Estimation of Store Choice Model with Endogenous Shopping Bundles

Hyunchul Kim*
Sungkyunkwan University

April 2014

Abstract

This paper examines the effects of measurement errors in household shopping plans and price beliefs on the estimates of price elasticities in retail outlet choice. Previous studies parsimoniously use realized purchases as a proxy for unobserved shopping lists and assume homogenous price expectation across all households and over the entire sample period. This measurement error in forecasting expected basket costs introduces attenuation bias in store-level price elasticities toward zero. I show that the bias is reduced with my improved measures of expected basket costs by estimating shopping lists of households and by constructing household-, time-, store-, and goods-specific price expectations.

*This paper is based on various chapters of my doctoral dissertation. I am grateful to my advisor, Amil Petrin, for his continuous advice and support at the University of Minnesota. I would like to thank Pat Bajari, Tom Holmes, Kyoo-il Kim, Om Narasimhan, Elena Pastorino, Maria Ana Vitorino, and Joel Waldfogel for their many useful comments and suggestions. I am also indebted to the seminar participants at the University of Minnesota, the Pennsylvania State University, the University of Arkansas, Sungkyunkwan University, and Yonsei University, as well as conference participants at the 2014 IIOC. All remaining errors are mine. Correspondence: Hyunchul Kim, Department of Economics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, 110-745, Korea, Tel: +82 2 760 0147, E-mail: hchkim@skku.edu

1 Introduction

In this paper I estimate a consumer level demand model for store choice in the supermarket industry. The store choice of grocery shoppers has two distinctive features compared to other consumer choice problems. First, store choice decisions involve bundle purchase behavior. Consumers decide on which stores to visit depending on their shopping plans, which are characterized by the goods and quantities they intend to buy at a store. Second, consumers face price uncertainty before they actually visit a store. Shoppers may acquire certain information on prices from out-of-store advertising such as newspaper inserts or weekly circulars on sales, but most of the shelf prices are unknown a priori.[1] Shoppers, therefore, rely on price expectation for their store choice.

Household level shopping lists and price beliefs are typically not observed by researchers. Previous store choice studies used realized purchases as a proxy for the unobserved shopping lists and assumed homogeneous price expectations across households and over the entire sample period. This parsimonious approach introduces measurement errors in the expected basket costs of each store trip and thus creates attenuation biases in the elasticities of expected basket costs. My main contribution in this study is to consistently estimate a household level store choice model using improved measures of the expected basket costs. To do this, I model household level shopping baskets and construct household-, store-, and time-specific price expectations. To my knowledge no prior work examines the potential biases caused by unobserved household shopping lists and price beliefs in the estimation of store choice.

To improve the measure of unobserved shopping lists, I proceed in the following two steps. In the first stage, I specify how households determine which goods to buy and how much of them to buy before they visit a store.[2] I use a continuous-time duration model to estimate the purchase

[1] Price information from advertising is limited in the context of store choice decisions. Bodapati and Srinivasan (2006) document that only a small fraction of customers are influenced by price advertising in their store choices. Moreover, the data show that the proportion of products with advertising is less than 10 percent of all the products sold by stores in any given week. The revenue-weighted proportion is 22 percent, but the increased revenue of advertised products is not solely attributed to price advertising because it also reflects instantaneous price effects at stores.

[2] Instead of planning on which goods to purchase, shopping lists may be assumed to be defined at a narrower level such as a brand or a brand-size combination. Shopping lists at these levels may be more tenable if strong brand loyalty prevails in grocery shopping because in that case shoppers would rather plan on buying specific products. However, the data show that brand loyalty is fairly weak for many goods. In addition, assuming a

probability for each good, which is specified as a function of the relevant states of each household such as consumption and inventory levels and shopping behavior for each good. The quantity choice in planning a shopping list is estimated by fitting the realized quantity for each of the basket goods as a function of expected prices and other relevant variables. In the second stage, I estimate a household level store choice model conditioned on the shopping baskets predicted in the first stage. Expected basket costs based on the expected shopping baskets of the first stage and other store attributes (e.g., transportation costs, service quality, floor space, and parking space) enter a utility function for each store visit and households choose a store that gives the highest utility from shopping at the store for the expected shopping basket.

The consideration of potential discrepancies between observed and planned shopping baskets in store choice studies has been introduced by Bell, Ho, and Tang (1998). In their seminal study of household level store choice, they assume that households specify shopping lists at the product level as opposed to at the goods level (e.g., a shopping list might consist of 144 oz Pepsi cans and 12 oz Kellogg's Special K, instead of soft-drinks and cereal goods). They use a discrete choice setting to estimate the probability that each product actually purchased was included in the *ex ante* shopping list as a function of consumption, inventory level, and expectation errors in prices.[3] In the current study, I define the basket composition model at the goods level, which allows for heterogeneity in inter-purchase time for each good as well as consumption and inventory levels and other relevant state variables on goods purchase decisions.

Given that the idiosyncrasy in purchase patterns across households and across goods is salient in grocery shopping, using a duration model setup that takes into account heterogeneity in inter-purchase time seems more appropriate than a standard discrete choice model for estimating goods purchase probabilities. Moreover, the basket composition model in this study allows for quantity choices for the basket goods, whereas Bell, Ho, and Tang assume that expected quantities are identical to realized quantities.

product level shopping list requires a consideration of brand choice problems for each good in specifying basket composition and this adds a substantial computational burden.

[3]In their purchase incidence estimation, consumption and inventory levels (and the coefficient variants across different customer segments) are the only source of heterogeneity since expected price in their model is a simple average of store prices over the entire sample period and thus does not capture the differences in price beliefs across households.

Considering shopping lists not only mitigates the measurement error problem, but it also allows me to circumvent the complication of estimating store choices in a discrete choice setup, which arises from dealing with the size and complexity of choice sets (see Katz (2007) and Pakes (2010)). The choice set in a store choice problem not merely consists of alternative stores, but it also contains alternative shopping baskets (what and how much to buy) depending on which store is chosen. Therefore, the choice set for each store trip should include all the combinations of alternative stores and the shopping baskets corresponding to each store. This makes the store choice estimation in a discrete choice setup almost intractable. Specifying the expected shopping baskets in the first stage allows me to sidestep the need to deal with such overwhelming size and complexity of the choice sets.

To deal with unobserved price knowledge, I develop household level price expectations for each good as a function of most recent trips to the store. It is well established in the marketing literature that households glean price information from external advertising or by retrieving memories of store prices. I assume that price expectation builds on store prices experienced from past shopping trips and purchases. Price expectation constructed this way not only captures short-term fluctuations in store prices but also reflects heterogeneity in price beliefs across households depending on their shopping behavior.[4]

To preview the results, I find that misspecifying household price expectation and shopping lists leads to substantial biases in store-level own price elasticities toward zero. Particularly, measurement errors in expected basket costs are most attributable to ignoring heterogeneity in price expectation across households. To determine the shopping list at the product level as in Bell, Ho, and Tang (1998) somewhat reduces such biases even without heterogeneity in price beliefs, but the own price elasticities on average are almost seven times smaller in magnitude than those based on my approach.

This paper is, to the best of my knowledge, the first attempt to evaluate the effects of heterogeneity in price expectation on store choice. There is extensive marketing literature on price perception and reference prices in the context of brand choice. Most of this literature,

[4] I do not allow for the differences among customers in information acquisition or memory recall processes which may depend on heterogeneity in price sensitivities.

starting from Monroe (1973) and Winer (1986), suggests that price information gathered from various sources, combined with previously observed or paid prices, is integrated into expected or reference prices. More recently, Erdem, Imai, and Keane (2003) study the role of expectation of future prices in brand and quantity choices of forward-looking consumers. On the other hand, relatively little has been studied on the effects of price expectation in store choice decisions (see Mazumdar, Raj, and Sinha, 2005).

