THE UNIVERSITY OF CHICAGO

ESSAYS IN CONSUMER CHOICE AND CONSUMER DEMAND

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECONOMICS

BY
JOONHWI JOO

CHICAGO, ILLINOIS
MARCH 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful to my advisors, Ali Hortaçsu, Jean-Pierre Dubé, Pradeep Chintagunta, and Doron Ravid, for their guidance and support. Lawrence Y. Chung has been a mentor in the rigor of mathematics for more than fifteen years. I am greatly indebted to his mentorship and friendship. I also gratefully acknowledge the financial support of Korea Foundation for Advanced Studies during my doctoral study at the University of Chicago.

The Nielsen-Kilts and Dominick's datasets used in this study were obtained from James M. Kilts Center for Marketing at the University of Chicago Booth School of Business. This dissertation includes figures calculated (or derived) based on data from The Nielsen Company (U.S.), LLC and marketing databases provided by the Kilts Center for Marketing at the University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are mine and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

ABSTRACT

In chapter 1, I study sales of larger packages with quantity surcharge. Sales of larger packages with quantity surcharges occur often in the consumer packaged goods industry. This phenomenon poses a challenge to rationalizing consumer behaviors because the same amount of an identical product can be bought at a cheaper price. I present evidence that consumers lose a considerable amount of money by purchasing quantity surcharged larger packages. I develop and estimate a structural econometric model that combines (i) rationally inattentive consumers with (ii) the address model of consumer demand in the product characteristics space. By simulating consumer demand using model parameter estimates, I decompose the contribution of information friction and preference heterogeneity over package sizes on sales of larger packages with quantity surcharges. The estimated model predicts that only 40% of sales of larger packages with quantity surcharges can be attributed to information friction. I suggest revenue-improving, nonlinear pricing schemes that preserve consumer welfare at the current level. Under the pricing schemes, retailers can raise their revenues by up to 18%, and the corresponding sales of larger packages with quantity surcharge triples. As a methodological contribution, I state and prove the theorem that allows estimating the Rational Inattention (RI) model as if estimating an augmented logit model.

In chapter 2 (coauthored with Ali Hortaçsu), we study the relation between logit demand systems and constant elasticity of substitution (CES) demand systems. We develop a characteristics based demand estimation framework for the Marshallian demand system obtained by solving a budget-constrained constant elasticity of substitution (CES) utility maximization problem. From our Marshallian CES demand system, we derive the same market share equation of Berry (1994); Berry, Levinsohn, and Pakes (1995)'s characteristics based logit demand system. Furthermore, our CES demand estimation framework can accommodate zero predicted and observed market shares by separating intensive and extensive margins, and allows a semiparametric estimation strategy that is flexible regarding the distribution of unobservable product characteristics. We apply the framework to scanner data on cola sales, where we show estimated demand curves can be upward sloping if zero market shares are not accommodated properly.

CHAPTER 1

BUYING A LARGER PACKAGE WITH QUANTITY SURCHARGE: INFORMATION FRICTION OR PREFERENCE HETEROGENEITY?

1.1 Introduction

When two or more package sizes of a product are offered to consumers, quantity surcharges exist if the unit price of a larger package is higher than that of a smaller package. Conventional wisdom in the nonlinear pricing literature asserts that increasing marginal prices over quantities cannot be optimal because of diminishing marginal utility, product deterioration, storage cost, and so on. Increasing marginal prices are considered even impossible to implement when consumption can be split into smaller units because consumers will switch to consume several small units, which is the case in the consumer packaged goods industry.[1] Consumers who intend to buy one larger package can switch to buying a few smaller packages at a lower total price when larger packages are priced with quantity surcharge. However, retailers price larger packages with quantity surcharges and consumers buy them. The primary goal of this research is to explain why.

Sales of larger packages with quantity surcharges are pervasive. My model-free evidence suggests that 6.25% of dollar sales were sales of larger packages with quantity surcharge in the entire laundry detergent sales of the United States; consumers purchased a larger package even when a smaller, cheaper alternative of the same product was available on the shelf. In monetary terms, sales of larger laundry detergent packages with quantity surcharge amounts to 168 million U.S. dollars yearly, and the corresponding consumer "losses" in terms of opportunity costs are as large as 18 million U.S. dollars.[2] Since my data cover one-third

1. See, e.g., Wilson (1997) for a survey on the nonlinear pricing literature.

2. Details on the data used, method to identify and calculate the sales of larger packages with quantity

to half of the entire yearly laundry detergent sales in the United States, a simple projection yields that yearly sales and consumer "losses" subject to quantity surcharges amount to 340-500 million U.S. dollars and 36-54 million U.S. dollars, respectively. These dollar amounts are only for liquid and powder laundry detergents. Sales and consumer "losses" caused by buying larger packages with quantity surcharges across product modules are much larger.

Two questions arise immediately. Were consumers aware that they could switch to smaller packages with a lower total cost? If they were aware but still bought larger packages, why did they? By conducting small-scale surveys following consumer purchases, early research, including Nason and Della Bitta (1983); Manning, Sprott, and Miyazaki (1998), documented that approximately half of consumers were aware of the presence of quantity surcharges on items they bought. Consumers who purchased larger packages with quantity surcharges and were aware of it must have done so because the larger packages yielded higher utility. For example, larger packages may be more convenient to transport and use than several smaller packages, consumers might have environmental concern of avoiding packaging waste, or they may reduce shopping instances.

Related closely to these two questions, existing explanations of the sales of larger packages with quantity surcharges can be categorized into either of the following hypotheses: information friction or preference heterogeneity over package sizes. The former assumes that consumers who buy larger packages with quantity surcharges are unaware of smaller, cheaper alternatives on the shelves, and the latter assumes that some consumers inherently prefer larger packages. Each hypothesis can rationalize consumers' behaviors of purchasing quantity surcharged larger packages. As Nason and Della Bitta; Manning, Sprott, and Miyazaki document using post-purchase, small-scale consumer surveys, the reality lies somewhere between what the two hypotheses imply. Contrarily, two recent studies, Clerides and Courty (Forthcoming); Kim (2016), construct structural econometric demand models that *a*

surcharges, definition of the consumer loss in the context of quantity surcharge will be given lather in Sections 1.3 to 1.5.

priori attribute the cause of the sales of larger packages with quantity surcharges solely as information friction.

If the sales of larger packages with quantity surcharges are due to the friction in the information, sellers have better opportunities to exploit the consumers' misinformation. If they are due to the preference heterogeneity over package sizes, then the sellers should tailor the pricing scheme more towards the consumers' heterogenous preferences. I find the contribution of information friction and preference heterogeneity over package sizes by building and estimating a structural model of consumer choice and consumer demand, rather than relying on small-scale, post-purchase surveys. If purchase of larger packages with quantity surcharges is attributed to information friction, at least one smaller, cheaper alternative should yield the higher per-unit utility. Since the consumption utility is never observed directly, decomposing the contribution of each hypothesis requires consumption utility estimates of each available alternative from a structural model of consumer choice and consumer demand.

A consumer choice and consumer demand model developed for this purpose should (i) accommodate both information friction and preference heterogeneity over package sizes, (ii) separate actual consumption utilities from informational choice probability shifters, such as promotions, and (iii) explicitly consider the purchase of multiple packages. The model I develop consists of an alternative choice part that addresses (i) and (ii), and a quantity choice part that addresses (iii). The alternative choice part combines the rationally inattentive agent in line with Sims (2003, 2006) and the Hotelling-type utility specification on the product characteristics space. By a rationally inattentive agent, I refer to an agent who optimizes over what and how much information to learn ex-ante, where the information cost is measured by Shannon entropy. The rational inattention model conceptually separates the consumption utility from the informational effects such as promotion that also shifts the alternative choice probability. To address consumer heterogeneity over package sizes, I introduce consumer-specific ideal points over package sizes in the consumption utility specification. Different consumers might prefer different package sizes because they are heterogeneous in transportation costs or storage space, some might want to reduce shopping instances, etc. The Hotelling-type ideal point utility specification is suitable to model such consumer heterogeneity because it captures individual-level preference heterogeneity over package sizes. The quantity choice part allows consumers to purchase multiple packages of an item, reflecting that the same amount of the same product yields the same utility to consumers if preference over package sizes does not exist.

I estimate the demand model using the liquid and powder laundry detergent module of the Nielsen-Kilts data from 2012. Two laundry detergent items that differ only in their package sizes should be treated as the same product during analysis of quantity surcharges. Identifying a product with different package sizes for a large number of products, however, is difficult. Most extant papers that study consumer packaged goods industry share the limitation of identifying a product by its universal product code (UPC), and therefore each different package size of the identical product is regarded as a distinct product.[3] I overcome this limitation by hand-coding each product label for all 9817 UPCs in the liquid and powder laundry detergent modules.

Using estimated model parameters, I calculate the consumption utilities of each alternative, which allows me to decompose the contributions of information friction and preference heterogeneity over package sizes on the sales of larger packages with quantity surcharges. In order for a purchase of larger package with quantity surcharge to be attributed due to information friction, at least one smaller and cheaper package should yield a higher utility than the purchased larger package. A smaller package that has both a lower per-ounce price and a higher per-ounce utility than the purchased package could be found in only 40% of the sales of larger packages with quantity surcharges instances; for the remaining 60%, consumers

3. See Allenby, Shively, Yang, and Garratt (2004) for a survey.

purchased the larger and more expensive packages because they actually yielded a higher consumption utility. In other words, only 40% of the sales of larger packages with quantity surcharges can be attributed to information friction. Then I turn to explain why retailers would want to price products with quantity surcharge, by suggesting revenue-improving nonlinear pricing schemes that preserve consumers' expected welfare at the current level. The suggested nonlinear pricing schemes tailor towards the consumer preferences over package sizes. Retailers can increase revenues by up to 18% under such pricing schemes, and the corresponding sales of larger packages with quantity surcharges more than triples. The results suggest that retailers would want to use nonlinear pricing schemes that involves quantity surcharges to increase their revenue.

This paper contributes to the literature that documents and studies quantity surcharges, by suggesting a novel decomposition attributing the sales of larger packages with quantity surcharges to information friction and preference heterogeneity over package sizes. In doing so, I develop a structural econometric consumer demand model that applies the rational inattention framework to microdata, which is the first in the literature. The rational inattention theory literature assumes the prior distribution as known, and solve for the unconditional choice probabilities. Implementing the framework to data proceeds in the reverse direction: the unconditional choice probabilities are identified from data first, and then the existence of the prior distribution that is consistent with the unconditional choice probabilities should be shown. This paper contributes to the rational inattention literature by giving a formal proof on the existence of the prior distribution that is consistent with the unconditional choice probabilities backed out from data, which enables applied researchers to estimate the rational inattention model as if estimating an augmented logit model (Theorem 1.6.1 of Section 1.6).

The remainder of this paper is organized as follows. Section 1.2 summarizes related literature. Section 1.3 describes the data. Section 1.4 illustrates the concept of quantity surcharges, and Section 1.5 presents model-free evidence of the existence and causes of sales of larger packages with quantity surcharges. Section 1.6 presents a structural econometric model that accommodates both information friction and preference heterogeneity over package sizes. Section 1.7 presents estimation results, and Section 1.8 the model prediction, and conducts counterfactual exercises on the welfare-improving pricing scheme that uses quantity surcharges. Section 2.8 concludes.

1.2 Related Literature

This paper relates to at least six streams of literature: (i) consideration set formulation, (ii) consumer searches, (iii) rational inattention in discrete choices, (iv) pure characteristics demand models, (v) address models of demand, and (vi) consumers' quantity decisions and quantity surcharges in the consumer packaged goods industry. (i) through (iv) relate to information friction, and (v) to the preference heterogeneity hypothesis when modeling consumer choice and consumer demand.

A common feature of conventional demand models is consumers' full information regarding alternatives. Under full information, consumers compare the utility of all alternatives in the choice set, and solve the corresponding utility maximization problem. Such full-information assumptions are often too restrictive and unrealistic, and researchers therefore introduce friction in the information. Due to friction in the information, consumers might consider only a subset of available alternatives, which is referred to as the consideration set. Consideration sets, from which consumers choose an alternative, can be very different from the set of all alternatives. Extant consideration set formulation literature typically models consumer choices as a two-stage process, in which the first-stage is the consideration set formulation, and the second the alternative choice among the formed consideration set. Another stream of literature that accommodates information friction is consumer search. Consumer search theory was developed to explain the price dispersion of the homogeneous products (Stigler (1961); Diamond (1971); Reinganum (1979); Carlson and McAfee (1983); Stahl II (1989)). In the context of a consideration set, consumer searches provide a concrete microfoundation for how consideration sets are formed (Honka (2014); Palazzolo and Feinberg (2015); Honka and Chintagunta (2017); Honka, Hortaçsu, and Vitorino (Forthcoming)). The alternative choice model of the present study uses the promotions as a pure information shifter, which is analogous to Metha, Rajiv, and Srinivasan (2003).

A difficulty common during empirical application of consideration set formulation or consumer searches is that the availability of direct observations on the consideration set or search stream is crucial. Empirical applications of consumer searches on differentiated products were pioneered and developed by Hortaçsu and Syverson (2004); Ellison and Ellison (2009); de Los Santos, Hortaçsu, and Wildenbeest (2012); Honka, Hortaçsu, and Vitorino (Forthcoming), including many others, commonly exploiting at least partial observations on consumers' search streams. Observations on the search streams are, however, rare in the context of offline retail store shopping, and can be obtained only through after-purchase surveys. In an offline retail store shopping context, it is more typical that only choice data are available to a researcher. As Horowitz and Louviere (1995) demonstrate, choice data alone might not have identifying power on consideration sets or search streams. To circumvent this problem, researchers impose specific behavioral assumptions on the consideration set formulation (Roberts and Lattin (1991); Allenby and Ginter (1995); Mitra (1995); Siddarth, Bucklin, and Morrison (1995); Chiang, Chib, and Narasimhan (1999); Metha, Rajiv, and Srinivasan (2003); Zhang (2006); Goeree (2008); van Nierop, Bronnenberg, Paap, Wedel, and Franses (2010); Terui, Ban, and Allenby (2011)), or during the search stage (Seiler (2013); Moraga-González, Sándor, and Wildenbeest (2015); Pires (2015); Gentry (2016)). Specific behavioral assumptions can be prone to model misspecification.

The approach I take to accommodate information friction is fundamentally different and novel in demand-estimation literature. I build on the rational inattention framework pioneered by Sims (2003, 2006), extended to discrete choice by Matějka and McKay (2012, 2015); Caplin, Dean, and Leahy (2016); Caplin, Leahy, and Matějka (2016); Fosgerau, Melo, d Palma, and Shum (2017). A common feature in rational inattention literature is that the major source of stochasticity in the model is information friction, and the cost of information is proportional to reduced Shannon (1948) entropy from information acquisition.[4] This paper contributes to rational inattention literature as it is the first application of rational inattention framework to consumer demand estimation using microdata. In particular, this paper is the first to give a formal proof the existence of the prior distribution that is consistent with any given unconditional choice probabilities identified from data. With the existence result, applied researchers can estimate the rational inattention model in a way that is very similar to estimating a random utility model. This paper also contributes to pure characteristics demand estimation literature developed by Bajari and Benkard (2005); Berry and Pakes (2007); Hortaçsu and Joo (2017) by suggesting a new microfoundation on the characteristics space demand model that does not require additive idiosyncratic random utility shocks. Most extant empirical demand models that accommodate information friction combine McFadden (1974, 1978, 1989); McFadden and Train (2000)'s random utility maximization (RUM) framework with either consideration set formulation or consumer search.

The ideal-point alternative utility specification that I employ relates closely to the spatial demand model of the Hotelling style in the product characteristics space. The model is often referred to as the address demand model in the literature. Thomadsen (2005); Davis (2006); Houde (2012) estimate address demand models in the context of geographical locations of a store or firm. Anderson, de Palma, and Thisse (1989, 1992); Feenstra and Levinsohn (1995);

4. A recent general equivalence result from Fosgerau, Melo, d Palma, and Shum (2017) suggests that nearly all discrete choice random utility maximization models can be rationalized as a class of rational inattention models by showing the existence of the corresponding *generalized entropy* cost function. However, the cost function loses the invariance property with respect to the breakdowns of multiple intermediate choice steps when it departs from the Shannon entropy.

Goettler and Shachar (2001) suggest methods to identify and estimate the consumer demand models with ideal points in the characteristics space.

Lastly, this paper contributes to the literature that study quantity surcharge, quantity choice, and nonlinear pricing over package sizes in the consumer packaged goods industry. Some authors document suggestive and model-free evidence (Widrick (1979); Manning, Sprott, and Miyazaki (1998); Sprott, Manning, and Miyazaki (2003); Shreay, Chouinard, and McCluskey (2016)), or formulate a simple economic model to explain why quantity surcharges occur (Joseph, Subramaniam, and Patil (2013)). Few studies rationalize the behaviors of consumers who buy larger packages with quantity surcharges. Two recent papers, Clerides and Courty (Forthcoming); Kim (2016), explain quantity surcharges by estimating a consumer demand system, *a priori* attributing the sales of larger packages with quantity surcharges solely to the consumer misinformation. In contrast, the current paper estimates a fully structural consumer demand model that disentangles information friction and preference heterogeneity over package sizes. To study quantity surcharges, consumer utility must be modeled in a way that explicitly considers consumers' quantity decisions. By allowing for stochastic quantity choices, the way I model consumers' quantity decision contributes to the literature that studies quantity choice in the packaged goods industry (Allenby, Shively, Yang, and Garratt (2004)). The counterfactual revenue-improving, consumer welfare preserving nonlinear pricing schemes over package sizes using the estimated structural demand model are in line with Chintagunta, Dubé, and Singh (2003); Khan and Jain (2005).

1.3 Data

I use Nielsen-Kilts Homescan Consumer Panel data and Retail Measurement System scanner data. The product modules I focus on are liquid and powder laundry detergents. I focus analyses on laundry detergents because they are necessities in nearly every household, and there are no alternative categories for them.[5] Datasets were acquired through James M. Kilts Center for Marketing at the University of Chicago Booth School of Business.

1.3.1 Terminology

Throughout the manuscript, I use the term *item* interchangeably with UPC, and *product* to refer to the unique combination of brand and functional attributes, except package size. Using this terminology, there can be many items for the same product, which differ only in their package sizes. The term *per-unit price* or *unit price* always denotes the price per 1 ounce, 1 pound, etc. The term *per-package price* denotes the price for one package. Depending on context, the term *consumer loss* from the sales of larger packages with quantity surcharges, calculated in terms of "opportunity costs," might be inappropriate, particularly when a consumer decides to buy a larger package knowing that a smaller, cheaper alternative is available on the shelf. It is more appropriate in that case to interpret the behavior as the consumer having a higher willingness to pay for the larger package, not a *loss* from information friction or inattention. However, I consistently use the term *consumer loss* because I could not find a better alternative.

1.3.2 Identifying the Same Product with Different Package Sizes

To compare unit prices of the same product across package sizes, it was necessary to identify the unique product from the UPC and its abbreviated product descriptions. I describe how I identified identical products across package sizes below.

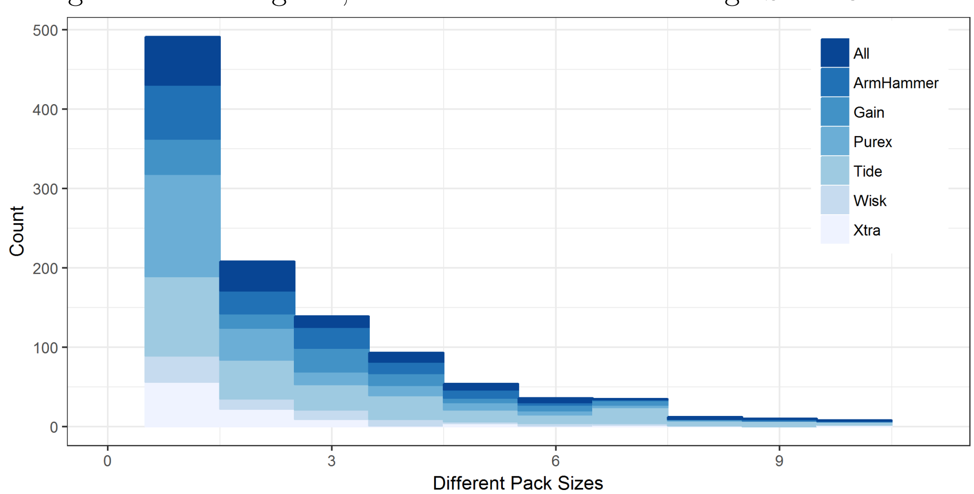
5. 85% of panel households purchased at least one laundry detergent item during 2012.

Table 1.1: Summary of Observed Product Characteristics

Characteristics	Obs.	Characteristics	Obs.	Characteristics	Obs.
liquid	6905	high efficiency	2065	baking soda	46
powder	2912	oxi-clean	167	plant based	31
fabric softener	568	colorsafe	483	low sudsing	72
Febreze	204	soft	667	low Chlorine	4
all temperature	401	unscented	985	low Sulfate	360
bleach	1684	sensitive skin	63	low Phosphorous	779
stain remover	98	baby	171	tablets	207
deep clean	120	pre-treater	2	sheet	29
ultra	3915	wrinkle reducer	4	refill	86
$n\times$concentrated	5227	enzyme	137		

Observations are the counts of UPCs of the corresponding characteristics. 9817 different UPCs are observed in the data.

Figure 1.1: Histogram, Number of Different Package Sizes Offered



This figure plots the histogram of number of package sizes of 1101 different products in seven large laundry detergent brands. 610 products offer more than two different package sizes. The effective sample size is 2915. The horizontal axis is truncated at 10.

Many product attribute abbreviations in the raw data are not standard, and thus, I manually coded them to identify unique products. I often searched and matched the UPCs with external databases to decode abbreviations. I classify the functional product characteristics of laundry detergents as in Table 1.1. Each criterion listed in Table 1.1 has the same label in the same brand, but it might have different labelings across brands. Therefore, I manually classify the functional product characteristics of laundry detergents to make the labeling consistent.[6] In addition to differences in their formulas, laundry detergents have different scents. I separately code and classify the scents because consumers are likely to recognize the same laundry detergent formula with different scents as different products. There were 273 scents across all brands.

The product attributes classification was sufficient to capture most product descriptions in the raw data. I identify the same product that differs only in its package size by brand, functional attributes, and scent. Figure 1.1 shows how many packages of a single product were offered. More than half of the products offered at least two packages.[7]

There are many smaller brands and a few large brands in the market. Nielsen classifies many brands as "control brands," which correspond to private/store brands. The names of private/store brands are not recorded, and they are typically not found in any of the UPC databases. Therefore, I dropped private brands when identifying the same product with different package sizes. I treat smaller brands, which had fewer than 10 UPCs, as private labels because most of the smaller brands offered only one package size for a product, and therefore they did not contain information for my purpose of identifying unit price schedules

6. For example, some brands refer to the oxygen cleaning formula as "oxi-clean," while another brands refer it as "active oxygen" or "oxifoam." They mean essentially the same functional characteristic having different names. Therefore I classify "oxi-clean, active oxygen, oxyfoam" as "oxi-clean." Similar classification occurs for most other product characteristics displayed in Table 1.1.

7. There is a discrepancy with data used by Clerides and Courty (Forthcoming) and my data. They analyze a supermarket chain level dataset of Dutch laundry detergent sales, and note that 99% of the products provided fewer than two package sizes. In contrast, only 63% of the products provided one or two package sizes in my data.

of the same product across package sizes.

1.3.3 Summary of Household Demographics

Table 1.2 summarizes demographics data, and how much of the sample was matched with scanner data. I constructed the variables in the "Loss from Quantity Surcharge" tab, explained in Section 1.4.

Table 1.2: Household Demographics and Laundry Detergent Purchase Patterns: Summary Statistics

	Obs.	Mean	St. Dev.	Min	Max
Projection Factor	52042	1950	2396	147	14974
Income Ratio to FPL (2012)	34191	3.844	2.054	0.093	10.072
Income Ratio to FPL (2010)	52042	3.795	2.094	0.061	10.388
Living in an Apartment	52042	0.109	0.311	0	1
Non-working Spouse	52042	0.515	0.500	0	1
Household Size	52042	2.419	1.290	1	9
Married, Living Together	52042	0.635	0.482	0	1
Any of Household Head Has a College Degree	52042	0.527	0.499	0	1
Any of Household Head is Employed	52042	0.717	0.451	0	1
No Child	52042	0.779	0.415	0	1
No. of Different Packs / Year	52042	6.059	6.624	1	176
Yearly Purchase (oz)	52042	491	482	0.510	15193
Yearly Purchase ($)	52042	31	28	0	596
Average Package Size	52042	98	59	0.510	572
No. of Different UPC / Year	52042	3.422	2.725	1	36
No. of Different Products / Year	52042	3.180	2.522	1	34
Unit Price Display is Enforced by State Law	52042	0.364	0.481	0	1
Matched with Scanner Data	21599				
Loss from Quantity Surcharge					
Baseline	1082	1.794	3.099	0.000	53
Packs s.t. $\leq$1/2 Size were Available	223	2.540	2.512	0.000	16
Excluding Smaller Packs $\geq$ 5% Discount/Feat.	658	1.068	2.031	0.010	28
Excluding Smaller Packs $\geq$ 10% Discount/Feat.	786	1.284	2.739	0.000	53
Excluding Smaller Packs $\geq$ 20% Discount/Feat.	917	1.509	2.794	0.000	53

This table presents the summary statistics of the 2012 consumer panel data. The summary statistics presented in this table are not weighted by the projection factors.

The projection factor is the multiplier for the demographics, multiplying the entire panel by which yields the entire U.S. demographic composition of the panel year. Thus, I use the projection factor as sample weights whenever possible. Income is recorded at the household level. It is reasonable to adjust income with respect to household size, and inflation. To adjust for household size and inflation, I calculate the income ratio to the federal poverty level of the corresponding year, and use it as the income variable.[8] Household income recorded in the raw data lagged by two years. When Nielsen surveys demographics of participating households prior to the beginning of the panel year, it asks the income level of the previous year at the point of survey. Therefore, the current year's income level can be obtained only for households that participated three consecutive years. For panel year 2012, the current year's income was available for only 66% of the samples. The correlation coefficient of the current income and two-year lagged income was 0.817. To avoid dropping approximately one-third of the sample, I used the two-year lagged income ratio to the federal poverty level as a proxy for the panel year's income.

1.4 Sales and Consumer Losses from Buying Larger Packages with Quantity Surcharges

I explain the concept of sales and consumer losses of larger packages with quantity surcharges, which is used throughout the remainder of the paper. Quantity surcharges are identified when the unit price of a larger package is higher than that of a smaller counterpart. Naturally, sales of larger packages with quantity surcharges can be identified only when two package sizes of the same product are available on the shelf. Explained in Section 1.3.2, I identify a laundry detergent product at the finest level. Brands, functional attributes,[9] and scents

8. U.S. Department of Health and Human Services announces the federal poverty level every year. Each year's federal poverty level is composed of two items: (i) poverty level for the first person and (ii) amount to add for each additional person in the household. With the household size in hand, it is straightforward to calculate the ratio of household income to the federal poverty level.

9. Including the form of package. For example, Tide original liquid laundry detergent and Tide pods original liquid laundry detergent are distinct products.

must be identical for two laundry detergent items to be identified as the same product. The only attribute that can differ in the same product is package size.

I use different criteria to identify the purchase of larger packages with quantity surcharges. A *Baseline* is identified by buying a larger package even when any smaller package that was cheaper in terms of the per-ounce price was available in the same week/store pair. *Packages s.t.* $\leq$*1/2 Size were Available* only considers the alternatives that are at most as large as half the size of the item. Using this criterion, a consumer buying a 100oz package priced at \$0.1 per ounce when a 75oz package priced at \$0.09 per ounce available on the shelf is not identified as buying a larger package with quantity surcharges. This criterion was added to guard against the argument that a consumer might want only 100oz of a product, not as much as 150oz. Analogously, the *Excluding Smaller Packs* $\geq$ *5% Discount/Featured* variable considers only the smaller alternatives such that they are not displayed/featured and the prices of the alternative are at least 95% of the mean price over the sample period in the same store/UPC pair.

Items with temporary price discounts or displayed/featured are salient to consumers, and thus the cost of acquiring information on such items should be smaller. When a consumer still purchases a larger package with quantity surcharges even when a smaller package is offered with temporary promotions, it is more likely that the consumer prefers the larger package strongly. Hence, the preferential hypothesis applies better in this case. When a consumer is identified as purchasing a larger package with quantity surcharges after excluding items with temporary promotions from the consumer's alternative set, information friction hypothesis is more likely to apply.

Consumer loss is calculated in terms of opportunity cost. For example, the *Baseline/Loss from Quantity Surcharge* variable is:

$$\mathbf{1}\,(\text{Smaller Packages are Available and Cheaper}) \times \text{Pack Size} \times \Delta\text{Unit Opportunity Cost}, \tag{1.1}$$

where ΔUnit Opportunity Cost is defined as the difference between the per-ounce price of the item purchased and that of the cheapest option among the smaller packages available in the same store/week. Calculations of consumer loss according to different criteria of identifying quantity surcharges are similar. The following example illustrates how I identified sales of larger packages with quantity surcharges.

Table 1.3: Example: Modes of Identifying Quantity Surcharge

Size (oz)	Unit Price ($)	$\geq$ 5% Discount/Featured	$\geq$ 10% Discount/Featured
150	0.08	O	X
100	0.1	O	X
75	0.09	O	X
50	0.11	X	X
25	0.095	O	O

Table 1.3 presents a hypothetical laundry detergent product on a shelf with different package sizes and their corresponding unit prices. Assume that a consumer purchased the 100oz package. The *Baseline Loss* is $100 \times (0.1 - 0.09) = \$1$, *Loss, Packs such that* $\leq$*1/2 Available* is $100 \times (0.1 - 0.095) = \$0.5$ because 75oz pack is not considered a smaller alternative in this case. *Loss, Excluding Smaller Packs* $\geq$ *5% Discount / Featured* is 0 because the only smaller alternative is the 50oz package, and *Loss, Excluding Smaller Packs* $\geq$ *10% Discount / Featured is* $100 \times (0.1 - 0.09) = \$1$. Whether the purchased item was under promotion does not affect the calculation.

