_{22})}{\det(R)}, \quad (4)$$

where *det* is a determinant of the following matrices. R_{11} is the correlation matrix for the specific categorical variable, R_{22} is the correlation matrix for the remaining regressors excluding the constant, and R is the correlation matrix for all set of regressors excluding the constant. To make GVIFs comparable across dimensions, it has been suggested that the GVIF metric should be raised to (1/(2*DF)) power, where DF is the number of categorical variable levels excluding the reference level (Fox & Monette, 1992). Furthermore, the usual VIF rule of thumb interpretation, that is, no value should exceed the threshold value of 10 (Schollenberg, 2012; Edenbrandt et al., 2018), can be applied after squaring the $GVIF^{1/(2\times DF)}$.

In hedonic food product pricing applications, testing for the autocorrelation is frequently overlooked. Even so, in this study the Breusch-Godfrey test for autocorrelation (Breusch & Godfrey, 1978) was employed. The Durbin-Watson test is another popular alternative, but it can only identify the first-order autocorrelation, hence the Breusch-Godfrey test was adopted instead.

Lastly, possible violation of the homoscedasticity assumption was aimed to be observed by employing a spread-location scatter plot which plots square root standardized residuals versus fitted values and conducting the Breusch-Pagan test (Breusch & Pagan, 1979). As the test has been used in many hedonic pricing applications to detect the presence of the heteroscedasticity (e.g., Kim & Chung, 2011; Schollenberg, 2012; Edenbrandt et al., 2018; Giombi et al., 2018; Fernández et al., 2019) it was chosen for this study.

3.3 The Box-Cox and goodness of fit

Following the previous literature (Bimbo et al., 2016; Ballco & De-Magistris, 2018; Giombi et al., 2018; Fernández et al., 2019), the single parameter Box-Cox transformation technique was used as an indicator towards the most appropriate dependent variable transformation. That is, lambda's value was rounded to the nearest five tenths number which was then used as the price variable's exponent. The dependent variable went through the transformation in the following way:

$$P^{(\lambda)} = \begin{cases} P^{\lambda}, & \text{if } \lambda \neq 0; \\ \ln P, & \text{if } \lambda \approx 0. \end{cases} \quad (5)$$

Furthermore, as is traditional in hedonic pricing research, adjusted R^2 was used to determine how well the model explains the variation in the price. Additionally, F-statistic with 22 and 33682 degrees of freedom and its p-value were chosen to measure the significance of the regression. Moreover, one more metric, the standard deviation of the residuals (RMSE), was chosen as it was automatically included in R-program's regression output.

The proper specification was chosen after comparing the results obtained from the Box-Cox transformation, diagnostic tests, goodness of fit, and regression significance metrics. This specification was then brought to the daily estimation to address the question whether marginal prices change over time. The results of this estimation will be presented only in a graphical format. However, whether possible price effect differences between the days were significant were examined with a statistical test. The Shapiro-Wilk normality test was undertaken to determine whether or not the samples followed a normal distribution, and the F-test was used to examine variance homogeneity. The conclusion was that neither sample followed a normal distribution but had equal variances. Therefore, Wilcoxon's Rank Sum test was performed to test if the price effect changes between the days were statistically significant.

3.4 Data collection

The data were collected from S-Group's e-commerce site foodie.fi by utilizing the web scraping technique. S-group was chosen as it is the largest retailer in Finland with a market share of 46 percent (Ministry of Economic Affairs and Employment, 2021). Additionally, it is the 91st largest food retailer globally (Deloitte, 2021). The scraping process began on 9th of November 2021 and ended on 23rd of February 2022. In total, 402 products were scraped daily throughout the 107-day period. However, a certain amount of data cleaning was necessary and therefore the final dataset size was reduced to 315 products and the total sample size was 33705 observations.

The web scraping of the data was implemented in Python 3.10 and the scraping process itself was performed with Python's Scrapy framework. The scraping program within this framework is called a Spider and it contains predefined libraries allowing to develop user-written algorithms to locate the

information on the web page with an exact identifier (e.g., an HTML element). The algorithm is then used to parse the desired information from websites based on the domain, xpath and css expressions, as well as on additional rules provided by the user. The specifying scraping program used in this study was built in the following way:

1. The Scrapy package was imported to Python and a class *FoodieSpider*, which acts as a container for all the objects, was created.
2. Urls' of the pages of the four ready-to eat bread categories, fresh bread, bakery, dry bread, and bagel, were provided for the class.
3. The first function *parse* was defined. The function uses a for-loop with a css selector such that the scraper follows each product's website link. When an individual link was followed, this function called the next function presented below.
4. The second defined function, *parse_products,* extracted desired information from the product page. Foodie website is structurally divided into content boxes which store certain amounts of information. Information from these boxes were collected by first identifying each box's content division element or div tag. The exact information was then identified by other div tags as well as span, a, h1, h2, and p tags. After all the information in the content boxes was identified, the program continued to step 6.
5. The scraped content boxes of the website are presented in Figure 1 below. As can be noticed large amount of information is stored in a table format. Each cell of the tables is ordered by numbers. Many products, however, had varying amounts of information available and hence trying to extract fat content of one product, for example, with a number two may collect carbohydrate content when scraping some other product. Therefore, the tables had to be scraped entirely with xpaths inside for-loops resulting in lots of unnecessary data.
   5.1. Xpath and css syntaxes are a bit different but can identify the same information and data between these is compatible. However, where the css selector is more accurate, xpath can navigate forward and backward inside tables resulting in more readable output in this case.
6. All the previously created dictionaries were then passed into *yield* keyword which is similar to *return* but does not exit the function and continues with the *parse_products* function's for-loop (step 4).
7. For the first time, the program was run from the command center or terminal of the computer with the command 'scrapy crawl foodie -o 'name of the category %(time)s.csv'. This

command stored scraped data to a working directory without any delays and timestamped each file.

8. Lastly, to achieve the benefits of automation the project was set up to Scrapy Cloud in a way that it was scheduled to run remotely each day at 9:30 GMT+2. However, sometimes running the Spider from this platform caused unknown errors which required the program to be run manually.

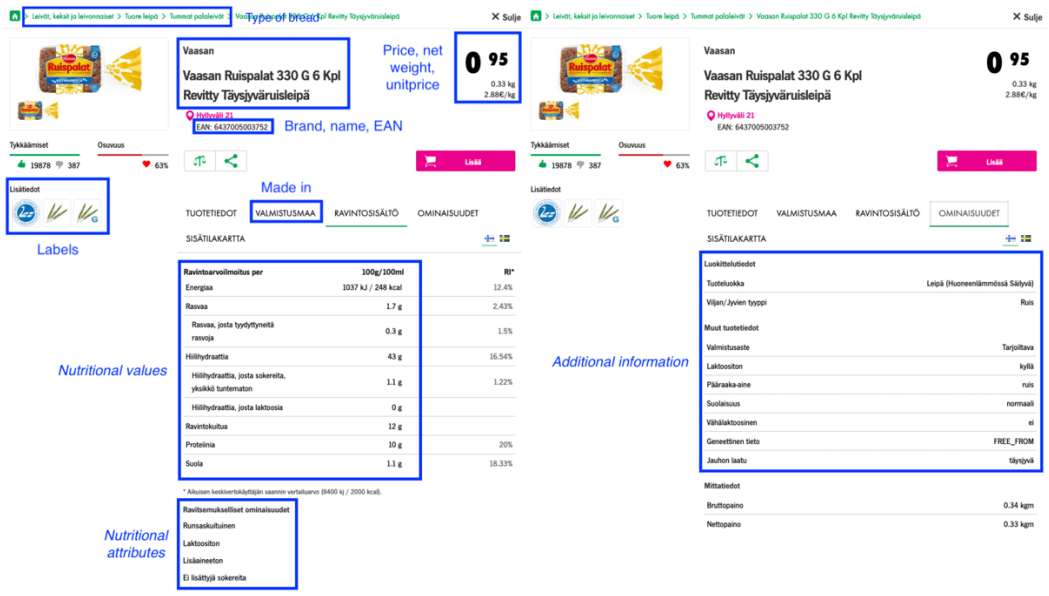

Figure 1. Scraped elements for each product.

In Figure 1, the type of bread element represents the four categories for which specific urls were used when building the Spider. The labels element consists of different product labels. The "nutritional values element" contains nutritional values per 100 grams and the "Nutritional attributes" element represents nutritional claims. Each product's name and EAN was scraped for identification purposes only. The "Additional information element" consists of the main ingredients, flour quality, genetical information (GMO), stage of serving, whether the product is sliced, and whether it is organic.

3.4.1 Variable selection

As discussed previously, the scraping process resulted in lots of unnecessary data and therefore elements and variables that were not essential for the research were removed. Many nutritional

attributes and product labels ended up removed as they lacked importance in the literature, product category, or Finnish markets. For example, shelly fish content in bread is not necessarily the characteristic consumers are basing their decisions on. Table 4 in the appendix contains a full description of each nutritional attribute and label as well as a short justification as to why they were discarded or kept. The decision whether a variable was kept or discarded was based on how significant it appeared to be in the bread and hedonic pricing literature. Furthermore, the preservation of some characteristics was based on their essential presence in the Finnish bread market, as well as on the author's conclusions on the significance of the variable.

- **Importance in the bread literature:** The most frequently appearing characteristics in the bread literature were salt, fiber and gluten content indicators as well as different cereals and bread types (Dewettnick et al., 2008; Nganje et al., 2008; Ginon et al., 2009; Rødbotten et al., 2013; Thunström & Nordstöm, 2015; Trieu et al., 2015; Antúnez et al., 2016).
- **Importance in the hedonic pricing literature:** Let us recall the discussion of the significant general product characteristics in the literature review. The significant attributes were labeling statements, packaging material and size, brand, and other attributes such as organic and origin.
- **Author's decisions and the Finnish bread market:** Customers' desire to pay for product characteristics is dependent on the traits that are easily perceivable (Giombi et al., 2018; Edenbrandt et al., 2018). Therefore, nutritional values were replaced by nutritional labels as they have more appealing layout, and their implications are more convenient to understand than those of nutritional values. For the same reason flour quality was removed. Moreover, based on the Finnish bread market discussion, origin was discarded due to presence of the Good From Finland and Key Flag labels which were combined into one dummy. Genetical information and stage of serving were removed due to being similar for all products. Furthermore, all information regarding lactose was removed due to its unimportance for this particular product category. No-additives label was removed as it was considered not to add any value for the analysis, and cereal label was removed as its content was bland.