This paper also contributes to the literature on store choice problems in the supermarket industry. Smith (2004) examines merger effects with the estimated substitution patterns in a discrete and continuous choice setting and Beresteanu, Ellickson, and Misra (2006) estimate store choices using market level data and study the welfare effects of entry and competition in a dynamic framework. More recently, Katz (2007) estimates a store choice model using moment inequality methods with household level data. The main difference between his work and mine is in dealing with the bundle-choice behavior in store choice. The moment inequality approach circumvents the need for dealing with the aforementioned complexity of choice sets by subtracting out all basket-related terms in a utility function except for those of interest. In the current study, I address the same problem by specifying the shopping basket composition in the first stage of the estimation.

The paper most similar to mine is Bell, Ho, and Tang (1998). They also use a discrete choice setup to estimate store choices allowing for the shopping basket composition. However, in their model a shopping list is defined at the product level (as opposed to the goods level in the present study) assuming strong brand loyalty. Their approach leaves out possible substitution among brands and sizes inside stores when actual prices are observed. Bell, Ho, and Tang (1998) also consider price uncertainty and assume that customers have information on the price distribution, particularly the average of store prices, as documented by Alba, Broniarczyk, Shimp, and Urbany (1994), Lal and Rao (1997), and Ho, Tang, and Bell (1998). Their approach to specifying price expectation does not capture heterogeneity in price beliefs among households.

The rest of the paper is organized as follows. Section 2 outlines the demand model. In Section 3, I show how I model the expected basket costs. Section 4 describes the data. The estimation results are presented in Section 5. Section 6 concludes.

2 Demand Model

In this section I introduce a demand specification of household level store choice. Following the characteristics approach of Lancaster (1971), the demand system of store choice posits that demand patterns are determined by preferences for store attributes. The utility of household i that plans to shop for basket b from visiting store j at time t is decomposed into basket-related and non-basket-related components in an additive form as follows

$$U_{ijbt} = V_1(Exp_{ijbt}, W_{ijbt}, Z_{it}) + V_2(X_{ijt}, Z_{it}) + \varepsilon_{ijt}, \tag{1}$$

where Exp_{ijbt} is expected spending on shopping bundle b, W_{ijbt} denotes other basket-specific store attributes, and Z_{it} is a vector of demographic characteristics of the household. X_{ijt} denotes the store attributes that are not related to the goods included in the shopping basket. ε_{ijt} is the random part of the utility that captures the idiosyncratic taste of household i for store j.

The specification of $V_1(\cdot)$ and $V_2(\cdot)$ in the utility equation (1) is

$$U_{ijbt} = \beta_i Exp_{ijbt} + \gamma_i Dist_{ij} + \sum_l \alpha_{jl} Z_{itl} + \varepsilon_{ijt}, \tag{2}$$

where $Dist_{ij}$ is the shortest driving distance between store j and the residence of household i.[5] With household level information on store choice decisions along with demographic characteristics, the demand model parameters are allowed to capture heterogeneity in consumer tastes for the store attributes. Specifically, I allow store attributes to interact with demographic variables as follows

$$\begin{aligned} \beta_i &= \beta_0 + \sum_l \beta_l Z_{itl}, \\ \gamma_i &= \gamma_0 + \sum_l \gamma_l Z_{itl}. \end{aligned} \tag{3}$$

The coefficients differ across consumers with different demographic characteristics. For example, if Z_{it} includes household income, the price coefficient for a household of the base income group is β_0 and that for a household in income group g is $\beta_0 + \beta_g$.

[5] In the utility function (2) the basket-related part only consists of expected basket cost. To check the robustness of this specification, I discuss an alternative specification with other basket-related variables. Particularly, I add preferences for a variety of products of each basket good. See Appendix A.

I also include store-specific fixed effects, which also differ by demographic characteristics. Specifically, the coefficient α_{jl} accounts for preference for the unobserved characteristics of store j that is common across the households that share the l-th demographic characteristic. Using the store-specific fixed effects to allow for unobserved store attributes is in line with Berry, Levinsohn, and Pakes (2004) and Goolsbee and Petrin (2004).In their models, the price coefficient for the base demographic group is subsumed in the fixed effect term and the estimation requires a two-step procedure to identify price parameters. In contrast, in the current study, the variation in the expected basket costs across households and shopping baskets allows the model to identify price parameters for each group separately from the fixed effects.

The random error ε_{ijt} in equation (2) then accounts for the unobserved store attributes not captured by the store fixed effects and for the unobserved consumer heterogeneity that the distribution of consumer tastes characterized by the observed demographics may not capture. The probability that household i with planned shopping list b chooses store j is

$$s_{ijbt} = \int_{A_{ijbt}} dF(\varepsilon_{it}),$$

where $A_{ijbt} = \left\{\varepsilon_{it} : U_{ijbt} > U_{ij'bt},\ \forall j' \neq j\right\}$ and $F(\cdot)$ is the distribution function of ε_{ijt}. The error term ε_{ijt} is assumed to be an $i.i.d.$ error with a Type I extreme value distribution. This assumption reduces the model to a logit model that allows a closed form solution for the integration over ε_{ijt}.

The logit distributional assumption greatly simplifies the estimation in discrete choice model settings by making the complicated integrals tractable when evaluating the likelihood function. However, it is well-known that the undesirable properties of the demand model arising from this distribution assumption yield unreasonable substitution patterns.[6] The standard approach to this problem is to use the nested multinomial logit or the random coefficients model. In this study, however, the reliance of the model on the error term when the model fits the observed store choices would not be that strong by virtue of using the richness of variation in store attributes

[6]For a discussion of the assumption and its implications on substitution patterns, see Hausman and Wise (1978), Berry, Levinsohn, and Pakes (1995), Petrin (2002), Bajari and Benkard (2003), and Berry, Levinsohn, and Pakes (2004).

across households in detailed micro data (Note, in particular, that expected basket costs are household–time–store specific.). Moreover, I avoid ruling out important substitution patterns in store choice by weakening the restriction on the distribution of the coefficients. Specifically, I allow consumer tastes for store attributes and store fixed effects to vary with demographic characteristics.

3 Model of Expected Bundle Cost

Expected basket cost entering into the utility function in the demand model is determined by an *ex ante* (or planned) shopping list and household price expectation for the goods included in the shopping list. Since neither of these components is observable in the data, the estimation of store choice problem should be preceded by specifying *ex ante* shopping lists and price expectations. An *ex ante* shopping list is characterized by the set of goods and the quantities of the goods that a shopper plans to purchase before visiting a store. The choice of these *ex ante* shopping lists is specified in two stages, as follows. In the first stage, the set of goods included in a shopping list is determined as a function of inter-purchase spells and other relevant states such as inventory and consumption level for each good. In the second stage, given this set of goods in the list, households make a decision on the purchase quantity for each good as a function of expected prices and other state variables.

The choice of goods purchases in the first stage is specified by a continuous-time duration model accounting for heterogeneity across households in goods purchase behavior. The composition of basket goods in a shopping list is characterized by the probability of each good's being included in the shopping list. Conditional on these estimated probabilities, the quantity of each good that a customer plans to buy is predicted based on the regression of realized quantities on actual prices, brand fixed effects, and other relevant variables. In specifying both of the basket components (goods and quantity), the choices are assumed to be made independently among different goods.[7]

[7] This assumption is straightforward for the goods that are irrelevant to each other such as milk and laundry detergent. However, there may exist a set of goods for which their purchase and consumption are interdependent, such as hotdog buns and ketchup goods. In my data there is only one case of such a dependency in purchase decisions (toothbrush and toothpaste), and I ignore its potential problems in this study.

For unobserved price expectation, the model posits that households develop price knowledge based on their experiences from past shopping visits. Specifically, I assume that the expected price is a weighted average of the past prices that a household has observed or paid during previous store trips. Price expectation defined in such a way depends on good purchases and shopping patterns of each household. The richness of household level information on shopping trips and purchases allows me to construct household-, time-, store-, and good-specific price expectation.