1.5 Model-free Evidence

I quantify the sales of larger packages with quantity surcharges and present various suggestive evidence on each modeling hypothesis. For the information friction hypothesis, I offer evidence that exogenous shifters on information friction change the sales of larger packages with quantity surcharges. For the preference heterogeneity hypothesis, I demonstrate that each consumer sticks with a specific package size within a moderate period, suggesting that consumer preferences for package sizes are stable.

1.5.1 Quantifying the Sales of Larger Packages with Quantity Surcharge

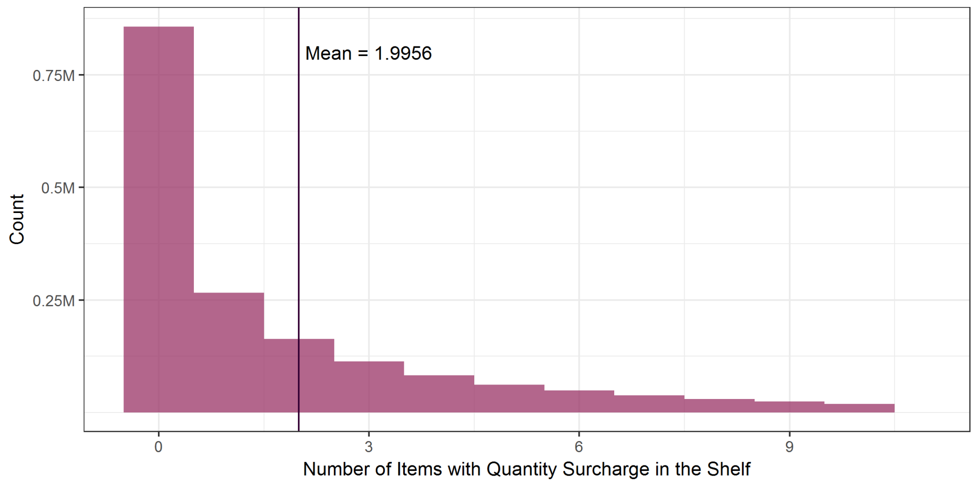
Whether sales of larger packages with quantity surcharges are observed in data, and how serious it is, is the first question addressed using real-world data. I present evidence that quantity surcharges are often observed in data, quantifying how much the sales of larger packages with quantity surcharges occur in laundry detergent modules in 2012 Nielsen-Kilts scanner data. Nielsen documents that the scanner data covers 32% to 55% of all grocery stores and supermarkets in the United States, for which it records every sales transaction. Therefore, total sales, sales subject to quantity surcharges, and consumer losses presented in this subsection represent one-third to half of yearly corresponding sales or consumer losses in the country.[10]

Items with quantity surcharges are common. When a consumer walks into a supermarket, the chances that there is more than one laundry detergent with quantity surcharges are approximately 51%. On average, there are two items on a store shelf that are subject to quantity surcharges (i.e., a smaller package with a lower unit price is available). Figure 1.2 is a histogram of the number of items with quantity surcharges in each store/week pair.

I quantify sales of larger packages with quantity surcharges in terms of dollar amounts. Figure 1.3 illustrates the extent of the sales of larger packages with quantity surcharges in terms of millions of U.S. dollars. The entire pie in the left panel represents the total sales of any laundry detergent during 2012. Approximately 6.25% of total sales, which is as large as 168 million dollars, were the sales of larger packages with quantity surcharges. For those

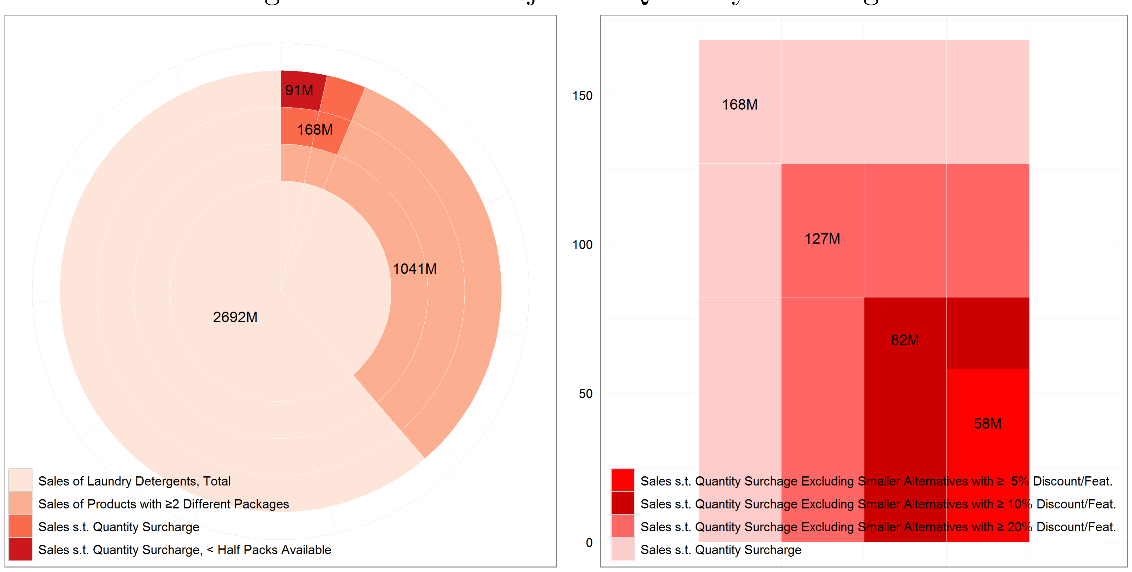
10. Albeit the availability of the projection factor in the consumer panel data, I prefer the scanner data in identifying and quantifying the quantity surcharge because the effective sample size is much larger than that of the consumer panel data.

Figure 1.2: How Many Items with Quantity Surcharge Can be Found in the Shelves?



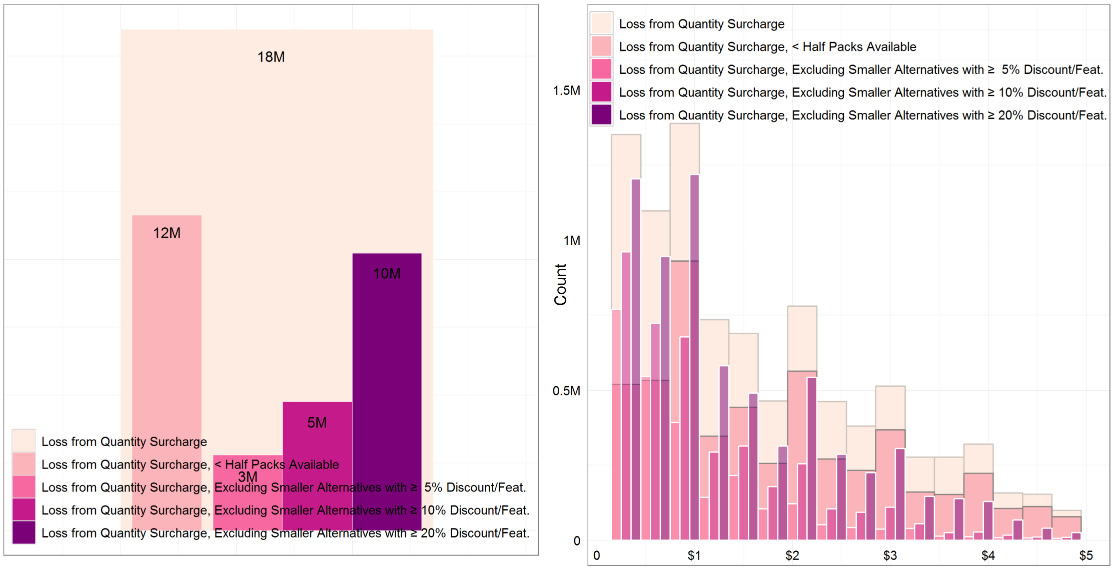
The figure plots the histogram of the number of items with quantity surcharge in each unique store/week pair. The horizontal axis is truncated at 10. The data source is the Nielsen-Kilts scanner data of the year 2012.

Figure 1.3: Sales Subject to Quantity Surcharge



The left panel of Figure 1.3 plots the total sales of laundry detergents, the total sales of laundry detergents that have more than two different package sizes, the sales of laundry detergents subject to any quantity surcharge, and the sales of laundry detergents subject to quantity surcharge when a package smaller than half of the purchased size was available. The right panel of Figure 1.3 plots the decomposition of the sales subject to quantity surcharge by different temporary price discount/promotion status criteria.

Figure 1.4: Consumer Loss from Purchasing Larger Packages with Quantity Surcharge



The left panel of Figure 1.4 plots the yearly sum of the consumer loss from purchasing larger packages of laundry detergents with quantity surcharge. The right panel of Figure 1.4 is the histogram of consumer loss from purchasing each larger package with quantity surcharge. "M" denotes million U.S. dollars. The data source is the Nielsen-Kilts scanner data of the year 2012.

sales, an identical product with a smaller package and lower unit price was available on the shelf. When I restrict that the smaller package is at most one-half the size of the item, 91 million dollars are still subject to the sales of larger packages with quantity surcharges.[11][12]

I illustrate that the sales of larger packages with quantity surcharges change with respect to different modes of identifying regular prices in the right panel. I sequentially exclude the smaller alternatives with temporary price discounts when calculating sales subject to quantity surcharges. Criteria used to identify temporary price discounts can also be understood as criteria that identify items priced at regular prices. 58 to 127 million dollars of sales were

11. Consumers may not want to substitute, for example, one 100oz package with two 75oz packages which sums up to 150oz.

12. In quantifying the sales of larger packages with quantity surcharge, a concern on the accurateness of the spot prices might arise because the Nielsen week definition may not exactly match with the individual store's pricing periods. Although the effect would averaged out when it is aggregated, I also calculated the sales of larger packages with quantity surcharge using the Dominick's data, in which the week definition matches exactly with the pricing period of the supermarket chain. 7.30% of the chainwide yearly sales of laundry detergent are the sales of larger packages with quantity surcharge in Dominick's data.

identified as purchasing a larger package with quantity surcharges when smaller alternatives were not under any promotion. In terms of a ratio, it amounts to one-third to two-thirds of the sales of larger packages with quantity surcharges.[13]

I now turn to consumer losses from buying larger packages with quantity surcharges. Figure 1.4 plots the yearly sum of consumer losses from purchasing larger packages with quantity surcharges (left panel), and its histogram (right panel). For example, the sum of consumer losses in terms of opportunity costs, which amounts to 18 million U.S. dollars in the left panel, is the area of the histogram marked with the same label in the right panel. Consumer loss is calculated using the formula introduced in Section 1.4. When a consumer buys a larger package with quantity surcharges, she loses $0.74 to $1.81, on average, for each purchased package; the consumer could have saved $0.74 to $1.81 by simply switching to a few smaller packages.

1.5.2 Information Shifters and Sales of Larger Packages with Quantity Surcharges

I examine whether exogenous information shifters affect the sales of larger packages with quantity surcharges. Information shifters used here are the state unit price display regulations and the items that are featured or displayed. The evidence presented in this subsection suggests that sales of larger packages with quantity surcharges decrease when there is less information friction or when more attention is paid by the consumers.

Some states mandate retailers to display the unit prices along with the item prices. 19 states mandate displaying the unit prices, 32 states do not.[14] Regulations on the display

13. Leaner criterion of identifying temporary price discount directly relates to more items with temporary promotions. Ruling out more alternatives with temporary promotions in identifying quantity surcharge means the identified purchases of larger packages with quantity surcharge relates more to the information friction hypothesis.

14. States that enforces displaying unit prices are AR, CA, CT, DC, FL, HI, MA, MD, MS, MT, NH, NJ, NV, NY, OR, RI, VA, VT, WV.

of unit prices is an exogenous variation in information. Therefore, I divide the sample by states that do and do not require retail stores to display unit prices, and calculate the sales of larger packages with quantity surcharges. This variation is weak for two reasons. First, many retailers display unit prices voluntarily in states in which it is not required. Second, consumers can still calculate the unit prices even when the unit price is not displayed. Another exogenous variation in information friction that I use is displayed or featured items, which does not necessarily involve temporary price discounts. When an item is displayed or featured in a corresponding week and store, it obtains consumers' attention much more easily. Displayed or featured items are reported only in part of the data, and therefore I separately calculate the sales of larger packages with quantity surcharges for items with missing display and feature information.

Table 1.4 and 1.5 summarize the sales of larger packages with quantity surcharges by unit price display regulations, and whether an item was displayed or featured, respectively. Consider Table 1.4. Comparing the percentage sales of *Baseline* sales of larger package with quantity surcharges between states with and without unit price regulations, sales of larger packages with quantity surcharges amounts to 6.09% of total laundry detergent sales in states in which unit price display is enforced, whereas it increases to 6.39% in states in which unit price display is not enforced. The increment is about 5%. Since the data cover at least one-third of all laundry detergent sales in the United States, the differences could be taken as nearly exact, suggesting that variation in information affects the sales of larger packages with quantity surcharges. However, the tendency reverts weakly for sales of larger packages *Excluding Smaller Packages with ≥5%* and *≥10% Discounted/Featured.* I attribute this to the within-column difference, the percentage point of the third row drops much more from the second row in the "Not Mandatory" tab than in the "Mandatory" tab.

Table 1.5 demonstrates more differences in the percentage sales of larger packages with quantity surcharges across whether an item was displayed or featured. When an item is

Table 1.4: Sales of Larger Packages with Quantity Surcharge by Unit Price Display Regulations

	Unit Price Display				All States	
	Mandatory		Not Mandatory			
	Sales	% Sales	Sales	% Sales	Sales	% Sales
Baseline	$72M	6.09%	$96M	6.39%	$168M	6.25%
≤ Half Packs Available	$35M	2.94%	$56M	3.70%	$91M	3.37%
Excl. Smaller Packs ≥ 5% Disc/Feat	$27M	2.29%	$31M	2.06%	$58M	2.16%
Excl. Smaller Packs ≥ 10% Disc/Feat	$37M	3.15%	$45M	2.98%	$82M	3.05%
Excl. Smaller Packs ≥ 20% Disc/Feat	$56M	4.69%	$71M	4.74%	$127M	4.72%
Total Sales	$1185M	100%	$1507M	100%	$2692M	100%

Table 1.4 summarizes the sales of larger packages with quantity surcharge by the states that has the regulation to display the unit prices.

Table 1.5: Sales of Larger Packages with Quantity Surcharge by Displayed / Featured

	Displayed / Featured					
	Yes		No		Missing	
Corr. w/ Item is Priced w/ QS by Retailer	−0.049		−0.035		0.056	
Corr. w/ Package Size	−0.004		0.001		0.002	
	Sales	% Sales	Sales	% Sales	Sales	% Sales
Baseline	\$5.0M	3.53%	\$18.8M	8.43%	\$144.5M	6.21%
$\leq$ Half Packs Available	\$1.9M	1.33%	\$11.5M	5.13%	\$77.3M	3.32%
Excl. Smaller Packs $\geq$ 5% Disc/Feat	\$1.0M	0.97%	\$4.5M	2.03%	\$52.6M	2.26%
Excl. Smaller Packs $\geq$ 10% Disc/Feat	\$1.2M	1.17%	\$5.9M	2.65%	\$75.1M	3.23%
Excl. Smaller Packs $\geq$ 20% Disc/Feat	\$1.5M	1.50%	\$8.0M	3.57%	\$117.6M	5.05%
Total Sales	\$141.3M	100%	\$223.4M	100%	\$2327.3M	100%

Table 1.5 the sales of larger packages with quantity surcharge by whether the item is displayed or featured.

displayed or featured, it attracts consumers' attention more easily, and it is less likely for a consumer to buy a larger package with quantity surcharges if the consumer does not prefer a larger package over its smaller counterpart.

The first row, *Corr. w/ Item is Priced w/ QS by Retailer,* presents the correlation coefficients of the Display / Feature status with the indicator variable representing whether an item is priced with quantity surcharge by the retailer. I constructed the indicator variables for each Display / Feature status (Yes, No, Missing), and also for whether an item is sold with quantity surcharges. Then I calculated the correlation coefficients not weighting by the sales quantity. Correlation coefficients are close to zero in all three tabs, suggesting that promotion status itself is not correlated systematically with whether an item is priced with quantity surcharge or not by the retailer. The correlations presented in this row guard against the argument that suppliers using a lot of promotions may price the products with quantity surcharge less (or more). Similarly, the second row, *Corr. w/ Package Size*, presents the correlation coefficients of the Display / Feature status with the package size. Correlations are all very close to zero, showing that there is no correlation between the package size and promotion status.

Turning to the sales amount, however, the differences across columns are substantial. When an item is displayed or featured, only 3.53% of sales amount is subject to quantity surcharges, whereas it increases to 8.43% when an item is not displayed or featured. It suggests an interpretation that less attention from consumers associates with more sales of larger packages with quantity surcharges, supporting the hypothesis that sales of larger packages with quantity surcharges can be attributed at least partially to the inattention of consumers. The *% Sales* Column in *Missing* tab is very similar to that of *All States* tab in Table 1.4, suggesting that whether information about display or feature is missing is not subject to systematic sample selection.

Table 1.6: Conditional and Unconditional Switching Probabilities

Unconditional Prob.		Conditional Prob.	
Pr (Size Switch)	0.45		
Pr (Any UPC Switch)	0.77		
Pr (Product Switch)	0.67	Pr (Size Switch—Product Switch)	0.57
Pr (Brand Switch)	0.44	Pr (Size Switch—Brand Switch)	0.64

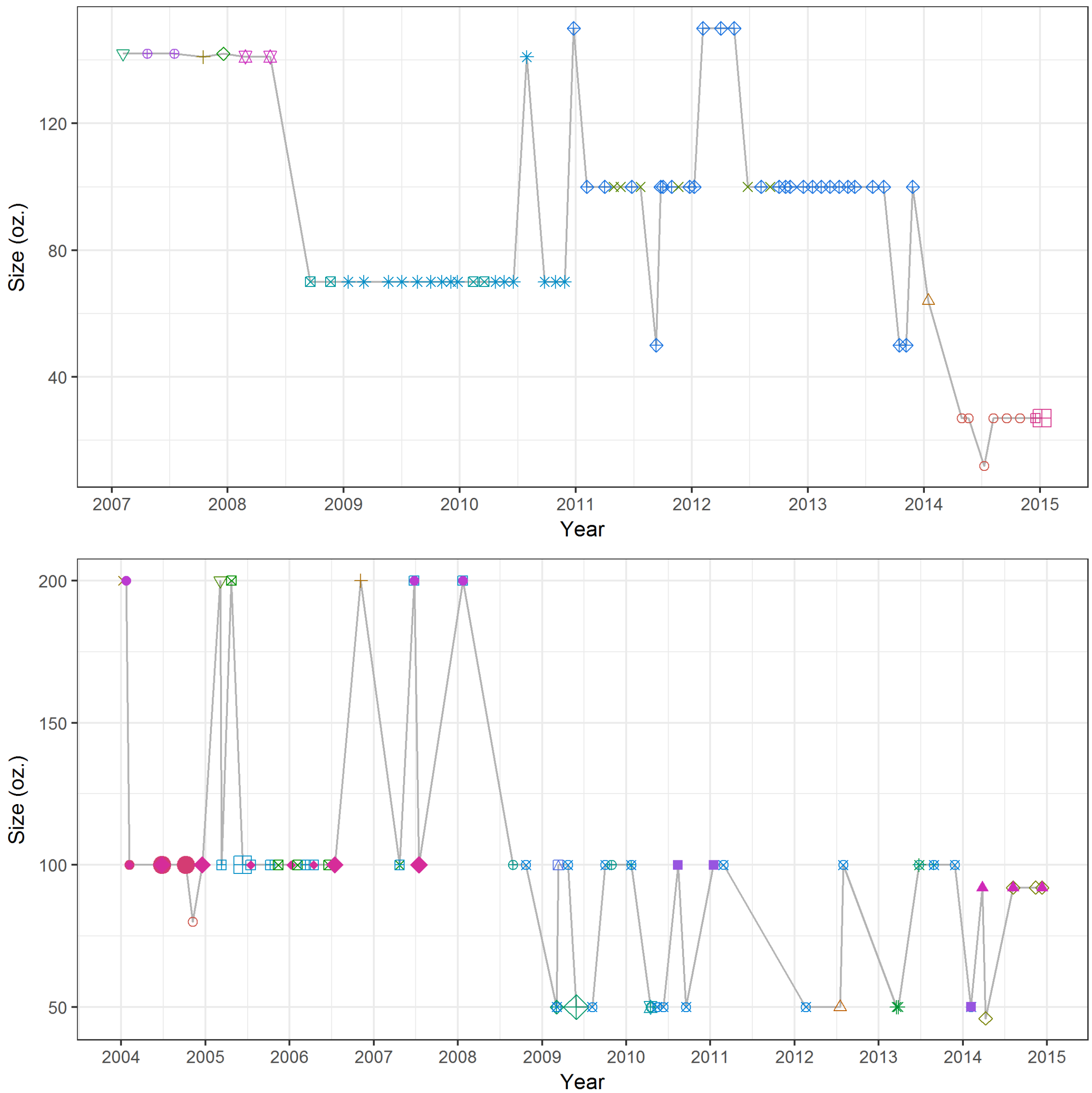
Table 1.6 presents the switching probabilities of households over time. Size switch is identified by purchasing a package that differs more than 20% from the previously purchased package. The switching probabilities are calculated after weighting for the sample projection factors. The effective sample size used to calculate the switching probabilities is 2582836 with 130922 households.

1.5.3 Households' Package Size Decisions Over Time

Another central premise in the modeling is that households have stable preferences for package sizes over a moderate period. I first present stylized package size decision patterns of households, and summarize package size switching probabilities. Package size decision patterns presented in this subsection show that households stick with specific package sizes over a moderate period.

Figure 1.5 plots the time series of two households' package size decisions. Different shapes of the points represent different products. In the top panel, the household buys 140oz packages from 2007 to mid-2008, and then switches to 75oz packages and stays there until the beginning of 2011. Then it switches to 100oz packages for three years, and ends up at around 28oz packages from the beginning of 2014. The bottom panel is noisier, yet the tendency is necessarily similar. From 2004 to 2008, the household oscillate around 150oz, buying 100oz and 200oz packages. After 2009, it switches to oscillate around 75oz, buying 50oz and 100oz packages. Preferences for package sizes are stable for a moderate period, such as a year, whereas there might be preference switches over a longer term. There is also frequent product-switching, even when a household sticks with the same package size. I plotted the same figure for every household in the panel, and similar patterns could be observed in the majority of households.

Figure 1.5: Individual Household Package Size Decisions Over Time



The figures plot the individual households' package size decisions over time. Different shapes of the points within a panel represent different products. The same shape across panels does not necessarily represent the same product. Sizes of the points are proportional to the number of packages purchased at the corresponding shopping instance. Top panel is the time series of the household ID 30012358, bottom panel is of the household ID 2019739.

Table 1.7: Cumulative Probability of Returning to the Same Package Size Within n Weeks

Weeks (n)	10	20	30	40	50	52
Cumulative Probability	0.64	0.70	0.73	0.75	0.76	0.77

Table 1.7 presents the cumulative probability of purchasing the same package size within the given period of time. Size switch is identified by purchasing a package that differs more than 20% from the previously purchased package. The switching probabilities are calculated after weighting for the sample projection factors. The effective sample size used to calculate the switching probabilities is 2582836 with 130922 households.

To provide more systematic evidence of the stability of household preferences over package sizes, I calculate the size switching probabilities and the probabilities that a household returns to the same size within a period. Table 1.6 summarizes the calculated switching probabilities. Households keep purchasing the same or similar sizes in 55% of shopping instances.[15][16] Even when they switch the product or brand, they stick to the same package size in 43% of shopping instances for product switching, and 36% for brand switching. Table 1.6 suggests that households have strong inertia when choosing package sizes during consecutive shopping instances.

Shown in Figure 1.5, households often return to the same package size after deviating just once or a few times from package sizes they used to purchase. To quantify temporary deviations, I calculate the cumulative probability of returning to the same package size within a certain period in Table 1.7. Households return to the same package size within a year in 77% of shopping instances. Considering that the remaining 23% of shopping instances contain a case in which a household tries a different package size just once or twice, and permanently returned to the package size they used to purchase, the effective returning probability within

15. The probability that the households stay in the same package size is $1 - \Pr(\text{Size Switch}) = 1 - 0.45 = 0.55$. The same applies for the conditional choice probabilities.

16. A household buying, for example, 100oz package and then 87oz package, can be considered to purchase virtually the same package size in a row. Therefore, size switching here is identified by purchasing a package that differs more than 20% from the previously purchased package size.

a year is much higher than 77%.[17]

1.6 The Model

I develop an econometric model of consumer choice that accommodates both information friction and preference heterogeneity over package sizes. The utility specification I develop has two components, per-ounce utility from the alternative choice and total utility from the quantity choice. Although I stick to the single choice assumption across items during the alternative choice stage, I allow consumers to purchase multiple packages of the same item.

Even if a consumer has full information on the unit price of each item in the choice set, disparate consumers might have different preferences over package sizes. Some might prefer one larger package over several of its smaller counterparts because it is more convenient to carry home and use, or it might reduce shopping instances, whereas others might prefer several smaller packages because the consumers have small storage shelves or they cannot use the product before it spoils.[18] My utility specification departs from the usual linear utility specification of product attributes. I allow for household-specific ideal point on package size, reflecting the prominent pattern observed in data that an individual household tends to keep purchasing a similar package size over time. The utility specification allows for the household-level heterogeneity of the preferences over package sizes, and thus accommodates the heterogeneity on the willingness to pay for different package sizes of the same product.

I summarize the choice procedure of consumers below. Precise definitions of notations that are not given here appear in subsequent subsections.

---

17. The year 2011 of the top panel in Figure 1.5 contains an example. Around July of 2011, the household deviated to 50oz package just once from 100oz packages that the household used to purchase in a row. The deviation from 100oz to 50oz package is counted as the returning purchase because the household immediately returns to 100oz package. On the other hand, however, the switch from 50oz to 100oz is considered to be not returning to 50oz within a year.

18. Even the laundry detergent, which is considered to be semi-durable, perishes after half to one year after opening.

1. A consumer walks into a store. She observes (i) the set of alternatives $\mathcal{J}_i$, (ii) the items and products in $\mathcal{J}_i$ that the consumer purchased before, and (iii) the items in promotion.

   (a) Using (i) through (iii), the consumer forms a prior belief $Q_i(\mathbf{u}_i)$ on the utility level of each alternative on the shelf.

   (b) Promotion and previous purchase history play roles as "action shifters" in forming the prior over $\mathbf{u}_i := \{u_{i,j}\}_{j\in\mathcal{J}_i}$, but they do not affect the level of the utility from consuming an alternative.

2. The consumer optimally chooses the learning strategy $P_i(\mathbf{s}_i|\mathbf{u}_i)$, which is a distribution of signals $\mathbf{s}_i$ conditional on each possible true utility vector $\mathbf{u}_i$.

3. Costly learning occurs (i.e., the consumer draws a signal from the chosen signal distribution, and updates his belief over the utility in a Bayesian fashion). The cost of learning is proportional to the entropy difference of the prior and posterior belief distributions.

4. The consumer chooses a package that maximizes utility according to the updated posterior belief.

5. For the chosen package, the consumer chooses the optimal number of packages to buy subject to the transportation cost.

This consumer-learning procedure might appear excessively complicated, but consumer learning in this context includes any action that a consumer must take to distinguish the identities of different products, such as recognizing the label of an item and calculating unit prices. This type of information-processing is necessary during each shopping instance.

The alternative choice model I employ eventually yields a choice probability expression that is very similar to the simple multinomial logit model in the discrete choice random utility maximization framework. The model nevertheless has a few conceptual advantages over its discrete choice random utility counterpart. First, the model provides a clear source of stochasticity, which is information friction. Since the information is costly, consumers might deviate from the alternative that yields the highest utility. The probabilistic expression from the rational inattention model is derived directly from economic and information theory. In the logit model, random utility shocks are interpreted as idiosyncratic utility shocks that are fully known to the agents, introduced to rationalize the data by simply introducing some stochasticity. Second, it separates utility components from information components. In discrete choice random utility models, the promotion or purchase history indicators are typically incorporated into the utility specification, though it is difficult to believe that advertisement, promotion, or previous purchase history shifts the utility level of consuming an alternative. In my model, consumers might prefer choosing an alternative because it is less costly to learn about the utility that the alternative provides, not because the alternative provides the highest utility. The model also gives a new interpretation to the degree of stochasticity μ_i, which is observationally equivalent to the standard deviation parameter in the Type-I extreme value distributed random utility shock term in the simple multinomial logit model when $\pi_{i,j} = 1/|\mathcal{J}_i|$ for all j.

The augmented logit form of the choice probability is driven by the functional form of the Shannon entropy difference. Shannon entropy measures uncertainty of a rational agent's belief distribution naturally in that it satisfies the following three axioms: (i) continuity, (ii) monotonically increasing over the number of alternatives when the belief distribution is uniform, and (iii) invariant to the breakdowns of choices to multiple intermediate steps. Shannon shows that entropy is the only form of the uncertainty measure that satisfies (i) through (iii). See Appendix 1.B.3 for a further discussion of a comparison of alternative choice models to the simple multinomial logit model.