3.5 Data cleaning

3.5.1 Variable grouping

For this project, it was important to categorize and group some variables in a way that was useful for generalization and did not needlessly hinder the model. Also, requiring the model to distinguish between very similar variables could force it to over-rely on other product specific characteristics. This mainly concerned brands and bread types as there were 48 different brands and 18 types of bread in the raw data. Following the discussion in the literature review, brands were grouped such that Fazer, Vaasan, and the brands they own were assigned under *Market leader*. S-groups own brands, Kotimaista and Rainbow, were assigned under *Store brand,* and all other brands were assigned under *Other.*

All bakery products were allocated to the *bakery* category. All rusks, crackers, crisps, and other sorts of dry bread were assigned under *drybread*. *Dark bread* category includes rye bread, some of the round loaf of bread, and bread baked at low heat. *White bread* includes toast, bread rolls, buns, bagels, white oat and barley bread, thin unleavened bread, and some of the round loaf of bread. All other types were assigned to *uncategorized.* As a side note, based on author's experience, this sort of bread type grouping matches S-groups' offline store shelf positioning quite well. Moreover, daily observations were aggregated to average monthly levels as there were not any major price changes between days in the data.

3.5.2 Missing values

Investigating the data revealed that many products had missing values. Unfortunately, it was not sure whether the information was not visible on the website because it had not been uploaded there or that such information was not even part of the product. Therefore, a distinction between these two cases needed to be made. As the focus was on studying pricing mechanisms of e-commerce, for product information to be considered as complete it was assumed that if a product had less than two missing mandatory nutritional values, it was kept in the analysis, and the presence of other characteristics was validated by searching manufacturers' websites with the EAN code. According to the Food

Information Regulation, nutrition labeling must contain information of energy (kcal), fat, saturated fat, carbohydrate, sugar, protein, and salt contents (Regulation (EU) No. 1169/2011).

Figure 2 shows the number of products with missing nutritional information after removing the products which did not have the sufficient nutrition information. The black dots represent which nutrition's value is missing and the lines connecting the dots indicate combinations of missing values. The upper bars represent the count of each dot vertically, and left bars represent the count of dots horizontally. In this sense, the upper bars show individual nutrition's missing value count, and the left bars consider the total count of missing values for each nutrition factor. As the number of NAs in the dataset at this point was relatively low it was easy and fast to fill missing values by searching from manufacturers' websites with EAN codes.

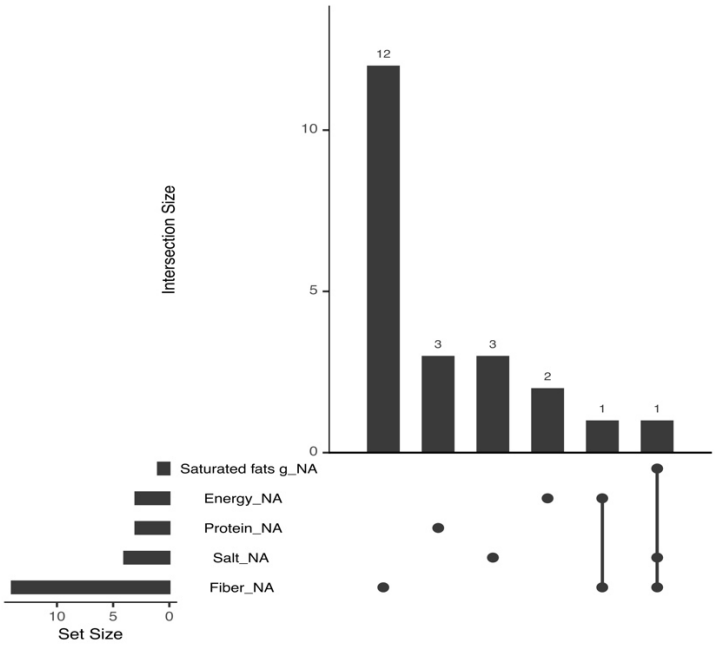

Figure 2. Number of products with missing nutritional values.

For some unknown reason, the foodie website design changed in the middle of the scraping process, which removed some product information completely from the latest days. Therefore, product characteristics except the price for those days were simply copied from the first day information.

3.5.3 Outliers

The OLS estimations in hedonic price analysis may be distorted by the presence of outliers, which may lead to biased results. Diewert (2003) even stated that the presence of outliers is one of the most important issues needed to be solved before running hedonic regressions as such data points attract the OLS hyperplane towards their, often special, characteristics. Detecting and removing outliers is a difficult process and, hence, some have estimated outlier robust hedonic price regressions (see e.g., Janssen et al., 2001). However, in this study outliers were identified with an Inter Quantile Range (IQR) criterion (Tukey, 1977). IQR is a difference between the first and the third quartiles and according to Tukey (1977), a data point is identified as an outlier if it lies below the first or above the third quartile by a factor 1.5 times the IQR. The top row of Figure 3 represents continuous variables with dots indicating the outliers identified by the IQR criterion. The bottom row shows the same variables after the identified outliers were removed (note the difference in the scale of the vertical axes).

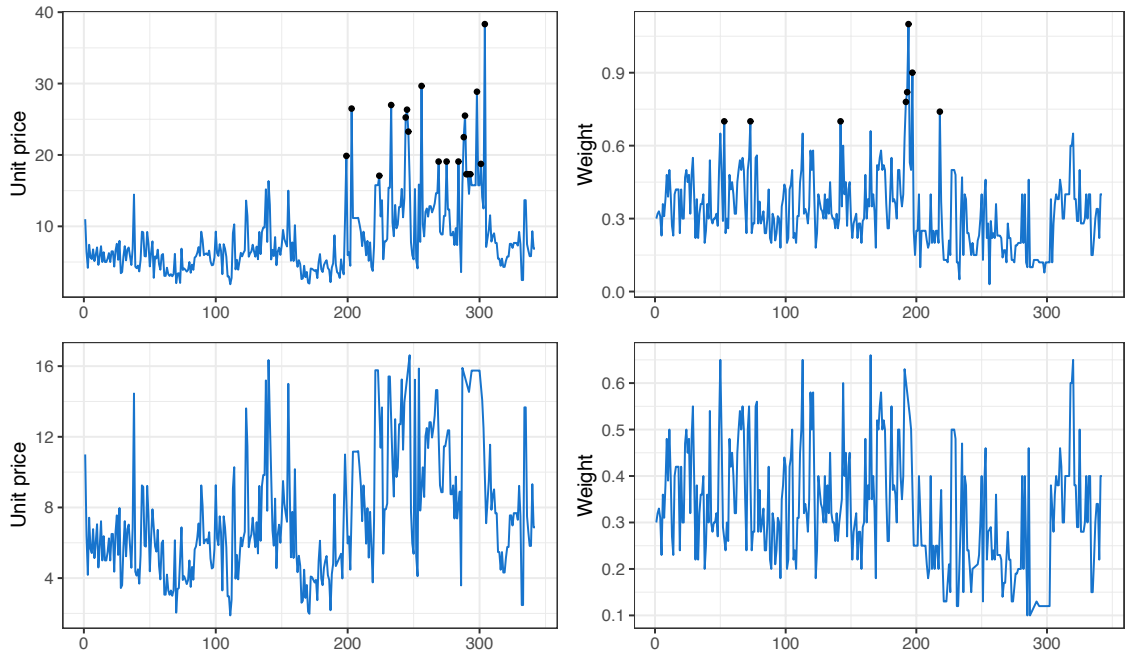

Figure 3. Continuous variables, unit price and weight, before (top row) and after (bottom row) removing outliers. Dots indicate data points that lie outside specified interval.

In overall, 28 outlier products (8% of the total) were removed. It is unfortunate substantial share, but it was noticeable that the removed observations contained unusual bread types, with most being more bread snacks rather than regular bread products.

3.6 Data summary

After removing the outliers and unusable data, the dataset contained 315 products and their selected characteristics. These characteristics are classical objective characteristics as the subjective characteristics were not available on this specific e-commerce platform. Table 3 presents all the characteristics that were chosen for the analysis, their mean, standard deviations, minimum and maximum values. For categorical and dummy variables, the mean value naturally represents the relative frequency of that specific variable receiving the value one. It is noticeable that there exists large variability in bread price from 1.85 €/kg to 16.62 €/kg, with a mean of 7.44 €/kg. The price range was quite wide even after removing the outliers, but this was mainly due to the fact that the data included many different categories of bread.

Table 3. Data definition and summary.

Variable	Type	Description	Mean	SD	Min	Max
Price	Continuous	Unit price	7,44	3,48	1,85	16,62
Weight	Continuous	Weight in kg	0,33	0,12	0,1	0,66
Brand	Categorical					
Market leader		Reference level	0,43	0,49	0	1
Other			0,54	0,5	0	1
Store brand			0,03	0,18	0	1
Main ingredient	Categorical					
Wheat		Reference level	0,25	0,43	0	1
Oat			0,2	0,4	0	1
Multigrain			0,08	0,27	0	1
Barley			0,02	0,12	0	1
Rye			0,32	0,47	0	1
Other			0,13	0,34	0	1
Organic	Dummy	1=organic, 0 otherwise	0,07	0,25	0	1
Type of bread	Categorical					
White bread		Reference level	0,37	0,48	0	1
Dark bread			0,22	0,42	0	1
Drybread			0,3	0,46	0	1
Bakery			0,03	0,18	0	1
Uncategorized			0,07	0,25	0	1
Reduced salt label	Dummy	1=Reduced salt label, 0 otherwise	0,05	0,21	0	1
High-fiber label	Dummy	1=High fiber label, 0 otherwise	0,49	0,5	0	1
Strong salt label	Dummy	1=Strong salt label, 0 otherwise	0,11	0,31	0	1

Gluten free label	Dummy	1=Gluten-free label, 0 otherwise	0,07	0,25	0	1
Domesticity label	Dummy	1=Domestic label, 0 otherwise	0,52	0,5	0	1
Heart label	Dummy	1=Health benefit badge, 0 otherwise	0,11	0,31	0	1
Time	Categorical					
Nov-21		Reference level			0	1
Dec-21					0	1
Jan-22					0	1
Feb-22					0	1

Half of the 315 bread studied consisted of other brands (54%), while the remaining included market leader brands (43%) and the S-Group's own brands (3%). The products' weight ranged between 100 and 660 grams with a mean of 330 grams. One out of three breads had rye (32%) as their main ingredient, shortly followed by wheat (25%), oat (20%), and other ingredients (13%), while the remaining 10 percent included multigrain and barley. Only 7 percent of the breads were organic. Bread type was most frequently white (37%), dry (30%) or dark (22%) while the remaining 10 percent included bakery and uncategorized types. As for the labels, 52 percent contained a *domesticity label*, and 19% a *heart* label. As for the nutritional labels, only 5 percent featured a reduced salt label (≤0.9 grams of salt per 100 grams), 49% a high-fiber label (≥6 grams of fiber per 100 grams), 11% a strong salt label (≥1.2 grams of salt per 100 grams), and 7% a gluten free label. Nov-21, or November 2021, is the reference level of the time dummy variable. Dec-21 and Jan-22 were the only whole months of 31 days. The length of data collection in Nov-21 was 22 days and Feb-22 was 23 days.