Given *ex ante* shopping lists and price expectation, households choose which store to visit based on the store attributes, both those built on the expected shopping basket and those that are not related to the shopping basket. Store choice decision is specified at the household level in a discrete choice setting as described in the previous section. In the store choice model, each household chooses the store that maximizes the utility from shopping for the expected shopping basket. In the following subsections I provide further details on the three components that govern expected basket costs.

3.1 Price Expectation

First, households form price expectations based on the past prices from their most recent trips to stores, which accounts for heterogeneity in price knowledge and its dependence on each household's shopping behavior. Second, given that each store sells a vast array of different products and customers do not typically remember the prices of all these products given time constraints for each shopping trip, I assume that the past prices of products are only kept in memory or retrievable from memory when customers have purchased the goods categories into which those products fall.[8] Last, memories of past prices fade over time, and thus, price expectation for each product is closest to the prices from the most recent trips to the store.

Based on these behavioral assumptions of the formation of price expectation, expected price

[8] If shoppers recollect the prices of all products regardless of whether they were shopping for the good during a store trip, price expectation becomes almost homogeneous among households since most households visit stores as frequently as once per week. For example, if two customers visit the same store almost every week, their price beliefs for each good will be almost identical even though they have purchased different goods for each trip. In this case, heterogeneity in price expectation stems only from the frequency of store visits no matter which goods they have purchased in past shopping trips.

is defined as the weighted average of past prices at the store where the weights exponentially decline with remoteness in time from the current period. Formally, the expected price of product k of good c for household i is written as follows

$$\bar{p}_{ijkt} = \sum_{\tau<t} w(\tau; t, \theta) d^s_{i\tau}(j) d^c_{i\tau}(c) p_{jk\tau}, \tag{4}$$

where $p_{jk\tau}$ is the actual price of product k at store j in time τ, $d^s_{i\tau}(j)$ is the indicator of store visit which equals 1 if the household visits store j and 0 otherwise, and $d^c_{i\tau}(c)$ is the good purchase indicator which equals 1 if the household buys any product of good c at τ and 0 otherwise. The weight function is specified as

$$w(\tau; t, \theta) = \frac{\exp(-\theta(t-\tau))}{\sum_{\tau'<t} exp(-\theta(t-\tau'))}$$

where θ is the coefficient of memory decay. A high (positive) value of θ puts greater weight on more recent prices, and the coefficient of zero equally weights the past prices (thus, price expectation becomes a simple average of the past prices). Price expectation specified in this manner not only captures the short-term variation of store prices by placing larger weights on more recent prices, it also accounts for heterogeneity in shopping behavior since it is based on the goods purchases and the store choices of each household in the past.[9]

This approach to dealing with unobserved price beliefs is a generalization of the price expectation used by Bell, Ho, and Tang (1998) and Ho, Tang, and Bell (1998). They assume that consumers have common knowledge on the price distribution for each store up to the first and second moments. Then, they define the expected price of each product as the average of the store prices over the data period. Their specification of price expectation implies that shoppers visit each store frequently enough and acquire price information from various sources to the extent that the price process is known up to the first moment of the price distribution. In such a way, their price expectation reflects a long-term variation in prices including future prices as

[9]Since past prices are embedded in price expectation only when the good is bought, customers have no price knowledge if they have never purchased the good at the store (that is, $d^c_{i\tau}(c) = 0$ for any $\tau < t$). In this case, I set the initial price expectation based on the past prices assuming they still observed the prices during their trips (albeit did not purchase the goods).

well as past prices. In my setting of price expectation given in equation (4), their expected price is written as

$$\bar{p}_{ijkt} = \sum_{\tau<\infty} w(\tau; t, \theta = 0) p_{jk\tau}.$$

That is, $d^s_{i\tau}(j) = d^c_{i\tau}(c) = 1$ for any j and c where the memory decay coefficient in the weight function is set as zero. Note that in this case price expectation is identical among all households and thus leaves out heterogeneity in price beliefs.[10]

3.2 Goods Purchase Probability

The choice of goods purchases consists of a set of sequential purchase incidences in which consumers visit stores to buy some products of a specific good at randomly selected times. I characterize the composition of basket goods as the probability of good purchase incidence (so called "hazard" in the duration model literature) as a function of the length of the elapsed time since the most recent good purchase. The hazard rate is also determined by other state variables that directly or indirectly affect purchase decisions at a given point of time. In what follows, I briefly introduce the duration model of goods purchases, adopting the standard notations in the duration model literature (e.g., Cox (1972), Heckman and Singer (1986), Lancaster (1990), and Kalbfleisch and Prentice (2002), among others).

Let T be a random variable of the duration of a state, or the elapsed time since the most recent purchase. Conditional hazard $h(\tau)$ is the instantaneous probability of leaving the no-purchase state (i.e., the occurrence of a new purchase incidence) after time τ. Given x_{it}, a vector of observed state variables for household i at time t, the hazard rate can be written as

$$h_i(\tau|x_{it}, \mu) = \lim_{\triangle \to 0} \frac{Pr[\tau \leq T < \tau + \triangle | T \geq \tau, x_{it}, \mu]}{\triangle}, \tag{5}$$

where μ is a vector of parameters. Using the specification proposed by Cox (1972), the condi-

[10] I loosen their assumptions and estimate with the price expectation only assuming $d^c_{i\tau}(c) = 1$ for any c. That is, households memorize prices even when they did not purchase the goods during a store trip. The estimated coefficients for expected basket costs with this price expectation become positive. The estimation results for this case will be provided upon request.

tional hazard (5) is given as

$$h_i(\tau|x_{it}, \mu) = h_{i0}(\tau)\phi(x_{it}, \mu). \tag{6}$$

The conditional hazard is assumed to be proportional to the two components on the right-hand side. The first component $h_{i0}(\tau)$ is baseline hazard as a function of no-purchase spell τ only, and the second component $\phi(x_{it}, \mu)$ is a function of observed state variables x_{it} at time t. The baseline hazard $h_{i0}(\tau)$ accounts for heterogeneity in shopping behavior across households (particularly the frequency of goods purchases). In equation (6), the parameters in the second component of the hazard is estimated by specifying the functional form of $\phi(\cdot)$ whereas the household-specific baseline hazard is not parameterized. Specifically, for the parameterized part, I use $\phi(x_{it}, \mu) = \exp(x'_{it}\mu)$.

For each grocery good, the time-variant state variables x_{it} include information on consumption and inventory levels, purchases of other goods, and individual households' shopping patterns and demographics.[11] Since the consumption and inventory levels are not directly observed in the data, I construct the variables based on the observed purchases (quantity and frequency) of the good for each household considering the specific features of each good such as average shelf lives and approximate storage costs.[12] More details on how I construct the consumption and inventory variables are provided in Appendix B.

Including the variables of consumption level and purchases of other goods in x_{it} allows for exogenous shocks to the need for purchasing the good of interest. For example, a household that recently consumes or purchases a large amount of other goods may be more likely to buy the good of interest soon after a previous purchase. The variables for shopping behavior are constructed based on the observed grocery shopping patterns of each household, and these include the average frequency and dollar spending of recent shopping visits, preference for weekend shopping, the average time interval between goods purchases, and seasonal dummies.

In my model, goods purchase decisions are not affected by the expected prices of the products.

[11] The variables to control for shopping patterns are generated by the household level data, and these are based on overall grocery shopping behavior (not just related to the good of interest).

[12] It may be natural to consider the consumption level as a decision variable in store choice as in Hendel and Nevo (2006). However, since store choices involve a wide variety of goods and the consumption decision is an object that should be understood in a dynamic framework, it is not straightforward to allow consumption to be endogenously determined. For this reason, I assume an exogenous consumption level in a static model.