1.6.1 Consumer Utility and Alternative Choice with Costly Learning

Let $\mathcal{J}_i = \{0, 1, ..., J_i\}$ denote the set of alternatives available to consumer i. $\mathcal{J}_i$ contains the outside option, which is denoted by option 0. Let $j\,(\in \mathcal{J}_i)$ denote an alternative, φ_j the package size, p_j the per-oz price, and $\mathbf{x}_j$ characteristics of item j, including brand dummies. Let r_i be the ideal point of consumer i's package size. Each alternative j denotes an item that is available on the shelf. $j'\,(\neq j)$ can denote the same product with j that has a different package size. Define the per-ounce utility of consumer i purchasing item j as:

$$\begin{aligned} u_{i,j} &:= \alpha - p_j\beta_1 + \mathbf{x}_j'\boldsymbol{\beta}_2 - \delta_1\rho\left(\varphi_j, r_i\right) \qquad \text{for } j \neq 0 \\ u_{i,0} &:= 0, \end{aligned} \tag{1.2}$$

where $\rho\left(\varphi_j, r_i\right)$ returns the distance of the purchased package's size from the consumer's ideal point. Intercept term α, which is to be normalized as 1, can be interpreted as the mean relative utility of consuming any laundry detergent in comparison to the outside option. If $\delta_1 = 0$, a consumer is indifferent with different package sizes of the same product. The utility is specified in a way that the same amount of the same product yields the same utility if $\delta_1 = 0$. I model the utility from all other product characteristics other than the package size to be linear in the variables because package size is the only continuous attribute in the laundry detergent data. Most other attributes are binary.[19]

The choice structure of the consumer is as follows. The consumer, who remembers which items and products she purchased in the last year, walks into the store. She immediately perceives the choice set $\mathcal{J}_i$, the promotion status. Let $\mathbf{D}_i := \left\{\mathbf{d}_{i,j}\right\}_{j\in\mathcal{J}_i}$ and $\boldsymbol{\eta}_i := \left\{\eta_{i,j}\right\}_{j\in\mathcal{J}_i}$ denote the observed and unobserved shifters of the prior belief on the utility $\mathbf{u}_i := \left\{u_{i,j}\right\}_{j\in\mathcal{J}_i}$,

19. The income effect is ignored in this utility specification because only a tiny portion of household consumer packaged goods spendings, and annual income, are spent on laundry detergents. On average, money spent on laundry detergents takes only 0.69% of annual household spendings on consumer packaged goods, and 8.19% of household annual income is spent on consumer packaged goods. Reflecting the fact that the income effect would be ignorable, I assume it away and proceed with discrete choice-based consumer demand modeling.

respectively. $\mathbf{d}_{i,j}$ includes the dummies of the previous purchase history and promotion status. This information, acquired without cost, results in a prior belief, which I denote $Q_i(\mathbf{u}_i) := Q(\mathbf{u}_i|\mathcal{J}_i, \mathbf{D}_i, \boldsymbol{\eta}_i)$. It is costly to learn the true utility $u_{i,j}$ that each alternative j provides. Let $\mathbf{s}_i := \{s_{i,j}\}_{j\in\mathcal{J}_i}$ be the vector of signals on the true utility. During the first stage, the consumer chooses the learning strategy, which is represented by the signal distribution $P_i(\mathbf{s}_i|\mathbf{u}_i)$[20] conditional on each possible true utility level $\mathbf{u}_i \left(\in \mathbb{R}^{|\mathcal{J}_i|}\right)$. During the second stage, the consumer draws signal $\mathbf{s}_i$ from $P_i(\mathbf{s}_i|\mathbf{u}_i)$, and then chooses alternative j out of $\mathcal{J}_i$ such that:

$$j = \arg\max_{j\in\mathcal{J}_i} \left\{ \int u_{i,j} P_i(d\mathbf{u}_i|\mathbf{s}_i) \right\}, \tag{1.3}$$

where $P_i(\mathbf{u}_i|\mathbf{s}_i)$ is the posterior belief after drawing signal $\mathbf{s}_i$.[21] Drawing signals in this context can be understood, for example, as recognizing the label of a product or comparing unit prices between items.

During the first stage, the consumer decides on the information strategy that maximizes the *ex-ante* expected utility net of the information cost. The information strategy is specified by the conditional distribution of signals $P_i(\mathbf{s}_i|\mathbf{u}_i)$, conditioned on each possible true value of $\mathbf{u}_i$. On the signal draws, the consumer formulates posterior belief $P_i(\mathbf{u}_i|\mathbf{s}_i)$ using Bayes' rule. To formalize, the first-stage problem is:

$$\max_{P_i\in\Delta\left(\mathbb{R}^{2|\mathcal{J}_i|}\right)} \int_{\mathbf{u}_i}\int_{\mathbf{s}_i} \left\{ \max_{j\in\mathcal{J}_i} \int u_{i,j} P_i(d\mathbf{u}_i|\mathbf{s}_i) \right\} P_i(d\mathbf{s}_i|\mathbf{u}_i)\, Q_i(d\mathbf{u}_i) - c_i(P_i(\mathbf{u}_i|\mathbf{s}_i))$$

$$s.t. \quad Q_i(\mathbf{u}_i) = \int_{\mathbf{s}_i} P_i(d\mathbf{s}_i, \mathbf{u}_i) \qquad \forall \mathbf{u}_i \in \mathbb{R}^{|\mathcal{J}_i|}. \tag{1.4}$$

20. With a little abuse of notation, I denote $P_i(\mathbf{s}_i|\mathbf{u}_i) = P_i(\mathbf{s}_i|\mathbf{u}_i; \mathcal{J}_i, \mathbf{D}_i, \boldsymbol{\eta}_i)$, $P_i(\mathbf{u}_i|\mathbf{s}_i) = P_i(\mathbf{u}_i|\mathbf{s}_i; \mathcal{J}_i, \mathbf{D}_i, \boldsymbol{\eta}_i)$, and $\Pr_i(i \text{ Chooses } j|\mathbf{u}_i) = \Pr(i \text{ Chooses } j|\mathbf{u}_i; \mathcal{J}_i, \mathbf{D}_i, \boldsymbol{\eta}_i)$.

21. I am imposing the single choice assumption in the alternative choice. In my data, households bought more than two different laundry detergent UPCs in less than only 12% of the shopping trips, which I take as separate shopping instances in the estimation.

Cost function $c_i(P_i(\mathbf{u}_i|\mathbf{s}_i))$ is defined as proportional to the mutual information of signal and utility $\mathcal{I}(\mathbf{s}_i;\mathbf{u}_i)$, or equivalently, $E_{\mathbf{s}_i}[D_{KL}(P_i(\mathbf{u}_i|\mathbf{s}_i)\|Q_i(\mathbf{u}_i))]$, the expected Kullback-Leibler divergence of the prior $Q_i(\mathbf{u}_i)$ and the posterior $P_i(\mathbf{u}_i|\mathbf{s}_i)$. Formally,

$$c_i(P_i(\mathbf{u}_i|\mathbf{s}_i)) := \lambda_i \mathcal{I}(\mathbf{s}_i, \mathbf{u}_i), \tag{1.5}$$

where λ_i is the unit information cost.[22] For the sake of notational convenience, I denote $\mu_i := \lambda_i^{-1}$ as the reciprocal of individual i's unit information cost.

In this setup, the consumer can acquire any kind of information in the form of signal. The only behavioral restriction is that the signal becomes more expensive as it becomes more precise, where precision is measured by the mutual information of the signal and the true utility. If the signal $\mathbf{s}_i$ does not give any information about the true consumption utility $\mathbf{u}_i$, the cost of information is zero, but the decision becomes imprecise. When the signal gives more precise information about $\mathbf{u}_i$, the decision becomes more precise, but the cost of information becomes larger. Given this tradeoff, the consumer optimizes over the conditional distribution of signal $P_i(\mathbf{s}_i|\mathbf{u}_i)$ *ex-ante* to maximize the expected utility net of the information cost.

It is shown in Matějka and McKay (2015) that (1.5) can be formulated as the entropy

22. When densities exist,

$$\mathcal{I}(\mathbf{s}_i;\mathbf{u}_i) = \varrho(Q_i(\mathbf{u}_i)) - E_{\mathbf{s}_i}[\varrho(P_i(\mathbf{u}_i|\mathbf{s}_i))]$$

holds, where

$$\begin{aligned}\varrho(Q_i(d\mathbf{u}_i)) &:= -\int [\ln Q_i(\mathbf{u}_i)]\, Q_i(d\mathbf{u}_i)\\ \varrho(P_i(d\mathbf{u}_i|\mathbf{s}_i)) &:= -\int [\ln P_i(\mathbf{u}_i|\mathbf{s}_i)]\, P_i(d\mathbf{u}_i|\mathbf{s}_i)\\ E_{\mathbf{s}_i}[\varrho(P_i(d\mathbf{u}_i|\mathbf{s}_i))] &= -\iint [\ln P_i(\mathbf{u}_i|\mathbf{s}_i)]\, P_i(d\mathbf{u}_i|\mathbf{s}_i)\, P_i(d\mathbf{s}_i).\end{aligned}$$

(1.5) is proportional to prior and posterior uncertainties expressed in terms of the Shannon entropy difference.

difference of the prior and posterior choice probabilities, as:

$$\lambda_i \left\{ -\sum_{j\in\mathcal{J}_i} \pi_{i,j} \ln \pi_{i,j} + \int_{\mathbf{u}_i} \sum_{j\in\mathcal{J}_i} \Pr_i\left(i \text{ Chooses } j|\mathbf{u}_i\right) \ln \Pr_i\left(i \text{ Chooses } j|\mathbf{u}_i\right) Q_i\left(d\mathbf{u}_i\right) \right\},$$

where $\pi_{i,j}$ is the unconditional choice probability defined by

$$\pi_{i,j} := \int \Pr_i\left(i \text{ Chooses } j|\mathbf{u}_i\right) Q_i\left(d\mathbf{u}_i\right).$$

Then, they show that the solution for the problem described in (1.3)-(1.5) predicts the posterior choice probability after the costly consumer learning as:

$$\Pr_i\left(i \text{ Chooses } j|\mathbf{u}_i\right) = \frac{\pi_{i,j} \exp\left(\mu_i u_{i,j}\right)}{\sum_{j'\in\mathcal{J}_i} \pi_{i,j'} \exp\left(\mu_i u_{i,j'}\right)} \quad \text{a.s.} \tag{1.6}$$

I sketch the derivation of (1.6) in Appendix 1.B.1.

For empirical tractability, I specify further structures on (1.6). First, I model that the prior belief distribution $Q_i\left(\mathbf{u}_i\right) \equiv Q\left(\mathbf{u}_i; \mathcal{J}_i, \mathbf{D}_i, \boldsymbol{\eta}_i\right)$ is affected only by the choice set $\mathcal{J}_i$ and the dummies $\mathbf{D}_i$, which represent the previous purchase history and promotion status. I parametrize $\pi_{i,j} \equiv \pi_j\left(\mathcal{J}_i, \mathbf{D}_i, \boldsymbol{\eta}_i\right)$ as proportional to $\exp\left(\mathbf{d}'_{i,j}\boldsymbol{\gamma} + \eta_{i,j}\right)$, where $\mathbf{d}'_{i,j}\boldsymbol{\gamma}$ captures the observed action shifter, $\eta_{i,j}$ captures the unobserved action shifter. The specification leads to the following multinomial logit form:

$$\pi_{i,j} = \frac{\exp\left(\mathbf{d}'_{i,j}\boldsymbol{\gamma} + \eta_{i,j}\right)}{\exp\left(\mathbf{d}'_{i,0}\boldsymbol{\gamma}\right) + \sum_{j'\in\mathcal{J}_i\setminus 0} \exp\left(\mathbf{d}'_{i,j'}\boldsymbol{\gamma} + \eta_{i,j'}\right)}. \tag{1.7}$$

Previous purchase history and promotions $\mathbf{d}_{i,j}$ play a role as a choice action shifter, but do not affect the utility that consumer i gets from consuming an alternative j. Next, I model consumers' costs of information that depend on demographics. To be specific,

$$\mu_i := \exp\left(\mathbf{w}_i' \boldsymbol{\theta}\right)$$

where $\mathbf{w}_i$ is the vector of demographics that does not include the constant term.

In Corollary 2 of Matějka and McKay (2015), it is shown that for $\left\{\pi_{i,j}\right\}_{j\in\mathcal{J}_i}$ such that $\pi_{i,j} > 0$, the equality constraints of (1.4) leads to the condition that the prior distribution $Q_i(\cdot)$ must satisfy the following fixed point equation:

$$\int_{\mathbf{u}_i} \frac{\pi_{i,j} \exp\left(\mu_i u_{i,j}\right)}{\sum_{j'\in\mathcal{J}_i} \pi_{i,j'} \exp\left(\mu_i u_{i,j'}\right)} dQ_i\left(\mathbf{u}_i\right) = \pi_{i,j}. \tag{1.8}$$

In order for an estimate of the prior choice probabilities $\left\{\pi_{i,j}\right\}_{j\in\mathcal{J}_i}$ to be justified as consistent with the rational inattention theory, there has to exist a prior distribution $Q_i(\cdot)$ over the utilities that yields $\left\{\pi_{i,j}\right\}_{j\in\mathcal{J}_i}$. In the rational inattention theory literature, the prior distribution on the payoff $Q_i(\cdot)$ is assumed to be known, and then the unconditional choice probability $\pi_{i,j}$'s are solved so as to satisfy the set of fixed-point equations (1.8). The direction taken here is the reverse, because the goal is to estimate the utility and information parameters using the choice data.

The following theorem justifies that any combination of $\left\{\mu_i, \left\{\pi_{i,j}\right\}_{j\in\mathcal{J}_i}\right\}$ backed out from choice data can be rationalized as a rationally inattentive agent's choice. It asserts that, for any given $\mu_i > 0$ and prior choice probabilities $\left\{\pi_{i,j}\right\}_{j\in\mathcal{J}_i}$, there exists a prior distribution $Q_i(\cdot)$ that yields $\left\{\pi_{i,j}\right\}_{j\in\mathcal{J}_i}$. Therefore, the prior choice probabilities backed out directly from a choice data can be justified as resulting from a rationally inattentive agent's choice.

Theorem 1.6.1. *(Existence of Prior Distribution) Fix i. Let $J_i\,(\in\mathbb{N}) \geq 2$ and let $\mathcal{J}_i = \{0, 1, ..., J_i\}$. For each $j \in \mathcal{J}_i$, let $\pi_{i,j} > 0$ such that $\sum_{k\in\mathcal{J}} \pi_{i,k} = 1$ be given. Let $\mu_i > 0$ be also given. Then, there exists a probability measure Q_i over $\mathbb{R}^{J_i}$ such that, for each $j\,(\in\mathcal{J}_i) \neq 0$, the following (i) and (ii) hold:*

(i) For each $j\ (\in \mathcal{J}_i) \neq 0$,

$$\pi_{i,j} = \int \frac{\pi_{i,j} \exp\left(\mu_i u_{i,j}\right)}{\sum_{k\in\mathcal{J}_i} \pi_{i,k} \exp\left(\mu_i u_{i,k}\right)} Q_i\left(d\mathbf{u}_i\right). \tag{1.9}$$

(ii) For each $j \in \mathcal{J}_i$, $Q_i\left(\left\{\mathbf{u}_i \in \mathbb{R}^{J_i} : u_{i,j} > u_{i,k}\ \forall k \neq j\right\}\right) > 0$.

Proof. See Appendix 1.A. □

(i) is the necessary condition given in Corollary 2 of Matějka and McKay (2015), and (ii) is the condition that requires the constructed prior distribution $Q_i(\cdot)$ is indeed coherent with $\pi_{i,j} > 0$ for all $j \in \mathcal{J}_i$. I show the existence by constructing a probability density function $q_i : \mathbb{R}^{J_i} \to (0,1)$ corresponding to Q_i, of which the support is $\mathbb{R}^{J_i}$. Although the class of density functions I construct here is a mixture resulting from the Dirichlet distribution, there can be other possibilities.

The specifications above lead to the following likelihood expression of consumer i purchasing alternative j:

$$\begin{aligned}
&E_{\boldsymbol{\eta}_i}\left[\Pr_i\left(i \text{ Chooses } j|\mathbf{u}_i\right)\right] \\
&= \int \frac{\exp\left(\mathbf{d}'_{i,j}\boldsymbol{\gamma} + \eta_{i,j} + \mu_i\left[\alpha - p_j\beta_1 + \mathbf{x}'_j\boldsymbol{\beta}_2 - \delta_1\rho\left(\varphi_j, r_i\right)\right]\right)}{1 + \sum_{j'\in\mathcal{J}_i\backslash 0}\exp\left(\mathbf{d}'_{i,j'}\boldsymbol{\gamma} + \eta_{i,j'} + \mu_i\left[\alpha - p_{j'}\beta_1 + \mathbf{x}'_{j'}\boldsymbol{\beta}_2 - \delta_1\rho\left(\varphi_{j'}, r_i\right)\right]\right)} dF\left(\boldsymbol{\eta}_i; \varsigma\right).
\end{aligned} \tag{1.10}$$

I take the alternative choice probability (1.10) as the first building block of the full model likelihood. In the estimation, I treat $\eta_{i,j}$ as random effects that follow the Gaussian distribution with variance ς^2 for tractability, and replace (1.10) with its simulation analogue and estimate model parameters using maximum simulated likelihood (Gouriéroux and Monfort (1990); Lee (1992)).

1.6.2 Quantity Choice

After the consumer chooses which item to buy, she chooses how many packages to buy. Let φ_j denote the package size of item j. At each shopping instance, consumers decide on product j, package size φ_j, and number of packages n_j to buy. $q_j := n_j\varphi_j$ is the quantity of item j purchased in ounces. The utility of consumer i purchasing n_j packs of item j is:

$$\bar{v}_{i,j}\left(n_j\right) := q_j u_{i,j} - \delta_2 q_j^2 \tag{1.11}$$

$$= n_j\varphi_j \left\{\alpha - p_j\beta_1 + \mathbf{x}_j'\boldsymbol{\beta}_2 - \delta_1\rho\left(\varphi_j, r_i\right)\right\} - \delta_2\left(n_j\varphi_j\right)^2. \tag{1.12}$$

The transportation cost term $\delta_2\left(n_j\varphi_j\right)^2$ is introduced to avoid $n_j = \infty$ at the optimum. Without the term, it is optimal for the consumer to choose $n_j = \infty$ whenever $u_{i,j}$ is positive. δ_2 represents the unit transportation cost, or diminishing marginal utility of consuming alternative j.

Since n_j must be an integer, $\bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j - 1\right)$ and $\bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j + 1\right)$ should hold at the optimum. This leads to the following inequalities if the quantity choice is fully deterministic:

$$\varphi_j u_{i,j} - \delta_2\left(2n+1\right)\varphi_j^2 \leq 0 \leq \varphi_j u_{i,j} - \delta_2\left(2n-1\right)\varphi_j^2.$$

To accommodate the shifters in quantity decision that are not observable to the econometrician such as optimization errors, define an unobservable $\epsilon_{i,j} \sim \text{iid } \mathcal{N}\left(0, \sigma^2\right)$, where:

$$\varphi_j u_{i,j} - \delta_2\left(2n+1\right)\varphi_j^2 + \epsilon_{i,j} \leq 0 \leq \varphi_j u_{i,j} - \delta_2\left(2n-1\right)\varphi_j^2 + \epsilon_{i,j}.$$

$\Pr\left(\bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j - 1\right), \bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j + 1\right)\right)$, the conditional probability of purchasing n_j packages given alternative j is chosen, becomes:

$$\Pr\left(\varphi_j u_{i,j} - \delta_2\left(2n_j+1\right)\varphi_j^2 \leq \epsilon_{i,j} \leq \varphi_j u_{i,j} - \delta_2\left(2n_j-1\right)\varphi_j^2\right). \tag{1.13}$$

Quantity choice probability (1.13) is the second building block of the likelihood expression.

1.6.3 Likelihood and Inference

I consider panel projection factors in the likelihood. Let ω_i be the projection factor of the household. The pseudo-likelihood of observing the consumer choice data is then:

$$\begin{aligned}
&\prod_{i,j} E_{\boldsymbol{\eta}_i}\left[\Pr_i\left(i \text{ Chooses } j|\mathbf{u}_i\right)\right]^{\omega_i \mathbf{1}(i \text{ Chooses } j)} \\
&\times \Pr\left(\bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j-1\right), \bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j+1\right)\right)^{\omega_i \mathbf{1}(\text{Purchased } n_j \text{ Packs})}.
\end{aligned} \tag{1.14}$$

The asymptotic covariance matrix is slightly different from the usual maximum likelihood estimator. Sample weights must be considered when calculating the asymptotic covariance matrix. Let ∇ denote the gradient operator, and

$$\begin{aligned}
\nabla_{i,j} := \nabla \Big[\omega_i \Big\{&\ln E_{\boldsymbol{\eta}_i}\left[\Pr_i\left(i \text{ Chooses } j|\mathbf{u}_i\right)\right] \\
&+ \ln \Pr\left(\bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j-1\right), \bar{v}_{i,j}\left(n_j\right) \geq \bar{v}_{i,j}\left(n_j+1\right)\right)\Big\}\Big]
\end{aligned}$$

the gradient for each sample evaluated at the optimum. The asymptotic covariance matrix formula for the estimator is:

$$\mathbf{V} = \boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}, \tag{1.15}$$

where

$$\begin{aligned}
\boldsymbol{\Omega} &:= -\sum_{i,j} \nabla\left(\nabla'_{i,j}\right) \\
\boldsymbol{\Delta} &:= \sum_{i,j}\left(\nabla'_{i,j}\nabla_{i,j}\right).
\end{aligned}$$

I use (1.15) in the following sections for the inference on model parameter estimates.

1.7 Estimation and Results

1.7.1 Estimation Data and Empirical Strategy

I use the matched sample of the panel and scanner data to estimate model parameters. Consumer panel data records only purchased items by participating households throughout the year, and thus lacks information on the choice set of households. To address this problem, I match scanner data with consumer panel data. Many shopping trips do not have retailer and store information. Approximately 40% of trip data were matched with scanner data. I dropped unmatched shopping trip samples during the following analysis using consumer panel data. For sample year 2012, I have 249,187 choice samples. 65,879 observations are matched with the scanner data (i.e., store and week information were available in the panel data), and the scanner data have a record of the same item sold during the corresponding week in the same store. The previous year's purchase history of a household is also required to incorporate the previous purchase history variable during estimation. Since only about 80% of households participated in the panel two years in a row, I have 53,590 choice observations to estimate the preferred model specification that has the history of the previous year's laundry detergent purchases. The weighted average number of alternatives that a consumer faces during each shopping trip is about 86. Therefore, to estimate the model for the 53,590 choice observations, the sample size that I need to consider is as large as 4.6 million.

A few remarks about generating the estimation data are necessary. I exploit the panel structure to find the proxy for the ideal point r_i of the package size. Consumer i's ideal point is taken as the mean of the purchased packages throughout the year. I use $\rho\left(\varphi_j, r_i\right) = \sqrt{\left(\varphi_j - r_i\right)^2}$ as the distance measure between the consumer's ideal point r_i and the alternative j's package size φ_j.[23] Since I do not impose any restriction on $-\delta_1$ during estimation, testing the alternative hypothesis that $-\hat{\delta}_1$ is negative against the null hypothesis

23. I tried different distance measures such as $(\varphi_j - r_i)^2$. The estimation results were robust to the specification of distance measure.

that $-\hat{\delta}_1 \geq 0$ proxies whether the ideal point modeling assumption is appropriate. Regarding promotion status indicators, I consolidate the *Deal* flag in the panel data and the *Featured/Displayed* flags in the scanner data. Since the set of the same UPCs is a strict subset of the set of the same product, I create the variable *Same Product (Different UPC) Purchased Within 1 Year* to be mutually exclusive to the variable *Same UPC Purchased Within 1 Year* for ease of interpretation. To improve the numerical stability of the estimator, I consolidate some similar product characteristics that do not appear often in the sample. Five variables were construed for the consolidated product characteristics variables: *Oxi-clean/Baking soda*, *Unscented/Sensitive/ Baby*, *Low Chlorine/Low Sulfate/Low Phosphorus*, and *Tablet/Sheet*. Binary indicator variables for the seven large brands were also included in product characteristics.

All stochasticity of the alternative choice model is attributed to information friction. Stated differently, identification of the model parameters relies on the assumption that information varies across shopping instances, whereas individual consumer's preference is stable over time.

1.7.2 Model Parameter Estimates and Interpretation

I present model parameter estimates and their interpretations in this subsection. Table 1.8 presents model parameter estimates for the preferred model specification, Figure 1.6 illustrates the estimated marginal density of the information sensitivity parameter μ_i, and Figure 1.7 the proxy for consumers' ideal point r_i. The joint distribution of (μ_i, r_i) fully characterizes the observed heterogeneity of demographics, given the structure of the demand model.

I begin interpretation of the model parameter estimates by recalling the utility specifica-

Table 1.8: Model Parameter Estimates

Utility Parameter $\left(-\hat{\beta}_1, -\hat{\delta}_1, \hat{\boldsymbol{\beta}}_2, -\hat{\delta}_2, \hat{\sigma}^{-1}\right)$				Information Parameter $(\hat{\boldsymbol{\gamma}}, \hat{\varsigma})$	
Per-oz Price $\left(-\hat{\beta}_1\right)$	-0.0601^{*}	Stain Remover	0.0227^{***}	Displayed / Featured / Deal	5.9930^{***}
	(0.0341)	/ Deep Clean	(0.0044)		(0.0445)
$\sqrt{(\varphi_j - r_i)^2}\left(-\hat{\delta}_1\right)$	-0.0016^{***}	Unscented /	-0.0046^{**}	Same UPC	3.1626^{***}
	(0.0001)	Sensitive / Baby	(0.0022)	Purchased Within 1 Year	(0.0285)
Powder	-0.0007	Low Cl / S / P	0.0071^{**}	Same Product (Different UPC)	2.0553^{***}
	(0.0030)		(0.0036)	Purchased Within 1 Year	(0.0420)
Fabric Softener	-0.0009	Tablet / Sheet	0.0391^{***}	sd $(\eta_{i,j})$ $(\hat{\varsigma})$	0.0017
	(0.0112)		(0.0063)		(0.0033)
Febreze	-0.0019	All	0.0152^{***}		
	(0.0032)		(0.0036)	Information Cost Parameter $\hat{\boldsymbol{\theta}}$	
All Temperature	0.0073	Arm&Hammer	0.0259^{***}	'10 Income Ratio to FPL	0.0960^{***}
	(0.0045)		(0.0038)		(0.0116)
Bleach	0.0004	Gain	-0.0207^{***}	Apartment	0.2106^{***}
	(0.0035)		(0.0039)		(0.0634)
Ultra	0.0031	Purex	0.0277^{***}	Non-working Spouse	0.5993^{***}
	(0.0020)		(0.0037)		(0.0577)
$n\times$Concentrated	0.0006	Tide	-0.0140^{***}	Household Size	0.2441^{***}
	(0.0013)		(0.0033)		(0.0179)
High Efficiency	0.0103^{***}	Wisk	-0.0110^{**}	Head Employment	0.9847^{***}
	(0.0018)		(0.0046)		(0.1412)
Oxi-Clean	-0.0062^{*}	Xtra	0.0700^{***}	Head College Degree	-0.0077
/ Baking Soda	(0.0036)		(0.0094)		(0.0474)
Colorsafe	0.0026	Total oz^2 $\left(-\hat{\delta}_2\right)$	-0.0042^{***}	Married,	-0.3440^{***}
	(0.0039)		(0.0001)	Living Together	(0.0597)
Soft	-0.0090	$\hat{\sigma}^{-1}$	0.0240^{***}	No Child	0.8426^{***}
	(0.0107)		(0.0003)		(0.0923)
Choice Obs.	53590	Sample Size	4617623	$\bar{\mu} = \overline{\exp\left(\mathbf{w}_i'\hat{\boldsymbol{\theta}}\right)}$	12.6324

*p<0.1; **p<0.05; ***p<0.01

This table summarizes the model parameter estimation results from the maximum simulated likelihood estimation. $\bar{\mu} = \overline{\exp\left(\mathbf{w}_i'\hat{\boldsymbol{\theta}}\right)}$ is calculated after weighting for the panel projection factor. Due to the utility specification of the model, μ_i should be multiplied to the utility parameter coefficients $\left(-\hat{\beta}_1, -\hat{\delta}_1, \hat{\boldsymbol{\beta}}_2, -\hat{\delta}_2, \hat{\sigma}^{-1}\right)$ to find the effective magnitudes. Comparing the magnitudes of $\left(-\hat{\beta}_1, -\hat{\delta}_1, \hat{\boldsymbol{\beta}}_2, -\hat{\delta}_2, \hat{\sigma}^{-1}\right)$ with the promotion parameter $\hat{\boldsymbol{\gamma}}$ estimates is only sensible after multiplying μ_i's to $\left(-\hat{\beta}_1, -\hat{\delta}_1, \hat{\boldsymbol{\beta}}_2, -\hat{\delta}_2, \hat{\sigma}^{-1}\right)$. Federal Poverty Level (FPL) of 2010 is \$10,840 for the first person in the household, with \$3,740 for each additional person. The data used for the estimation is the matched sample of laundry detergent purchase in the 2012 Nielsen-Kilts consumer panel data and scanner data, that also participated in the household panel in 2011. Standard error estimates are in the parentheses.