Based on the general attribute discussion in the literature review and variable selection in this chapter, we expect the following signs for the estimated coefficients. Reference levels of the categorical variables are in the parenthesis.

Table 4. Expected signs of the coefficients.

Variable	**Expected sign**
Brand (Market leader)	
Other	+
Store brand	-
Weight	-
Main ingredient (Wheat)	
Oat	+
Multigrain	-

Other	+
Barley	+
Rye	+
Organic	+
Type of bread (White bread)	
Bakery	+
Dark bread	+
Dry bread	+
Uncategorized	+
Reduced salt label	-
High fiber label	-
Strong salt label	+
Gluten-free label	+
Domesticity label	+
Heart label	+

4 Results

This chapter presents the main empirical results and is separated into two sections. The first section describes how a preferred specification of the empirical model was selected and presents several diagnostic methods to assess the robustness of that model. The results of the selected regression are then provided in the second section. That section also consists of performing the daily regression with the functional form specified in the first section.

4.1 Functional form selection

Three OLS hedonic price regression models were estimated to perform the specification search. These models were linear, semi-logarithmic, and double logarithmic. For each specification results of the diagnostic methods are presented in this section.

The Box-Cox transformation with a single parameter was used to assist the selection of the appropriate functional form. The derived lambda value of 0.18 is closer to zero than to 0.5, hence the natural logarithmic transformation of the price was favored. Next, the metrics for goodness of fit and regression significance were examined. The findings are displayed in Table 4 below. Both the semi- and double logarithmic specifications exhibited slightly greater adjusted R^2 values of 0.657 and 0.667 than the linear specification (0.656). Additionally, the linear specification had the largest regression prediction error, or RMSE, of 2.04, compared to semi- and double logarithmic specifications, 0.273 and 0.268, respectively. Furthermore, the overall significance of the regressions was measured, and the corresponding F-statistic with 22 and 33682 degrees of freedom is shown in Table 4. The F-statistic was the highest for the double logarithmic form (3071) compared to semi-logarithmic (2927) and linear forms (2923).

Table 4. Adjusted R^2, residual standard error, root mean square error, and F-statistic for each specification. Asterisk indicates statistical significance at 1% level.

Metric	Linear	Semi-logarithmic	Double logarithmic
Adj.-R^2	0,656	0,657	0,667
RMSE	2,04	0,273	0,268
F(22,33682)	2923*	2927*	3071*

As stated earlier, statistical inference from OLS regressions is not completely trustworthy if the five assumptions underlying the methodology are not met. Consequently, the test of those assumptions was undertaken. The linearity of the data was examined with Figure 4 which visualizes the residuals against the fitted values of the models. Following the fluctuating red line of the average residual values for a given fitted value demonstrates that the data relationships were not exactly linear in any of the specifications. Especially with the linear specification, it was evident that the value of the bread was overestimated for the smaller and larger predictions. For logarithmic specifications, residuals were somewhat evenly distributed around the zero line.

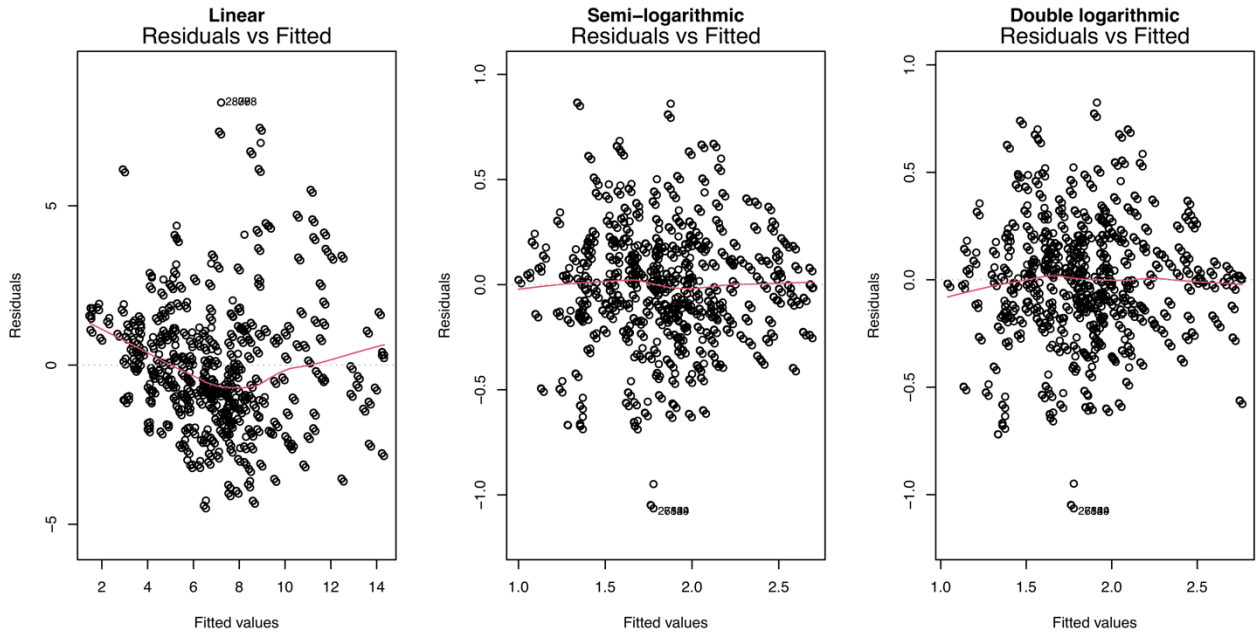

Figure 4. Residuals against fitted values of the models plots.

As observed in the residuals versus fitted values plot provided in Figure 5 below, residual normality was met broadly with every specification. As for the linear specification, the right tail was longer than the left one, and the residuals appeared to be somewhat positively skewed. For both logarithmic transformations, residuals were distributed symmetrically around the zero mean.

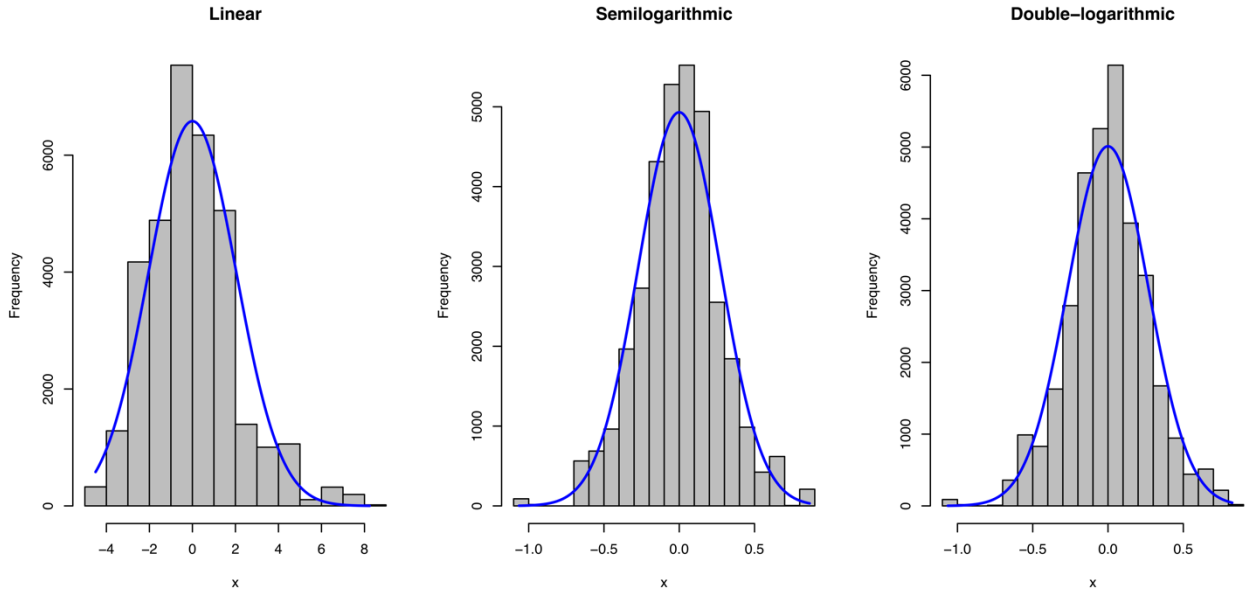

Figure 5. Distribution of the regression residuals.

GVIF metric did not indicate any problems with multicollinearity as the highest value in each specification was 1.72 for the high-fiber label. Similar conclusion could be drawn from Figure 6 which represents the correlations between the regressors that have higher than 0.3 absolute correlation according to the Pearson's correlation coefficient. The highest positive correlation was between dark bread and rye ingredient (0.58) and the lowest negative correlation was between domestic label and other brands (-0.39).

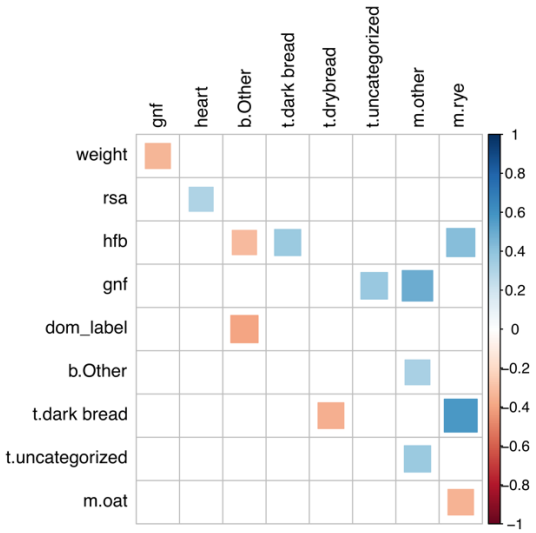

Figure 6. Correlations between regressors. t.=type of bread, m.=main ingredient, b.=brand.

The subsequent objective was to detect heteroscedasticity through the scale-location plot and the Breusch-Pagan test. On the basis of Figure 7, it can be inferred that every model suffered from heteroscedasticity. To avoid violating the assumption of homoscedasticity, the red line should be approximately horizontal. These findings were confirmed by the Breusch-Pagan test that resulted in test statistics of 5027 for the linear, 3632 for the semi-logarithmic, and 3681 for the double-logarithmic specifications. All the test statistics had p-values much lower than the threshold value of 0.05, i.e., they were highly statistically significant. Therefore, the null hypothesis of homoscedasticity was rejected. It may be determined that the residual variance was the largest for the linear specification, followed by the double logarithmic form.

Afterwards, the Breusch-Godfrey test was employed to examine the final assumption of no autocorrelation. The test produced statistically significant test statistics for all specifications. The linear specification had the least significant test statistic (3067), or the highest p-value of the three, followed by the double-logarithmic (3436) and semi-logarithmic (3602) specifications.