I assume that the effects of price changes on the need or consumption of each good is negligible or none. For example, people do not increase their average consumption level of carbonated beverage or laundry detergent or purchase them more frequently when they expect low prices. In addition, the effects of expected prices on goods purchases will be only tenuously identified in my case because price expectation is updated after households purchase the goods and thus the expected prices do not change during no-purchase periods.[13]

Specifying a shopping list at the goods level instead of at the product level makes the model flexible enough to allow substitution between the products within each good to take place inside stores. If households have strong loyalty to specific brands or sizes of each good, they will always plan to buy specific products, and thus a product-level shopping list would be appropriate. However, it is well documented in both the economics and the marketing literature that the consumer choice of brand or size is determined to a large extent by in-store promotions. Therefore, it seems reasonable to define a shopping list at the goods level. In Bell, Ho, and Tang (1998), a planned shopping list is specified at the product level and consists of the products that were actually purchased at the store. This approach leaves out the possibility that the purchased products were not contained in the shopping list and did not influence store choice decisions.

3.3 Quantity of Purchased Goods

The third component that determines expected bundle cost is the quantities of the goods that shoppers plan to purchase. Households choose the purchase quantity for each basket good based on their price expectation and other relevant states such as shopping frequency, consumption and inventory level for the good. Since neither the price expectation nor the planned purchase quantity is observed in the data, I infer how much consumers plan to buy based on how they choose the quantities at stores. Then, in predicting the expected quantity, the shelf prices that households observe inside a store are replaced by their price expectation. Specifically, the

[13] If price expectation reflects price information obtained from the sources other than goods purchases such as price advertising, expected prices will evolve during no-purchase periods, and this would affect the timing of the upcoming goods purchases. However, this paper ignores the effects of price advertising on store choices for the reason mentioned above.

expected quantity is estimated by regressing the realized quantity on actual price taking into account brand and household fixed effects, seasonal dummies, and demographic variables.

To deal with the potential issue of endogeneity in prices, actual prices in the regression are instrumented by the average prices of each product in the same supermarket chain stores in nearby cities. As in Nevo (2001), if the stores of the same supermarket chain have a similar cost structure (thus correlated with the actual price) and if the city-specific valuation of the price is independent across different cities, the prices at the same chain stores in other cities can be valid instrument variables.[14]

Bell, Ho, and Tang (1998) also acknowledge that the realized quantity may differ from the expected quantity in a shopping list. But they argue that there is only a slight improvement in their estimation results when considering the discrepancy between realized and planned quantity. There can be two reasons for the limited role of quantity choice in their study. First, if the estimation of quantity choice poorly fits the data, the prediction of planned quantity is not accurate. Second, this may be because expected prices in their model do not allow for heterogeneity in price knowledge as discussed above.

3.4 Expected Bundle Cost

Given the three components of a shopping list as described above, the expected basket cost is defined as the sum of the expected spending on each basket good. For each basket good, the expected spending is the average of the expected costs of the products of the good. Then, the expected bundle cost is the weighted average of the expected spending on each good with the weight of the good purchase probability. Formally, the expected basket cost can be written as

$$Exp_{ijbt} = \sum_{c \in \mathcal{C}(b)} h_i^c(\tau | x_{it}, \mu) \sum_{k \in \mathcal{G}_{ic}} \frac{\bar{p}_{ijkt} E(q_{ijkt} | \bar{p}_{ijkt}, s_{it})}{n(\mathcal{G}_{ic})}, \tag{7}$$

where $\mathcal{C}(b)$ is the set of goods contained in shopping basket b, $h_i^c(\tau | x_{it}, \mu)$ is the hazard rate for good c after a duration of time τ since the most recent good purchase, and $n(\mathcal{G}_{ic})$ is the number

[14] As pointed out by Erdem, Imai, and Keane (2003), endogeneity in prices of frequently purchased consumer goods may be more attributable to the omitted variables such as consumer inventories than to aggregate demand shocks. Since I control for inventory level in my quantity estimation, I use the average prices in nearby cities for dealing with the endogeneity problems that may arise from unobserved demand shocks.

of products in the consideration set $\mathcal{G}_{ic}$. $E(q_{ijkt}|\bar{p}_{ijkt}, s_{it})$ is the expected quantity for product k at store j as a function of expected price $\bar{p}_{ijkt}$ and other variables denoted by s_{it}.[15]

In calculating the expected spending for each good, I assume that households only consider the products that they can possibly buy at stores. Based on this assumption, the expected cost only accounts for the products in the consideration set of each household for the good. The consideration set $\mathcal{G}_{ic}$ is defined as the products that the household ever purchased in the sample period. For example, if a household has only bought Coke, Pepsi, and Dr Pepper, the expected spending of soft-drink good for this household is defined by the expected costs of these three brands. Since the household level data cover five years, the consideration set defined in such a way comprehensively captures the substitute brands and sizes for each household.

Restricting the set of products to a consideration set for each household provides another source of heterogeneity in price expectation. The variation in price expectation arising from this is substantial because the grocery stores carry a vast array of products for each good and their pricing strategies may differ across these products among competing stores. For example, if the prices of Pepsi and Mountain Dew products are relatively low and those of Coke and Dr Pepper are high in a store compared to other stores, depending on which products a household considers buying, the expected spending on soft-drink good for the store can vary among households.

Note that the expected basket cost, Exp_{ijbt}, has substantial variation across households, stores, and time. The sources of such variations are multifold. First, the price expectation for each product builds on the household's shopping history. Second, the consideration set for each good differs across households depending on their choices of brands and sizes. Third, heterogeneity in the decisions of the expected quantity also creates variations in the expected spending for each good across households.[16] Last, the good purchase probability predicted based on the shopping behavior of each household is another source of variation in the expected basket costs.

Table 1 shows the expectation errors in expected basket costs across households. Expectation

[15]The purchase quantity of each product has a common measure within the same good and the price is defined for each measure unit. For example, the measure of quantity for milk good is gallon and the price is defined for each gallon.

[16]Note that the expected quantities of basket goods are estimated with household fixed effects.

error is defined as the root mean square of the difference between expected cost and actual spending for the basket goods. The actual basket spending is based on actual store prices, which households do not observe *a priori*. Expected basket cost in the full model is based on all the sources of heterogeneity across households as described above, whereas the non-heterogeneity model does not allow for household-specific shopping histories and consideration sets. The full model in the table shows that frequent shoppers have relatively smaller expectation errors than infrequent shoppers. It can be inferred that price expectation becomes more up to date from the frequent past store trips. The expectation errors based on the non-heterogeneity model are larger than those in the full model, on average, for all household groups. More importantly, the variations in expectation errors across different household groups does not have the same pattern as in the full model.[17]

4 Data

I use scanner data collected by IRI.[18] The data set is in two parts, as follows. The first data set is store level data containing weekly store sales and the second is household level data containing the weekly purchases and store visits of individual households. The data sets cover seven years from January 2001 to December 2007 and include 30 grocery goods, which consist of 17 food and beverage goods (e.g., carbonated drinks, coffee, peanut butter) and 13 non-food household goods (e.g., laundry detergent, facial tissue, razor). The data are drawn from seven stores in a small city in Massachusetts and the sample stores belong to four different supermarket chains. The store level data include weekly price and quantity for each product sold in each store at the UPC (universal product code) level. The household level data have information on individual purchases of each household at each store. I restrict the sample to about 2,000 customers who report their purchases persistently enough to satisfy the minimal reporting requirements set by

[17]Note that the expectation errors in the full model for the third group are somewhat smaller than those for the fourth group (i.e., they are not exactly proportional to the shopping frequency). The expectation errors presented in Table 1 may not fully capture heterogeneity in price expectation across households because the shopping frequency is merely one of the various shopping patterns that characterize each household.

[18]Further details on the data sets are provided in Bronnenberg, Kruger, and Mela (2008). I thank IRI for making this data set available. All analysis based on the data in this paper is by the author and not by IRI.