Figure 1.6: Density of μ_i

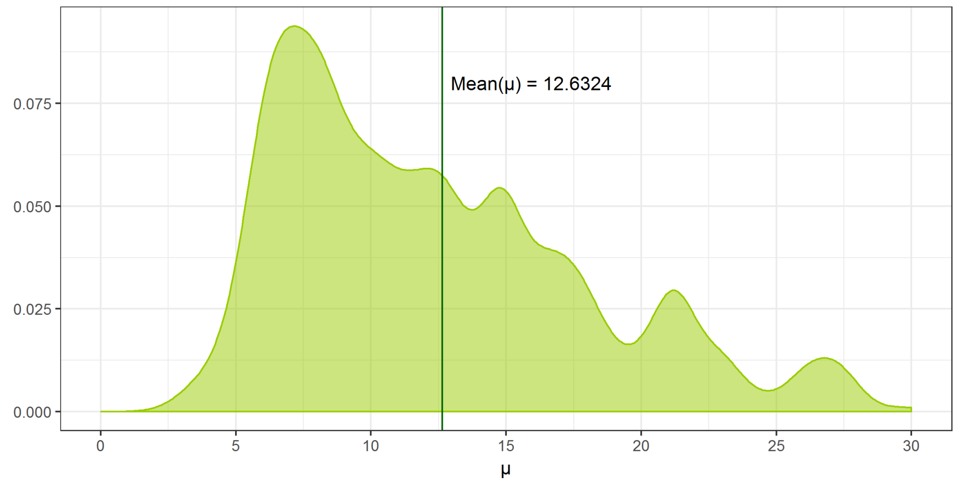


Figure 1.6 illustrates the estimated marginal density of μ_i, corresponding to the $\hat{\boldsymbol{\theta}}$ estimates in Table 1.8. The density is calculated after weighting for the panel projection factor. The horizontal axis is truncated at 30.

Figure 1.7: Density of r_i

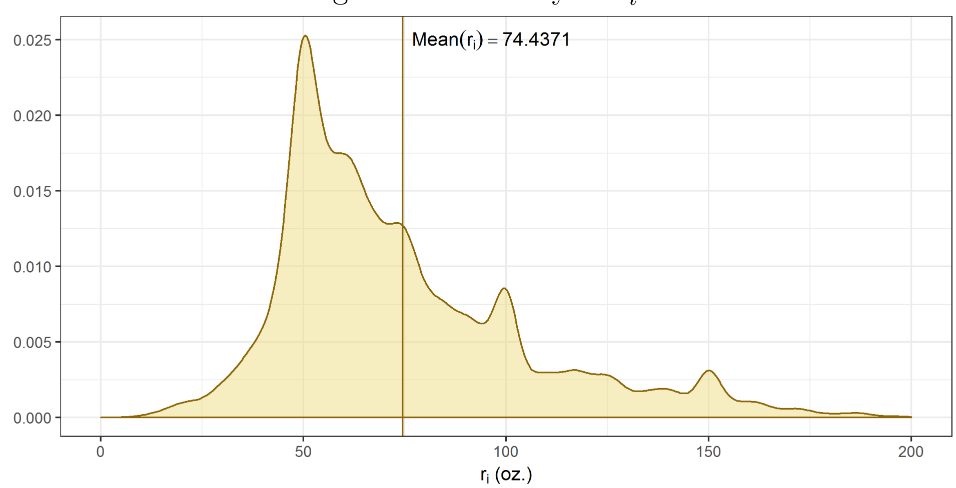


Figure 1.7 illustrates the marginal density of r_i. The density is calculated after weighting for the panel projection factor. The horizontal axis is truncated at 200.

tion. The utility of consumer i purchasing n_j packages of the alternative j is:

$$\bar{v}_{i,j}\left(n_j\right) = n_j\varphi_j\left(1 - p_j\beta_1 + \mathbf{x}_j'\boldsymbol{\beta}_2 - \delta_1\sqrt{\left(\varphi_j - r_i\right)^2}\right) - \delta_2\left(n_j\varphi_j\right)^2. \tag{1.16}$$

1 is the normalized level of consuming 1 ounce of laundry detergent in comparison to the outside option. The magnitudes of $\left(-\hat{\beta}_1, -\hat{\delta}_1, \hat{\boldsymbol{\beta}}_2, -\hat{\delta}_2, \hat{\sigma}^{-1}\right)$ estimates in Table 1.8 are relative to 1.

Model parameters of interest are the coefficients of *Per-oz Price* and $\sqrt{\left(\varphi_j - r_i\right)^2}$, both of which have the expected signs, where $-\hat{\delta}_1$ is negative and highly statistically significant, suggesting that the ideal point model was an appropriate modeling assumption.[24] This requires caution in interpreting the promotion parameter $\hat{\boldsymbol{\gamma}}$. Although *Displayed/Featured/Deal* and *Same UPC/Product Purchased Within 1 Year* are indicator variables, the size of the coefficient estimates should not be directly compared to those of the utility parameter estimates. The correct comparison to the utility parameters $\left(-\hat{\beta}_1, -\hat{\delta}_1, \hat{\boldsymbol{\beta}}_2\right)$ is made only after scaling $\hat{\boldsymbol{\gamma}}$ with $1/\mu_i$. Since μ_i varies with each individual, I present the mean estimates of μ_i in the bottom of the table, and its estimated density in Figure 1.6. Promotion and previous purchase history still have strong effects after dividing by $\bar{\mu}$, which are 0.47, 0.25, and 0.16, respectively, but not as drastic as they appear to be without scaling for information sensitivity.

According to the alternative choice probability expression that I derive in equation (1.10), utility parameter estimates $\left(-\hat{\beta}_1, -\hat{\delta}_1, \hat{\boldsymbol{\beta}}_2\right)$ are multiplied by μ_i, which varies over i. Households with larger μ_is have smaller costs to acquire information about each alternative, and thus they respond more sensitively to the utility from the product attributes in the alternative choice. When $\mu_i \to \infty$, the household simply knows the utility of every alternative. There is then no uncertainty in the alternative choice, and a promotion or previous pur-

24. The coefficient of *Per-oz Price* turns out to be only statistically significant at the 10% level. It is because the price variable is correlated with the previous purchase history variables, which absorb some of the relevant variations.

chase history does not play any role as the alternative choice probability shifter. Higher income, living in an apartment, non-working spouse, household size, head employment, and not having a child reduce information costs for households, inducing them to respond more sensitively to the utility of consuming an alternative. Unlike a typical search theoretic model, a coherent time cost of information story does not explain the coefficient estimates of $\hat{\boldsymbol{\theta}}$ well.

1.8 Model Predictions and Counterfactuals

1.8.1 Decomposing the Contribution of Information Friction and Preference Heterogeneity in Package Sizes

Using model parameter estimates from the previous section, I calculate consumption utility prediction of each alternative. The utility predictions allow me to attribute the cause of each larger package purchase with quantity surcharges, whether due to information friction or preference heterogeneity over package sizes. Buying a larger package with quantity surcharges cannot be attributed solely to information friction because a larger package might yield higher utility for a consumer when she prefers a larger package. A consumer who buys a larger package with quantity surcharges can be attributed to information friction only when there is a smaller item that yields a higher utility to the consumer, with a lower unit price.

Consumers might have different willingness to pay for different package sizes of an identical product. Consider the alternative choice problem specified in Section 1.6.1. The per-ounce utility of item j is specified by:

$$u_{i,j} := \alpha - p_j\beta_1 + \mathbf{x}_j'\boldsymbol{\beta}_2 - \delta_1\rho\left(\varphi_j, r_i\right).$$

Suppose items j and j' are the same products that differ only on their package sizes. To achieve a specific level of utility $\bar{u}$, per-ounce prices p_j and $p_{j'}$ should be different for different package sizes φ_j and $\varphi_{j'}$, given that r_i is fixed for individual i and $\mathbf{x}_j = \mathbf{x}_{j'}$. This implies that some portion of the sales of larger packages with quantity surcharges occur simply because larger packages yield higher utility for the consumer. To separate out the contribution of information friction, smaller packages that yield higher per-ounce utility should be removed when calculating sales of larger packages with quantity surcharges. Since the utility is never directly observed to an econometrician, using estimates and predictions from the model is necessary to find which package size yields higher utility for a consumer.

Tables 1.9 and 1.10 presents the decomposition of sales of larger packages with quantity surcharges using the predicted utility from model parameter estimates. Table 1.9 uses choice data, and Table 1.10 predicted choice from the model. Columns (1) in each table are the dollar and percentage sales of larger packages with quantity surcharges, calculated using the same formula as in Section 1.4. When calculating Columns (2), quantity surcharges were identified using only the smaller alternatives that yielded a higher per-ounce utility than the alternative. Columns (3) are the fractions of Columns (2) in comparison to Columns (1).

Consider Column (1) in Table 1.10. In comparison to Column (1) in Table 1.9, the model accurately predicts the sales of larger packages with quantity surcharges in terms of dollars. However, the model over-predicts total sales of laundry detergents, and therefore the percentage prediction of the sales of larger packages with quantity surcharges is lower than the choice data. Columns (2) and (3) of Tables 1.9 and 1.10 are the primary results of this subsection. 36.92% of *Baseline* sales of larger packages with quantity surcharges can be attributed to information friction when choice data are used, which increases to 42.25% when the predicted choice is used. Only about 40% of the sales of larger packages with quantity surcharges can be attributed to information friction, and they amount to about 2% of total sales of laundry detergent. The remaining 60% of sales of larger packages with quantity surcharges can be attributed to consumers having higher willingness to pay for larger packages.

Table 1.9: Decomposition Using the Actual Choice Data

	(1) Sales s.t. QS		(2) Sales s.t. QS by Info. Friction		(3)
	$ Sales	% Total Sales	$ Sales	% Total Sales	(2)/(1)
Baseline	$43.0M	6.01%	$15.9M	2.22%	36.92%
$\leq$ Half Packs Available	$19.4M	2.71%	$5.2M	0.73%	26.85%
Excluding Smaller Packs $\geq$ 5% Discount/Feat	$22.0M	3.07%	$8.7M	1.21%	39.55%
Excluding Smaller Packs $\geq$ 10% Discount/Feat	$26.9M	3.76%	$10.1M	1.42%	37.65%
Excluding Smaller Packs $\geq$ 20% Discount/Feat	$37.4M	5.23%	$13.8M	1.93%	36.85%
Total Sales	$715.6M	100%	$715.6M	100%	

This table decomposes the sales of larger packages with quantity surcharge using the model prediction of the utility using the actual choice data. Column (2) calculated the sales of larger packages with quantity surcharge considering the smaller alternatives with higher per-ounce utility. The sample used to calculate the actual sales and the model prediction same with the estimation sample. The sales amounts are calculated after weighting for the household projection factor.

Table 1.10: Decomposition Using the Predicted Choice

	(1)		(2)		(3)
	Sales s.t. QS		Sales s.t. QS by Info. Friction		
	\$ Sales	% Total Sales	\$ Sales	% Total Sales	(2)/(1)
Baseline	\$47.4M	4.44%	\$20.0M	1.88%	42.25%
$\leq$ Half Packs Available	\$25.5M	2.39%	\$10.2M	0.95%	39.94%
Excluding Smaller Packs $\geq$ 5% Discount/Feat	\$20.2M	1.90%	\$8.7M	0.81%	42.83%
Excluding Smaller Packs $\geq$ 10% Discount/Feat	\$27.4M	2.57%	\$11.1M	1.04%	40.66%
Excluding Smaller Packs $\geq$ 20% Discount/Feat	\$39.0M	3.65%	\$16.3M	1.53%	41.84%
Total Sales	\$1067.5M	100%	\$1067.5M	100%	

This table decomposes the sales of larger packages with quantity surcharge using the model prediction of the utility using the predicted choice data. Column (2) calculated the sales of larger packages with quantity surcharge considering the smaller alternatives with higher per-ounce utility. The sample used to calculate the actual sales and the model prediction same with the estimation sample. The sales amounts are calculated after weighting for the household projection factor.

1.8.2 Counterfactual Experiment: Revenue Improving, Consumer Welfare Preserving Retailer Pricing Schemes

The previous subsection addressed the question “why do consumers buy larger packages with quantity surcharge?,” and suggested an answer that consumers did not know a smaller, cheaper alternative that will even yield a better utility for them in around 40% of the instances. In this subsection, I turn to retailer’s perspective and at least indirectly address the question “why suppliers may want to price products with quantity surcharge?” Retail price of an item can be affected by various factors, some of which are decision variables of the retailers while others are not. The former include the retailer-level promotion and retailer margins, and the latter include the wholesale prices, land prices, labor cost, and so on. Ideally, one could obtain the cost data at each stage of the supply chain and study the product-line pricing behaviors of producers, wholesalers, and the retailers, and then solve for the optimal (unit) pricing scheme over package sizes. Because such detailed data is generally not available, I take an indirect approach to examine, *cetris paribus*, how the retailers could improve the revenue holding the consumer welfare fixed,[25] by tweaking the unit prices over package sizes from the current level. The constraint that the expected consumer welfare should not be harmed is similar to Chintagunta, Dubé, and Singh (2003). It turns out that retailers can raise revenues substantially while preserving the expected consumer welfare, and the corresponding sales of larger packages with quantity surcharge increase substantially.

The setup that I consider is as follows. Retailers know only the estimated demand model parameters presented in Table 1.8 of Section 1.7.2, along with joint distribution of consumer demographics. Noted in Section 1.7.2, consumer heterogeneity of the demand model’s boils down to the joint distribution of (μ_i, r_i), reducing the dimension of the problem substantially.

25. Laundry detergent is a necessity, and retailers can easily improve revenue by simply raising prices uniformly. I also simulated retailers’ revenues without imposing consumer expected welfare constraints. Retailers can raise revenues nearly indefinitely by increasing prices uniformly. Expected consumer welfare decreases substantially.

Retailers solve the nonlinear programming problem by finding the price multiplier function over package sizes to increase revenues, holding expected consumer welfare fixed.[26] The problem I consider is similar to the second-degree price discrimination since retailers cannot distinguish individual consumers' type to price discriminate.[27]

Let $q_j^* := n_j^* \varphi_j$ be the optimal quantity of alternative j when j is chosen.[28] Retailers' expected revenue is:

$$ER(\boldsymbol{\kappa}) = \int \sum_{j \in \mathcal{J}_i} \Pr\left(i \text{ chooses } j; p_j \kappa\left(\varphi_j\right)\right) q_j^*\left(p_j \kappa\left(\varphi_j\right)\right)\left\{p_j \kappa\left(\varphi_j\right)\right\} dF_i, \tag{1.17}$$

where F_i is the demographics distribution, and non-negative function $\kappa\left(\varphi_j\right)$ is the unit price multiplier function. $\kappa\left(\varphi_j\right)$ is a function of package size φ_j, which returns the multiplier of the per-ounce price over package sizes. All retailers in the market are assumed to employ the same price multiplier function $\kappa\left(\varphi_j\right)$.[29] Consumer i's expected consumer welfare is:

$$EU(\boldsymbol{\kappa}) = \int \sum_{j \in \mathcal{J}_i} \Pr\left(i \text{ chooses } j; p_j \kappa\left(\varphi_j\right)\right) \bar{v}_{i,j}\left(q_j^*; p_j \kappa\left(\varphi_j\right)\right) dF_i, \tag{1.18}$$

where

$$\bar{v}_{i,j}\left(q_j^*; p_j \kappa\left(\varphi_j\right)\right) = u_{i,j} q_j^*\left(p_j \kappa\left(\varphi_j\right)\right) - \delta_2\left\{q_j^*\left(p_j \kappa\left(\varphi_j\right)\right)\right\}^2, \tag{1.19}$$

and $u_{i,j}$ is the per-ounce utility of alternative j. Terms $\Pr\left(i \text{ chooses } j; p_j \kappa\left(\varphi_j\right)\right)$ and $u_{i,j}$ are household-specific because they involve information sensitivity μ_i and household-level

26. I consider retailers' revenues, not profit, because I do not have marginal cost or markup estimates of retailers.

27. Retailers can take the estimated information parameters $(\hat{\boldsymbol{\gamma}}, \hat{\varsigma})$ as fixed when they adjust unit prices. This is a direct implication of the assumption that consumers' prior distribution $Q_i(\mathbf{u}_i)$ is affected only by the "free" information that consumers have before costly information acquisition, not the utilities of consuming an alternative. If $Q_i(\mathbf{u}_i)$ is also affected by consumption utilities, $(\hat{\boldsymbol{\gamma}}, \hat{\varsigma})$ should be recalculated whenever changes to consumption utilities are made.

28. For tractability, I ignore integer constraints that consumers can purchase only integer numbers of packages while finding the optimal $\boldsymbol{\kappa}$. This can be justified using the law of large numbers argument. However, I consider integer constraints when I evaluate the performance of suggested pricing schemes.

29. For example, wholesale price change might induce such price change that is uniform to all the retailers.

Table 1.11: Revenue, Quantity Sales, and Consumer Welfare Change Over Pricing Schemes Relative to the Baseline ψ_0

Price Schedule	Revenue ($) (1)	Quantity Sales (oz) (2)	Consumer Welfare (3)	Sales s.t. QS ($) (4)	Sales s.t. QS Info Only ($) (5)
ψ_0	100.0%	100.0%	100.0%	100.0%	100.0%
ψ_1	106.4%	100.0%	99.9%	231.8%	200.5%
ψ_2	109.3%	100.0%	99.9%	260.4%	226.9%
ψ_3	112.3%	100.0%	99.9%	288.4%	253.6%
ψ_4	115.2%	100.0%	99.8%	315.7%	282.7%
ψ_5	118.0%	100.0%	99.8%	340.4%	313.1%
($/oz) of ψ_0	$1.068B	10.787B oz.	5.095B	$47.40M	$20.03M

This table summarizes the simulated consumer behaviors under the counterfactual pricing schemes. The rows "ψ_0"-"ψ_5" represent the relative ratio to the baseline model prediction ψ_0, and the row "($/oz) of ψ_0" represents the absolute level of the baseline model prediction ψ_0.

ideal point r_i.

Let $\overline{ER}_0$ and $\overline{EU}_0$ be the current model predictions of expected retailer revenue and expected consumer welfare, respectively. Let $\{\psi_b\}_{b=1}^{B}$ be an increasing multiplier sequence for retailer expected revenue. Retailers sequentially solve the following nonlinear programming problem:

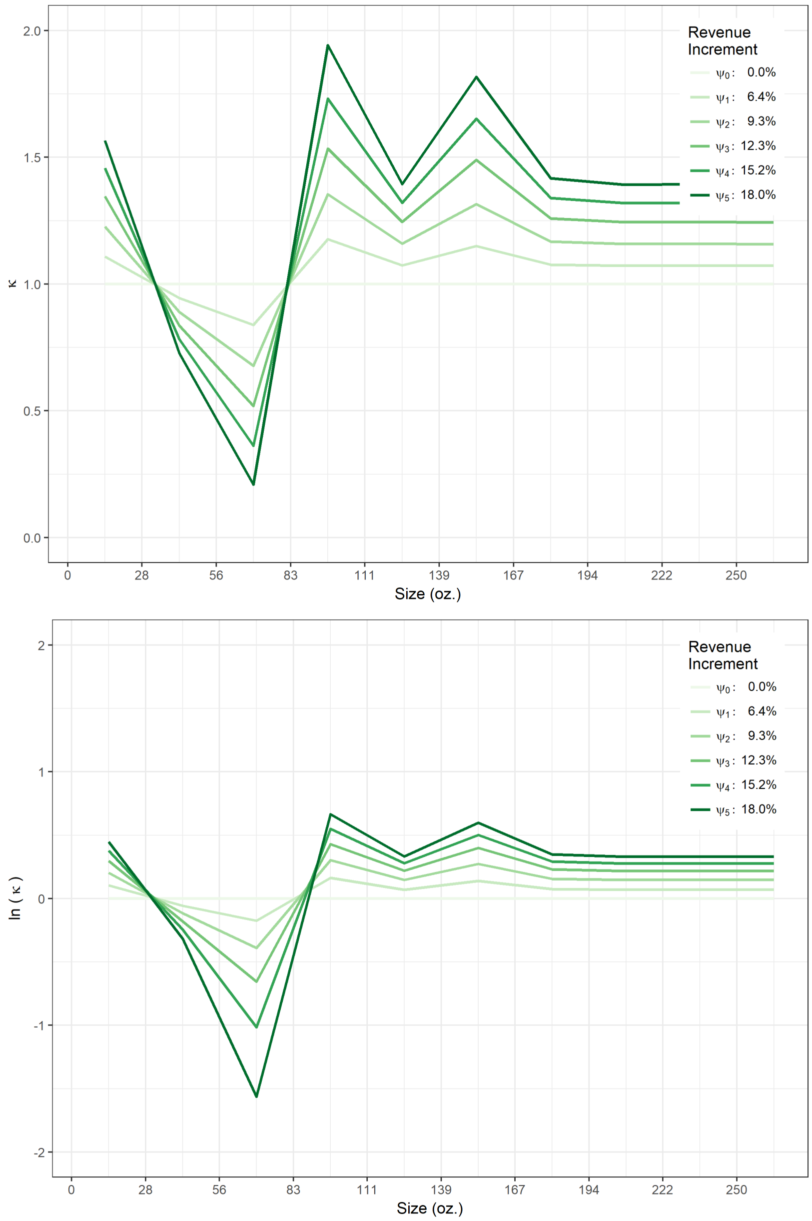
$$\max_{\{\boldsymbol{\kappa}(\varphi_j)\}} ER(\boldsymbol{\kappa}) \tag{1.20}$$

$$s.t.\ \overline{EU}_0 \leq EU(\boldsymbol{\kappa})$$

$$\psi_b \overline{ER}_0 \leq ER(\boldsymbol{\kappa}) \leq \psi_{b+1}\overline{ER}_0. \tag{1.21}$$

Counterfactual model predictions become less precise as deviations in the adjusted pricing scheme from the original pricing scheme becomes larger. Therefore, I sequentially impose the constraint (1.21) to bound the changes of $\boldsymbol{\kappa}$ from the original pricing scheme.

Figure 1.8: Welfare-improving Price Adjustment Over Package Sizes



Notes. (i) These figures illustrate the revenue increasing, consumer welfare preserving surcharge/discount schemes ψ_b over package sizes. (ii) The top and the bottom panels represent the same pricing schemes $\psi_0 - \psi_5$, where the only difference is the scale of the vertical axes. The vertical axes of the top panel and the bottom panel are $\kappa(\varphi_j)$ and $\ln(\kappa(\varphi_j))$, respectively. The horizontal axes represent the package sizes. (iii) ψ_0 is the baseline model prediction where $\kappa(\varphi_j) = 1$ for all package sizes.

Table 1.12: Sales of Larger Packages with Quantity Surcharge as % Total Sales, Change Over Pricing Schemes

Price Schedule	Sales s.t. QS (% Total) (6)	Sales s.t. QS Info Only (% Total) (7)	Ratio of (7)/(6) (8)
ψ_0	4.44%	1.88%	42.25%
ψ_1	9.67%	3.54%	36.55%
ψ_2	10.58%	3.89%	36.82%
ψ_3	11.41%	4.24%	37.15%
ψ_4	12.17%	4.61%	37.83%
ψ_5	12.80%	4.98%	38.85%

This table summarizes the simulated consumer behaviors under the counterfactual pricing schemes. The percentage points presented in Columns (6) and (7) of Table 1.12 are the ratio to the total dollar sales.

Figure 1.8 and Tables 1.11 and 1.12 present results from exercise (1.20) across a series of ψ_b schedule. Figure 1.8 illustrates how retailers should adjust per-ounce prices over package sizes to achieve 0% to 18% increment in revenue, holding expected consumer welfare fixed. The horizontal axes represent package sizes φ_j, and vertical axes the magnitude of $\kappa\left(\varphi_j\right)$ and $\ln\left(\kappa\left(\varphi_j\right)\right)$, respectively. The top and the bottom panels represent the same pricing schemes, except that the scales of the vertical axes differ. Each line corresponds to each ψ_b schedule, where darker lines achieve higher revenues. The shape of the price adjustment scheme suggests that retailers should lower prices of the 28 to 83 ounce packages, and raise the prices on all other packages.

Tables 1.11 and 1.12 summarize outcomes from simulated consumer behaviors under each counterfactual pricing schemes ψ_0-ψ_5, presented in Figure 1.8. Rows ψ_0-ψ_5 of Table 1.11 are the percentage ratios relative to ψ_0, and row *($/oz) of* ψ_0 is the absolute level of the baseline model prediction ψ_0. Columns (1) through (3) represent Expected Revenue, Expected Quantity Sales, and Expected Consumer Welfare change over price schedules ψ_0-ψ_5, respectively. Expected retailer revenue increases by 18%, whereas the expected quantity sales and expected consumer welfare remain nearly the same. The fact that expected quantity sales are constant implies that increased revenue can be converted directly to increased profit of retailers. Expected consumer welfare drops by at most 0.2%, which is attributed to imposing integer constraints when I simulate consumer behaviors. Columns (4) through (8) relate to the sales of larger packages with quantity surcharges. Columns (4) and (5) indicate that dollar sales of larger packages with quantity surcharges more than triple over ψ_0-ψ_5. Even after adjusting for increased revenue, it jumps from 4.44% to 12.80% of total sales, which is nearly three times of the baseline. Column (8) implies that the ratio of sales of larger packages with quantity surcharges that can attributed to only information friction does not vary much.

Results presented above partially explain why retailers would price products with quantity surcharges. Retailers can increase revenues substantially in a way that does not harm consumer welfare on average by employing a nonlinear pricing scheme over package sizes more aggressively. The intuition for this welfare-improving, nonlinear pricing scheme is to discount the prices of package sizes that many consumers prefer, and raise the prices of other package sizes to raise overall revenue. The shape of the unit price multiplier presented in Figure 1.8 is roughly the opposite of the density of the ideal points presented in Figure 1.7 in Section 1.7.2. It explains retailers' motivation to discount package sizes for which competition is intense. As a result of such nonlinear pricing adjustments, sales of larger packages with quantity surcharges also increase substantially.

1.9 Conclusion

I shed light on the existence and extent of sales of larger packages with quantity surcharges using a combination of detailed consumer panel data and national-level scanner data. Two competing hypotheses explain consumers' behavior of buying larger packages with quantity surcharges – information friction and preference heterogeneity over package sizes. To decompose and disentangle the contribution of each effect, I develop and estimate a structural econometric model of consumer choice and consumer demand that accommodates both hypotheses. Using model parameter estimates, I decompose the contribution of information friction and preference heterogeneity over package sizes on the sales of larger packages with quantity surcharges. Only about 40% of the sales of larger packages with quantity surcharges can be attributed to information friction. I then suggest a retailer revenue-improving, nonlinear pricing schedule that preserves consumers' expected welfare. Retailers can increase their revenues substantially by implementing nonlinear pricing schemes over package sizes, which lead to higher sales of larger packages with quantity surcharges.

This paper focuses on consumer choice and demand. Developing a model that explains the supplier side systematically would be desirable. There are different, competing explanations on the supplier side regarding when and why they price smaller packages lower in terms of per-unit price. Sellers might lower prices of smaller packages intentionally or inadvertently. Quantity surcharges are likely the result of an intentional behavior of retailers when competition is intense for a specific smaller package size. However, quantity surcharges are more likely to be a result of an inadvertent behavior of retailers when tied to a temporary price discount of a specific package size. Although the results presented in this paper have implications for the former, a systematic and thorough approach is required to explain retailers' pricing behaviors with quantity surcharges.

Another constraint of this study is that laundry detergents are a semi-durable, frequently purchased product category. Sales of larger packages with quantity surcharges are universal, and occur often with perishable and less frequently purchased product categories. Nevertheless, quantifying the extent of sales, consumer losses, and attribution to either information friction or preference heterogeneity is an empirical question, and the answer might differ substantially by product category. I leave such topics and extensions to future research.

1.A Proof of Theorem 1.6.1

Drop the index i for the sake of notational convenience. Let $\mathcal{J} = \{1, 2, ..., J\}$. For each $k \in \mathcal{J}$, let $\pi_k > 0$ be given. Let $\mu > 0$ be given. Normalize $u_J = 0$ as the utility of the outside option. Let $\tilde{\mathbf{u}} := (u_1, u_2, ..., u_{J-1})$. Define a mapping $D : \mathbb{R}^{J-1} \to \mathbb{R}^J$ where the j'th element of the mapping is

$$\begin{aligned} \{D(\tilde{\mathbf{u}})\}_j &= \frac{\pi_j \exp(\mu u_j)}{\sum_{k \in \mathcal{J}} \pi_k \exp(\mu u_k)} \\ &=: y_j. \end{aligned}$$

This mapping is globally invertible because it is continuous and strictly increasing, and the range is $\mathcal{A} := \left\{\mathbf{y} \in \mathbb{R}^J : 0 < y_1, y_2, ..., y_J < 1 \text{ and } \sum_{k=1}^{J} y_k = 1\right\}$. $\mathcal{A}$ is the interior of the unit simplex in $\mathbb{R}^J$, which is a $J-1$ dimensional object. Note that $y_J = 1 - \sum_{k=1}^{J-1} y_k$ because of the normalization constraint, and $\pi_J = 1 - \sum_{k=1}^{J-1} \pi_k$ by construction because $\{\pi_k\}_{k \in \mathcal{J}}$ is a set of unconditional choice probabilities.