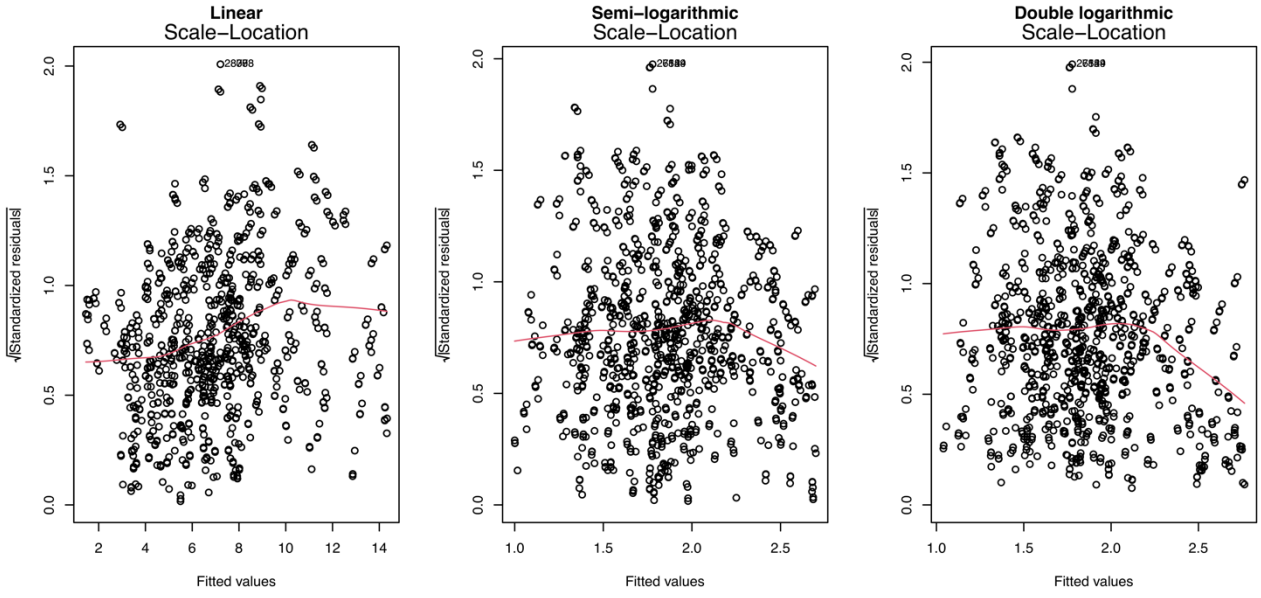

Figure 7. The scale-location plot to detect heteroscedasticity.

As all of the specifications violated the OLS assumptions, particularly the absence of autocorrelation and homoscedasticity, the most appropriate model for the analysis was selected based on the overall significance, goodness of fit, and the Box-Cox transformation metrics, as well as the choices made by previous studies. Both the semi-logarithmic and double logarithmic specifications fitted the data

better than the linear form, and their F-statistics were larger and more significant. Moreover, the Box-Cox transformation supported the logarithmic transformations which indeed satisfied the assumptions of linearity and normality of the residuals better than the linear model. Furthermore, due to the assumed nonlinear relationship between the price and attributes, the theory and literature lean towards logarithmic transformations. As a result, the linear specification was omitted from the study. There were no significant differences between the two logarithmic specifications, and neither performed considerably better than the other. Since the semi-logarithmic specification had more statistically significant variables and is more consistent with past work, it was it was selected for this study.

To address the violated assumptions, a common variance-covariance matrix estimation procedure in the literature was followed. As both heteroscedasticity and autocorrelation were of concern, HAC variance-covariance matrix estimators were specified according to Newey and West (1994) using an automated bandwidth selection procedure and pre-whitening with a first-order Vector Autoregressive filter. Pre-whitening is the process of fitting time series models to the original data in order to minimize autocorrelation, and bandwidth is a measure of how closely density should match distribution. The robust standard error estimation was implemented with R's *sandwhich* package (Zeileis, 2004).

4.2 Results of the regressions

The results of the OLS estimation using the semi-logarithmic form are provided in Table 5 below. The dependent variable in the regression model was the unit price (henceforth price) of bread. The price of bread was explained with 12 variables. Let us recall that in the case of the semi-logarithmic functional form, the coefficients of the independent variables are associated with a $\left(e^{\beta_k} - 1\right) \times 100$ percentage change in the price as the variable increases by one unit. Table 5 displays these estimated elasticity values in the second column and Figure 8 summarizes them visually. Let us also recall that the intercept, or average bread, defined a bread product offered under the Fazer or Vaasan brand names. The product's main ingredient was conventional wheat, and it was identified as white bread that did not possess any labels. Time reference level was November 2021. The reference levels of the categorical variables are specified inside the parentheses in the first column.

The third column in Table 5 reports the robust standard errors of the coefficients. The closer the standard error is to zero, the more precise the coefficient's estimate. The asterisks after the standard errors represent the statistical significance level, with one asterisk indicating significance at the five percent level, two asterisks indicating significance at the one percent level, and three asterisks indicating significance at the 0.1 percent level.

Table 5. Semi-logarithmic OLS regression results with HAC standard errors and significance levels.

Variable	Coefficient	Std. Error	Significance
Intercept	8,82	0,011	***
Brand (Market leader)			
Other	25,22	0,008	***
Store brand	-25,92	0,015	***
Weight	-79,44	0,015	***
Main ingredient (Wheat)			
Oat	10,21	0,004	***
Multigrain	-7,57	0,005	***
Other	14,28	0,003	***
Barley	11,36	0,011	***
Rye	7,14	0,003	***
Organic	1,69	0,008	*
Type of bread (White bread)			
Bakery	60,59	0,007	***
Dark bread	-6,30	0,006	
Dry bread	31,34	0,007	***
Uncategorized	20,65	0,006	***
Reduced salt label	10,16	0,006	***
High-fiber label	-8,30	0,004	***
Strong salt label	9,34	0,006	***
Gluten-free label	15,39	0,007	***
Domesticity label	4,81	0,004	***
Heart label	-20,92	0,006	***
Time (Nov-21)			
Dec-21	0,20	0,004	
Jan-22	0,18	0,004	
Feb-22	1,65	0,005	***

As can be noted from Table 5, December 2021, and January 2022 were the only variables that were not statistically significant at the previously indicated levels. Furthermore, organic was the only

variable with statistical significance at the 5% level, whereas all other variables were significant at the 0.1% level.

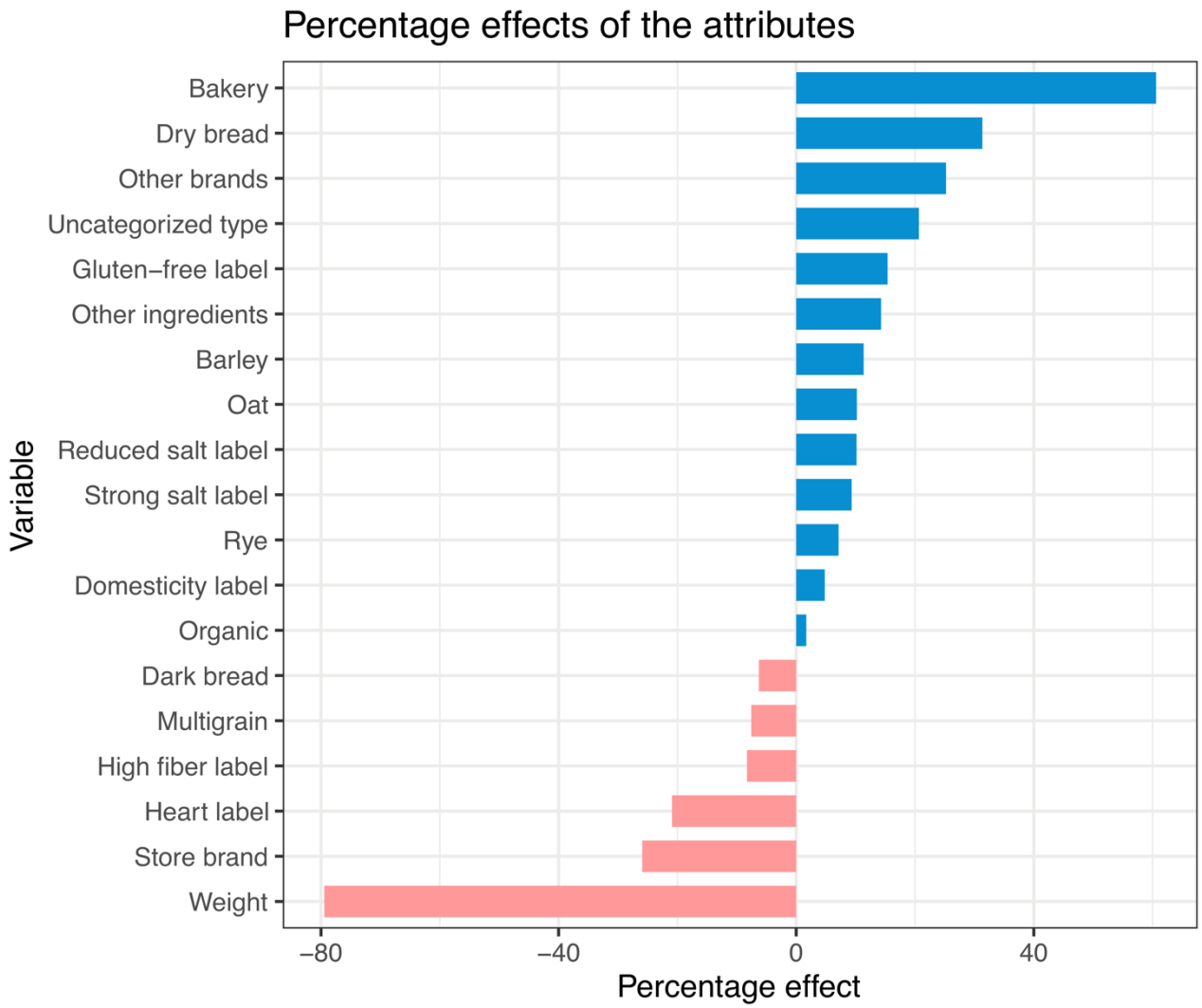

Figure 8. Results of the hedonic regression.

The first thing that stands out from the results is the negative 25.92 percent pricing effect of the store brand. Other brands that were not market leaders or owned by S-Group had prices that were 25.22 percent higher than the average bread. In terms of the main ingredient, oat raised the price by 10.21 percent, while barley increased by 11.36 percent and rye by 7.14 percent. Moreover, there was a 7.57 percent negative association between multigrain and price. Furthermore, when the product's weight increased by a single unit, the price decreased by 79.44 percent.

