Lu, Anna

Conference Paper

Inference of Choice Sets: Application to Grocery Retailing

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Empirical IO, No. A02-V1

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Lu, Anna (2016) : Inference of Choice Sets: Application to Grocery Retailing, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Empirical IO, No. A02-V1, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/145683

Inference of Choice Sets: Application to Grocery Retailing

February 2016

Preliminary and Incomplete.

Abstract

In the empirical literature on grocery retailing it is typically assumed that all consumers choose from the full set of products in the market. We develop an approach to formally test this assumption. Our test can be applied to individual-level purchase data. Unlike previous approaches it does not require stated data on choice sets; instead, it relies on only widely available cost shifter data. We show an application to the German retail market for milk and find that the model of homogeneous, full choice sets is outperformed by a model in which consumers consider only the products of the retailer they are currently shopping in.

JEL: D12, L00, L13

1 Introduction

The literature on demand estimation in empirical industrial organisation commonly assumes that all consumers choose from all products in the market. However, market frictions such as search costs, transport costs or limited information can create heterogeneity in the so-called *choice set*, i.e. the set of products that a consumer considers for purchase.[1] If such heterogeneity is ignored, demand estimates may be biased. This affects, in particular, policy evaluations because they require demand estimates as an input.

What makes it difficult to incorporate choice sets into demand estimation is the fact that the econometrician typically observes only which choices were made but not which alternatives were considered. In consumer panel data for instance, one can observe which retail chains a household chose to visit but not which retail chains it considered. The standard assumption is that households consider all chains in the market. However, this retailer choice set could be smaller because it may be shaped by other things, e.g. by which chains were visited in the past.

The aim of this paper is to infer information on unobserved choice sets using only observed choices. In order to do so, we model shopping as a two-stage process. Households have a fixed set of retail chains in their shopping rotation which can be regularly updated and from which households choose a retail chain on each shopping trip. In the second step, they select a final product from the assortment of their chosen retail chain. We estimate demand based on this model and use supplemental information on marginal

[1] In the marketing literature, the term "consideration set" is more common. For the sake of consistency, we will stick to the term "choice set" throughout this paper.

cost-shifters to test for households' updating frequency of the retailer choice set.

We aim to make two contributions. Firstly, we are - to our best knowledge - the first to use cost-shifters for the inference of choice sets. More specifically, we employ a menu approach which allows us to test discretized scenarios against each other. The central assumption is that retailers observe true consumer choice sets,[2] and factor this information into their retail prices. Our test procedure is as follows: For each scenario we estimate demand and use it to recover marginal costs. We regress these estimated marginal costs on observed cost shifters and use a non-nested model selection test to determine the regression with the best fit. The idea behind this is that the better a demand scenario describes the true choice process, the better will estimates of demand and marginal costs be. This "goodness" of the estimates is then captured in how well estimated marginal costs can be explained by observed marginal cost shifters.

In our second contribution, we apply our model to the German retail market for fluid milk. We test three scenarios against each other: a) Consumers choose from all products at all retailers; b) consumers choose from all products at the retailer they visited in the past month and c) consumers choose from all products at the retailer they are currently visiting. We find that scenario c) performs significantly better than the other two scenarios. Under scenario a), i.e. under the standard assumption of homogeneous full choice sets, retailer margins will be underestimated by 7%.

[2] This assumption is motivated by large retailer spending on market research. For example, more than 9.6 billion US dollars were spent on market research in the US in 2012 (**?**).

One advantage of our approach are the relatively limited data requirements. Firstly, we need micro purchase data. This type of data is provided by a number of market research companies and has become increasingly available in the past decades. Secondly, we need data on cost shifters which, on an aggregate level, are widely available. The power of our test will however be larger when firm- or product-level cost shifters are used. The analysis in this paper may be limited to a particular market and product category, but our method translates well to other retailing markets.

The remainder of this paper is organized as follows: First, we give a brief overview of the related literature. We develop our model in section 3 and describe our data, the German milk market and patterns in consumer behavior in section 4. The identification strategy and the estimation results are explained in section 5. We conclude in section 6.

2 Related Literature

The literature on demand estimation in Industrial Organization has primarily focused on cases in which it is assumed that all consumers choose from all products in the market (e.g. **???**). However, this assumption has been under scrutiny both in economics and marketing research, and there is evidence that umption may not be innocuous: Falsely assuming full, homogeneous choice sets can strongly bias estimated demand elasticities, mark-ups, profits and the impact of marketing control variables (**?????**).

The literature has developed two main strands which tackle this problem in different ways. The first strand of the literature aims to explicitly model choice set formation, typically as a two-stage process:[3] In the first stage,

[3] A recent literature departs from the two-stage approach and develops econometric tools to estimate demand under choice set misspecification. For example, **?** uses an approach

consumers choose their choice set, i.e. the products they consider for purchase. In the second step, they choose the final product from their choice set.

One of the most prominent examples in this literature is by **?** who looks at heterogeneity in product awareness in the U.S. market for personal computers. She lets brand awareness be a function of socio-demographic characteristics and exposure to advertising and finds that when limited awareness is ignored, mark-ups are underestimated by 75%. Similarly, other papers in the literature model choice set formation as a function of socio-demographics and marketing instruments, for instance in the context of advertising (**???**) or search costs (**???**).