Take any $\chi > 0$ and define $\pi_k^\chi := \chi \pi_k$. Consider the density q^D of Dirichlet distribution over $\mathcal{A}$ with concentration parameters $(\pi_1^\chi, \pi_2^\chi, ..., \pi_J^\chi)$, given by:

$$q^D\left(y_1, y_2, ..., 1 - \sum_{k=1}^{J-1} y_k\right) = \frac{1}{B(\pi_1^\chi, \pi_2^\chi, ..., \pi_J^\chi)} \prod_{k=1}^{J} y_k^{\pi_k^\chi - 1},$$

where

$$B(\pi_1^\chi, \pi_2^\chi, ..., \pi_J^\chi) := \frac{\prod_{k=1}^{J} \Gamma(\pi_k^\chi)}{\Gamma\left(\sum_{k=1}^{J} \pi_k^\chi\right)}$$

and $\Gamma(\cdot)$ is the Gamma function. Let $\tilde{\mathbf{y}} := (y_1, y_2, ..., y_{J-1})$. It is known for Dirichlet distributions that, for $1, 2, ..., J-1$,

$$\int_{\mathcal{A}} y_j q^D(\tilde{\mathbf{y}})\, d\tilde{\mathbf{y}} = \frac{\pi_j^\chi}{\sum_{k \in \mathcal{J}} \pi_k^\chi}.$$

Therefore,

$$\int_{\mathcal{A}} y_j q^D(\tilde{\mathbf{y}})\, d\tilde{\mathbf{y}} = \frac{\pi_j}{\sum_{k\in\mathcal{J}} \pi_k} = \pi_j \tag{1.22}$$

$$\int_{\mathcal{A}} q^D(\tilde{\mathbf{y}})\, d\tilde{\mathbf{y}} = 1. \tag{1.23}$$

Applying the change of variables $y_j = \{D(\tilde{\mathbf{u}})\}_j$ respectively on (1.22) and (1.23) yields:

$$\int_{\mathbb{R}^{J-1}} \{D(\tilde{\mathbf{u}})\}_j \left|\det(\nabla D(\tilde{\mathbf{u}}))\right| q^D(D(\tilde{\mathbf{u}}))\, d\tilde{\mathbf{u}} = \pi_j \quad \forall j \in \mathcal{J}$$

$$\int_{\mathbb{R}^{J-1}} \left|\det(\nabla D(\tilde{\mathbf{u}}))\right| q^D(D(\tilde{\mathbf{u}}))\, d\tilde{\mathbf{u}} = 1.$$

[30] Define

$$q(\tilde{\mathbf{u}}) := \left|\det(\nabla D(\tilde{\mathbf{u}}))\right| q^D(D(\tilde{\mathbf{u}})) \tag{1.24}$$

as the density of the distribution $Q(\cdot)$. Because $D : \mathbb{R}^{J-1} \to \mathcal{A}$ is globally invertible and q^D has a full support over $\mathcal{A}$, q has a full support over $\mathbb{R}^{J-1}$, which is a sufficient condition for (ii). Q.E.D.

The necessary conditions for q^D over $\mathcal{A}$ are (i) (1.23) holds and (ii) $E\left[y_j\right] = \pi_j\ \forall j \in \mathcal{J}$. The restrictions that (ii) impose is only on the first moments. Therefore, there can be other possibilities to construct such probability density q^D, and, by using (1.24), any density q^D that satisfies (i) and (ii) be converted to the density $q(\tilde{\mathbf{u}})$ that we are looking for. It is not necessary that q^D has to be a Dirichlet density.

The proof given here applies to the more general setup presented in Fosgerau, Melo, d Palma, and Shum (2017). The only requirement for the exactly same argument given above to work is that the mapping $D(\cdot)$ in Fosgerau, Melo, d Palma, and Shum to be globally invertible and continuously differentiable. It is indeed the case because their generator function $\mathbf{S}(\cdot)$ and its inverse mapping $\mathbf{H}(\cdot)$ are continuously differentiable.

30. This is a slight abuse of notation as we are fixing $u_J = 0$.

1.B Details on the Model, Likelihood, Estimation, and Counterfactual Simulation

1.B.1 Sketch on the Derivation of the Alternative Choice Likelihood

I sketch the derivation of equation (1.6) for completeness. The sketch is essentially a rephrasing of the relevant parts in Matějka and McKay (2015), to which readers interested in the full proof should refer. To derive (1.6), for each $\mathbf{u}_i \left(\in \mathbb{R}^{|\mathcal{J}_i|}\right)$, the optimal signal distribution $P_i\left(\mathbf{s}_i|\mathbf{u}_i\right)$ has the support of at most $|\mathcal{J}_i|$ points. The argument is that when more than two distinct signals associate with one action, the signal distribution can be reformulated with another distribution that generates the same expected payoff with a lower cost. In other words, the signal from the optimal learning strategy is of the form "choose j because I believe it will yield the maximum utility." It follows immediately that conditional distributions $P_i\left(\mathbf{s}_i|\mathbf{u}_i\right)$ and $P_i\left(\mathbf{u}_i|\mathbf{s}_i\right)$ can be equivalently formulated in terms of $P_i\left(\mathbf{s}_i = j|\mathbf{u}_i\right)$ and $P_i\left(\mathbf{u}_i|\mathbf{s}_i = j\right)$, respectively. Therefore, the expected payoff is:

$$
\begin{aligned}
&\int_{\mathbf{u}_i}\int_{\mathbf{s}_i}\left\{\max_{j\in\mathcal{J}_i}\int u_{i,j}P_i\left(d\mathbf{u}_i|\mathbf{s}_i\right)\right\}P_i\left(d\mathbf{s}_i|\mathbf{u}_i\right)Q_i\left(d\mathbf{u}_i\right)\\
&=\int_{\mathbf{u}_i}\sum_{j\in\mathcal{J}_i}\left\{\int_{\mathbf{u}_i}u_{i,j}P_i\left(d\mathbf{u}_i|j\right)\right\}P_i\left(j|\mathbf{u}_i\right)Q_i\left(d\mathbf{u}_i\right)\\
&=\sum_{j\in\mathcal{J}_i}\left\{\int_{\mathbf{u}_i}u_{i,j}P_i\left(d\mathbf{u}_i|j\right)\right\}\left\{\int_{\mathbf{u}_i}P_i\left(j|\mathbf{u}_i\right)Q_i\left(d\mathbf{u}_i\right)\right\}\\
&=\sum_{j\in\mathcal{J}_i}\int_{\mathbf{u}_i}u_{i,j}P_i\left(d\mathbf{u}_i|j\right)\pi_{i,j}\\
&=\sum_{j\in\mathcal{J}_i}\int_{\mathbf{u}_i}u_{i,j}P_i\left(j|\mathbf{u}_i\right)Q_i\left(d\mathbf{u}_i\right) \qquad (1.25)\\
&=\sum_{j\in\mathcal{J}_i}\int_{\mathbf{u}_i}u_{i,j}\mathrm{Pr}_i\left(i \text{ Chooses } j|\mathbf{u}_i\right)Q_i\left(d\mathbf{u}_i\right). \qquad (1.26)
\end{aligned}
$$

(1.25) is by Bayes' rule, and (1.26) is because $P_i(\mathbf{s}_i = j|\mathbf{u}_i) = \Pr_i(i \text{ Chooses } j|\mathbf{u}_i) \equiv \Pr(i \text{ Chooses } j|\mathcal{J}_i, \mathbf{D}_i, \mathbf{u}_i)$ under this form of the optimal strategy.

Problem (1.4) can then be restated as:

$$\max_{\{\Pr_i(i \text{ Chooses } j|\mathbf{u}_i)\}_{j\in\mathcal{J}_i}} \sum_{j\in\mathcal{J}_i} \int_{\mathbf{u}_i} u_{i,j} \Pr_i(i \text{ Chooses } j|\mathbf{u}_i) Q_i(d\mathbf{u}_i)$$
$$-\lambda_i \left\{ -\sum_{j\in\mathcal{J}_i} \pi_{i,j} \ln \pi_{i,j} + \int_{\mathbf{u}_i} \sum_{j\in\mathcal{J}_i} \Pr_i(i \text{ Chooses } j|\mathbf{u}_i) \ln \Pr_i(i \text{ Chooses } j|\mathbf{u}_i) Q_i(d\mathbf{u}_i) \right\}$$
$$s.t. \begin{cases} \Pr_i(i \text{ Chooses } j|\mathbf{u}_i) \geq 0 & \forall j \in \mathcal{J}_i, \forall \mathbf{u}_i \in \mathbb{R}^{|\mathcal{J}_i|} \\ \sum_{j\in\mathcal{J}_i} \Pr_i(i \text{ Chooses } j|\mathbf{u}_i) = 1. \end{cases}$$

The information cost function part is also modified using the symmetry of mutual information. Setting up a Lagrangian and solving for first-order conditions yields the desired conclusion.

1.B.2 Normalization and Identification

I informally discuss identification of model parameters, identified up to the scale normalization. To see why, fix η_i, normalize $u_{i,0} = 0$, $\alpha = 1$, and then take the log-ratios, which leads to:

$$\ln\left(\frac{\Pr(i \text{ Chooses } j; \eta_i)}{\Pr(i \text{ Chooses } 0)}\right) = \ln\left(\frac{\pi_{i,j}\exp(\mu_i u_{i,j})}{\pi_{i,0}}\right) \tag{1.27}$$
$$= \ln\left(\frac{\exp\left(\eta_{i,j} + \mathbf{d}'_{i,j}\boldsymbol{\gamma}\right)}{\exp(0)}\right) + \mu_i\left(1 + p_j\beta_1 + \mathbf{x}'_j\boldsymbol{\beta}_2 - \delta_1\rho(\varphi_j, r_i)\right) \tag{1.28}$$
$$= \eta_{i,j} + \mathbf{d}'_{i,j}\boldsymbol{\gamma} + \mu_i\left(1 + p_j\beta_1 + \mathbf{x}'_j\boldsymbol{\beta}_2 - \delta_1\rho(\varphi_j, r_i)\right). \tag{1.29}$$

[31][32] The remaining arguments rely on a typical, nonlinear M-estimator identification, which assumes the existence of a global maximizer of likelihood.[33] The same normalization that $\alpha = 1$ identifies σ^{-1} in $\Pr\left(v_{i,j}\left(n_j\right) \geq v_{i,j}\left(n_j - 1\right), v_{i,j}\left(n_j\right) \geq v_{i,j}\left(n_j + 1\right)\right)$.

1.B.3 Comparison to the Simple Multinomial Logit Model

I describe the discrete choice random utility counterpart of the model that I described in Section 1.6 when all alternatives are *a priori* homogeneous. I demonstrate that simple multinomial logit demand models are identified only after imposing the same location and scale normalization as in Section 1.B.2.

Consider a typical discrete choice demand model in which consumer i chooses one alternative j out of alternative set $\mathcal{J}$ such that:

$$j = \arg\max_{k \in \mathcal{J}} \left\{u_{i,k} + \epsilon_{i,k}\right\},$$

where $\epsilon_{i,k} \sim$ iid Gumbel$\,(a, b)$. The Gumbel distribution, often called the Type-1 Extreme Value distribution, has both the location and scale parameters. Its cumulative distribution function is:

$$F\left(\epsilon\right) = \exp\left(-\exp\left(-\frac{\epsilon - a}{b}\right)\right).$$

31. Define $\left\{u_{i,j}^0\right\}_{j \in \mathcal{J}_i}$ as

$$\begin{aligned} u_{i,j}^0 &:= \alpha + p_j \beta_1^0 + \mathbf{x}_j' \boldsymbol{\beta}_2^0 - \delta_1 \rho\left(\varphi_j, r_i\right) \qquad \text{for } j \neq 0 \\ u_{i,0}^0 &:= 0. \end{aligned}$$

Dividing $u_{i,j}^0$ by $\alpha > 0$ for all $j \in \mathcal{J}_i$, and then defining $\beta_1 := \beta_1/\alpha$. $\boldsymbol{\beta}_2 := \boldsymbol{\beta}_2^0/\alpha$, $\delta_1 := \delta_1^0/\alpha$, and $\delta_2 := \delta_2^0/\alpha$ yields the same normalization. Since the consumer chooses the alternative that yields the maximum expected utility conditioning on his belief, dividing $u_{i,j}^0$ by $\alpha > 0$ for all $j \in \mathcal{J}_i$ does not affect choice structures. Therefore, the solution to the consumer's problem is invariant with respect to this normalization.

32. The discrete choice random utility counterpart of my model also imposes a similar location and scale normalization. See Appendix 1.B.3 for details.

33. Due to the term $\exp\left(\mathbf{w}_i' \boldsymbol{\theta}\right)$ inside the exponentials, the log-likelihood is not globally concave regarding model parameters, making the problem computationally intensive.

Parameterize $u_{i,k} := \alpha + p_k\beta_1 + \mathbf{x}_k'\boldsymbol{\beta}_2$. The individual choice probability of i choosing j is:

$$\Pr\left(i \text{ Chooses } j\right) = \frac{\exp\left(\frac{a+\alpha+p_j\beta_1+\mathbf{x}_j'\boldsymbol{\beta}_2}{b}\right)}{\sum_{k\in\mathcal{J}}\exp\left(\frac{a+\alpha+p_k\beta_1+\mathbf{x}_k'\boldsymbol{\beta}_2}{b}\right)}.$$

It is straightforward that model parameters $(\alpha, \boldsymbol{\beta})$ can be identified only after normalizing the location and scale parameters. The usual choice in the discrete choice random utility demand models is setting $a = 0$ and $b = 1$, but different choices are possible. Especially when $\mathcal{J}$ includes the outside option, which is denoted by option 0, normalizing $u_{i,0} := 0$ following the convention of Berry (1994); Berry, Levinsohn, and Pakes (1995) is a popular choice in the literature.

CHAPTER 2

SEMIPARAMETRIC ESTIMATION OF A CES DEMAND SYSTEM WITH OBSERVED AND UNOBSERVED PRODUCT CHARACTERISTICS[1]

2.1 Introduction

Constant elasticity of substitution (CES) preferences, often called Dixit-Stiglitz-Spence preferences, have been used extensively to analyze markets with product differentiation in macroeconomics and international trade literature since Spence (1976); Dixit and Stiglitz (1977); Anderson (1979); Krugman (1980); Feenstra (1994). However, recent analyses of the demand of differentiated products in empirical industrial organization literature since Berry (1994); Berry, Levinsohn, and Pakes (1995) have been based on a different microfoundation – the discrete choice random utility model in the product characteristic space. We reconcile these approaches of differentiated products demand estimation by adding a flexible "quality kernel" to CES preferences. The quality kernel is a non-negative function that maps observed and unobserved product characteristics to the marginal utility multiplier of consuming one unit of a product. Adding the quality kernel to CES preferences allows us to (i) derive the identical predicted market share equation as that of Berry (1994); Berry, Levinsohn, and Pakes (1995) by incorporating observed and unobserved product characteristics into CES preferences, and (ii) accommodate zero predicted and observed market shares directly by embedding an intensive and an extensive margin of choice. We demonstrate how to semiparametrically estimate a product-differentiated demand model with data that has a multitude of zero observed market shares.

Demand estimation is a central problem in industrial organization, and the major workhorse of demand estimation using market data in recent empirical industrial organizational litera-

1. This chapter is coauthored with Ali Hortaçsu.

ture was developed by Berry (1994); Berry, Levinsohn, and Pakes (1995). One of the fundamental breakthrough that was made by Berry (1994); Berry, Levinsohn, and Pakes (1995) was to (re)introduce the characteristic space approach in demand estimation, which dates back to Lancaster (1966), and combine the approach with the discrete choice problem. In Berry (1994); Berry, Levinsohn, and Pakes (1995)'s demand estimation framework, a product is defined as a bundle of observed and unobserved product characteristics. A consumer can choose up to one product that yields the highest utility among her finite choice set, or can decide to buy nothing. A consumer's (dis)utility of consuming a product is comprised of the utility from price, observed product characteristics, unobserved product characteristics, and idiosyncratic utility shock. The individual choice probability equation is derived from the distributional properties of the idiosyncratic utility shock, which is assumed to follow the Type-I extreme value distribution. Individual choice probabilities are taken as equal to the predicted quantity shares of the individual demand, aggregation of which is taken as the predicted quantity market shares. We refer to demand models based on these microfoundations as logit demand models, which provide a tractable method of estimating differentiated product demand systems by reducing the dimension of parameters to be estimated.

Drawback of such tractability is that the logit demand models impose strong assumptions that are often unrealistic. Several features of logit demand were criticized in the literature, including single-choice assumptions, lack of income effects in the derived demand system, impossibility of accommodating zero shares, and ignoring dynamics (Nevo (2000); Ackerberg, Benkard, Berry, and Pakes (2007); Reiss and Wolak (2007)). We tackle the first three problems by developing a new microfoundation for the logit demand estimation frameworks based on the Marshallian CES demand system. On the other hand, researchers argue that CES preferences are inflexible to incorporate observed and unobserved product characteristics, and therefore are inappropriate to analyzing microdata (Aguirregabiria and Nevo (2013); Nevo (2011)). By deriving the market share equation of Berry (1994); Berry, Levinsohn, and Pakes (1995) from a Marshallian CES demand system, we show that the CES demand system can be just as flexible as the logit demand system in analyzing microdata.

Our first contribution to the literature is provision of a concrete link between the CES demand system and Berry (1994); Berry, Levinsohn, and Pakes (1995)'s homogeneous/random coefficients logit demand system. We do so by developing a flexible and empirically tractable method of directly incorporating observed and unobserved product characteristics into the "taste parameter" of the CES preferences, and then deriving the same predicted quantity individual/market share equation of Berry (1994); Berry, Levinsohn, and Pakes (1995) from the resulting Marshallian CES demand system. Early studies by Anderson, de Palma, and Thisse (1987, 1992) point out similarities between the CES and logit demand systems without product characteristics. Our predicted market shares equivalence result is an extension of Anderson, de Palma, and Thisse (1987, 1992), in the context of Berry (1994); Berry, Levinsohn, and Pakes (1995)'s characteristics based demand estimation framework. Incorporating the taste parameter in CES preferences dates back to at least Spence (1976); Anderson (1979), and recent empirical trade literature derives the estimating equation for the trade expenditure shares using the taste parameters (see Eaton and Kortum (2002); Waugh (2010) among others). Analogously, Einav, Knoepfle, Levin, and Sundaresan (2014) incorporate sales tax indicator and distance from the seller, which are the seller-consumer specific characteristics, into the taste parameter of the CES preferences. However, to the best of our knowledge, none in the literature model the taste parameter of the CES preferences directly as a mapping from the observed and unobserved product characteristics to develop a general empirical framework for demand estimation.

The equivalence result we establish helps connect the divergent demand modeling approaches utilized in the empirical industrial organizational literature and macroeconomics / international trade literature. For the empirical industrial organizational literature, it provides an additional appealing microfoundation for extant homogeneous/random coefficient logit demand estimation frameworks. For the macroeconomics/international trade literature, it allows users of CES demand systems to apply the identification results and estimation methods developed for logit demand frameworks, such as Berry (1994); Berry, Levinsohn, and Pakes (1995); Dube, Fox, and Su (2012); Berry and Haile (2014).

Our second contribution is the development of a direct method that accommodates zero predicted and observed market shares. Accommodating zero market shares in the data has been a major difficulty in the demand estimation literature for the past two decades, since Berry (1994); Berry, Levinsohn, and Pakes (1995). Discrete choice frameworks with additive idiosyncratic errors with unrestricted support inherently do not allow for zero individual choice probabilities. Individual choice probabilities are treated as predicted individual quantity shares, aggregated over homogeneous or heterogeneous individuals and equated with observed market shares for identification and estimation of model parameters. In logit demand models, additive idiosyncratic shocks are distributed as an i.i.d. Type-I extreme value. In such a case, the numerator of the individual choice probability is the exponential of the alternative utility's deterministic part. Provided that an alternative yields any utility higher than negative infinity, the alternative must have a strictly positive predicted market share. However, zero observed market shares are often observed in data. Thus, in practice, researchers simply drop samples with zero observed market shares, or add a small, arbitrary number to zero observed market shares. These *ad hoc* measures cause biases in estimates. We argue that selection in a consumer's choice set must be considered for identification and estimation of model parameters. The choice set selection drives the conditional expectation of unobserved product characteristics that are conditioned on instruments to be non-zero and highly likely to be positive. The usual generalized method of moments estimation yields price coefficient estimates that are biased upward when this choice set selection process is ignored.

To accommodate the zero predicted and observed market shares, we provide a microfoundation for the selection-correction estimation equation *à la* Heckman (1979), by embedding both extensive and intensive margins on the quality kernel. Modeling both margins in a utility maximization problem was introduced and developed by Dubin and McFadden (1984); Hanemann (1984); Chiang (1991); Chintagunta (1993); Nair, Dubé, and Chintagunta (2005), commonly relying on a single-choice assumption. We extend the idea to the Marshallian CES demand system, and model a consumer's choice as a two-stage decision process. During the first stage, a consumer engages in a buy-or-not decision on the alternatives, resulting in the consumer's choice set. During the second, the consumer chooses how much to buy of each product in the choice set. The first stage determines the extensive margin, and the second stage the intensive margin of the consumer. If an alternative is not included in the choice set, it naturally leads to zero predicted and observed market shares. By modeling intensive and extensive margins separately, we provide an empirically tractable estimation framework that accommodates zero predicted and observed quantity market shares. Our two-stage strategy to accommodate the zero predicted and observed market shares relates closely to the consideration set literature in marketing (see, e.g., Roberts and Lattin (1991); Ben-Akiva and Boccara (1995); Jedidi, Kohli, and DeSarbo (1996); Mehta, Rajiv, and Srinivasan (2003); Gilbride and Allenby (2004) among many others), where the first stage is consideration set formulation, and the second stage is a choice among the consideration set. We apply the strategy in the context of the Marshallian CES demand system, to accommodate the zero predicted and observed market shares. As for the contexts where the single choice assumption is more plausible, we also provide the microfoundation for the same selection-correction estimation equation in the context of the logit demand model.

Gandhi, Lu, and Shi (2013) rationalize zero observed market shares differently, regarding such shares as measurement errors of strictly positive predicted market shares, and provide a partial identification result of model parameters. The difference between their research and ours is that we rationalize zero predicted and observed market shares, whereas they allow only observed market shares to be zero. Nevertheless, their Monte-Carlo simulations and empirical applications suggest similar implication to ours; when samples with zero market shares are dropped, price coefficient estimates are biased upward. In international trade literature, Helpman, Melitz, and Rubinstein (2008) develop another method that relates closely to ours in the context of gravity models. They use a gravity model with endogenous censoring of trade volumes, and their structural approach to handling zero trade flows is similar to ours. However, their approach is fully parametric in that they assume the Gaussian error term, whereas our approach is semiparametric as we do not specify the distribution of unobservables in our preferred specification. We employ the Klein and Spady (1993) estimator for the first-stage estimator, demonstrating how the distribution-free efficient semiparametric estimator for the binary response model can be easily applied to the demand estimation problem with a multitude of zero predicted and observed market shares. Furthermore, in our empirical example, we provide evidence that the distribution of the unobservable product characteristics is far from Gaussian.

Our approach can be viewed as a hedonic, or pure characteristics, model of demand, in that we do not require i.i.d. random utility shocks on the utility specification. Recent developments on hedonic demand estimation frameworks were made by Bajari and Benkard (2005); Berry and Pakes (2007), of which the former relates more closely to our study. Bajari and Benkard investigate a general hedonic model of demand with product characteristics, focusing on local identification and estimation of model parameters. For global identification when a product space is continuous, they specify Cobb-Douglas preferences. Our study extends their Cobb-Douglas specification to the more flexible CES preferences specification that can also accommodate zero predicted and observed market shares.

2.2 CES Demand System with Observed and Unobserved Product Characteristics

2.2.1 Specification of the CES Demand System

We consider a differentiated product market denoted by subscript t, composed of homogeneous consumers with a CES preference. We begin by focusing on homogeneous consumers. The extension to product markets comprised of heterogeneous consumers, with each consumer allowed to have disparate utility parameters, is considered in Section 2.3.1. The utility from a product category is:

$$u\left(\left\{q_{j,t}, \mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right\}_{j\in\mathcal{J}_t}\right) := \left(\sum_{j\in\mathcal{J}_t} \left\{\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)\right\}^{\frac{1}{\sigma}} q_{j,t}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}. \quad (2.1)$$

Set $\mathcal{J}_t$ is a set of alternatives in the category, which might include the numeraire that represents the outside option. $q_{j,t}$ is the quantity of product j consumed in market t. $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)$, defined by the quality kernel, is a non-negative function of observed and unobserved product characteristics. $\mathbf{x}_{j,t}$ and $\mathbf{w}_{j,t}$ are vectors of product j's characteristics in market t, which are observable to the econometrician. $\xi_{j,t}$ and $\eta_{j,t}$ are scalars that represent utility from product j's characteristics that are unobservable to the econometrician. $\mathbf{w}_{j,t}$ and $\eta_{j,t}$ are extensive margin shifters that a consumer considers whether to buy the product. $\mathbf{x}_{j,t}$ and $\xi_{j,t}$ are intensive margin shifters that determine the level of utility when a consumer buys a product. $\mathbf{w}_{j,t}$ and $\mathbf{x}_{j,t}$ might have common components, but we can require exclusion restriction on $\mathbf{w}_{j,t}$ for semiparametric identification when the extensive margin matters. In such a case, $\mathbf{w}_{j,t}$ must contain at least one component that is not in $\mathbf{x}_{j,t}$. We explain identification conditions further in Section 2.4. The observed extensive margin shifter, $\mathbf{w}_{j,t}$, might contain the price $p_{j,t}$ or a nonlinear function of $p_{j,t}$.

The quality kernel $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)$, introduced in equation (2.1), is critical to our framework. Researchers conventionally employ taste parameters or utility weights in places we put the quality kernel. The quality kernel, taste parameters, and utility weights can be commonly interpreted as multipliers on the (marginal) utility of consuming a product. The quality kernel is a straightforward extension of such conventions that allows us to incorporate observed and unobserved product characteristics directly into a consumer's utility. The quality kernel also allows the possibility of explicitly separating intensive and extensive margins. This feature accommodates zero predicted and observed market shares in model parameters.

The representative consumer's budget-constrained utility maximization problem, the solution of which is the Marshallian demand system, is:

$$\max_{\{q_{j,t}\}_{j\in\mathcal{J}_t}} \left(\sum_{j\in\mathcal{J}_t} \left\{ \chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) \right\}^{\frac{1}{\sigma}} q_{j,t}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \qquad s.t. \qquad \sum_{j\in\mathcal{J}_t} p_{j,t} q_{j,t} = y_t. \tag{2.2}$$

The Marshallian demand system is:

$$q_{j,t} = y_t \left\{ \frac{\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) p_{j,t}^{-\sigma}}{\sum_{k\in\mathcal{J}_t} \chi\left(\mathbf{x}_{k,t}, \xi_{k,t}, \mathbf{w}_{k,t}, \eta_{k,t}\right) p_{k,t}^{1-\sigma}} \right\} \qquad \forall j \in \mathcal{J}_t, \tag{2.3}$$

which leads to the predicted quantity market shares expression:

$$\begin{aligned} \pi_{j,t} &\equiv \frac{q_{j,t}}{\sum_{k\in\mathcal{J}_t} q_{k,t}} \\ &= \frac{\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) p_{j,t}^{-\sigma}}{\sum_{k\in\mathcal{J}_t} \chi\left(\mathbf{x}_{k,t}, \xi_{k,t}, \mathbf{w}_{k,t}, \eta_{k,t}\right) p_{k,t}^{-\sigma}}. \end{aligned} \tag{2.4}$$

Equation (2.4) is what we call the CES demand system with observed and unobserved product characteristics. The demand system (2.4), which is in the form of the predicted quantity market shares, is our primary interest because the same predicted quantity market share expression from Berry (1994); Berry, Levinsohn, and Pakes (1995) can be derived by imposing a further structure on the quality kernel, $\chi(\cdot)$. (2.4), a system of predicted quantity market shares, imposes only $\#(\mathcal{J}_t) - 1$ constraints on the Marshallian demand system, $\mathbf{q}_t$, in (2.3). Only when combined with the budget constraint $\sum_{j\in\mathcal{J}_t} p_{j,t}q_{j,t} = y_t$ can the Marshallian demand quantities, $\mathbf{q}_t$, be uniquely pinned down for a given price vector, $(\mathbf{p}_t, y_t) \in \mathbb{R}^{\#(\mathcal{J}_t)+1}$.

For the invertibility of the demand system, (2.4), we consider the subset $\mathcal{J}_t^+ \, (\subseteq \mathcal{J}_t)$, such that $\pi_{j,t} > 0$ for all $j \in \mathcal{J}_t^+$. The demand system specified by $\{\pi_{j,t}\}_{j\in\mathcal{J}_t^+}$ satisfies the connected substitutes conditions from Berry, Gandhi, and Haile (2013), and is thus invertible. Invertibility of the demand system implies that σ, the elasticity of substitution, is identified. If we impose suitable structures on $\chi(\cdot)$, such as monotonicity with index restriction, the structural parameters of $\chi(\cdot)$ are also identified. We investigate the specific functional forms of $\chi(\cdot)$ in Section 2.3.

2.2.2 Properties of the CES Demand System and Comparison with the Logit Demand System

We now explain the properties of the CES demand system (2.4). We begin with the Marshallian and Hicksian own and cross price elasticities of the demand system. Let $b_{j,t}$ and $\pi_{j,t}$ be the budget and quantity share of product j in market t, respectively.[2] Denote $\varepsilon^M_{jc,t}$ and $\varepsilon^H_{jc,t}$ by the Marshallian and Hicksian cross price elasticities between alternatives j and c, respectively. If $\mathbf{w}_{j,t}$ does not include the prices or function of the prices as its component, we have the following simple closed-form formulas for the Marshallian and the Hicksian own

2. We use the term budget share and expenditure share exchangeably.

and cross price elasticities:

$$\begin{aligned}\varepsilon^M_{jj,t} &= -\sigma + (\sigma - 1)\, b_{j,t} \\ \varepsilon^M_{jc,t} &= (\sigma - 1)\, b_{c,t} \\ \varepsilon^H_{jj,t} &= -\sigma\left(1 - b_{j,t}\right) \\ \varepsilon^H_{jc,t} &= \sigma b_{c,t}, \end{aligned} \tag{2.5}$$

and the income elasticity is 1.[3][4] These elasticities can be easily calculated given that σ is identified. From these elasticity expressions, it can be immediately noticed that a version of the independence of irrelevant alternatives (IIA) property holds; the substitution pattern depends solely on the budget shares of corresponding products. The price elasticities of the CES demand system should not be derived based on the quantity market shares as in the logit demand models.[5] Price elasticities in the logit demand models when the mean utility

3. In calculating the elasticities in practice, observed budget shares can be used in place of $b_{j,t}$ and $b_{c,t}$.

4. If $\mathbf{w}_{j,t}$ includes the prices or a function of the prices so that the extensive margin is affected by the price changes, then the simple closed-form expressions for the own and cross elasticities cannot be derived. In practice, the corresponding price elasticities can be calculated using simulations.