The impact of dark bread on the price was -6.30% while bakery-type bread increased the price by 60.59 percent. In addition, 31.34 and 20.65 percent, respectively, of the price determination was influenced positively by dry and uncategorized bread types. Organic bread was associated with a positive price increase of 1.69%.

A product with the reduced salt label resulted in a price increase of 10.65 percent. However, the strong salt content label also led to a price increase of 9.34%. The inclusion of a high-fiber label was associated with an 8.30% price reduction. The gluten-free label increased the price by 15.39% and the domesticity label resulted in 4.81% price increase. The inclusion of the Heart label led to a 20.92 percent price decrease.

The December 2021 and January 2022 dummies were statistically insignificant, therefore their 0.20 and 0.18 percent influence on the price cannot be determined with certainty. Nevertheless, the February 2022 dummy was statistically significant, showing higher prices (+ 1.65%) in that month.

After the pooled regression, the semi-logarithmic hedonic price function was estimated for each individual day of the data. The results are being summarized in a graphical format in Figure 9 that visualizes the percentage price effects (elasticities) for each attribute. The 89th day in the figure, when the largest changes occurred, is the 6th of February 2022. In order to determine if the changes were statistically significant, the daily estimation coefficients were compiled into a single sample and then split into two samples on day 90 as on that day the largest changes occurred. The Shapiro-Wilk normality test was conducted to test if the samples followed a normal distribution. The test rejected normality of both samples with p-values significantly lower than the significance level of 0.05. The F-test of variance homogeneity concluded that the variances between samples were equal. Therefore, Wilcoxon's Rank Sum test was used to test whether the price effect change on the 90th day was statistically significant.

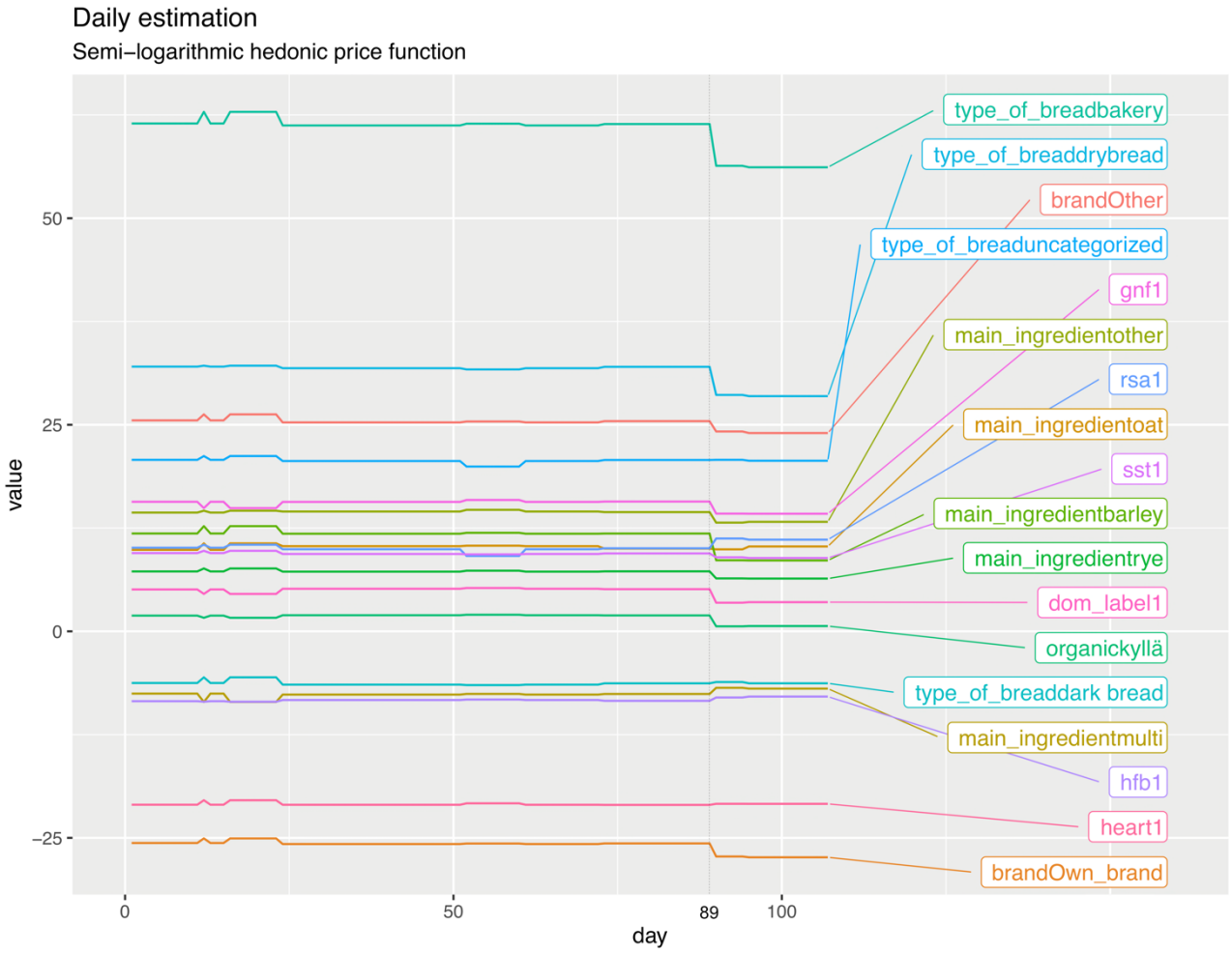

Figure 9. Percentage price effects of the attributes for each day of the data.

Wilcoxon's Rank Sum test resulted in p-value of 0.39 therefore failing to reject the null hypothesis of equal means. Additionally, the test was separately conducted for bakery and uncategorized type of bread, which had the largest changes according to Figure 8. In contrast to samples including all the characteristics, for these two individual characteristics the Wilcoxon's Rank Sum test indicated statistically significant price differences with p-values of 0.00 and 0.04, respectively.

5 Discussion

The main purpose of this study was to fill some identified gaps in the research literature on hedonic pricing of food products. The examination of the main research question expanded the current hedonic pricing literature by focusing on e-commerce, bread, and the Finnish market. Based on the results, the first null hypothesis that bread attributes do not significantly affect the online prices of bread can be rejected. All the estimated coefficients, besides the monthly dummy, are statistically significant at the five percent or better level.

Although this is not unexpected, the results provide specific information on the determinants of bread prices. The findings reveal that bakery-type bread have the strongest correlation with price. This outcome may be mostly attributable to the fact that good-tasting bakery products have been a major component of Finnish bread consumption patterns in recent years (Ministry of Economic Affairs and Employment, 2021). Additionally, bakery-type breads are more expensive to produce than average bread because they are baked fresh each morning. This result was consistent with the results of Bruschi et al. (2015) who, however, focused only on young Russian consumers and utilized data collected with experimental auctions.

Dry and uncategorized types of bread were significantly more expensive than the average bread. The primary reason for this is the significance of stomach-friendly products for Finns (Ministry of Economic Affairs and Employment, 2021). Besides containing a substantial proportion of oat and rye cereals that Finns consider to be the healthiest and the most stomach-friendly (Ministry of Economic Affairs and Employment, 2021), the majority of these types of bread had the label indicating gluten-free product. It is popular that even customers without celiac disease favor gluten free products nowadays, resulting in a significant increase in willingness to pay for gluten free labeled products.

Furthermore, the negative association between the dark bread dummy and price was unexpected as Finns consume most dark rye bread in the world, and the product has high cultural value, as seen by its designation as Finland's national food. However, the negative correlation with price may be a result of white bread, particularly oat and barley bread, gaining market share at the expense of dark bread. In 2020, according to the Ministry of Economic Affairs and Employment (2021), the market share of white bread was 55%, while the market share of dark bread was 45%. The rising demand for white bread is attributable to the fact that oat and barley bread offer the same health benefits as rye

but are easier on the stomach as they are naturally gluten free. Moreover, barley and oat bread have been demonstrated to be a good aid in weight loss programs, and their beta glucan concentration decreases cholesterol levels, among other benefits (Rødbotten et al., 2015). The question whether it is cheaper to producer dark bread than white bread remains unanswered and thus any producer point of view cannot be given here.

In hedonic pricing analyses of other food products, which include Schollenberg (2012) and Ballco and De-Magistris (2018), it is common to include multiple different types of the product into the model. This has not been the case for bread research as previous bread studies have focused on one or two bread types (e.g., Ginon et al., 2009; Bruschi et al., 2015; La Croix et al., 2015; Di Vita et al., 2016). Therefore, this study was the first of its kind to include multiple bread types into a single analysis.

Given the preceding ingredient arguments, it is not surprising that oat and barley ingredients contributed to price positively. Additionally, although the dark bread type had a negative effect on pricing and rye was the primary ingredient in 83 percent of dark bread, rye was positively associated with price. This finding may be related to that 38% of rye was the main ingredient of stomach-friendly dry bread, which generally had a positive impact on the price. Furthermore, there was a negative correlation between multigrain and price, which may be due to the fact that customers may link this type of healthier bread with an unpleasant flavor (Muth et al., 2013; Tønnesen et al., 2022). However, the term multigrain gives customers a mistaken idea that it is always a healthier alternative. In many instances, multigrain goods featured far less fiber than, for instance, rye or oat breads.

Moreover, it was evident that bread produced mainly from other ingredients, such as pumpkin seeds, typically possessed stomach-friendly characteristics, which had a positive effect on the pricing. It was also obvious that market leading brands sold bread that did not contain cereals, indicating that other ingredients provide profitable product differentiation strategies for reaching consumers. There appeared to be no research available on bread ingredients other cereals. Additionally, earlier bread research appeared to focus on consumer acceptability of a particular cereal (Rdbotten et al., 2015; Teuber et al., 2016), but this study considered four cereal varieties in addition to multigrain and other ingredients.

Brand was found to have a significant impact on the pricing of the product. In accordance with the literature, products sold under private label were cheaper (Schollenberg, 2012; Muth et al., 2013; Szathvary & Trestini, 2014; Muñoz et al., 2015; Edenbrandt et al., 2018). Other brands, however, had higher prices compared to the market leaders, which is partly in line with the study of Muth et al. (2013). This result may be explained by market-leaders' economies of scale, which result in lower production costs and cheaper products. Due to their large production, market leading brands have the ability to compete on price against private label brands. As an illustration, market leading Fazer's revenue was considerably over 200 million euros greater than that of the largest other brand in 2020. Additionally, Fazer had ten times as many employees as the largest other brand in 2020 (Ministry of Economic Affairs and Employment, 2021). However, despite the fact that Fazer and Vaasan have a strong brand image and their bread are among the most popular in Finland, Finns may value small and medium-sized enterprises (SME) and local production, resulting in price increases for other brands. Additionally, SMEs may supply specific niches for which they can get a premium.