IRI.[19]

Since the price variable in the household level data is the weekly average of store prices, it may not be the same as the price that was actually paid by the customer. This may generate measurement errors in prices if customers redeemed retailer coupons or the store prices change within a week. However, cases in which coupons are offered by stores are only about 0.05 percent in the store level data. Therefore, it would be more often the case that a discrepancy between the prices in the data and the actually paid price occurs when the shelf prices vary within the week.

I have information on demographic characteristics for each household, such as income, age, home ownership, dummy variables for single male and single female, number of children, and an indicator for full-time working female. I also have information on the location of each household with latitude and longitude.[20] I computed the shortest driving distance between each household and stores using the Google Maps API with the location information. Using the location information is unique to my data set. Most previous studies on store choice approximate the trip distance from the zip code or census block information. Given that the conventional wisdom says that the spatial distribution of stores is one of the key determinants in store choice, using accurate information on trip distance is critical in estimating store choice consistently.

The geographic area where the sample customers and stores are located is a small urban area. This area is appropriate for the study of store choice because the consumers in this market do not have non-traditional grocery stores nearby, such as supercenters or warehouse clubs which are not included in my data. The non-traditional stores provide not only grocery products but also a wide variety of general merchandise goods. The presence of such stores can make a store choice study complicated because the purpose of grocery shopping can be confounded witg shopping for non-grocery products, and this is not distinguished without observing the purchases of those products. The supermarket chains included in my data sets are all traditional supermarkets, and the closest supercenter supermarket (Walmart or Target) is 19.6 miles from the sample

[19]The persistency of reporting is evaluated by IRI every year. I only include the households who meet the reporting requirements continuously for the full participating period. For example, if a customer meets the criteria in 2003 and 2005 but does not in 2004, I dropped the customer from the sample.

[20]For confidentiality reasons, the exact location of each household is disguised by a trifling error (about 0.1 mile).

customers on average, whereas the average distances between the sample customers and stores in the data lie between 2.9 and 5.6 miles.

Table 2 provides descriptive statistics of the household level data. The frequency of store visits is very high, with the average consumers making shopping trips almost once per week.[21] But there is a large variation in the trip frequency across households. For example, shopping frequency is negatively correlated with dollar spending per trip (correlation is -0.5), implying that shoppers with large shopping baskets tend to visit stores less often than those with small baskets. Consumers visit about four different stores per year on average. But the Herfindahl–Hirschman Index (HHI) of store choice, which measures the extent to which the store choice of each household is concentrated among the stores, amounts to 0.5, based either on number of visits or dollar spending. This suggests that, on average, store choices are concentrated on nearly two different stores. Households buy three different goods, on average, for each store visit.

Figure 1 depicts the distribution of the store HHI of households, computed based on the number of store visits and dollar spending. It shows that the concentrations of store choices substantially differ among households. For most households, neither spending nor store visits is concentrated on one or two stores.[22] Examining the extent to which various factors in store choice affect switching behavior is left to an empirical question.

Transportation costs are crucial in store choice decisions. In this study, I assume that transportation cost is a linear function of the driving distance between the residences of the households and the stores. Figure 2 shows the choice of stores by the rank in distance among alternative stores. Both in the number of store visits and dollar spending, the nearest store is the most popular choice. However, given that the difference in distance between stores in two consecutive ranks is only half a mile on average, it is not obvious to what extent the distance affects store choices.

[21] The frequency of store visits might be somewhat misleading because multiple visits to a store in a week are not distinguishable in the data. The supplementary data set that contains all the store visits made by the sample customers show that about 70 percent of visits are made only once to a store in a week. Considering that the trip data cover all the store trips made for any goods, the case of multiple visits would be much less problematic for 30 goods.

[22] Alternatively, the HHI can be calculated based on the choices of supermarket chains instead of stores. The distribution of the chain HHI is barely different from those based on the store choices.

Most households do not visit all the sample stores during the sample period. The reason why they never visit certain stores is not obvious. It may be because the utilities from visiting those stores do not exceed those from visiting other stores they usually choose or because they do not know about the stores and thus those stores are simply not in their choice set. Including these never-visited stores in the choice set therefore may lead to inconsistent estimates of demand parameters. For example, suppose a household has little knowledge about a store for some reason and never visited the store, but the model includes this store in the household's consideration set. If the prices at this store are low enough that the utility from visiting this store would be higher than those from other stores, the estimated price elasticities will be biased toward zero. For this reason, I exclude these never-visited stores from the choice set of each household.[23]

Table 3 presents the summary statistics of the store level data. According to the non-disclosure agreement with IRI, the chain names of the sample stores are disguised. Columns 2 and 3 show the percentage frequency of price discounts that are larger than 5 percent and the number of different products carried by each store. There is a substantial variation in price promotions across stores and the variation is much larger for some goods than others. It is notable that price promotions and assortment sizes differ across stores within the same chain. The average driving distance between the sample customers and each store ranges from 3 miles to 5.6 miles.

5 Estimation Results

5.1 Basket Estimation

In the first stage, the probability of a good purchase in a shopping plan is estimated separately for each good. The estimated models include log of inventory and consumption levels, recent shopping patterns, and demographic information. The variables to control for recent shopping patterns are trip frequency, spending per trip, weekend-shopping preference, and intervals between recent goods purchases. Demographic variables include log of household income, family

[23] Alternatively, if a store loyalty variable is included such that it accounts for habitual choice behavior based on store choices in a certain length of an initial period of data (e.g., Bell, Ho, and Tang (1998)), this variable would absorb this biased selection in store choices.

size, and indicators of marriage, pet ownership, and house ownership. I also add the recently purchased quantities of other goods to control for unobserved exogenous shocks to the good purchase of interest.

Table 4 reports the estimates of goods purchases for five selected goods. The coefficients represent the impact of each variable on the good purchase probability. Most of the variables are statistically significant and the signs are intuitive. The negative coefficient of inventory level implies that, if households have a large amount of the good in their pantry, they wait a long time for its next purchase. Higher consumption is associated with a shorter purchase interval. Shoppers with a high visit frequency and large dollar spending for each trip tend to have a high probability of the upcoming purchase. If the average inter-purchase interval in recent shopping is long, the purchase probability becomes low. A large volume size of recent purchases being associated with a low probability of good purchases implies that the purchase quantity in each trip is more strongly related to the average consumption shocks than to stock-piling behavior.

Goodness of fit, summarized by pseudo R-squared, varies across the goods. The duration model fits the data relatively poorly for the goods with high-frequency purchases such as soft-drinks and milk. The inter-purchase time for these goods is typically as short as one week. If the inter-purchase time is often less than a week in the actual shopping pattern (whereas the time unit of the data is a week), the variation in purchase duration captured by the data would be subtle.

Figure 3 shows survivor functions for different levels of household inventory for blades and milk, holding other covariates fixed at their mean values. The survivor function $S(\tau|x)$ is the probability that the duration until the next purchase exceeds τ given x. The survivor functions are monotonically decreasing with the elapsed time since the most recent purchase. The declining speed of survivor functions varies among different goods since the purchase frequencies are different. For example, the state of no-purchase for a blade can continue for more than five months with a positive probability, whereas survivor functions for milk reduce to zero after only about three months since the previous purchase. More importantly, a higher level of household inventory at the moment of the recent purchase is associated with a higher survival probability at any point in time. This implies that the households wait longer until the next purchase with

a higher level of inventory level.[24]

Table 5 reports the estimation of quantity choice for three goods. Since most goods are sold in various sizes of packages, the price variable in the estimation is rescaled to a standard size for each good. The price coefficients are negative for all goods, and the magnitudes of the coefficients in absolute terms increase when dealing with the endogeneity problems. Allowing for brand and household fixed effects substantially improves the fit of the quantity estimation.