5.

$$\begin{aligned}\frac{\partial \ln \pi_{j,t}}{\partial \ln p_{j,t}} &= \frac{\partial\left(\ln q_{j,t} - \ln\left(\sum_{k\in\mathcal{J}_t} q_{k,t}\right)\right)}{\partial \ln p_{j,t}} \\ &= \frac{\partial \ln q_{j,t}}{\partial \ln p_{j,t}} - \frac{\partial \ln\left(\sum_{k\in\mathcal{J}_t} q_{k,t}\right)}{\partial \ln p_{j,t}} \\ &= \varepsilon^M_{jj,t} - \frac{\partial \ln\left(\sum_{k\in\mathcal{J}_t} q_{k,t}\right)}{\partial \ln p_{j,t}} \\ &\neq \frac{\partial \ln q_{j,t}}{\partial \ln p_{j,t}}\end{aligned}$$

The term $\frac{\partial \ln \pi_{j,t}}{\partial \ln p_{j,t}}$ is the Marshallian price elasticity only when $\sum_{k\in\mathcal{J}_t} q_{k,t}$ is constant, which is the case for the logit demand models. See Appendix 2.A for the details.

is log-linear in prices are given by:

$$\begin{aligned}\varepsilon_{jj,t}^{L} &= -\alpha\left(1-\pi_{j,t}\right)\\ \varepsilon_{jc,t}^{L} &= \alpha\pi_{c,t}.\end{aligned} \tag{2.6}$$

[67] The expressions (2.6) parallel the Hicksian price elasticities of the CES demand system. The only difference to the Hicksian price elasticities (2.5), derived from the CES demand system, is that the multiplied terms of the log-price coefficient, α, are comprised of quantity market shares, not budget shares.

Because we derive the demand system from the budget-constrained CES utility maximization problem, the duality between the Marshallian and Hicksian demand functions holds. The Slutsky equation follows, and thus we can decompose the substitution and income effect more naturally. The Slutsky equation in the elasticity form is:

$$\varepsilon_{jc,t}^{M} = \varepsilon_{jc,t}^{H} - \varepsilon_{j,t}^{I} b_{c,t}.$$

Because $\varepsilon_{j,t}^{I} = 1$ in the CES demand system, the income effect depends solely on budget shares, which is a considerable limitation. However, there are at least two advantages over the discrete choice counterpart. First, although the numeraire can be included in the consumer's choice set, $\mathcal{J}_t$, it is unnecessary in our CES demand system. In contrast, inclusion of the numeraire in the choice set is necessary in the logit demand system to induce an income effect, especially when the income is not a direct argument of the discrete choice utility or

6. When the mean utility is linear in prices, the elasticity expression becomes:

$$\begin{aligned}\varepsilon_{jj,t}^{L} &= -\alpha p_{j,t}\left(1-\pi_{j,t}\right)\\ \varepsilon_{jc,t}^{L} &= \alpha p_{c,t}\pi_{c,t}.\end{aligned}$$

See Section 2.3.1 for a further discussion on the functional form of the mean utilities in the logit demand model.

7. In calculating the elasticities in practice, observed quantity shares can be used in place of $\pi_{j,t}$ and $\pi_{c,t}$.

it is canceled out.[8] In such a case, the price increase of an alternative leads to consumers switching to only other alternatives in the choice set. The magnitude of the income effect is in a sense determined *a priori* by the researcher in logit demand models because the income effect depends primarily on quantity market shares of the numeraire. The size of the share of the numeraire is often arbitrarily assumed or imposed by a researcher in practice. Second, the income effect depends on budget shares, not quantity shares, in the Marshallian CES demand system. In logit demand models, the income effect of a product with a small budget share and a large quantity share is large, which is even more unrealistic.

2.3 The Exponential Quality Kernel

So far, we have not restricted the quality kernel, $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)$. In principle, χ can be any non-negative function. Under this weak restriction, the demand system specified by predicted market shares (2.4) can be identified locally, as investigated by Bajari and Benkard (2005). However, nonparametric estimation of a locally identified demand system places a considerable burden on the data and computational power, which is often impractical. Locally identified parameter values are often uninformative regarding counterfactual analyses, and alternatively, we can impose further structures on the consumer utility from product characteristics. We focus on the exponential quality kernel with an index restriction. This specific functional form deserves a special attention for two reasons. First, by using this functional form, we can derive the same individual choice probability equation of the homogeneous and random coefficient logit models of demand from the CES demand system developed in the previous section. Second, this functional form simplifies the estimation problem substantially because the estimation equation reduces to a log-linear form. We use the exponential quality kernel to propose a tractable, semiparametric estimation method that accommodates zero predicted and observed market shares.

8. For example, the case when the mean utility is linear in income net of price.

2.3.1 Nesting the Homogeneous and Random Coefficient Logit Models of Demand

We show that the predicted quantity market share expressions of the homogeneous and random coefficient logit models of demand can be derived from (2.4) by choosing a functional form of the quality kernel, $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)$. Suppose that $\mathbf{x}_{j,t} = \mathbf{w}_{j,t}$, $\xi_{j,t} = \eta_{j,t}$, $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}\right) > 0$, $\pi_{j,t} > 0$, and let $\mathbf{x}_{j,t}$ be exogenous for all j, t. We do not require an exclusion restriction in this setup because the predicted quantity shares are positive for every alternative. Let $\mathcal{J}_t$ contain the numeraire, denoted by product 0, and normalize $p_{0,t} = 1$.[9] Taking the ratios of products j and 0, and taking the logarithm of equation (2.4), yields:

$$\ln\left(\frac{\pi_{j,t}}{\pi_{0,t}}\right) = -\sigma \ln\left(p_{j,t}\right) + \ln \chi\left(\mathbf{x}_{j,t}, \xi_{j,t}\right) - \ln \chi\left(\mathbf{x}_{0,t}, \xi_{0,t}\right). \tag{2.7}$$

We normalize $\mathbf{x}_{0,t} = \mathbf{0}$, $\xi_{0,t} = 0$, and let $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}\right) = \exp\left(\mathbf{x}'_{j,t}\boldsymbol{\beta} + \xi_{j,t}\right)$. (2.7) then becomes:

$$\ln\left(\frac{\pi_{j,t}}{\pi_{0,t}}\right) = -\sigma \ln\left(p_{j,t}\right) + \mathbf{x}'_{j,t}\boldsymbol{\beta} + \xi_{j,t}. \tag{2.8}$$

(2.8) coincides with the estimation equation of the homogeneous logit model of demand, except that in (2.8), $\ln\left(p_{j,t}\right)$ is used in place of $p_{j,t}$, which is a convention in the literature.[10] The log of price should be used in (2.8) because it is inherited from the consumer's budget constraint. In contrast, we observe that $\ln\left(p_{j,t}\right)$ can also be used in place of $p_{j,t}$ in the utility specification of the logit demand system; by substituting $\ln\left(p_{j,t}\right)$ with $p_{j,t}$ in the linear utility specification in the logit demand model, the estimation equation of the proposed CES demand system lines up exactly with that of the homogeneous logit demand system. We take this substitution with the log of prices as a simple scale adjustment in the linear

9. We emphasize that $\mathcal{J}_t$ might not contain a numeraire for our CES demand system. In such a case, product 0 can be considered any alternative in $\mathcal{J}_t$, and all estimation equations that follow should be adjusted in terms of differences between j and 0.

10. See Appendix 2.A for the derivation of homogeneous and random coefficient logit models.

utility specification of the logit demand model. The predicted market share equation of the random coefficient logit model of demand developed by Berry, Levinsohn, and Pakes (1995) can be derived similarly. Let i denote an individual, and suppress the market subscript t temporarily. For the sake of notational simplicity, let $\phi_j := \ln p_j$. We specify the quasi-linear utility of the random coefficient logit model of demand as:

$$u_{i,j} = -\alpha_i \phi_j + \mathbf{x}_j' \boldsymbol{\beta}_i + \xi_j + \epsilon_{i,j}.$$

In contrast, the individual quantity share expressions of the CES demand system (2.4) become:

$$\pi_{i,j} = \frac{\chi_i \left(\mathbf{x}_j, \xi_j\right) \exp\left(-\sigma_i \phi_j\right)}{\sum_{k=0}^{J} \chi_i \left(\mathbf{x}_k, \xi_k\right) \exp\left(-\sigma_i \phi_k\right)} \tag{2.9}$$

$$= \frac{\exp\left(-\sigma_i \phi_j + \mathbf{x}_j' \boldsymbol{\beta}_i + \xi_j\right)}{\sum_{k=0}^{J} \exp\left(-\sigma_i \phi_k + \mathbf{x}_k' \boldsymbol{\beta}_i + \xi_k\right)}, \tag{2.10}$$

where the second equality follows by specifying $\chi_i \left(\mathbf{x}_j, \xi_j\right) = \exp\left(\mathbf{x}_j' \boldsymbol{\beta}_i + \xi_j\right)$. Note that (2.10) is nearly identical to the individual choice probability equation obtained by Berry, Levinsohn, and Pakes (1995).[11] The predicted market share equation is obtained by aggregating these individual quantity shares over i.

Discussions in the current subsection provide the microfoundation and justification for international trade and macroeconomics literature, based on the CES demand system, to use differentiated products demand estimation methods developed in empirical industrial organizational literature since Berry (1994); Berry, Levinsohn, and Pakes (1995); Nevo (2001). After model parameters are estimated, price and income elasticities can be calculated ac-

11. The only structural difference is the correlation structure of the individual heterogeneity; we must assume that $Cov\left(\sigma_i, \boldsymbol{\beta}_i\right) = \mathbf{0}$. As those cross-correlations are often assumed to be zero in practice when estimating the random coefficient logit model of demand (see Dube, Fox, and Su (2012)), we do not consider the restriction a serious limitation.

cording to equation (2.5), and the welfare analyses can be conducted correspondingly.

However, discrete choice differentiated product demand estimation literature imposes a critical restriction, which is necessary when inverting the individual quantity share, $\pi_{j,t} > 0$, for all j, t.[12] The restriction is inevitable in logit demand models, which assume additive idiosyncratic shocks on preferences distributed with unrestricted support. The most important example in the literature is additive i.i.d. Type-I extreme value distributed shocks. Individual choice probabilities derived from the assumption must have exponential functions in the numerators of choice probabilities. Zero quantity market shares are often observed in data, which are equated with predicted market shares for identification and estimation of model parameters. The flexibility of the quality kernel, $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)$, in our model allows us to accommodate zero predicted market shares by embedding a buy-or-not decision of the consumer, which determines extensive margins. We now illustrate how to accommodate zero predicted and observed market shares directly.

2.3.2 Accommodating Zero Predicted and Observed Market Shares: Separating Intensive and Extensive Margins

We restrict attention to homogeneous consumers again, and let $\mathbf{x}_{j,t} \neq \mathbf{w}_{j,t}$, $\eta_{j,t} \neq \xi_{j,t}$. We let $\mathcal{J}_t$ contain the numeraire for convenience of illustration, and normalize $p_{0,t} = 1$. The predicted market shares equation of the proposed CES demand system is:

$$\pi_{j,t} \quad = \quad \frac{\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) \exp\left(-\sigma\phi_{j,t}\right)}{\sum_{k\in\mathcal{J}_t} \chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) \exp\left(-\sigma\phi_{k,t}\right)}. \tag{2.11}$$

The expression (2.11) allows zero predicted market shares of product j by letting $\chi = 0$ for some subset of the product characteristic space where $\left(\mathbf{w}_{j,t}, \eta_{j,t}\right)$ lives on. By taking the

12. For a detailed discussion on share inversion, see Berry, Gandhi, and Haile (2013).

ratio $\pi_{j,t}/\pi_{0,t}$, we obtain a reduced form of the demand system (2.11) as:

$$\frac{\pi_{j,t}}{\pi_{0,t}} = p_{j,t}^{-\sigma} \frac{\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)}{\chi\left(\mathbf{x}_{0,t}, \xi_{0,t}, \mathbf{w}_{0,t}, \eta_{0,t}\right)}. \tag{2.12}$$

If $\mathcal{J}_t$ does not include the numeraire, any product with a strictly positive market share can be considered a reference product, denoted by product 0. All arguments in the current and subsequent sections remain valid provided that statistical independence of the observable and unobservable product characteristics across products can be assumed. This assumption implies that product characteristics are uncorrelated across products, which is consistent with many extant demand estimation frameworks, including Berry (1994); Berry, Levinsohn, and Pakes (1995). For tractability during identification and estimation, we consider the following functional form with an index restriction:

$$\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) = \mathbf{1}\left(\left\{\gamma + \mathbf{w}_{j,t}'\boldsymbol{\delta} + \eta_{j,t} > 0\right\}\right) \exp\left(\alpha + \mathbf{x}_{j,t}'\boldsymbol{\beta} + \xi_{j,t}\right), \tag{2.13}$$

where $\mathbf{1}(\cdot)$ is an indicator function. Employing this quality kernel is equivalent to assuming a certain structure on the consumer's choice. A consumer initially considers the utility from product characteristics represented by $\mathbf{w}_{j,t}'\boldsymbol{\delta} + \eta_{j,t}$. If the utility exceeds the threshold $-\gamma$, the consumer decides to buy the product. Then $\left(\phi_{j,t}, \mathbf{x}_{j,t}, \xi_{j,t}\right)$ is considered, which affects the amount of consumption $q_{j,t}$. In contrast, if the utility does not exceed the threshold $-\gamma$, the consumer decides not to buy the product, and thus, $q_{j,t} = \pi_{j,t} = 0$. We emphasize that $\mathbf{w}_{j,t}$ can contain the raw price, $p_{j,t}$, or other endogenous variables provided that the corresponding instruments are available to the researcher.

2.4 A Semiparametric Estimation Framework with Exponential Quality Kernel and Zero Market Shares

We provide a semiparametric estimation framework for the CES demand system with exponential quality kernel that accommodates zero predicted and observed market shares. The estimation method we provide includes two stages. During the first stage, parameters that determine extensive margins are estimated using the efficient semiparametric estimator developed by Klein and Spady (1993), and during the second, parameters that determine intensive margins are estimated, correcting for price endogeneity and selectivity bias caused by a consumer's choice set selection. The second-stage estimator that we use was developed by Ahn and Powell (1993); Powell (2001). When zero market shares are not observed in the data, one can proceed with existing demand estimation frameworks developed by Berry (1994); Berry, Levinsohn, and Pakes (1995) to estimate model parameters. The first-stage estimation framework illustrated in this section allows only exogenous covariates for the observed extensive margin shifter, $\mathbf{w}_{j,t}$. We chose this framework because of the availability of data and efficiency.[13] If a researcher wants to include endogenous variables such as prices in the extensive margin shifters, $\mathbf{w}_{j,t}$, the researcher can proceed with the method developed by Blundell and Powell (2003, 2004) or Rothe (2009) during the first stage. They provide semiparametric estimation frameworks for binary choice models with endogenous covariates.

We assume the existence of instruments for prices, such that $E\left[\xi_{j,t}|\mathbf{z}_{j,t}\right] = 0$, where $\mathbf{z}_{j,t}$ might include $\mathbf{x}_{j,t}$. Let $d_{j,t} = \mathbf{1}\left(\left\{\gamma + \mathbf{w}'_{j,t}\boldsymbol{\delta} + \eta_{j,t} > 0\right\}\right)$. It is well documented in the literature that $E\left[\xi_{j,t}|\phi_{j,t}, \mathbf{x}_{j,t}\right] \neq 0$, and it is highly likely to be positive. Consequently, when prices are not instrumented, upward-sloping demand curves are often estimated. The same intuition applies when a consumer's choice set selection is ignored and samples with zero

13. Although our data include information on product availability to consumers, even when sales in a corresponding week/store pair were zero, they do not include prices in corresponding weeks/stores without sales. Thus, we could not contain the endogenous variable $p_{j,t}$ in $\mathbf{w}_{j,t}$ during first-stage estimation. Characteristics of data used in our empirical application are discussed in Section 2.7.

observed market shares are simply dropped during estimation. Even after instrumenting for prices, $E\left[\xi_{j,t}|\mathbf{z}_{j,t}\right]=0$ does not imply that $E\left[\xi_{j,t}|\mathbf{z}_{j,t},d_{j,t}=1\right]$ is zero. $E\left[\xi_{j,t}|\mathbf{z}_{j,t},d_{j,t}=1\right]$ is likely to be positive because consumers select products with high $\eta_{j,t}$ during the first-stage choice set decision, and $\eta_{j,t}$ is likely to be positively correlated with $\xi_{j,t}$. Thus, dropping samples with zero observed market shares during estimation biases price coefficients upward, which can even yield positive price coefficients. Imputing zero observed market shares with some small positive numbers during estimation can cause an even more serious problem in that the direction of the bias is unpredictable.

We normalize $\phi_{0,t}\equiv\ln p_{0,t}=0$, $\xi_{0,t}=\eta_{0,t}=0$, $\mathbf{w}_{0,t}=\mathbf{0}$, and $\mathbf{x}_{0,t}=\mathbf{0}$. Under the choice of $\chi(\cdot)$ specified in (2.13), (2.12) simplifies to:

$$\frac{\pi_{j,t}}{\pi_{0,t}}=\mathbf{1}\left(\left\{\gamma+\mathbf{w}'_{j,t}\boldsymbol{\delta}+\eta_{j,t}>0\right\}\right)\exp\left(-\sigma\phi_{j,t}+\mathbf{x}'_{j,t}\boldsymbol{\beta}+\xi_{j,t}\right), \tag{2.14}$$

which is the econometric model that we identify and estimate in this section. A consumer buys product j if $\gamma+\mathbf{w}'_{j,t}\boldsymbol{\delta}+\eta_{j,t}>0$. For the sample with $d_{j,t}=1$, demand system (2.14) further reduces to:

$$\ln\left(\frac{\pi_{j,t}}{\pi_{0,t}}\right)=-\sigma\phi_{j,t}+\mathbf{x}'_{j,t}\boldsymbol{\beta}+\xi_{j,t}.$$

However, the conditional expectation $E\left[\xi_{j,t}|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right]$ is not zero anymore, which leads to the sample selection problem. Several methods to estimate parameters of the sample selection models have been proposed in the literature under different assumptions.[14] We follow Heckman (1979), who imposes a conditional mean restriction. By taking the conditional expectation, we have:

$$E\left[\ln\left(\frac{\pi_{j,t}}{\pi_{0,t}}\right)|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right]=-\sigma\phi_{j,t}+\mathbf{x}'_{j,t}\boldsymbol{\beta}+E\left[\xi_{j,t}|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right]. \tag{2.15}$$

14. For example, Powell (1984, 1986); Blundell and Powell (2007) propose a least absolute deviation type estimator under the conditional quantile restriction, and Honoré, Kyriazidou, and Udry (1997) propose symmetric trimming under the symmetricity assumption of error terms.

Ahn and Powell (1993); Powell (2001); Newey (2009) propose two-stage $\sqrt{N}$-consistent estimators for the model parameters of (2.15). We use the pairwise differenced weighted least squares estimator from Ahn and Powell (1993); Powell (2001), which corrects for the endogeneity of $\phi_{j,t}$ using instruments during the second stage. During the first stage, $\boldsymbol{\delta}$ should be estimated. A few estimators are available for this semiparametric binary choice model, among which we use the method from Klein and Spady (1993), which achieves asymptotic efficiency. During the second stage, parameters $(\sigma, \boldsymbol{\beta})$ from the following linear equation are estimated:

$$E\left[\ln\left(\frac{\pi_{j,t}}{\pi_{0,t}}\right)|\mathbf{z}_{j,t}, \mathbf{w}_{j,t}, d_{j,t}=1\right] = -\sigma\phi_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta} + \lambda\left(1 - G_\eta\left(-\mathbf{w}'_{j,t}\boldsymbol{\delta}\right)\right), \tag{2.16}$$

where $\lambda(\cdot)$ is an unknown smooth function. For semiparametric identification of $(\sigma, \boldsymbol{\beta})$, term $\lambda\left(1 - G_\eta\left(-\mathbf{w}'_{j,t}\boldsymbol{\delta}\right)\right)$ must not be a linear combination of $\left(\phi_{j,t}, \mathbf{x}_{j,t}\right)$; some component of $\mathbf{w}_{j,t}$ must be excluded from $\left(\phi_{j,t}, \mathbf{x}_{j,t}\right)$. We impose the following assumptions on the data-generating process for the $\sqrt{N}$-consistency and asymptotic normality of our proposed estimator.

Assumption 2.4.1. *The vector of observed product characteristics* $\mathbf{w}_{j,t}$ *is exogenous.*

Assumption 2.4.2. $\mathbf{w}_{j,t}$ *contains at least one component that is not included in* $\mathbf{x}_{j,t}$.

Assumption 2.4.3. $\eta_{j,t}$ *is independent of* $\mathbf{w}_{j,t}$ *with* $E\left[\eta_{j,t}|\mathbf{w}_{j,t}\right] = 0$, $\eta_{j,t}$ *is i.i.d. over* j *and over* t, *and the conditional distribution of* $\eta_{j,t}$ *given* $\mathbf{w}_{j,t}$ *has the full support over* $\mathbb{R}$ *with bounded first derivatives.*

Assumption 2.4.4. $\mathbf{w}_{j,t}$ *and* $\mathbf{x}_{j,t}$ *contain at least one continuous variable. Furthermore, the conditional distribution of the continuous variable conditioned on* $d_{j,t}$ *and other exogenous variables is sufficiently smooth.*

Assumption 2.4.5. *Denote* $\mathbf{r}_j := (\phi_j, \mathbf{x}_j)'$. *Denote* $g_{(\cdot)}(w) := E\left[\cdot|\mathbf{w}_{j,t}'\boldsymbol{\delta} = w\right]$ *and* $f(w)$ *be the density of* $\mathbf{w}_{j,t}'\boldsymbol{\delta}$. *Then,* $f(w)$, $g_{\mathbf{r}_{j,t}}(w)$, $g_{\mathbf{z}_{j,t}}(w)$, $g_{\eta_{j,t}}(w)$, $g_{\mathbf{z}_{j,t}\mathbf{w}_{j,t}'}(w)$, *and* $g_{\mathbf{z}_{j,t}\mathbf{r}_{j,t}'}(w)$ *are sufficiently smooth on their supports.*

Assumption 2.4.6. *There exists a set of instruments* $\mathbf{z}_{j,t}$ *such that* $\dim(\mathbf{z}_{j,t}) \geq \dim(\phi_{j,t}, \mathbf{x}_{j,t})$, $\xi_{j,t} \perp\!\!\!\perp \phi_{j,t}|\mathbf{z}_{j,t}$, *and* $E[\xi_{j,t}|\mathbf{z}_{j,t}] = 0$.[15]

Assumption 2.4.7. *The parameter vector* $(\sigma, \alpha, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta})$ *lies in a compact parameter space, with the true parameter value lying in the interior.*

In Assumptions 2.4.1 through 2.4.3, we impose the independence of observed and unobserved product characteristics, and homoskedasticity of unobservable product characteristics, $\eta_{j,t}$, that relate to extensive margins. However, we do not assume that unobserved product characteristics and prices are independent. We allow for endogeneity in prices, which should be considered during identification and estimation; prices can be a function of observed and unobserved product characteristics. Assumption 2.4.2 is the exclusion restriction, which is required for identification during the second stage. Note that a sufficient condition for Assumption 2.4.3 to hold is that $\eta_{j,t} \perp\!\!\!\perp \mathbf{w}_{j,t}$ and $\eta_{j,t}$ has a bounded and continuous density over the real line. Assumptions 2.4.4 and 2.4.5 are smoothness conditions, imposed for the suggested estimators to be well-behaved.[16] Assumption 2.4.6 is the standard instrument condition to correct for price endogeneity.[17] Assumption 2.4.7 is the usual compactness assumption.

We now describe first- and second-stage estimators. During the first stage, we estimate $\boldsymbol{\delta}$ using the efficient semiparametric estimator developed by Klein and Spady (1993). The

15. $\mathbf{z}_{j,t}$ may contain the exogenous components of $\mathbf{x}_{j,t}$.

16. Roughly, they require the existence of the higher-order derivatives for the respective conditional distribution and conditional expectation functions. See (C.6) of Klein and Spady (1993) and Assumption 5.7 of Powell (2001) for the exact conditions.

17. See, e.g., Nevo (2001) for a discussion of suitable instruments in practice.

estimator allows us to estimate parameters of binary choice models without having to specify the distribution of the unobservables. The insight is to replace the likelihood with its uniformly consistent estimates, and run the pseudo-maximum likelihood. The estimator is defined as:

$$\hat{\boldsymbol{\delta}} := \arg\max_{\boldsymbol{\delta}} \sum_{j,t} \left\{ \mathbf{1}\left(\pi_{j,t} > 0\right) \ln\left(1 - \hat{G}_{\eta}\left(-\mathbf{w}'_{j,t}\boldsymbol{\delta}\right)\right) + \mathbf{1}\left(\pi_{j,t} = 0\right) \ln\left(\hat{G}_{\eta}\left(-\mathbf{w}'_{j,t}\boldsymbol{\delta}\right)\right)\right\}, \tag{2.17}$$

where

$$\hat{G}_{\eta}\left(-\mathbf{w}'_{j,t}\boldsymbol{\delta}\right) = \hat{\tau}_{j,t} \frac{\sum_{k \neq j,t} \kappa\left(\frac{1}{h_n}\left(\mathbf{w}_k - \mathbf{w}_{j,t}\right)'\boldsymbol{\delta} + \iota_0\left(\boldsymbol{\delta}\right)\right)\left(1 - \mathbf{1}\left(\pi_{j,t} > 0\right)\right)}{\sum_{k \neq j,t} \kappa\left(\frac{1}{h_n}\left(\mathbf{w}_k - \mathbf{w}_{j,t}\right)'\boldsymbol{\delta} + \iota\left(\boldsymbol{\delta}\right)\right)}.$$

$\kappa(\cdot)$ is a fourth-order kernel, h_n is the bandwidth, and $\hat{\tau}_{j,t}, \iota_0(\boldsymbol{\delta}), \iota(\boldsymbol{\delta})$ are trimming sequences for small estimated densities.[18] During the second stage, we follow the method illustrated by Powell (2001). With an abuse of notation by suppressing the market index t and letting $\mathbf{r}_j = \left(\phi_j, \mathbf{x}_j\right)'$, the estimator is defined by the following weighted instrumental variable estimator:

$$\begin{aligned}\left(-\hat{\sigma}, \hat{\boldsymbol{\beta}}\right) = & \left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{\omega}_{i,j}\left(\mathbf{z}_i - \mathbf{z}_j\right)\left(\mathbf{r}_i - \mathbf{r}_j\right)'\right)^{-1} \\ & \times \left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{\omega}_{i,j}\left(\mathbf{z}_i - \mathbf{z}_j\right)\left(\ln\left(\frac{\pi_i}{\pi_0}\right) - \ln\left(\frac{\pi_j}{\pi_0}\right)\right)\right),\end{aligned} \tag{2.18}$$

where $\hat{\omega}_{i,j} = \frac{1}{h_n}\kappa\left(\frac{1}{h_n}\left(\mathbf{w}_i - \mathbf{w}_j\right)'\hat{\boldsymbol{\delta}}\right)$.[19][20] Intuition regarding the estimator suggests can-

18. We ignore these trimming sequences for technical and notational convenience from now on. Klein and Spady (1993) also note that the trimming does not affect the estimates in practice.

19. When the number of instruments is larger than that of explanatory variables, the projection matrix can be calculated beforehand to find the $\mathbf{z}_j$ vector. Efficiency loss might occur, but the estimator will be still $\sqrt{N}$-consistent and asymptotically normal.

20. The bandwidth sequence h_n should be such that $h_n \to 0$, $nh_n^6 \to \infty$, and $nh_h^8 \to 0$ as $n \to \infty$ in both

celing out the bias correction term, $\lambda\left(1-G_\eta\left(-\mathbf{w}_j'\boldsymbol{\delta}\right)\right)$; if $\mathbf{w}_i$ is the same as $\mathbf{w}_j$, term $\lambda\left(1-G_\eta\left(-\mathbf{w}_j'\boldsymbol{\delta}\right)\right)$ in (2.16) cancels out when differences are taken. Thus, more weights are placed on differenced terms that are close. The estimator is $\sqrt{N}$-consistent and asymptotically normal. For the closed-form covariance matrix formula and its consistent estimator, see Powell (2001). The following theorem summarizes discussions in this subsection thus far.

Theorem 2.4.1. *Under Assumptions 2.4.1 through 2.4.7,* $\left(-\hat{\sigma},\hat{\boldsymbol{\beta}}\right)$*, defined in (2.18), is* $\sqrt{N}$*-consistent and asymptotically normal.*

The semiparametric, log-linear estimation illustrated in this subsection requires an exclusion restriction on $\mathbf{w}_{j,t}$ to identify $(-\sigma,\boldsymbol{\beta})$; $\mathbf{w}_{j,t}$ cannot be a linear combination of $\left(\phi_{j,t},\mathbf{x}_{j,t}\right)$. This exclusion restriction can be circumvented by adding an interaction term or nonlinear transformation of a non-binary variable contained in $\left(\phi_{j,t},\mathbf{x}_{j,t}\right)$. For example, if one employs the method proposed by Blundell and Powell (2003, 2004), which accommodates endogenous variables during first-stage estimation, including raw prices, $p_{j,t}$, in $\mathbf{w}_{j,t}$ can be a viable choice. However, finding additional exogenous variables that affect only a consumer's buy-or-not decision is ideal. If one is willing to assume that $\eta_{j,t}$ is distributed as standard Gaussian, the classic Heckman correction estimator with instruments can be used, in which the inverse Mills ratio is added as an additional regressor. In that case, identification of model parameters is achieved by the non-linearity of the inverse Mills ratio, and therefore the exclusion restriction is unnecessary.