The organic attribute had a positive effect on the pricing of bread, which was consistent with the literature (Kim & Chung, 2011; Muth et al., 2013; Bimbo et al., 2016; Giombi et al., 2018; Ribeiro et al., 2019; Caso et al., 2020). However, only 6% of the bread were organic, backing up the fact that Finnish organic grain products are being demanded more in exports due to their high quality and the growing demand for plant-based products (Ministry of Economic Affairs and Employment, 2021). The modest but increasing market share of organic bread also implies that demand for organic bread is not as great as for other food categories, such as meat and dairy. In addition, organic bread itself is not necessarily any healthier than bread produced with conventional methods as health aspects of bread are heavily based on the amounts of salt and fiber, which may contribute to organic bread’s moderate demand. Moreover, the increase in product's weight decreased the unit price, as anticipated based on the studies of Nganje et al. (2008), Carlucci et al. (2013), Carlucci et al. (2014), Bimbo et al. (2016), and Caso et al (2020).

Additionally, the Key Flag and Good From Finland labels of domesticity were seen favorably. Numerous studies have identified domestic origin as a significant contributor to consumer WTP (Schollenberg, 2012; Muñoz et al., 2015; Fedoseeva, 2020), but the price increase may also be attributable to higher domestic manufacturing costs. It remains unclear whether a product’s domesticity in previous studies was indicated with a classical “made in” text or label.

Lastly, our study provides useful information regarding the value of nutrition and health labels. The value that consumers place on reduced salt content in bread has been abundantly studied (Antúnez et al., 2015; La Croix et al., 2015; Lobo & Ferreira, 2021). In line with our results, it has been shown that consumers are interested in low sodium content bread and value them positively (Di Vita et al., 2016). However, consumers' WTP for such products is limited as healthiness of a product is not always the main concern, and its importance may be dwarfed by that of taste (Dewettnick et al., 2008; Lähteenmäki et al., 2010; Bruschi et al., 2015; Di Vita et al., 2016). For instance, Thunström and Nordström (2015) and Teuber et al. (2016) found that when bread's taste was controlled, health labels and attributes lost their importance as a demand determinant. Consequently, as salt is one of the main components of bread's taste, the strong salt label also resulted in a price increase. However, it should be noted that consumers probably are not purely seeking to purchase bread with strong salt label which may be the case with reduced salt bread.

Contrary to the findings of Nganje et al. (2008) and Ginon et al. (2009), the relationship between the high-fiber label and price was negative (2009). However, Nganje et al. (2008) did not control for the taste component in their model, while Ginon et al. (2009) hypothesized that their results could have been different if they had considered regular and repeated purchases. Furthermore, Rødbotten et al. (2015) concluded that consumers may rate the same product differently if fiber information is indicated on the product's label as opposed to nutritional tables. Previously mentioned studies focused on the actual fiber content and its rating, but our research provided insight into bread scenarios in which taste is controlled for and the fiber information is given only with labels.

Additionally, although fiber enhances human health it also forms a part of bread's taste in which high fiber amounts may be associated with lesser taste and therefore decrease consumers' valuation, as discussed by Muth et al. (2013), and Tønnesen et al. (2022). Furthermore, lower WTP for the high-fiber labeled products supports the fact that a majority of Finns do not value enough fiber products in their diets. In this sense, it may be considered that taste is valued more than healthiness also among Finns, and hence salty and good tasting bread find their way into consumer shopping carts more regularly than high-fiber labeled products.

The finding of Heart label's negative connection to price contrasts with most previous studies, which include Bonnano (2016) and Szathvary and Trestini (2014). However, research about health claims has provided inconclusive results (Tønnesen et al., 2022). Lähteenmäki et al. (2010), for example,

showed that in Nordic countries health claims did not cause any highly positive inferences in perceived product quality. While health claims may also trigger the intuition that the healthier the product, the worse the taste, the impacts of health claims on consumers are also driven by many other factors than taste, and those factors are product and claim dependent (Tønnesen et al., 2022). According to Taloustutkimus (2020) 91% of the Finns recognize the Heart label and its general meaning, but it has been established that consumers are unwilling to pay for the claim when they are unfamiliar with it (Miklavec et al., 2015). Many of the products in our data set contained the Heart label, but the health claim itself or the meaning of the label was not found from the website. Additionally, the majority of the health claims that were found related to unimportant nutrients for bread, such as zinc. Moreover, some of the health claims suggested that the bread product was reducing, for example, the risk of coronary artery disease. Hence, a consumer without such disease may not value the Heart label to a great extent. The valuation of the Heart label, however, might be higher among consumers that have a strong health interest in food.

Additionally, other studied labels, Good From Finland, and the Key flag, were valued more than the Heart among Finns according to Taloustutkimus (2020), suggesting that manufacturers may have adopted a strategy of promoting only labels that consumers value the most. On the other hand, the cost of obtaining either of the domesticity labels is 1000-5900 euros per annum in the revenue class in which the brands under the scope were (Ruokatieto, 2022; Suomalaisen Työn Liitto, 2022). For the Heart label the cost is capped at 500 euros per annum (Sydänmerkki, n.d.) which implies that higher cost and thus higher marginal prices are associated with domesticity labels than the Heart label.

The assessment of the secondary research question extended the study of Fedoseeva (2020) which was the first of its kind. From Figure 8, it is noticeable that the attributes' percentage effects on the price of the product remained flat throughout the time. This visual interpretation was confirmed by the Wilcoxon's Rank Sum test. Therefore, the null hypothesis that the marginal prices of bread attributes do not significantly change during the data collection period cannot be rejected. It was evident that S-Group did not utilize the dynamic pricing nature of e-commerce in their pricing strategies of bread over the holidays. This reflects that even online retailing in Finland does not give room to differentiate prices more, i.e., prices converge. In fact, Hillen (2019) showed that price convergences between competitors are expected to increase with online retailing in some cases as consumers' search costs decrease. Consequently, Cavallo's (2018) conclusion that retailers tend to advertise online prices to offline shoppers is supported by our result.

Probably the same results to the first research question could have been obtained if the data had been extracted only on the first and last days. Still, it was noticeable that the most economically significant attributes were also those most affected by time. Despite this being in line with the analysis of Fedoseeva (2020), the implicit prices of only two bread attributes can be said to have varied significantly over time. These attributes, the bakery and dry bread dummies, were the only ones for which the individual Wilcoxon Rank Sum test rejected the null hypothesis of equal means between the pre and post periods of the 89th day. Given the global situation in 2022, the outcome of this estimation might be completely different if data were scraped once more.

Lastly, the developed web scraping program did not only make it possible to collect the data for this analysis but introduced the method for further hedonic price analysis of foods and other food market research. However, as shown in Figure 8 and Table 3 daily scraping seemed to add only little value. Therefore, for this specific study web scraping did not provide additional value, although it saved time and effort to observe bread prices and characteristics. On the one hand, prices and characteristics could have been observed within the same time that was spent on developing the program. On the other hand, the program was created in such a way that the user only needs to change the urls to scrape any other desired product category (or all of them). Furthermore, the author suspects that only minor changes are required to adapt the program for scraping the e-commerce platform of the K-Group.

Due to the nature of this study the length of the scraping period was short compared to, for example, Cavallo (2018) and Hillen (2020). Given the price rigidity of the retail sector, major price changes were not expected and consequently, it would probably have not changed the results significantly if data were scraped from the beginning of 2021 or 2020 to this day. However, the data collection period covered the Christmas holiday season when, if ever, short term price changes were expected to occur. Nonetheless, soaring inflation, a hot and dry season cutting crop yields in Europe, especially in France, and the cessation of Ukrainian grain exports will shift hedonic price functions of bread also in Finland. With this study, historical data prior to these situations are now available for post-analysis.

6 Conclusion

The main aim of this research was to explain bread price formation in Finland by investigating online bread prices and its product-level determinants. Web scraping was used to obtain bread attributes and prices from the S-group's website over a period of three months. For the empirical investigation, the semi-logarithmic hedonic price function was estimated using the OLS. The method enabled the analysis of the relationship between many qualities and bread price in terms of percentage effects, facilitating the interpretation of the results.

Moreover, the second aim was to estimate the hedonic pricing function for each day of data collection period to determine whether the relationships between qualities and price changed during the period over time. While the methodology of this thesis was compatible with prior research and basic econometric analysis, violations of the OLS assumptions, however, may have affected the results.

Based on the quantitative research, it was clear that Finnish customers are prepared to spend more for taste than for health features. This was evidenced by the negative pricing effects of the Heart and high-fiber labels while taste components, such as strong salt label and bakery type of bread, resulted in significant price increases. Other essential factors for consumers and producer differentiation strategies include domesticity and the stomach-friendliness of breads. Especially white and dry breads made of oat and barley ingredients find their way into consumer shopping carts more often than dark bread despite its cultural significance.

Additionally, Finnish consumers are almost indifferent regarding the production method (organic versus conventional). However, the moderate effect of the organic attribute on pricing may also be attributable to the fact that Finnish organic grain products are highly demanded in exports due to their high quality and the growing demand for plant-based products. Furthermore, the data reveal that market-leading brands compete with private labels for shelf space and, thus, have generally lower prices than other brands. However, other brands may as well supply specific niches resulting in higher prices.

Moreover, it was established that price rigidity exists in Finnish online retailing, as the percentage effects of the qualities on pricing did not change considerably during the data collection period. There may have occurred more price adjustments for more volatile goods than bread, but it appears that

online bread prices are utilized to advertise prices for offline shoppers and online grocery markets in Finland do not give more room to differentiate or discriminate prices between holiday seasons.

Lastly, although daily web scraping did not provide much additional value compared to less frequent data collection, the method was the only practical alternative for conducting this type of analysis. Now, research on hedonic pricing in food markets includes a web scraping application, and we have shown that anybody can utilize this effective method. Furthermore, an alternative way to obtain detailed product data in Finland for future studies has been demonstrated with this thesis.

6.1 Limitations and further research

6.1.1 OLS

While the simplicity of understanding and applicability of the OLS regression are undeniable benefits, it is sometimes unavoidable to bend its restrictive assumptions. If this is the case, using the OLS regression on such data causes the predictions, confidence intervals, and insights produced by the model to be ineffective and deceptive. Therefore, using the OLS to capture non-linear patterns in data is a tough, if not impossible. That is problematic in the context of hedonic pricing modeling as Rosen (1974) even in his early work mentioned that the relationship between the price and characteristics is likely to be nonlinear.