5.2 Store Choice Estimation

Based on the expected basket costs estimated in the first stage, I estimate the store choice model. The estimated parameters are reported in Table 6. In particular, I estimate the store choice model for different cases of expected bundle costs. First, the full model uses the expected basket cost based on all sources of heterogeneity, as described in Section 3.4. Then, I estimate the cases in which each of the heterogeneity sources is eliminated in the expected bundle costs. This allows me to examine which component of my expected bundle costs makes a difference in the estimation of store choice.

The first three columns show the results based on the choice-based price expectation.[25] The second and third columns are the cases for which the estimation of goods purchases or purchase quantity is removed in constructing the expected bundle cost. The next three columns are based on price expectation that does not (or only partially) allow for heterogeneity in price beliefs.[26] The price expectation in Column 4 is a simple average of store prices over the past one year, regardless of household shopping history. Column 5 is based on the price expectation, ignoring both household-specific shopping history and consideration set.[27] In this case, there is

[24]Survivor functions for other goods are consistent with these findings. The figures of the survivor functions for other goods will be provided upon request.

[25]The parameter in the weight function of price expectation is estimated based on the full model by a profiling estimation method. That is, I fix the parameter at a specific value and estimate the full model. Then, a number of iterations of this procedure with different values yields the estimate of the parameter. The estimated parameter is -.03.

[26]The estimations of non-heterogeneity in price expectation are also based on the estimation of good purchase probability and quantity of goods. The results are similar to the cases of choice-based price expectation when removing these two components of basket models.

[27]By construction, it is not possible to leave out only the consideration set because the past prices of the never-bought products are not defined when households consider the prices of all products instead of only those in their consideration sets.

no heterogeneity in price expectation and expected prices are all the same across households. The last column is also based on price expectation without heterogeneity, and a shopping list is defined at the product level as in Bell, Ho, and Tang (1998).[28]

The comparison of the first three columns shows that the estimation of goods purchases and quantity in the first stage makes a slight difference. When the quantity estimation is removed, the variation in price effects among different income groups becomes somewhat small compared to the other two cases of the choice-based price expectation. Columns 4 and 5 show that ignoring heterogeneity in price expectation substantially changes the magnitudes and signs of the coefficients for expected bundle cost. When price expectation has no variation across households in Column 5, the price coefficients become positive. When a shopping list is defined at the product level, price sensitivities for different age groups show a different pattern from the other cases. In this case, young shoppers are more price sensitive than the older shoppers whereas it is the other way around in other cases. In all the cases, households prefer closer stores and such disutility due to traveling distance is stronger for high income and older households.

Since the coefficients for basket costs given in Table 6 indicate the marginal changes in the latent utility, they need to be translated into the elasticities. Table 7 presents the own elasticities of expected basket cost for each store. The full model gives the largest magnitudes of own price elasticities but the first three columns are quite similar in magnitude. This implies that allowing for the good purchase probability or the quantity of purchased goods in modeling expected bundle costs only slightly changes the store choice estimation. Columns 4 and 5 show that, when ignoring heterogeneity in household price beliefs, own price elasticities are severely biased toward zero and even take a positive sign. The comparison between Columns 4 and 6 shows that using a product-level shopping list somewhat mitigates these biases in own elasticities but the magnitudes of own price elasticities are fairly small compared to the full model.

[28]In the replication of the work of Bell, Ho, and Tang (1998), the expected price is the average store price over one year whereas theirs is the average of the entire data period (two years). Therefore, this case focuses on the role of the goods-level shopping basket by comparing with Column 4.

6 Conclusions

In this paper, I propose a household-level store choice model that reduces the biases arising from mismeasuring or misspecifying the unobserved price expectation and shopping list of each household. I specify household-level price expectation based on shopping history and choice set of products of each household for each good. This captures the effects of the short-term variations of store prices on each household's price knowledge and thus allows for heterogeneity in shopping behavior. I also specify a flexible and realistic shopping list that accounts for potential substitution within a good that can take place inside a store.

My results indicate that the store-level own elasticities of expected basket cost are biased toward zero by as much as seven times when ignoring heterogeneity in price expectation. They also show that defining a shopping list at the product level somewhat mitigates the biases that stem from ignoring heterogeneity in price expectation but the own price elasticities are still substantially biased compared to the case of the basket-level shopping list.

Appendix

A Robustness check of the utility specification

To check the robustness of the specification of my utility equation, I estimate store choice with the following utility function

$$U_{ijbt} = \beta_i Exp_{ijbt} + \gamma_i Asrt_{ijbt} + \delta_i Dist_{ij} + \sum_l \alpha_{jl} Z_{il} + \varepsilon_{ijt},$$

where $Asrt_{ijbt}$ is the variety of products carried by the store for the basket goods. The product variety in the model depends on the shopping list because households only consider the goods they plan to purchase. Thus, $Asrt_{ijbt}$ is defined by the average of the log of the number of products for the basket goods. Households may rather value the *overall* product variety for all goods instead of the basket goods. But I found that the overall assortment size independent of a shopping list is not statistically significant. This is because the overall assortment size barely changes over time at each store and it is rather absorbed by the store fixed effects in the model.

Although including a variety of available choices in a utility function is not a standard approach in consumer choice problems, allowing for good-specific product variety may be regarded as a remedy for ignoring the complicated brand choices in store choice estimation. However, as shown in Table 8, the own elasticities of expected basket cost only slightly change when product variety is included in the utility function.

B Explanatory variables in the duration models

I first construct a data-driven consumption variable for each good and then define inventory based on consumption levels. Consumption and inventory are defined at the goods level.[29] The underlying assumption in constructing consumption variables is that each household consumes at a constant rate (or amount) for a set length of time. The amount of consumption may depend on various aspects of consumer heterogeneity, such as preference or special life events. Since

[29] Consumers may consume different brands for the same good and manage inventory at the brand level rather than at the good level. I will not allow brand-level inventories, and I assume good-specific inventory.

none of these heterogeneous shocks is observable, I use a parsimonious approach to estimating consumption rates based on purchase frequency and quantity. Specifically, the constant consumption rate is computed as an average of the quantity of the products for the good purchased over a set period of time. Consumption is inferred for each household at each time, so it captures both time effects and heterogeneity of individual consumers.[30] Purchase quantity and frequency are both associated with complex consumer behaviors, such as stock-piling and sales responses. Potential errors that can arise from neglecting to consider such consumer behavior in inferring the consumption level are believed to be mitigated by averaging and smoothing the consumption rates over a length of time specified for each good.

The length of time for averaging the consumption level is set differently for each good, considering good-specific features such as average purchase frequency, shelf life, and whether it is a necessity good. For example, consumption rates are averaged over a short period of time for goods that are perishable and purchased with a high frequency such as milk and yogurt. On the other hand, for the goods that are storable (possibly with low-frequency of purchases) and that people typically consume in a persistent and constant pattern (e.g., necessity goods) I smooth consumption over a long period (e.g., tissue, laundry detergent, and toothpaste).

Given the consumption level of each good, the inventory level at the beginning of each period is the residual amount in storage defined as follows:

$$I_{it} = \max\{0,\ I_{i,t-1} + Q_{i,t-1} - C_{i,t-1}\}$$

where I_{it} is the inventory level consumer i faces at time t, Q_{it} is the quantity of the purchase at t, and C_{it} is the consumption level. The purchases during the first six months in the raw data are used to generate the distribution of initial inventories.

[30] Ailawadi and Neslin (1998) also construct time-varying consumption rates as a (continuous and nonlinear) function of inventory in their empirical study. Their approach heavily depends on functional specification.

References

Ailawadi, K. L., and S. A. Neslin (1998): "The Effect of Promotion on Consumption: Buying More and Consuming It Faster," *Journal of Marketing Research*, 35(3), 390–398.

Alba, J. W., S. M. Broniarczyk, T. A. Shimp, and J. E. Urbany (1994): "The Influence of Prior Beliefs, Frequency Cues, and Magnitude Cues on Consumers' Perceptions of Comparative Price Data," *Journal of Consumer Research*, 21(2), 219–235.