2.5 Excursus: Derivation of the Selection-Correction Estimation Equation for the Logit Demand Model

In this section, we provide a two-stage model of consumer choice within the logit demand frameworks when zero market shares are present. The logit demand model with two-stage

the first and the second stage.

decision process can also lead to the same estimation equation derived from our proposed CES demand system, which is presented in Section 2.4. Although sticking to the logit demand frameworks might be less appealing because the intensive and extensive margins cannot be distinguished conceptually, the single-choice assumption can be more adequate in some contexts. In such contexts, the first-stage decision obtains an interpretation akin to the consideration set selection in the consideration set literature.

We show that the estimation equation (2.15) can be derived from the two-stage decision process from the logit demand model. We consider a representative consumer with a two-stage decision process. During the first stage, the consumer searches $\mathcal{J}_t$, which includes all possible alternatives. The consumer's choice set $\mathcal{J}_t^+$ is determined from the search. During the second stage, the consumer encounters the usual discrete choice decision problem over $\mathcal{J}_t^+$, that is, purchase the product that yields the highest utility. Let $\left(\mathbf{w}_{j,t}, \eta_{j,t}\right)$ be the variables that affect the first-stage choice set search, and $\left(\mathbf{x}_{j,t}, \xi_{j,t}\right)$ the variables that affect the second-stage discrete choice unconstrained utility maximization problem. Notice that these variables form an analogue of the notations used in Sections 2.3 and 2.4. The second-stage utility of the consumer is modeled as:

$$u_{i,j,t} = -\alpha \ln p_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta} + \xi_{j,t} + \epsilon_{i,j,t}.$$

[21][22] The representative consumer solves:

$$\max_{j \in \mathcal{J}_t^+} \left\{ u_{i,j,t} \right\}.$$

With the i.i.d. Type-I extreme value assumption on $\epsilon_{i,j,t}$'s, the individual choice probability becomes:

$$\Pr(i \to j|t) = \frac{\exp\left(-\alpha \ln p_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta} + \xi_{j,t}\right)}{\sum_{k \in \mathcal{J}_t^+} \exp\left(-\alpha \ln p_{k,t} + \mathbf{x}'_{k,t}\boldsymbol{\beta} + \xi_{k,t}\right)}.$$

$\Pr(i \to j|t)$ is the predicted quantity market share, $\pi_{j,t}$. During estimation, $\pi_{j,t}$ is equated with the observed market share, $s_{j,t}$. The inversion theorem of Berry (1994); Berry, Levinsohn, and Pakes (1995) applies. Again, the only difference is the moment condition; the conditional expectation of $\xi_{j,t}$ given the instruments and exogeneous variables is not zero and is highly likely to be positive. Thus, a correction term is needed for the selection of the choice set, which leads to the estimation equation:

$$E\left[\ln\left(\frac{\pi_{j,t}}{\pi_{0,t}}\right)|\mathbf{z}_{j,t}, \mathbf{w}_{j,t}, d_{j,t} = 1\right] = -\alpha \ln p_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta} + E\left[\xi_{j,t}|\mathbf{z}_{j,t}, \mathbf{w}_{j,t}, d_{j,t} = 1\right]. \quad (2.19)$$

(2.19) coincides with (2.15).

---

21. By not including the income y_i, we disregard the "indirect utility" interpretation of the alternative choice utility $u_{i,j,t}$ here. Early literature on the discrete choice consumer demand, which dates back to McFadden (1974, 1978, 1981); Dubin and McFadden (1984); Anderson, de Palma, and Thisse (1987, 1988, 1992), stick to the indirect utility interpretation of an alternative utility. To our understanding, the main intention of interpreting $u_{i,j,t}$ as an indirect utility was to place the discrete choice demand systems in the context of the Walrasian demand, especially because the price should not be a direct argument of the Walrasian utility function. However, discrete choice modeling has gained greater popularity and has been applied to a much wider context since McFadden's original works. The "modern" approach tend more to interpret $u_{i,j,t}$ as a direct utility of an alternative. Many recent research using the discrete choice demand estimation framework specify the mean utility as either linear in $-p_{j,t}$ or in $y_i - p_{j,t}$ so the income y_i cancels out over alternatives. See, for example, Berry (1994); Nevo (2001). Berry, Levinsohn, and Pakes (1995) suggest using $-p_{j,t}$ in their homogeneous coefficients utility specification and $\ln(y_i - p_{j,t})$ in their random coefficients utility specification. We note that y_i does not cancel out over the alternatives in the latter case.

22. In the context of the logit demand models, the utility can be linear, not log-linear, in prices. If one does not want to interpret the estimated parameters to be originating from a CES demand system, one can replace $\ln p_{j,t}$ with $p_{j,t}$.

2.6 Monte-Carlo Simulations

We simulate market data and back out model parameters to examine the finite-sample performance of the estimator that we proposed in the previous section. We compare the estimation result using our model to the estimation result of the logit demand model, which either drops the sample with zero observed market shares or imputes the zero observed market shares with a small positive number. The estimator we proposed in the previous section works well when the model is specified correctly.

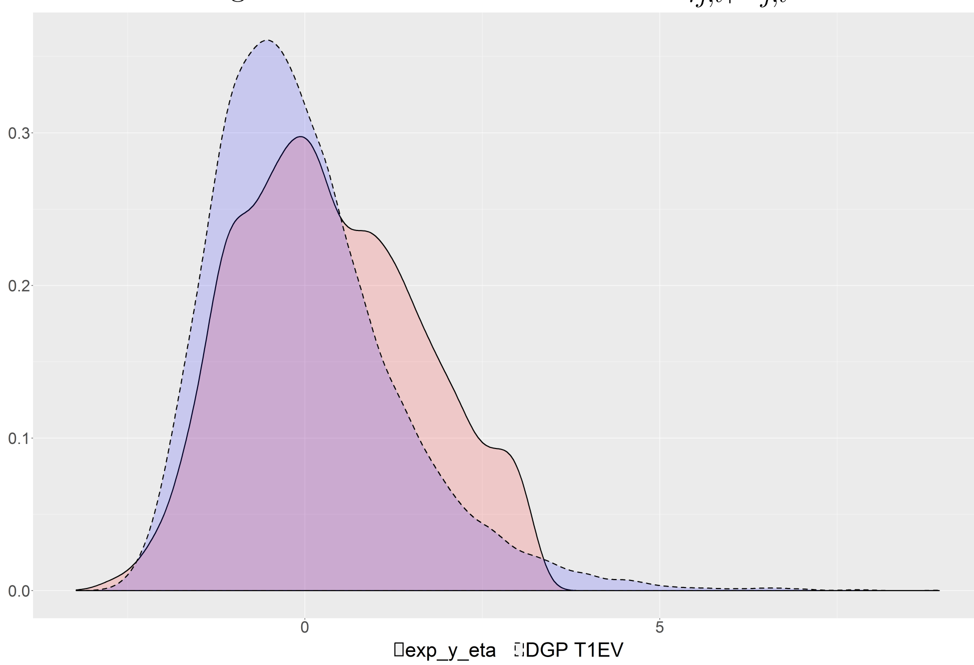
We first describe our data-generating process that satisfies the exclusion restriction. Each market, t, has two to five products, with the exact number of products in each market drawn randomly. The observed product characteristic vector, $\mathbf{w}_{j,t}$, includes three continuous components, one discrete component, and three brand dummies. One of the continuous components is excluded in $\mathbf{x}_{j,t}$. The first component, $w_{j,t}^{(1)}$, follows $\text{lognormal}\,(0,1)$, the second component, $w_{j,t}^{(2)}$, follows $\text{uniform}\,(1,5)$, the third, $w_{j,t}^{(3)}$, $\text{Poisson}\,(3)$, and the fourth, $w_{j,t}^{(4)}$, $\mathcal{N}\,(0,1)$. $w_{j,t}^{(4)}$ is excluded from $\mathbf{x}_{j,t}$. $\eta_{j,t}|\mathbf{w}_{j,t}$ follows the Type-I extreme value distribution with mean zero. Two instruments are employed for prices, which are proxies for cost shocks. Prices, $p_{j,t}$, which is an endogenous variable, is determined by $p_{j,t} = \psi\left(\mathbf{x}_{j,t}, \xi_{j,t}\right)$, where ψ is some (possibly) nonlinear function that is strictly monotonic in $\xi_{j,t}$. We specify ψ as:

$$
\begin{aligned}
&\psi\left(\mathbf{z}_{j,t}, \xi_{j,t}\right) \\
&= 2 + \frac{1}{50}\left(2z_{j,t}^{(1)} + 4z_{j,t}^{(2)} + 2x_{j,t}^{(1)} + x_{j,t}^{(1)}x_{j,t}^{(2)} - x_{j,t}^{(2)}x_{j,t}^{(3)} + 5x_{j,t}^{(4)} + 7x_{j,t}^{(5)} + 9x_{j,t}^{(6)} + 8\xi_{j,t}\right).
\end{aligned}
$$

We intentionally let the influence of the cost proxies, $z_{j,t}^{(1)}$ and $z_{j,t}^{(2)}$, to be fairly weak, which reflects common circumstances in practice. We calibrate the parameters as $\sigma = 2$, $\alpha = 1$, $\boldsymbol{\beta} = (1, -2, 1.5, 0.3, 0.2, 0.4)'$, $\gamma = \alpha$, and $\boldsymbol{\delta} = \frac{1}{4} \times (\boldsymbol{\beta}, 0.1)'$, and market shares are determined by (2.11).

Figure 2.1 depicts the estimated density of $\eta_{j,t}|\mathbf{w}_{j,t}$ from the first stage, and compares

Figure 2.1: Estimated Densities of $\eta_{j,t}|\mathbf{w}_{j,t}$



"exp_y_eta," the pink solid density, is the estimated density of $\eta_{j,t}|\mathbf{w}_{j,t}$ from the Klein-Spady model. "DGP T1EV," the blue dotted density, is the Type-I Extreme Value density that is used to generate the data. 10,000 sample draws are taken and plotted from the estimated density of the Klein-Spady model and the true Type-I Extreme Value density, respectively. The Klein-Spady model identifies the distribution of unobservables up to location and scale. Thus, we made the location and scale adjustment.

Table 2.1: Estimation Result of the Simulated Data

Estimation	(1) Our Model, K/S	(2) Our Model, Heckman	(3) Logit, Drop 0	(4) Logit, Impute 10^{-8}
Log prices (-2)	-1.969	-1.972	-1.287	-5.444
	(0.128)	(0.119)	(0.074)	(0.372)
$x_{j,t}^{(1)}$ (1)	0.807	1.017	0.886	1.536
	(0.014)	(0.011)	(0.007)	(0.041)
$x_{j,t}^{(2)}$ (-2)	-1.635	-2.045	-1.574	-3.884
	(0.028)	(0.035)	(0.013)	(0.050)
$x_{j,t}^{(3)}$ (1.5)	1.222	1.530	1.339	1.821
	(0.020)	(0.023)	(0.008)	(0.040)
$x_{j,t}^{(4)}$ (0.3)	0.251	0.290	0.366	-0.468
	(0.051)	(0.071)	(0.046)	(0.219)
$x_{j,t}^{(5)}$ (0.2)	0.140	0.169	0.266	-0.673
	(0.051)	(0.071)	(0.046)	(0.220)
$x_{j,t}^{(6)}$ (0.4)	0.338	0.435	0.458	-0.099
	(0.053)	(0.073)	(0.047)	(0.225)
D	5059	5059	5059	10500
N	10500	10500	5059	10500

Target values are in parentheses of corresponding items in the first column. The Estimation row specifies the method used during estimation. Column (1) is our proposed estimator, in which the first-stage propensity score was estimated using the Klein-Spady estimator. For Column (2), Probit was used for the first-stage propensity score estimation, and the inverse Mills ratio is added as an additional regressor during the second stage. Column (3) is the logit estimator with dropping the samples with zero observed market shares, and Column (4) is the logit estimator with imputing 10^{-8} in place of the zero observed market shares. Asymptotic standard error estimates appear in parentheses. D is the number of non-censored samples, and N is the effective sample size.

it with the distribution used for generating the data. Although the estimated density does not coincide perfectly with the exact density of the Type-I extreme value distribution, it preserves the approximate shape of the distribution. A larger sample is needed for the estimated densities to fit exactly with the distribution used during data generation.

Table 2.1 shows the estimation results of the simulated data. The "Estimation" row indicates the estimation method used. Column (1) is the correct quality kernel specification with our semiparametric estimator, and Column (2) is the correct quality kernel specification with the classical Heckman correction estimator assuming Gaussian error term in the first stage. Our estimator is successful in recovering the true parameters if the model is specified correctly. The estimator continues to be successful when we estimated the model using the classical Heckman correction estimator that assumes the joint normality of the error term distribution. Column (3) is the logit estimator where we drop the sample with zero observed market shares, and Column (4) is the logit estimator where we impute small positive numbers in place of zero observed market shares. Both dropping zeros and imputing small numbers in place of zeros bias the estimators substantially. The price coefficient is biased upward when the zero shares are dropped, whereas it is biased downward when a small number is imputed in place of zero shares. We also generated and estimated several other specifications, such as different error term distributions, functional forms of quality kernels, variables, pricing functions, etc. For brevity, we do not present all specifications here, but we note that results and implications presented in this section remain robust to these alternative specifications. Details on the estimation procedure appear in Appendix 2.B.

2.7 Empirical Example: Scanner Data with a Multitude of Zero Shares

We implement our proposed demand estimation framework using Dominick's supermarket cola sales scanner data. Data were obtained from the James M. Kilts Center for Marketing, University of Chicago Booth School of Business. The data contained weekly pricing and sales information for the Dominick's chain of stores from 1989 to 1997 for every universal product code (UPC) level product in 29 product categories. Promotion statuses and profitability of each unit sold were also recorded in the data. One shortcoming was that systemic records of product characteristics were unavailable, which we overcame by choosing cola sales data and hand-coding the product characteristics.

2.7.1 Data

We chose Dominick's data because they were ideal for illustrating the application of our framework for two reasons. First, Dominick's data contained information on which products were displayed on the shelves, even if a product did not sell in the corresponding week and store. This feature was necessary because we wanted the exact information on products that were in a consumer's choice set but were not chosen. Presented in Figure 2.2, approximately one-fourth of observations exhibited zero observed market shares. Second, Dominick's data contained information on average profit per unit sold. Combined with price data, we could back out the average cost per unit. Cost information is useful because an ideal instrument for prices when estimating consumer demand should proxy cost shocks. We avoided constructing instruments using indirect proxies for cost, which has been a major difficulty in demand estimation literature.

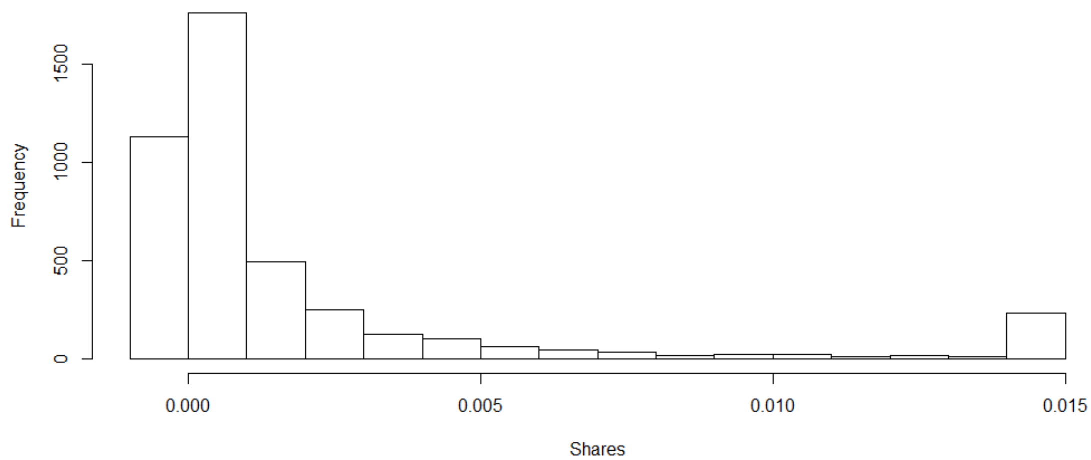
We focus on cola sales for several reasons. First, the cola market is a typical market of product differentiation, in which many brands with disparate tastes and packages competes. Among them, Coke and Pepsi, the two prominent brands, take the majority of market shares. Second, product characteristics were not coded separately in Dominick's data, but only category information such as "soft drinks" or "bottled juices." We had to extract the information from product descriptions truncated at 30 characters, for which cola was ideal because it had clearly labeled product characteristics. Finally, companies producing cola, and product characteristics of cola, have not changed much during the past few decades. Coke and Pepsi have been two leaders in the market. Diet, cherry-flavored, and caffeine-free colas are still sold in the market with considerable market shares in 2016, and in 1996. This feature made our analysis convenient, and the implications of analysis more realistic.

Dominick's data covered 100 chain stores in the Chicago area for 400 weeks, from September 1989 to May 1997. We chose the cross-section of week 391, which is the second week of March 1997. We used the cross-section data of a week because demand for soft drinks fluctuates in weeks with holidays or events such as the Super Bowl, and varies considerably by season. Therefore, we chose a week in March without any close holidays. As Dominick's experimented with prices across chain stores for the same product during the same week, we still have sufficient price variations after choosing a cross-section of data. Even after restricting the sample to a cross-section of one week, the sample size was as large as 4,300. We present summary statistics in Tables 2.2, 2.3, and 2.4.

We define individual products and markets naturally. An individual product was defined by its UPC, and a market by a store-week pair. This was the finest manner of defining a product and market that the data allowed, which resulted in a multitude of zero observed market shares. Illustrated in Figure 2.2, approximately one-fourth of products that were displayed on shelves did not sell.

We converted package prices and costs to per-ounce prices and costs. Dominick's did not record the price and cost of the week if sales of a product were zero in a corresponding week. Therefore, we could not include prices in $\mathbf{w}_{j,t}$, and proceeded only with other exogenous variables during first-stage estimation. When estimating the logit model while substituting the zero observed market shares with small numbers, we imputed missing prices and costs using other chain stores' prices and profits with the same product and promotion status. We had to compute market shares of outside options for both our model and the logit demand

Figure 2.2: Histogram of the Observed Market Shares



This figure plots the histogram of the observed quantity market shares for cola sales of week 391 (03/06/1997 to 03/12/1997) in Dominick's scanner data. Sample points larger than 0.015 is top-coded as 0.015. 216 out of 4356 (4.96%) sample points are top-coded. 1130 out of 4356 samples (25.94%) have zero market shares.

Table 2.2: Summary of the Product Characteristics of the Sample

	Frequency	Mean	Std
Diet	2163	0.497	0.500
Caffeine Free	1085	0.249	0.433
Cherry	151	0.035	0.183
Coke	365	0.084	0.277
Pepsi	2644	0.607	0.488
Promo	1751	0.402	0.490
Bottle Size	-	26.592	29.696
# Bottles per Bundle	-	12.436	9.667
# Stores	73	-	-
Uncensored Obs (D)	3226	-	-
Sample Size (N)	4356	-	-

Data are the cross-section of week 391 (03/06/1997 to 03/12/1997) in Dominick's scanner data.

Table 2.3: Per-ounce Price, Cost, Profitability, and Market Shares of Products in the Full Sample

	Mean	Median	Std	Min	Max
Per-ounce Prices ($)	0.020	0.024	0.014	0	0.042
Per-ounce Cost ($)	0.014	0.017	0.010	0	0.028
Profitability (%)	20.470	28.380	16.020	−98.550	58.620
Shares (%)	0.586	0.038	2.879	0	42.967

Data are the cross-section of week 391 (03/06/1997 to 03/12/1997) in Dominick's scanner data. The mean and median of price and cost were calculated including those zeros.

Table 2.4: Per-ounce Price, Cost, Profitability, and Market Shares of Products in the Non-censored Sample

	Mean	Median	Std	Min	Max
Per-ounce Prices ($)	0.027	0.026	0.008	0.005	0.042
Per-ounce Cost ($)	0.019	0.019	0.006	0.003	0.028
Profitability (%)	27.640	29.180	12.499	−98.550	58.620
Shares (%)	0.791	0.084	3.321	0.001	42.967

Data are the cross-section of week 391 (03/06/1997 to 03/12/1997) in Dominick's scanner data.

model.[23] When estimating market size, we assumed that an average person consumed 100 ounces of soft drinks a week,[24] and computed the size of the market using daily customer count data for each store in the chain.

2.7.2 Estimation, Result, and Discussion

We estimate our model using the method proposed in Section 2.4. We also estimate the model correcting for a consumer's choice set selection using the Probit as a first-stage estimator, with the Powell (2001) estimator and the simple Heckman selection correction estimator during the second stage. The simple Heckman estimator was implemented using the inverse Mills ratio as an additional regressor as usual. As a benchmark, we estimated

23. Although including the numeraire in a consumer's choice set was unnecessary in our model, we included it because we wanted to compare estimation results of our model with those of the logit model using the same setup.

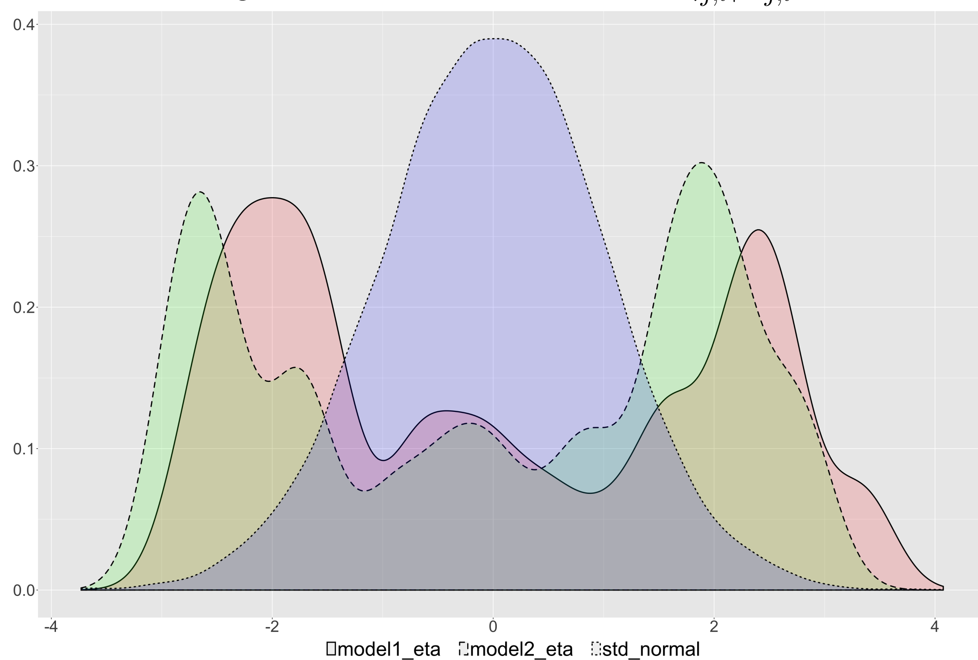
24. On average, Americans consume about 45 gallons of soft drinks a year. Source: http://adage.com/article/news/consumers-drink-soft-drinks-water-beer/228422/

the homogeneous logit model of demand, with different ways of handling the zero observed market shares: (i) dropping samples with zero observed market shares, and (ii) substituting zero observed market shares with small numbers. We also used the log of prices in the logit model to compare the magnitudes of coefficients. Mentioned previously, using the log of prices instead of raw prices represents a scale adjustment in the utility specification of the logit demand model.

We estimated two models with different specifications. In the baseline model (Model 1), $\mathbf{x}_{j,t}$ includes several product characteristics: bottle size, number of bottles per bundle, diet, caffeine-free, cherry flavor, Coke/Pepsi brand dummies. As an instrument of the per-ounce price, we used the per-ounce cost calculated from the profitability variable. For Model 1, we excluded promotion status from $\mathbf{x}_{j,t}$, and use it as a variable that satisfies the exclusion restriction. The exclusion assumption in this case reflects the informational hypothesis: promotions affect only consumers' information about a choice set, not the level of utility associated with consuming a certain product. For Model 2, we included the promotion statuses in $\mathbf{x}_{j,t}$, and used store-level demographics for variables included in $\mathbf{w}_{j,t}$ that were not included in $\mathbf{x}_{j,t}$: % Blacks and Hispanics, % college graduates, and log of the median income. The exclusion assumption of these variables reflects the preferential hypothesis of the extensive margin: a consumer who never buys a certain product will not become an inframarginal consumer regardless of other product characteristics.

The first-stage parameter estimation result for $\hat{\boldsymbol{\delta}}$ is shown in Table 2.5. Model 1 is the baseline model, with promotion statuses as excluded variables during second-stage estimation. Model 2 can be considered an additional robustness check, which uses the store-level demographics in the first stage. For Models 1 and 2, we estimated the Probit model for a benchmark, and for setting an initial value for the nonlinear optimizer to estimate the Klein-Spady model. The coefficient for bottle size was normalized to 1. We find that coefficient estimates from Probit estimation and Klein-Spady estimation are considerably different. We

Figure 2.3: Estimated Densities of $\eta_{j,t}|\mathbf{w}_{j,t}$



“model1_eta,” the pink solid density, is the estimated density of $\eta_{j,t}|\mathbf{w}_{j,t}$ from Model 1. “model2_eta,” the green dotted density, is the estimated density of $\eta_{j,t}|\mathbf{w}_{j,t}$ from Model 2. “std_normal,” the blue dotted density, is the standard normal density plotted for benchmark. The Klein-Spady model identifies the distribution of unobservables up to location and scale, and thus we made a location and scale adjustment of $E[\eta_{j,t}|\mathbf{w}_{j,t}] = 0$ and $Var(\eta_{j,t}|\mathbf{w}_{j,t}) = 4$. For Model 1 and 2, 10,000 sample draws were taken from the density estimates of the Klein-Spady model, and the density of the drawn sample was then plotted. The density of 10,000 sample draws from the standard Gaussian distribution is plotted for comparison.

Table 2.5: First-stage Parameter Estimates $\hat{\boldsymbol{\delta}}$

	Model 1		Model 2	
	Probit	Klein-Spady	Probit	Klein-Spady
$\mathbf{w}_{j,t}$	(1)	(2)	(3)	(4)
Bottle Size	1	1	1	1
	(0.195)	(−)	(0.196)	(−)
# Bottles per Bundle	−1.431	−91.991	−1.453	−133.500
	(0.329)	(3.780)	(0.330)	(12.535)
Diet	18.342	−277.300	18.222	−515.187
	(4.818)	(12.238)	(4.822)	(50.981)
Caffeine Free	−13.619	149.733	−13.663	118.946
	(6.158)	(10.108)	(6.171)	(4.446)
Cherry	−53.619	19.536	−53.582	−22.960
	(13.821)	(4.700)	(13.805)	(4.620)
Coke	−41.949	−20.901	−41.915	91.351
	(10.192)	(3.607)	(10.209)	(4.797)
Pepsi	79.989	−70.255	79.904	−170.295
	(5.917)	(2.492)	(5.910)	(8.323)
Promo	135.449	396.440	135.534	601.361
	(8.679)	(18.234)	(8.686)	(53.123)
% Blacks and Hispanics	-	-	−29.891	−6.454
			(21.430)	(13.069)
% College Graduates	-	-	3.850	4.550
			(20.978)	(12.624)
Log Median Income	-	-	−16.888	26.399
			(93.441)	(43.820)
D	3226	3226	3226	3226
N	4356	4356	4356	4356

D is the number of non-zero market share observations, and N is the sample size. Asymptotic standard error estimates appear in parentheses. The unit of bottle size is liquid ounces. We normalized the coefficients of the Bottle Size variable to one.

Table 2.6: Implied Own Price Elasticities and Second-stage Parameter Estimates $\left(\hat{\sigma}, \hat{\boldsymbol{\beta}}\right)$

	Our Model, K/S		Our Model, Probit		Heckman Correction		Logit Model		
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Drop 0	10^{-8}	10^{-4}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mean Price Elasticity (CES)	−1.326	−1.369	−1.138	−1.139	−1.299	−1.290	-	-	-
Mean Price Elasticity (Logit)	−1.324	−1.367	−1.133	−1.135	−1.297	−1.287	4.777	7.703	4.726
Log Price ($-\sigma$ / $-\alpha$)	−1.331	−1.375	−1.140	−1.142	−1.304	−1.295	4.805	7.748	4.754
	(0.041)	(0.040)	(0.058)	(0.061)	(0.040)	(0.041)	(0.099)	(0.158)	(0.074)
Bottle Size	0.011	0.009	0.020	0.021	0.013	0.011	0.121	0.219	0.118
	(0.002)	(0.001)	(0.010)	(0.019)	(0.001)	(0.002)	(0.004)	(0.007)	(0.003)
# Bottles per Bundle	0.167	0.060	0.176	0.171	0.130	0.136	0.385	0.589	0.357
	(0.010)	(0.025)	(0.019)	(0.010)	(0.003)	(0.006)	(0.012)	(0.020)	(0.009)
Diet	0.361	0.207	−0.350	−0.316	−0.186	−0.219	0.943	2.241	0.940
	(0.072)	(0.124)	(0.160)	(0.112)	(0.045)	(0.056)	(0.118)	(0.206)	(0.097)
Caffeine Free	−1.369	−1.198	−1.104	−1.104	−1.072	−1.052	−2.157	−2.924	−1.859
	(0.058)	(0.054)	(0.178)	(0.116)	(0.056)	(0.063)	(0.131)	(0.229)	(0.108)
Cherry	−2.990	−2.927	−3.086	−3.085	−2.232	−2.139	−3.452	−5.367	−2.915
	(0.123)	(0.132)	(0.835)	(1.071)	(0.176)	(0.185)	(0.343)	(0.535)	(0.252)
Coke	−0.375	−0.181	−0.828	−0.729	0.063	0.071	0.593	1.151	0.472
	(0.103)	(0.115)	(0.255)	(0.549)	(0.102)	(0.113)	(0.252)	(0.393)	(0.186)
Pepsi	1.644	1.699	0.977	1.120	1.068	0.868	4.398	9.090	4.374
	(0.069)	(0.066)	(0.191)	(0.962)	(0.062)	(0.191)	(0.172)	(0.269)	(0.127)
Promo	-	0.671	-	0.168	-	−0.260	-	-	-
		(0.130)		(1.446)		(0.233)			
D	3226	3226	3226	3226	3226	3226	3226	4090	4090
N	4356	4356	4356	4356	4356	4356	3226	4090	4090

For columns "Our Model, K/S" and "Our Model, Probit," results from the Klein-Spady and Probit estimators in Table 2.5 were used for the first-stage estimator, respectively. Then, the pairwise differenced weighted instrumental variable estimator was used during the second stage. For the "Heckman Correction" columns, Probit was used during the first stage, and Heckman's selection correction estimator with the inverse Mills ratio as an additional regressor was used during the second stage. Row "Mean Price Elasticity (CES)" is the mean of the implied Marshallian own price elasticity over the sample for the CES demand system. Row "Mean Price Elasticity (Logit)" is the mean of the implied own price elasticity over the sample for the Logit demand system when the alternative utility is log-linear in prices. D is the number of non-zero market share observations, and N is the effective sample size. Asymptotic standard error estimates are in parentheses. (v) The unit of bottle size is liquid ounces. Because Dominick's did not record the price and cost when sales were zero, when estimating the 10^{-8} and 10^{-4} columns, we used average prices and costs of the same product with the same promotion statuses from other stores.

also plot the estimated conditional density of $\eta_{j,t}$ given $\mathbf{w}_{j,t}$ from each model in Figure 2.3. The estimated density of $\eta_{j,t}$ given $\mathbf{w}_{j,t}$ is not even unimodal, which is strong evidence that the unobservable product characteristic, $\eta_{j,t}$, does not follow a Gaussian distribution.