As a result, housing and real estate hedonic pricing research has established advanced machine learning techniques, such as random forests and artificial neural networks, to accommodate this. Neloy et al. (2019) examined the performance of hedonic regressions and machine learning approaches, concluding that machine learning methods were empirically superior. Using a random forest model, Hong et al. (2020) found that the average percentage difference between the projected and real pricing was 5.5%, but it was nearly 20% using the conventional OLS method. However, their data were of exceptionally high quality; hence, utilizing the decision tree model alone may not be practicable in most situations. Nonetheless, using these methods and comparing them to conventional regression models could result in more robust results.

However, the OLS regression is not the worst alternative which is evidenced by its popularity in the previous studies. The OLS also offers a good starting point for econometric analyses and therefore seemed well-suited for this study.

6.1.2 Web scraping

In the data collecting process, some ways to do things differently emerged. To begin with, more e-commerce platforms could have been scraped to achieve a wider representation of the marginal prices. However, as the retail sector in Finland is very centralized (Ministry of Economic Affairs and Employment, 2021), there were not many e-commerce platforms available for scraping although new operators, such as Oda, are entering to the market. Nevertheless, it was decided that without the prior experience of web scraping only one platform should be sufficient for this study.

Additionally, more careful planning should have been taken place before the scraping such that extracting too much data could have been avoided. Fortunately, the extracted data was in well-structured format which facilitated the data cleaning and processing. Moreover, because it was established early on that the data would be obtained via web scraping, the length of data collection period could have been longer.

Furthermore, the website layout characteristics and unique dynamics of online stores such as product liking and relevancy indicators as well as product picture availability could have been scraped as well to find more information regarding the pricing dynamics of e-commerce platforms. However, even in the extremely well-researched areas outside of the food industry, we found no consistent evidence that online price setting differs significantly from the offline price setting. As highlighted throughout this thesis, even less is known about online price setting for online grocery retail, and empirical studies with large sample sizes are rare. It is commonly believed that internet prices are more flexible than offline prices, permitting suppliers to shift prices frequently in response to market fluctuations (Hillen, 2019; Fedoseeva, 2020). On the consumer side, search costs have decreased due to the convenience of online price comparison. With this flexibility, it is anticipated that price conversions will increase. Indeed, Hillen's (2019) overview article offered findings demonstrating that online prices fluctuate more frequently and to a lesser extent than offline pricing. However, these data were acquired through online marketplaces and price comparison tools, which may not be representative

of online retailing as a whole. In this aspect, however, there is no conclusive evidence of price discrepancies between online and offline grocery stores.

Hillen's (2019) article also cited material that suggested the exact opposite conclusion. In other words, online prices do not converge more, but online markets allow for greater discrimination between consumer groups, holiday seasons, competitor's price setting, and prior demand. As a result, there is a need to examine online pricing strategies, for which web scraping is an effective technique of data collection. In the limitation context of this study, to examine price convergence, future studies should consider scraping price data from more than one retailer. Detecting the mentioned patterns and activities, however, requires very sophisticated web scraping programs that, among other things, pretend to log on from various IP-addresses with different profiles (Hillen, 2019).

As it has been shown that web scraping is an effective and practical method for obtaining detailed product data, the method should be utilized more also in the field of hedonic pricing research of food products. Whereas this study focused on, for instance, health claims as a single dummy, future research may break the claims into individual variables, as Muth et al. (2013) and Ballco and De-Magistris (2018) did. Furthermore, although our analysis included all the perceived product characteristics on the website, future research could focus on one specific characteristic, such as the organic one and its moderate effect. Lastly, future studies should consider utilizing web scraping and hedonic pricing method on other product categories as there does not exists similar studies than ours in Finnish markets and as the script is easily adaptable to other products in S-Group's e-commerce platform.

Food price formation is a complex and multifaceted phenomena, and its underlying mechanisms can always be investigated further. This study provided a solid foundation for further elaboration on price formation by exploring an advanced data collection method and the fundamental basis of food product price formation, with a particular focus on bread characteristics.

Acknowledgements

I would like to thank my supervisor, Prof. Xavier Irz, for his substantial contributions to my thesis while he was extremely busy with his own research and projects. He deserves full credits for suggesting the hedonic pricing framework and utilizing web scraping. In addition, his assistance was crucial for organizing the thesis and correcting the spelling.

References

Antúnez, L., Giménez, A., & Ares, G. (2016). A consumer-based approach to salt reduction: Case study with bread. *Food Research International*, *90*, 66–72. https://doi.org/10.1016/j.foodres.2016.10.015

Ballco, P., & De-Magistris, T. (2018). Valuation of nutritional and health claims for yoghurts in Spain: A hedonic price approach. *Spanish Journal of Agricultural Research*, *16*(2). https://doi.org/10.5424/sjar/2018162-12130

Bartik, T. J. (1987). The Estimation of Demand Parameters in Hedonic Price Models. *Journal of Political Economy*, *95*(1). https://www.jstor.org/stable/1831300

Bimbo, F., Bonanno, A., Liu, X., & Viscecchia, R. (2016). Hedonic analysis of the price of UHT-treated milk in Italy. *Journal of Dairy Science*, *99*(2), 1095–1102. https://doi.org/10.3168/jds.2015-10018

Bishop, K. C., & Timmins, C. (2011). *Hedonic Prices and Implicit Markets: Estimating Marginal Willingness to Pay for Differentiated Products Without Instrumental Variables*. http://www.nber.org/papers/w17611

Bonanno, A. (2016). A Hedonic Valuation of Health and Nonhealth Attributes in the U.S. Yogurt Market. *Agribusiness*, *32*(3), 299–313. https://doi.org/10.1002/agr.21448

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. http://www.jstor.org/stable/2984418

Breusch, T. (1978). Testing for Autocorrelation in Dynamic Linear Models. *Australian Economic Papers, 17*(31), 334-355.

Breusch, T., & Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica, 47*(5), 1287-1294. https://doi.org/10.2307/1911963

Bruschi, V., Teuber, R., & Dolgopolova, I. (2015). Acceptance and willingness to pay for health-enhancing bakery products - Empirical evidence for young urban Russian consumers. *Food Quality and Preference*, *46*, 79–91. https://doi.org/10.1016/j.foodqual.2015.07.008

Butler, R. V. (1982). The Specification of Hedonic Indexes for Urban Housing. *Land Economics*, *58*(1), 96–108. https://doi.org/10.2307/3146079

Carlucci, D., de Gennaro, B., Roselli, L., & Seccia, A. (2014). E-commerce retail of extra virgin olive oil: An hedonic analysis of Italian Smes supply. *British Food Journal*, *116*(10), 1600–1617. https://doi.org/10.1108/BFJ-05-2013-0138

Carlucci, D., Stasi, A., Nardone, G., & Seccia, A. (2013). Explaining Price Variability in the Italian Yogurt Market: A Hedonic Analysis. *Agribusiness*, *29*(2), 194–206. https://doi.org/10.1002/agr.21332

Caso, G., Freda, R., & Lerro, M. (2020) How Quality Attributes Contribute to Market Price? *Access to Success, 21*(176). 144-148.

Cassel, E., & Mendelsohn, E. C. (1985). The Choice of Functional Forms for Hedonic Price Equations: Comment. *Journal of Urban Economics, 18*(2), 132-142. https://doi.org/10.1016/0094-1190(85)90012-9

Cavallo, A. (2018). Scraped data and sticky prices. *The review of Economics and Statistics, 100*(1), 105-119. https://doi.org/10.7910/DVN/IAH6Z6

Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, *30*(2), 151–178. https://doi.org/10.1257/jep.30.2.151

Chen, B., Zhang, X., & Zhou, Q. (2021). Product differentiation and brand building: a hedonic analysis of yogurt price in China. *International Food and Agribusiness Management Review*, *24*(3), 481–498. https://doi.org/10.22434/IFAMR2020.0040

Chin, T. L., & Chau, K. W. (2003). A Critical Review of Literature on the Hedonic Price Model. *International Journal for Housing Science and Its Application, 27*(2), 145-165. https://www.researchgate.net/publication/255726402

Costanigro, M., & McCluskey J. J. (2011). Hedonic Price Analysis in Food Markets. The Oxford Handbook of the Economics of Food Consumption and Policy. 10.1093/oxfordhb/9780199569441.013.0007

Cropper, M. L., Deck. L. B., & McConnell. K. E. (1988). On the Choice of Functional Form for Hedonic Price Functions. The Review of Economics and Statistics, 70(4), 668-675. https://doi.org/10.2307/1935831

Deloitte. (2021). Global Powers of Retailing. Retrieved from https://www2.deloitte.com/global/en/pages/consumer-business/articles/global-powers-of-retailing.html

Dewettinck, K., van Bockstaele, F., Kühne, B., van de Walle, D., Courtens, T. M., & Gellynck, X. (2008). Nutritional value of bread: Influence of processing, food interaction and consumer perception. *Journal of Cereal Science*, *48*(2), 243–257. https://doi.org/10.1016/j.jcs.2008.01.003

Diamond, D. B., & Smith, B. A. (1985). Simultaneity in the Market for Housing Characteristics. *Journal of Urban Economics, 17,* 280-292. https://doi.org/10.1016/0094-1190(85)90051-8

Diewert, E. (2003). Hedonic Regressions: A Review of Some Unresolved Issues.

Di Vita, G., D'amico, M., Lombardi, A., & Pecorino, B. 2016. Evaluating trends of low sodium content in food: The willingness to pay for salt-reduced bread, a case study. *Agricultural Economics Review, 17*(2). https://doi.org/10.22004/ag.econ.262442

Edenbrandt, A. K., Smed, S., & Jansen, L. (2018a). A hedonic analysis of nutrition labels across product types and countries. *European Review of Agricultural Economics*, *45*(1), 101–120. https://doi.org/10.1093/erae/jbx025

Epple, D. (1987). Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products. *Journal of Political Economy,* 95(1), 59-80. http://www.jstor.org/stable/1831299

Fedoseeva, S. (2020). Dynamic willingness to pay and e-commerce: Insights from sparkling wine sector in Russia. *Journal of Retailing and Consumer Services*, *57*. https://doi.org/10.1016/j.jretconser.2020.102180

Fernández, J., Melo, O., Larraín, R., & Fernández, M. (2019). Valuation of observable attributes in differentiated beef products in Chile using the hedonic price method. *Meat Science*, *158*. https://doi.org/10.1016/j.meatsci.2019.107881

Finnish Bread Information. (2020). Tietoa leivästä. https://www.leipatiedotus.fi/tietoa-leivasta/tilastointi/leipomovalmisteiden-tuotantomaarat.html

Fox J., & Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistical Association, 87*, 178-183. https://doi.org/10.2307/2290467

Gallarga, I. (2001). On the economics of eco-labeling: a case study for Fair Trade coffee in the UK. Revised PhD thesis, University of Bath, Bath.