Bajari, P., and C. L. Benkard (2003): "Discrete Choice Models as Structural Models of Demand: Some Economic Implications of Common Approaches," Working Paper.

Bell, D. R., T. H. Ho, and C. S. Tang (1998): "Determining Where to Shop: Fixed and Variable Costs of Shopping," *Journal of Marketing Research*, 35(3), 352–369.

Beresteanu, A., P. B. Ellickson, and S. Misra (2006): "The Dynamics of Retail Oligopoly," Working Paper, Duke University.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

——— (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112(1), 68–105.

Bodapati, A. V., and V. Srinivasan (2006): "The Impact of Feature Advertising on Customer Store Choice," Mimeo, Stanford University.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008): "Database Paper: The IRI Marketing Data Set," *Marketing Science*, 27(4), 745–748.

Cox, D. R. (1972): "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, 34(2), 187–220.

Erdem, T., S. Imai, and M. P. Keane (2003): "Brand and Quantity Choice Dynamics Under Price Uncertainty," *Quantitative Marketing and Economics*, 1(1), 5–64.

Goolsbee, A., and A. Petrin (2004): "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 72(2), 351–381.

Hausman, J. A., and D. A. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46(2), 403–426.

Heckman, J. J., and B. Singer (1986): "Econometric Analysis of Longitudinal Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. Intriligator, vol. 3. Amsterdam: North-Holland.

Hendel, I., and A. Nevo (2006): "Measuring the Implications of Sales and Consumer Inventory Behavior," *Econometrica*, 74(6), 1637–1673.

Ho, T. H., C. S. Tang, and D. R. Bell (1998): "Rational Shopping Behavior and the Option Value of Variable Pricing," *Management Science*, 44(12), S145–S160.

Kalbfleisch, J., and R. L. Prentice (2002): *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley, Chap. x.

Katz, M. (2007): "Estimating Supermarket Choice Using Moment Inequalities," Unpublished Ph.D. dissertation, Harvard University.

Lal, R., and R. Rao (1997): "Supermarket Competition: The Case of Every Day Low Pricing," *Marketing Science*, 16(1), 60–80.

Lancaster, K. (1971): *Consumer Demand: A New Approach*. New York: Columbia University Press.

Lancaster, T. (1990): *The Econometric Analysis of Transitional Data*. Cambridge, UK: Cambridge University Press, Chap. x.

Mazumdar, T., S. P. Raj, and I. Sinha (2005): "Reference Price Research: Review and Propositions," *Journal of Marketing*, 69(4), 84–102.

Monroe, K. B. (1973): "Buyer's Subjective Perceptions of Price," *Journal of Marketing Research*, 10(1), 70–80.

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69(2), 307–342.

Pakes, A. (2010): "Alternative Models for Moment Inequalities," *Econometrica*, 78(6), 1783–1822.

Petrin, A. (2002): "Quantifying the Benefits of New Products: The Case of Minivan," *Journal of Political Economy*, 110(4), 705–729.

Smith, H. (2004): "Supermarket Choice and Supermarket Competition in Market Equilibrium," *Review of Economic Studies*, 71(1), 235–263.

Winer, R. S. (1986): "A Reference Price Model of Brand Choice for Frequently Purchased Products," *Journal of Consumer Research*, 13(2), 250–256.

Table 1: Errors in Expected Bundle Costs

	Expectation errors		Actual
Frequency of trips	Full model	Non-heterogeneity	basket costs
0< Frequency < 5	5.37	8.70	15.39
5≤ Frequency < 10	5.04	8.06	14.28
10≤ Frequency < 25	4.27	7.40	13.41
25≤ Frequency < 40	4.79	9.54	14.25
40≤ Frequency	4.40	9.45	18.29

Note: Expectation error is defined as the root mean square of the difference between expected cost and actual spending for each shopping basket. The full model reflects the sources of heterogeneity in price expectation, whereas non-heterogeneity model does not.

Table 2: Summary Statistics of Household-Level Data

	Mean	Median	Std	Min	Max
Number of trips per month	4.9	5	2.4	1	23
Number of visited stores	4.2	4	1.6	1	7
Store HHI (visits)	.48	.42	.22	.16	1
Store HHI (spending)	.56	.50	.23	.17	1
Distance per trip (mile)	3.2	2.9	2.1	.01	20.8
Weekly spending (30 goods)	20.83	15.54	18.71	.16	363.16
Weekly spending (all goods)	102.81	84.20	81.23	.13	1290.72
Number of goods per trip	3.1	2	2.3	1	22

Note: Store HHI of household i is the sum of the square of the share of visits to (or dollar spending in) each store. That is, $StoreHHI_i = \sum_s \left(y_{its}/\sum_{s'} y_{its'}\right)^2$, where y_{its} is the number of visits (or expenditure) of household i for store s at week t.

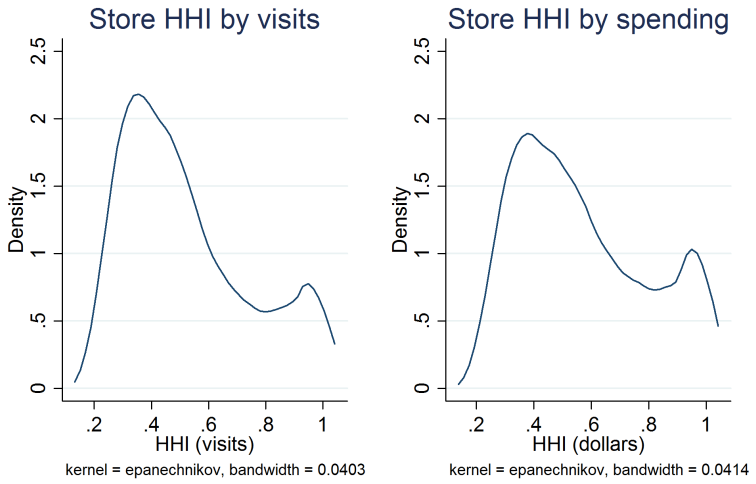

Figure 1: Store HHI by Visits and Spending

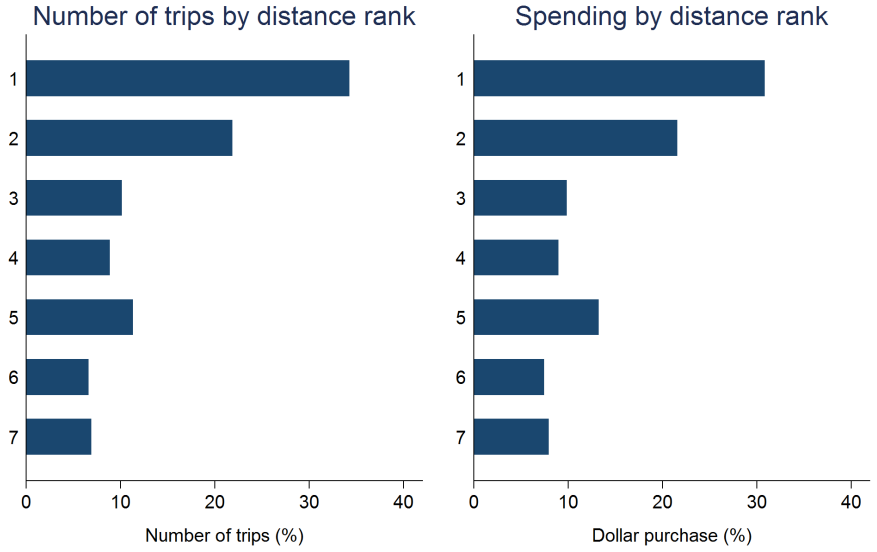

Figure 2: Distance and Store Choice

Table 3: Summary Statistics of Store-Level Data

	Chain	Price cuts (> 5 percent)	Assortment size	Average driving distance
Store 1	A	19.7%	2,346	3.3 mile
Store 2	A	16.9%	3,416	2.9 mile
Store 3	B	30.7%	6,177	4.2 mile
Store 4	B	27.0%	6,309	4.7 mile
Store 5	C	23.0%	4,890	5.6 mile
Store 6	C	24.3%	5,902	4.0 mile
Store 7	D	22.0%	5,598	2.9 mile