The primary estimation result is shown in Table 2.6. In the first two rows, we present the mean of the implied own price elasticities from the Marshallian CES and logit demand systems, respectively. In the logit demand models, coefficients of the log of prices were positive, and economically and statistically significant, even after instrumenting for prices using supplier side cost information. As a result, the own price elasticities are positive, meaning that an upward-sloping demand curve is estimated. In contrast, the log-linear estimation of our model with Klein-Spady first-stage estimator returned the expected signs and magnitudes for coefficients of the log of prices, and thus, the own price elasticities are negative and the estimated demand curves are downward-sloping. Estimators assuming a standard Gaussian distribution on unobservables performed well, despite estimated distributions of unobservables being far from Gaussian. We argue that such good performance of models assuming normality is due to the fact that estimated propensity scores from the Klein-Spady and Probit models correlate highly, with a correlation coefficient of about 0.7 for Models 1 and 2. We are unsure whether it can be generalized to a different dataset or market.[25]

Results provide strong evidence of a choice set selection process that has been ignored in demand estimation literature. Ignoring the choice set selection process of consumers biases the estimates, even resulting in an upward-sloping demand curve. Recall the estimation equation (2.15) under the exponential quality kernel:

$$E\left[\ln\left(\frac{\pi_{j,t}}{\pi_{0,t}}\right)|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right]=-\sigma\phi_{j,t}+\mathbf{x}'_{j,t}\boldsymbol{\beta}+E\left[\xi_{j,t}|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right].$$

25. Gandhi, Lu, and Shi (2013) also uses bath tissue data from Dominick's database, using a time series variation of a single chain store. Their estimated demand function is much more elastic than ours. For example, price coefficient estimates from simply dropping samples with zero market shares, which should be biased upward, remain negative. However, the implication they draw – that samples with zero observed market shares should not be simply dropped – is similar to ours.

Except for term $E\left[\xi_{j,t}|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right]$, the estimation equation is the same as that of the logit demand model when we dropped samples with zero observed market shares. Columns (1) (Our Model, K/S, Model 1) and (7) (Logit Model, Drop 0) should coincide exactly when term $E\left[\xi_{j,t}|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right]$ is zero, yet this was not the case. $E\left[\xi_{j,t}|\mathbf{z}_{j,t},\mathbf{w}_{j,t},d_{j,t}=1\right]$ is likely positive in our case because consumers select unobservables $\eta_{j,t}$ and observables $\mathbf{w}_{j,t}$, and $\eta_{j,t}$ correlates highly with $\xi_{j,t}$. Even after instrumenting for prices, price coefficient estimates are likely to be biased upward when samples with zero observed market shares are simply dropped. Imputing small numbers on zero observed market shares might cause a more serious problem – the direction of the bias cannot be predicted. In contrast to Table 2.1 in the previous section, Table 2.6 shows that imputing zero observed market shares with small positive numbers causes upward bias in price coefficient estimates. We cannot explain the direction of the bias when zeros are imputed.

2.8 Conclusion

We develop a semiparametric demand estimation framework based on the Marshallian demand function derived from the budget-constrained CES utility maximization problem. Our framework is sufficiently flexible to incorporate observed and unobserved product characteristics, and is compatible with the widely used homogeneous and random coefficient logit models of demand. The framework accommodates zero predicted and observed market shares with a reasonable microfoundation by separating intensive and extensive margins, and embedding both margins in a quality kernel. We account for selection of a consumer's choice set, which is unrecognized in the literature. If the choice set selection stage is ignored, estimates of price coefficients can be misleading not only regarding their magnitudes, but also their signs. We demonstrate that ignoring choice set selection can even result in upward-sloping demand curves. A direct extension of our study is a random coefficient demand estimation framework that can accommodate zero predicted and observed market shares.

When a representative agent is assumed, the own and cross price elasticities derived from our model exhibited unrealistic substitution patterns, as in the homogeneous logit demand model of Berry (1994). Overcoming such unrealistic substitution patterns was one of the most important motivations for development of a random coefficient logit model of demand by Berry, Levinsohn, and Pakes (1995). Although we provide the microfoundation for a random coefficient CES demand estimation framework, we do not develop identification and estimation of model parameters with random coefficients that can accommodate zero market shares. We leave that extension to future research.

2.A Derivation of the Logit Demand System

We illustrate the derivation of a homogeneous and random coefficient logit demand systems for completeness. The illustration in this section largely follows the original presentation of Berry (1994); Berry, Levinsohn, and Pakes (1995).

Let $j \in \mathcal{J}_t$, where $\mathcal{J}_t$ is a finite set of alternatives that must contain the numeraire. Individual i in market t solves the following discrete choice utility maximization problem:

$$\max_{j \in \mathcal{J}_t} \left\{ u_{i,j,t} \right\},$$

where the (indirect) utility of choosing alternative j in market t is:

$$u_{i,j,t} = \alpha_i \left(y_i - p_{j,t} \right) + \mathbf{x}'_{j,t} \boldsymbol{\beta}_i + \xi_{j,t} + \epsilon_{i,j,t}. \tag{2.20}$$

$\epsilon_{i,j,t}$ follows the i.i.d. Type-I extreme value distribution. Note that it is also legitimate to specify the utility as:

$$u_{i,j,t} = -\alpha_i \ln p_{j,t} + \mathbf{x}'_{j,t} \boldsymbol{\beta}_i + \xi_{j,t} + \epsilon_{i,j,t},$$

given that we stick to the direct utility interpretation of $u_{i,j,t}$ as individual i choosing alternative j in market t. The logarithm can be regarded as a scale adjustment on the level of disutility from prices.

The coefficients $(\alpha_i, \boldsymbol{\beta}_i)$ might vary over individuals, and are specified as:

$$\alpha_i := \alpha + \Pi_\alpha \mathbf{q}_i + \Sigma_\alpha v_{\alpha,i}$$
$$\boldsymbol{\beta}_i := \boldsymbol{\beta} + \boldsymbol{\Pi}_\beta \mathbf{q}_i + \boldsymbol{\Sigma}_\beta \mathbf{v}_{\beta,i},$$

where $\mathbf{q}_i$ is the demographic variable, $\mathbf{v}_i$ is the vector of a unit normal shock, $\left(\Pi_\alpha, \boldsymbol{\Pi}_\beta\right)$ is the correlation component between demographic variables and the corresponding coefficients, and $\left(\Sigma_\alpha, \boldsymbol{\Sigma}_\beta\right)$ represents the covariance structure of the shocks on the coefficients. The linear utility specification (2.20) becomes:

$$\begin{aligned}
& u_{i,j,t} \\
= \; & \alpha_i \left(y_i - p_{j,t}\right) + \mathbf{x}'_{j,t}\boldsymbol{\beta}_i + \xi_{j,t} + \epsilon_{i,j,t} \\
= \; & \alpha_i y_i - \left(\alpha + \Pi_\alpha \mathbf{q}_i + \Sigma_\alpha v_{\alpha,i}\right) p_{j,t} + \mathbf{x}_{j,t}\left(\boldsymbol{\beta} + \boldsymbol{\Pi}_\beta \mathbf{q}_i + \boldsymbol{\Sigma}_\beta \mathbf{v}_{\beta,i}\right) + \xi_{j,t} + \epsilon_{i,j,t} \\
= \; & \alpha_i y_i + \left(-\alpha p_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta} + \xi_{j,t}\right) - \left(\Pi_\alpha \mathbf{q}_i + \Sigma_\alpha v_{\alpha,i}\right) p_{j,t} + \mathbf{x}'_{j,t}\left(\boldsymbol{\Pi}_\beta \mathbf{q}_i + \boldsymbol{\Sigma}_\beta \mathbf{v}_{\beta,i}\right) + \epsilon_{i,j,t} \\
= \; & \alpha_i y_i + \left(-\alpha p_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta} + \xi_{j,t}\right) + \begin{pmatrix} -p_{j,t} & \mathbf{x}'_{j,t} \end{pmatrix} \left(\boldsymbol{\Pi}\mathbf{q}_i + \boldsymbol{\Sigma}\mathbf{v}_i\right) + \epsilon_{i,j,t} \\
=: \; & \alpha_i y_i + \delta_{j,t} + \mu_{i,j,t} + \epsilon_{i,j,t},
\end{aligned}$$

where $\delta_{j,t}$ is the mean utility of alternative j that is common to every individual in market t, and $\mu_{i,j,t}$ is the individual specific structural utility component. For the log-linear specification, one can simply replace the term $p_{j,t}$ with $\ln p_{j,t}$.

Given the assumption that $\epsilon_{i,j,t}$ follows i.i.d. Type-I extreme value distribution, the individual choice probability $\Pr(i \to j|t)$ becomes:

$$\Pr(i \to j|t) = \frac{\exp\left(\delta_{j,t} + \mu_{i,j,t}\right)}{\sum_{k \in \mathcal{J}_t} \exp\left(\delta_{k,t} + \mu_{i,k,t}\right)}.$$

This individual choice probability is taken as the individual predicted quantity share $\pi_{i,j,t}$. Given distributions of the demographics $F(\mathbf{z}_i)$ and of shocks on the preference parameter $F(\mathbf{v}_i)$, the predicted quantity market share of good j is aggregated as:

$$\begin{aligned}\pi_{j,t} &= \int\int \pi_{i,j,t} dF(\mathbf{z}_i)\, dF(\mathbf{v}_i) \\ &= \int\int \frac{\exp\left(\delta_{j,t} + \mu_{i,j,t}\right)}{\sum_{k\in\mathcal{J}_t} \exp\left(\delta_{k,t} + \mu_{i,k,t}\right)} dF(\mathbf{z}_i)\, dF(\mathbf{v}_i)\,.\end{aligned} \tag{2.21}$$

If $\alpha_i = \alpha$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, which implies that the preference is homogeneous across individuals, the model reduces to the homogeneous logit demand model.

By definition, the predicted market share $\pi_{j,t}$ is:

$$\pi_{j,t} := \frac{q_{j,t}}{\sum_{k\in\mathcal{J}_t} q_{k,t}}.$$

This system of predicted quantity market shares for $\#(\mathcal{J}_t)$ alternatives in a market t provides only $\#(\mathcal{J}_t) - 1$ restrictions on the system of quantity demand $\mathbf{q}_t$. An additional restriction is required, and Berry (1994); Berry, Levinsohn, and Pakes (1995) impose a fixed market size assumption to derive the quantity demand; denominator $\sum_{k\in\mathcal{J}_t} q_{k,t}$ is regarded as fixed at some level M.

2.B Implementation Details

We use the Gaussian kernel during first- and second-stage estimation. For tractability, higher-order kernels are not used. The bandwidth, h_n, for the Klein-Spady estimator is $h_n = \text{std}\left(\mathbf{w}_j' \hat{\boldsymbol{\delta}}_{\text{Probit}}\right) C_1 n^{-\frac{1}{7}}$. The rate $n^{-\frac{1}{7}}$ follows the original suggestion from Klein and Spady (1993). Bandwidth for the second-stage Powell (2001) estimator is $h_n = \text{std}\left(\mathbf{w}_j' \hat{\boldsymbol{\delta}}_{\text{KS}}\right) C_2 n^{-\frac{1}{7}}$, where $\hat{\boldsymbol{\delta}}_{\text{KS}}$ is the Klein-Spady estimator from the first stage. We use the tuning parameter $C_1 = C_2 = 1$ in Section 2.6 and $C_1 = C_2 = 0.5$ in Section 2.7.

We tried 100 randomly generated starting values in the first stage Klein-Spady estimation to guard against the argument that the optimization routine stopped at the local minima. The randomly generated initial values follow the distribution:

$$\mathcal{N}\left(\hat{\boldsymbol{\delta}}_{\text{Probit}}, \frac{1}{5}\text{diag}\left(\sqrt{\left|\hat{\boldsymbol{\delta}}_{\text{Probit}}\right|}\right)\right).$$

We also tried several tuning parameters to assess robustness. First-stage parameter estimates varied considerably regarding the choice of tuning parameters and bandwidth, whereas second-stage parameter estimates, which are our primary interest, were robust to the choice of bandwidth and initial values for nonlinear optimization.

For the simple Heckman estimator with endogeneity, we computed standard errors that account for the fact that the inverse Mills ratio is a generated regressor. Details on the covariance formula of the estimator can be found in Newey and McFadden (1994). Because finding the inverse Mills ratio is fast, one can also consider the bootstrapped standard errors for the Heckman estimator. Finally, we tried both IPOPT and KNITRO, which are state-of-the-art, derivatives-based, nonlinear optimizers for nonlinear optimization. Results were robust to choice of optimizer.

REFERENCES

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007): *Econometric tools for analyzing market outcomes*vol. 6A of *Handbook of Econometrics*, chap. 63, pp. 4171–4276. Elsevier.

Aguirregabiria, V., and A. Nevo (2013): *Recent developments in empirical IO*vol. 3 of *Advances in Economics and Econometrics*, chap. 2, pp. 53–122. Cambridge University Press.

Ahn, H., and J. L. Powell (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29.

Allenby, G. M., and J. L. Ginter (1995): "The effects of in-store displays and feature advertising on consideration sets," *International Journal of Research in Marketing*, 12, 67–80.

Allenby, G. M., T. S. Shively, S. Yang, and M. J. Garratt (2004): "A choice model for packaged goods: Dealing with discrete quantities and quantity discounts," *Marketing Science*, 23(1), 95–108.

Anderson, J. E. (1979): "A theoretical foundation for the gravity equation," *American Economic Review*, 69(1), 106–116.

Anderson, S. P., A. de Palma, and J.-F. Thisse (1987): "The CES is a discrete choice model?," *Economics Letters*, 24, 139–140.

——— (1988): "A representative consumer theory of the logit model," *International Economic Review*, 29(3), 461–466.

Anderson, S. P., A. de Palma, and J.-F. Thisse (1989): "Demand for differentiated products, discrete choice models, and the characteristics approach," *Review of Economic Studies*, 56(1), 21–35.

Anderson, S. P., A. de Palma, and J.-F. Thisse (1992): *Discrete choice theory of product differentiation*. The MIT Press, Cambridge, Massachusetts.

Bajari, P., and C. L. Benkard (2005): "Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach," *Journal of Political Economy*, 113(6), 1239–1276.

Ben-Akiva, M., and B. Boccara (1995): "Discrete choice models with latent choice sets," *International Journal of Research in Marketing*, 12, 9–24.

Berry, S. (1994): "Estimating discrete-choice models of product differentiation," *RAND Journal of Economics*, 25(2), 242–262.

Berry, S., A. Gandhi, and P. Haile (2013): "Connected substitutes and invertibility of demand," *Econometrica*, 81(5), 2087–2111.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica*, 63(4), 841–890.

Berry, S., and A. Pakes (2007): "The pure characteristics demand model," *International Economic Review*, 48(4), 1193–1225.

Berry, S. T., and P. A. Haile (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82(5), 1749–1797.

Blundell, R., and J. L. Powell (2007): "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*, 141, 65–83.

Blundell, R. W., and J. L. Powell (2003): *Endogeneity in nonparametric and semiparametric regression models*vol. 2 of *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, chap. 8, pp. 312–357. Cambridge University Press, Cambridge.

——— (2004): "Endogeneity in semiparametric binary response models," *Review of Economic Studies*, 71, 655–679.

Caplin, A., M. Dean, and J. Leahy (2016): "Rational inattention, optimal consideration sets and stochastic choice," *Working Paper*.

Caplin, A., J. Leahy, and F. Matějka (2016): "Rational inattention and inference from market share data," *Working Paper*.

Carlson, J. A., and R. P. McAfee (1983): "Discrete equilibrium price dispersion," *Journal of Political Economy*, 91(3), 480–493.

Chiang, J. (1991): "A simultaneous approach to the whether, what and how much to buy questions," *Marketing Science*, 10(4), 297–315.

Chiang, J., S. Chib, and C. Narasimhan (1999): "Markov chain Monte Carlo and models of consideration set and parameter heterogeneity," *Journal of Econometrics*, 89, 223–248.

Chintagunta, P. K. (1993): "Investigating purchase incidence, brand choice and purchase quantity decisions of households," *Marketing Scinece*, 12(2), 184–208.

Chintagunta, P. K., J.-P. Dubé, and V. Singh (2003): "Balancing profitability and customer welfare in a supermarket chain," *Quantitative Marketing and Economics*, 1(1), 111–147.

Clerides, S., and P. Courty (Forthcoming): "Sales, quantity surcharge, and consumer inattention," *Review of Economics and Statistics*.

Davis, P. (2006): "Spatial competition in retail markets: Movie theaters," *RAND Journal of Economics*, 37(4), 964–982.

de Los Santos, B., A. Hortaçsu, and M. R. Wildenbeest (2012): "Testing models of consumer search using data on web browsing and purchasing behavior," *American Economic Review*, 102(6), 2955–2980.

Diamond, P. A. (1971): "A model of price adjustment," *Journal of Economic Theory*, 3, 156–168.

Dixit, A. K., and J. E. Stiglitz (1977): "Monopolistic competition and optimum product diversity," *American Economic Review*, 67(3), 297–308.

Dube, J.-P., J. T. Fox, and C.-L. Su (2012): "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation," *Econometrica*, 80(5), 2231–2267.

Dubin, J. A., and D. McFadden (1984): "An econometric analysis of residential electric appliance holdings and consumption," *Econometrica*, 52(2), 345–362.

Eaton, J., and S. Kortum (2002): "Technologu, geography, and trade," *Econometrica*, 70(5), 1741–1779.

Einav, L., D. Knoepfle, J. Levin, and N. Sundaresan (2014): "Sales taxes and internet commerce," *American Economic Review*, 104(1), 1–26.

Ellison, G., and S. F. Ellison (2009): "Search, obfuscation, and price elasticities on the internet," *Econometrica*, 77(2), 427–452.

Feenstra, R. C. (1994): "New product varieties and the measurement of international prices," *American Economic Review*, 84(1), 157–177.

Feenstra, R. C., and J. A. Levinsohn (1995): "Estimating markups and market conduct with multidimensional product attributes," *Review of Economic Studies*, 62(1), 19–52.

Fosgerau, M., E. Melo, A. d Palma, and M. Shum (2017): "Discrete choice and rational inattention: a general equivalence result," *Working Paper*.

Gandhi, A., Z. Lu, and X. Shi (2013): "Estimating demand for differentiated products with error in market shares," Working Paper.

Gentry, M. (2016): "Displays, sales, and in-store search in retail markets," *Working Paper*.

Gilbride, T. J., and G. M. Allenby (2004): "A choice model with conjunctive, disjunctive, and compensatory screening rules," *Marketing Science*, 23(3), 391–406.

Goeree, M. S. (2008): "Limited information and advertising in the U.S. personal computer industry," *Econometrica*, 76(5).

Goettler, R. L., and R. Shachar (2001): "Spatial competition in the network television industry," *RAND Journal of Economics*, 32(4), 624–656.

Gouriéroux, C., and A. Monfort (1990): "Simulation based inference in models with heterogeneity," *Annales d'Économie et d Statistique*, 20/21, 69–107.

Hanemann, W. M. (1984): "Discrete/continuous models of consumer demand," *Econometrica*, 52(3), 541–561.

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–62.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating trade flows: Trading partners and trading volumes," *Quarterly Journal of Eonomics*, 123(2), 441–487.

Honka, E. (2014): "Quantifying search and switching costs in the US auto insurance industry," *RAND Journal of Economics*, 45(4), 847–884.

Honka, E., and P. K. Chintagunta (2017): "Simultaneous or sequential? Search strategies in the U.S. auto insurance industry," *Marketing Science*, 36(1), 21–42.

Honka, E., A. Hortaçsu, and M. A. Vitorino (Forthcoming): "Advertising, consumer awareness, and choice: Evidence from the U.S. banking industry," *RAND Journal of Economics*.

Honoré, B. E., E. Kyriazidou, and C. Udry (1997): "Estimation of Type 3 Tobit models using symmetric trimming and pairwise comparisons," *Journal of Econometrics*, 76, 107–128.

Horowitz, J. L., and J. L. Louviere (1995): "What is the role of consideration sets in choice modeling?," *International Journal of Research in Marketing*, 12, 39–54.

Hortaçsu, A., and J. Joo (2017): "Semiparametric estimation of a CES demand system with observed and unobserved product characteristics," *Working Paper*.

Hortaçsu, A., and C. Syverson (2004): "Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds," *Quarterly Journal of Economics*, 119(2), 404–456.

Houde, J.-F. (2012): "Spatial differentiation and vertical mergers in retail markets for gasoline," *American Economic Review*, 102(5), 2147–2182.

Jedidi, K., R. Kohli, and W. S. DeSarbo (1996): "Consideration sets in conjoint analysis," *Journal of Marketing Research*, 33(3), 364–372.

Joseph, K., R. Subramaniam, and V. Patil (2013): "The impact of consumption hassle on pricing schedules," *Managerial and Decision Economics*, 34, 1–14.

Khan, R. J., and D. C. Jain (2005): "An empirical analysis of price discrimination mechanisms and retailer profitability," *Journal of Marketing Research*, 42(4), 516–524.

Kim, H. (2016): "Quantity surcharge and consumer heterogeneity," *Working Paper*.

Klein, R. W., and R. H. Spady (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, 61(2), 387–421.

Krugman, P. (1980): "Scale economies, product differentiation, and the pattern of trade," *American Economic Review*, 70(5), 950–959.

Lancaster, K. J. (1966): "A new approach to consumer theory," *Journal of Political Economy*, 74, 132–157.

Lee, L.-F. (1992): "On efficienciy of methods of simulated moments and maximum simulated likelihood estimation of discrete response models," *Econometric Theory*, 8(4), 518–552.

Manning, K. C., D. E. Sprott, and A. D. Miyazaki (1998): "Consumer response to quantity surcharges: Implications for retail price setters," *Journal of Retailing*, 74(3), 373–399.

Matějka, F., and A. McKay (2012): "Simple market equilibria with rationally inattentive consumers," *American Economic Review: Papers and Proceedings*, 102(3), 24–29.

——— (2015): "Rational inattention to discrete choices: A new foundation for the multinomial logit model," *American Economic Review*, 105(1), 272–298.

McFadden, D. (1974): *Conditional logit analysis of qualitative choice behavior*chap. 4, pp. 105–142. Academic Press.

——— (1978): *Modelling the choice of residential location*vol. 3 of *Spatial interaction theory and planning models*, chap. 3, pp. 75–96. North-Holland Publishing Company.

——— (1981): *Econometric models of probabilistic choice*chap. 5, pp. 198–272, Structural analysis of discrete data with econometric applications. The MIT Press.

——— (1989): "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 57(5), 995–1026.

McFadden, D., and K. Train (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447–470.

Mehta, N., S. Rajiv, and K. Srinivasan (2003): "Price uncertainty and consumer search: A structural model of consideration set formation," *Marketing Science*, 22(1), 58–84.

Metha, N., S. Rajiv, and K. Srinivasan (2003): "Price uncertainty and consumer search: A structural model of consideration set formulation," *Marketing Science*, 22(1), 58–84.

Mitra, A. (1995): "Advertising and the stability of consideration sets over multiple purchase occasions," *International Journal of Research in Marketing*, 12, 81–94.

Moraga-González, J. L., Z. Sándor, and M. R. Wildenbeest (2015): "Consumer search and prices in the automobile market," *Working Paper*.

Nair, H., J.-P. Dubé, and P. Chintagunta (2005): "Accounting for primary and secondary demand effects with aggregate data," *Marketing Science*, 24(3), 444–460.

Nason, R. W., and A. J. Della Bitta (1983): "The incidence and consumer perceptions of quantity surcharges," *Journal of Retailing*, 59(2), 40–54.

Nevo, A. (2000): "A practitioner's guide to estimation of random-coefficients logit models of demand," *Journal of Economics and Management Strategy*, 9(4), 513–548.

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

——— (2011): "Empirical models of consumer behavior," *Annual Review of Economics*, 3, 51–75.

Newey, W. K. (2009): "Two-step series estimation of sample selection models," *Econometrics Journal*, 12, S217–S229.

Newey, W. K., and D. McFadden (1994): *Large sample estimation and hypothesis testing*vol. 4 of *Handbook of Econometrics*, chap. 36, pp. 2111–2245. Elsevier, Amsterdam.

Palazzolo, M., and F. Feinberg (2015): "Modeling consideration set substitution," *Working Paper*.

Pires, T. (2015): "Measuring the effects of search costs on equilibrium prices and profits," *Working Paper*.

Powell, J. L. (1984): "Least absolute deviations estimation for the censored regression model," *Journal of Econometrics*, 25, 303–325.

——— (1986): "Censored regression quantiles," *Journal of Econometrics*, 32, 143–155.

——— (2001): *Semiparametric estimation of censored selection models*chap. 6, pp. 165–196. Cambridge University Press, Cambridge.

Reinganum, J. F. (1979): "A simple model of equilibrium price dispersion," *Journal of Political Economy*, 87(4), 851–858.

Reiss, P. C., and F. A. Wolak (2007): *Structural econometric modeling: Rationales and examples from industrial organization*vol. 6A of *Handbook of Econometrics*, chap. 64, pp. 4277–4415. Elsevier.

Roberts, J. H., and J. M. Lattin (1991): “Development and testing of a model of consideration set composition,” *Journal of Marketing Research*, 28(4), 429–440.

Rothe, C. (2009): “Semiparametric estimation of binary response models with endogenous regressors,” *Journal of Econometrics*, 153, 51–64.

Seiler, S. (2013): “The impact of search costs on consumer behavior: A dynamic approach,” *Quantitative Marketing and Economics*, 11, 155–203.

Shannon, C. E. (1948): “A mathematical theory of communication,” *The Bell System Technical Journal*, 27, 379–423, 623–656.

Shreay, S., H. H. Chouinard, and J. J. McCluskey (2016): “Product differentiation by package size,” *Agribusiness*, 32(1), 3–15.

Siddarth, S., R. E. Bucklin, and D. G. Morrison (1995): “Making the cut: Modeling and analyzing choice set restriction in scanner panel data,” *Journal of Marketing Research*, 32(3), 255–266.

Sims, C. A. (2003): “Implications of rational inattention,” *Journal of Monetary Economics*, 50(3), 665–690.

——— (2006): “Rational inattention: Beyond the linear-quadratic case,” *American Economic Review*, 96(2), 158–63.

Spence, M. (1976): “Product selection, fixed costs, and monopolistic competition,” *Review of Economic Studies*, 43, 217–235.

Sprott, D. E., K. C. Manning, and A. D. Miyazaki (2003): “Grocery price setting and quantity surcharges,” *Journal of Marketing*, 67(3), 34–46.

Stahl II, D. O. (1989): “Oligopolistic pricing with sequential consumer search,” *American Economic Review*, 79(4), 700–712.

Stigler, J. J. (1961): “The economics of information,” *Journal of Political Economy*, 69(3), 213–225.

Terui, N., M. Ban, and G. M. Allenby (2011): “The effect of media advertising on brand consideration and choice,” *Marketing Science*, 30(1), 74–91.

Thomadsen, R. (2005): “The effect of ownership structure on prices in geographically differentiated industries,” *RAND Journal of Economics*, 36(4), 908–929.

van Nierop, E., B. Bronnenberg, R. Paap, M. Wedel, and P. H. Franses (2010): "Retrieving unobserved consideration sets from household panel data," *Journal of Marketing Research*, 47(1), 63–74.

Waugh, M. E. (2010): "International trade and income differences," *American Economic Review*, 100(5), 2093–2124.

Widrick, S. M. (1979): "Measurement of incidents of quantity surcharge among selected grocery products," *Journal of Consumer Affairs*, 13(1), 99–107.

Wilson, R. B. (1997): *Nonlinear Pricing*. Oxford University Press.

Zhang, J. (2006): "An integrated choice model incorporating alternative mechanisms for consumers' reactions to in-store display and feature advertising," *Marketing Science*, 25(3), 278–290.