Ginon, E., Lohéac, Y., Martin, C., Combris, P., & Issanchou, S. (2009). Effect of fibre information on consumer willingness to pay for French baguettes. *Food Quality and Preference*, *20*(5), 343–352. https://doi.org/10.1016/j.foodqual.2009.01.002

Giombi, K. C., Muth, M. K., & Levin, D. (2018). A comparative analysis of hedonic models of nutrition information and health claims on food products: An application to soup products. *Journal of Food Products Marketing*, *24*(7), 906–926. https://doi.org/10.1080/10454446.2018.1428259

Gorodnichenko, Y., Sherminov, V., & Talavera, O. (2018). Price Setting in Online Markets: Does It Click? *Journal of the European Association, 16*(6), 1764-1811. https://doi.org/10.1093/jeea/jvx050

Hillen, J. (2019). Web scraping for food price research. *British Food Journal*, *121*(12), 3350–3361. https://doi.org/10.1108/BFJ-02-2019-0081

Hon, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management, 24*(3), 140-152. https://doi.org/10.3846/ijspm.2020.11544

Janssen, C., Söderberg, B., & Zhou, J. (2001). Robust estimation of hedonic price models of price and income for investment property. Journal of Property Investment & Finance, 19(4), 342-360. https://doi.org/ 10.1108/EUM0000000005789

Kahn, S., & Lang, K. (1988). Efficient Estimation of Structural Hedonic Systems. *International Economic Review, 29*(1), 157-166. https://doi.org/10.2307/2526815

Kim, C., & Chung, C. (2011). Hedonic Analysis of Retail Egg Prices Using Store Scanner Data: An Application to the Korean Egg Market. *Journal of Food Distribution Research, 42*(3).

Koski, H. (2018). How Do Competition Policy and Data Brokers Shape Product Market competition? *ETLA Working Papers, 61. http://pub.etla.fi/ETLA-Working-Papers-61.pdf*

la Croix, K. W., Fiala, S. C., Colonna, A. E., Durham, C. A., Morrissey, M. T., Drum, D. K., & Kohn, M. A. (2015). Consumer detection and acceptability of reduced-sodium bread. *Public Health Nutrition*, *18*(8), 1412–1418. https://doi.org/10.1017/S1368980014001748

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy, 74*(2), 132-157. https://www.jstor.org/stable/1828835

Linneman, P. (1980). Some Empirical Results on the Nature of the Hedonic Price Function for the Urban Housing Market. *Journal of Urban Economics, 8*(1), 47-68. https://doi.org/10.1016/0094-1190(80)90055-8

Lobo, C. P., & Ferreira, T. A. P. de C. (2021). Hedonic thresholds and ideal sodium content reduction of bread loaves. *Food Research International*, *140*. https://doi.org/10.1016/j.foodres.2020.110090

Martínez-Garmendia, J. (2010). Application of hedonic price modeling to consumer packaged goods using store scanner data. *Journal of Business Research*, *63*(7), 690–696. https://doi.org/10.1016/j.jbusres.2009.05.002

Miklavec, K., Pravst, I., Grunert, K. G., Klopcic, M., & Pohar, J. (2015). The influence of health claims and nutritional composition on consumers' yoghurt preferences. *Food Quality and Preference, 43,* 26-33. http://dx.doi.org/10.1016/j.foodqual.2015.02.006

Ministry of Economic Affairs and Employment. (2021). Leipomoala – paikallisuudesta kansainvälisyyteen. *TEM Toimialaraportit 2021:5*. https://urn.fi/URN:ISBN:978-952-327-998-8

Mok, H. M. K., Chan, P. P. K., Kong, H., & Department, H. (1995). A Hedonic Price Model for Private Properties in Hong Kong. *Journal of Real Estate Finance and Economics*, *10*, 37-48. https://doi.org/10.1007/BF01099610

Muñoz, R. R., Moya, M. L., & Gil, J. M. (2015). Market values for olive oil attributes in Chile: A hedonic price function. *British Food Journal*, *117*(1), 358–370. https://doi.org/10.1108/BFJ-01-2014-0009

Muth, M. K., Zhen, C., Taylor, J., Cates, S., Kosa, K., Zorn, D., & Choiniere, C. (2013). The Value to Consumers of Health Labeling Statements on Breakfast Foods and Cereals. *Journal of Food Products Marketing*, *19*(4), 279–298. https://doi.org/10.1080/10454446.2013.724372

Neloy, A., Haque, H., & Islam, M. (2019). Ensemble learning based rental apartment price prediction model by categorical features factoring. https://doi.org/10.1145/3318299.3318377

Nganje, W., Kaitibie, S., Wachenhiem, C., Acquah, E. T., Matson, J., & Johnson, G. (2008). Estimating Price Premiums for Bread Marketed as "Low-Carbohydrate Bread". *Journal of Food Distribution Research, 39(*2). http://ageconsearch.umn.edu/record/55976/files/Nganje.pdf

Pakes, A. (2003). A Reconsideration of Hedonic Price Indexes with an Application to PC's. *The American Economic Review, 93*(5), 1578-1596. http://www.jstor.org/stable/3132143

Ribeiro, J. E., Gschwandter, A., & Revoredo-Giha, C. (2019). Estimation of a Hedonic Price Equation with instruments for Chicken Meat in the UK: Does the Organic Attribute matter? *Scholl of Economics Discussion Papers, 1911.* http://hdl.handle.net/10419/227803

Rødbotten, M., Tomic, O., Holtekjølen, A. K., Grini, I. S., Lea, P., Granli, B. S., Grimsby, S., & Sahlstrøm, S. (2015). Barley bread with normal and low content of salt; sensory profile and consumer preference in five European countries. *Journal of Cereal Science*, *64*, 176–182. https://doi.org/10.1016/j.jcs.2015.05.001

Rasmussen, D., & Zuehlke, T. (1990). On the choice of functional form for hedonic price functions. *Applied Economics, 22*(4), 431-438. https://doi.org/10.1080/00036849000000002

Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy, 82*(1), 34-55. https://www.jstor.org/stable/1830899

Ruokatieto. (2022). *Jäsenmaksut.* https://www.hyvaasuomesta.fi/yrityksille/hakuohjeet/jasenmaksu

Schollenberg, L. (2012). Estimating the hedonic price for Fair Trade coffee in Sweden. *British Food Journal*, *114*(3), 428–446. https://doi.org/10.1108/00070701211213519

Sintonen, M., Takala, K., Hellqvist, M., & Liikanen, J. (2021). Koronapandemia muuttaa maksutapoja pysyvästi. https://www.suomenpankki.fi/fi/Tilastot/maksuliiketilastot/

Statistics Finland. 2008. Toimialaluokitus 2008. https://www.stat.fi/fi/luokitukset/toimiala/?code=10710

Suomalaisen työn liitto. (2022). *Avainlippu.* https://suomalainentyo.fi/yrityksille/avainlippu/

So, H. M., Tse, R. Y. C., & Ganesan, S. (1997). Estimating the influence of transport on house prices: evidence from Hong Kong. *Journal of Property Valuation & Investment, 15*(1), 40-47.

Stock, J. H., & Watson M. W. (2019). *Introduction to Econometrics* (4th ed.). Pearson Education.

Symmank, C. (2019). Extrinsic and intrinsic food product attributes in consumer and sensory research: literature review and quantification of the findings. *Management Review Quarterly*, *69*(1), 39–74. https://doi.org/10.1007/s11301-018-0146-6

Sydänmerkki. (n.d.). Onko Sydänmerkki maksullinen? Retrieved June 10, 2022, from https://www.sydanmerkki.fi/maksaako-merkki-ja-miksi/

Szathvary, S., & Trestini, S. (2014). A Hedonic Analysis of Nutrition and Health Claims on Fruit Beverage Products. *Journal of Agricultural Economics, 65*(1), 505-517. https://doi.org/10.1111/1477-9552.12056

Taloustutkimus. (2020). *Brändien arvostus 2020. https://www.taloustutkimus.fi/ajankohtaista/uutisia/brandien-arvostus-2020.html*

Teuber, R., Dolgopolova, I., & Nordström, J. (2016). Some like it organic, Some like it purple and some like it ancient: Consumer preferences and WTP for value-added attributes in whole grain bread. *Food Quality and Preference*, *52*, 244–254. https://doi.org/10.1016/j.foodqual.2016.05.002

Thunström, L., & Nordström, J. (2015). Determinants of food demand and the experienced taste effect of healthy labels - An experiment on potato chips and bread. *Journal of Behavioral and Experimental Economics*, *56*, 13–20. https://doi.org/10.1016/j.socec.2015.02.004

Trieu, K., Neal, B., Hawkes, C., Dunford, E., Campbell, N., Rodriguez-Fernandez, R., Legetic, B., Mclaren, L., Barberio, A., & Webster, J. (2015). Salt reduction initiatives around the world - A Systematic Review of Progress towards the Global Target. *PLoS ONE, 10*(7). https://doi.org/10.1371/journal.pone.0130247

Tukey, J. (1977). *Exploratory data analysis* (17th ed.). Addison-Wesley Pub.

Ward, C., Lusk, J., & Dutton, J. (2008). Implicit Value of Retail Beef Product Attributes. *Journal of Agricultural and Resource Economics, 33*(3), 364-381.

World Health Organization. (2004). Global strategy on diet, physical activity and health. Retrieved from https://www.who.int/dietphysicalactivity/strategy/eb11344/strategy_english_web.pdf

Yang, T., & Dharmasena, S. (2020). Consumers preferences on nutritional attributes of dairy-alternative beverages: hedonic pricing models. *Food Science and Nutrition*, *8*(10), 5362–5378. https://doi.org/10.1002/fsn3.1757

Zeilis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software, 11*(10). 10.18637/jss.v011.i10

Appendices

Appendix 1. Scraped attributes

Table 1. Kept and discarded labels.

Variable	Description	Discarded	Reason
EBS	Contains fiber	x	HFB
NPR	No preservatives	x	Importance
RSA	Reduced salt		
MHE	Added yeast	x	Importance
NCS	No additive color	x	Importance
SNA	No added sugar	x	Importance
NDD	No additives		Importance
HFB	High fiber		
SST	Strong salt		
HEF	No additive yeast	x	Importance
HPR	Strong protein	x	Importance
HOF	Stong Omega-3	x	Importance
GNF	Gluten-free		
HUF	Stong unsaturated fats	x	Importance
SPR	Contains protein	x	Importance
SOF	NA	x	NA
LST	Low salt	x	SST
HIO	NA	x	NA
USW	Unsweetened	x	Importance
HVD	NA	x	NA
egg	Contains egg	x	Importance
milk	Contains milk	x	Importance
cereal	Contains cereals		Bland
glutein	Contains glutein	x	GNF
soy	Contains soy	x	Importance
domestic	Good From Finland		Combined
eu_organic	EU organic label	x	Organic
flag	The Key Flag		Combined
heart	Health claim		
nuts	Contains nuts	x	Importance
fish	Contains fish	x	Importance