Table 4: Estimation of Goods Purchases

	Blades	Soft-drink	Cereal	HH cleaners	Milk
Inventory (log)	-1.15***	-0.64***	-1.37***	-0.98***	-0.74***
Consumption rate (log)	0.63***	0.56***	0.67***	0.68***	0.66***
Trip frequency	0.03***	0.02***	0.01***	0.02***	0.01***
Trip dollars (log)	0.42***	0.11***	0.16**	0.31***	0.03***
Weekend shopping	-0.01*	0.01***	0.01**	0.01	0.01***
Purchase intervals	-0.05***	-0.13***	-0.12***	-0.06***	-0.29***
Volume size of purchases	0.42***	0.39***	0.63***	0.19***	0.32***
Unit price of purchases	0.10***	-0.01	0.02***	0.13***	0.02***
Demographics	Y	Y	Y	Y	Y
Other goods purchases	Y	Y	Y	Y	Y
Monthly dummy	Y	Y	Y	Y	Y
Pseudo R-squared	0.19	0.04	0.10	0.18	0.03
Observations	34,398	359,979	333,788	100,450	373,196

Note: Inventory level and volume size of purchases are normalized to the average quantity. Demographic variables include log income, family size, and indicators for marriage, pet ownership and house ownership.

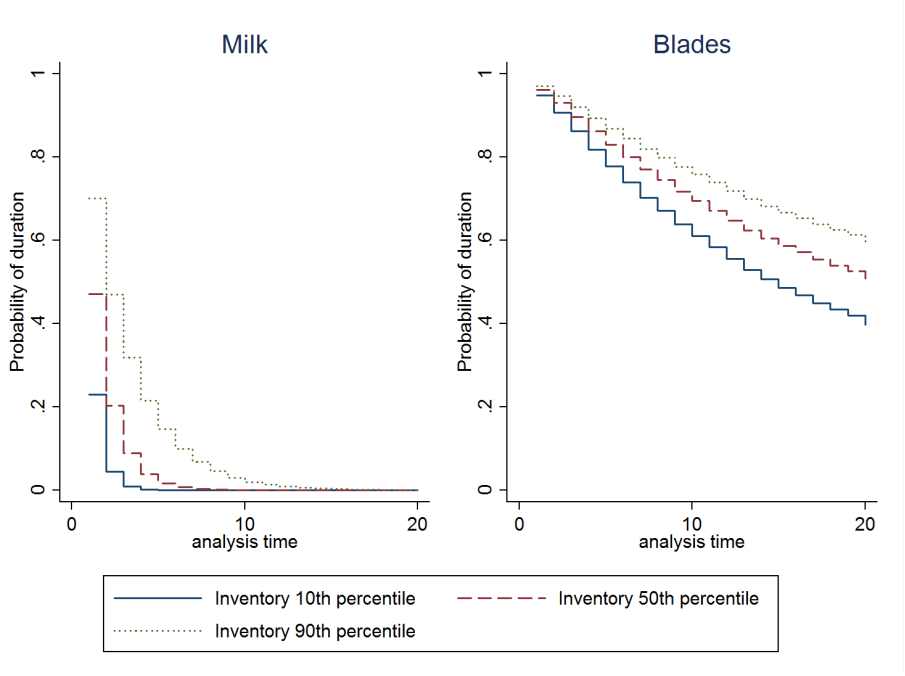

Figure 3: Inventory and Survivor Function

Table 5: Quantity Choice Estimation

	Coffee		Milk		Paper towels	
	OLS	IV	OLS	IV	OLS	IV
Price	-1.75***	-1.93***	-14.32***	-16.04***	-24.50***	-32.72***
Consumption	0.20***	0.20***	0.10***	0.10***	0.49***	0.46***
No Trip (good)	-0.22***	-0.22***	-0.20***	-0.18***	-0.66***	-0.63***
No Trip (overall)	0.59***	0.62***	-03.06***	-3.53***	0.83***	0.26***
Month dummy	Y	Y	Y	Y	Y	Y
Brand FE	Y	Y	Y	Y	Y	Y
Customer FE	Y	Y	Y	Y	Y	Y
R-squared	0.42	0.25	0.57	0.36	0.56	0.40

Table 6: Store Choice Estimation

	Choice-based price expectation			Non-heterogeneity price expectation		
	Full model	No quantity	No good purchase	No shopping history	No choice-set	Product level basket
Expected basket cost	-0.311***	-0.145***	-0.280***	-0.023**	0.167***	-0.034***
Income						
15k-25k	0.234***	0.019***	0.217***	0.057***	0.008	-0.006
25k-45k	0.105***	0.016***	0.092***	0.024**	-0.012*	0.029**
45k-65k	0.083***	0.001	0.072***	0.007	-0.036***	0.034***
65k+	0.143***	0.011*	0.128***	0.031***	-0.051***	0.066***
Age						
≤34	0.116***	0.101***	0.105***	-0.017	0.031**	-0.044***
35-54	0.111***	0.118***	0.099***	-0.004	0.008	-0.031***
55-64	0.027***	0.063***	0.025***	0.016**	-0.007	-0.059***
Distance	-0.334***	-0.336***	-0.334***	-0.335***	-0.336***	-0.379***
Income						
15k-25k	-0.044***	-0.045***	-0.044***	-0.043***	-0.043***	-0.034***
25k-45k	-0.057***	-0.058***	-0.056***	-0.057***	-0.056	-0.061**
45k-65k	-0.104***	-0.106***	-0.104***	-0.103***	-0.102***	-0.101***
65k+	-0.103***	-0.104***	-0.103***	-0.103***	-0.103***	-0.102***
Age						
≤34	0.152***	0.154***	0.152***	0.152***	0.151***	0.201***
35-54	0.081***	0.084***	0.080***	0.080***	0.080***	0.088***
55-64	0.068***	0.069***	0.068***	0.067***	0.067***	0.081***
Log likelihood	-658695.65	-658701.09	-658811.01	-660073.38	-658742.88	-519833.3
Observations	2,354,517	2,354,517	2,354,517	2,354,517	2,354,517	1,962,652

The base groups for income and age are those who earn less than 15 thousand and those who are older than 65, respectively. *** represents under the 1% significance level, and ** and * are under the 5% and 10% significance levels, respectively.

Table 7: Own Price Elasticities

	Choice-based price expectation			Non-heterogeneity price expectation		
	Full model	No quantity	No good purchase	No shopping history	No choice-set	Product level basket
S1-A	-6.891	-6.587	-6.731	1.385	8.398	-1.699
S2-A	-6.788	-5.995	-6.080	-.024	5.469	-1.411
S3-B	-7.618	-6.424	-6.931	.503	7.774	-.566
S4-B	-8.351	-7.908	-8.333	1.177	9.651	-.264
S5-C	-2.744	-1.708	-1.792	.026	2.582	-.352
S6-C	-1.201	-.869	-0.883	.052	1.728	-.229
S7-D	-1.206	-.799	-0.895	.062	1.472	-.268
Mean	-4.971	-4.327	-4.516	.454	5.296	-.684

Table 8: Robustness: Own Price Elasticities

	Choice-based price expectation Full model	
	With assortment size	Without assortment size
S1-A	-7.897	-6.891
S2-A	-7.750	-6.788
S3-B	-7.686	-7.618
S4-B	-7.466	-8.351
S5-C	-2.370	-2.744
S6-C	-1.294	-1.201
S7-D	-1.272	-1.206
Mean	-4.971	-5.105