In an important paper from marketing research, **?** provide experimental evidence that two-staged models perform well. In their online experiment, participants face a virtual supermarket in which they periodically choose a brand and indicate to which extent they considered the respective brands for purchase. The authors model a two-stage process of brand consideration and compare the model predictions to stated choice data. They find that the predicted consideration sets closely match stated choice sets.

The second strand of the literature uses supplemental information to enhance demand estimates. For example, **?** combine macro sales data with micro data on past purchases and assume that the choice set consists of the previously purchased brands. However, the authors note that this approach requires strong stationary assumptions, e.g. no product introductions. **?** combine survey data of stated choice sets with data of supermarket sales and advertising expenditures to show that both brand valuations and price

with lower and upper bounds, and **?** develop a parametric choice model that exploits across-time and across-choice variation. These papers constitute a strand of their own.

coefficients are biased downwards in the full model. **?** use data on vending machine stock-outs in order to look at heterogeneity in physical product availability; they find that when stock-outs are ignored, demand estimates and predicted sales and profits are significantly biased.

Our paper relates to both strands of the literature: We follow the approach of the first strand and estimate a two-stage decision process: Consumers first choose a retail chain, then they pick a product from the assortment of that retail chain. More importantly, we add to the second strand of the literature in that we combine micro-level purchase data with (macro-level) cost shifters. We use these cost shifters to predict marginal costs and match them with the marginal costs recovered from our demand estimation. This allows us to identify the best-performing choice set specification.

3 Model

In this section, we develop our model. First, we model the two-stage decision process of households as a two-stage mixed logit and derive their purchase probabilities. Then we present the pricing problem of firms and show how to obtain marginal costs. Lastly we describe how to test between different specifications of consumer behavior.

Two-stage decision process Each consumer i makes two consecutive choices. In the first stage, she chooses a choice set C_i from C possible choice sets. Each choice set C_i contains a different subset of the J products in the market. The utility from choosing choice set C_i is given by

$$U_{ict} = X_{ct}\gamma_i + \xi_{ct} + \eta_{ict} \tag{1}$$
$$i = 1,\ldots,I, \quad t = 1,\ldots,T \quad c = 1,\ldots,C$$

where X_{ct} is a vector of observed choice set characteristics, ξ_{ct} are unobserved characteristics of the choice set and η_{ict} is a vector of i.i.d. extreme-value I distributed shocks.[4]

In the second stage, the consumer chooses a product j from her choice set C_i. The corresponding utility is

$$U_{ijt} = \alpha_i p_{jt} + x_{jt}\beta_i + \xi_{jt} + \varepsilon_{ijt} \tag{2}$$
$$j \in C_i, \quad i = 0, \ldots, I, \quad t = 1, \ldots, T$$

where x_{jt} is a K-dimensional vector of observed product characteristics, ξ_{jt} are unobserved product characteristics, and p_{jt} denotes the price of product j at time t. ε_{ijt} is a zero-mean, i.i.d. extreme-value I distributed individual-specific random shock. Consumers can choose not to buy any of the J products. The utility from this outside option is

$$U_{i0t} = \alpha_i p_{0t} + x_{0t}\beta_i + \xi_{0t} + \varepsilon_{i0t} \tag{3}$$

Since the mean utility of the outside good is not identified, we normalize it to zero. The coefficient α_i is consumer i's marginal disutility of price, β_i is a K-dimensional vector of individual marginal utilities with respect to the K observed product characteristics.

All coefficients $\alpha_i, \beta_i, \gamma_i$ are allowed to contain a mean coefficient and a varying component. For instance, the individual price coefficient is

$$\alpha_i = \alpha + \sigma_\alpha \nu_i^\alpha, \quad \nu_i^\alpha \sim N(0,1) \tag{4}$$

[4] This distributional assumption is characteristic for logit models. It allows for closed-form probabilities and makes the model straightforward to estimate and interpret.

where α denotes the mean price response across all consumers and ν_i^{α} gives the random consumer specific taste variation with parameter σ_{α}.

Probabilities Let L_{ijt} be the probability of consumer i choosing product j conditional on the parameters α_i, β_i and γ_i. Using Bayes' rule, L_{ijt} can be computed as $\sum_c L_{ijt|c} L_{ict}$, where

$$L_{ict}(\gamma_i) = \frac{\exp(V_{ict})}{1 + \sum_{l=1}^{C} \exp(V_{ilt})} \tag{5}$$

$$L_{ijt|c}(\alpha_i, \beta_i) = \frac{\exp(V_{ijt})}{1 + \sum_{k=1}^{J_c} \exp(V_{ikt})} \tag{6}$$

Under the standard assumption of consumers choosing from all products in the market, L_{ict} is equal to 1 for all c. $L_{ijt|c}$ is zero if product j is not included in choice set c. For simplicity, we will notate the conditional distributions $f(\alpha_i|\alpha, \sigma^{\alpha}) f(\beta_i|\beta, \sigma^{\beta}) f(\gamma_i|\gamma, \sigma^{\gamma})$ as $f(\alpha_i), f(\beta_i)$ and $f(\gamma_i)$ in the following. The unconditional probability of observing the sequence of T choices made by consumer i is then

$$P_i(\alpha, \beta, \gamma, \sigma^{\alpha}, \sigma^{\beta}, \sigma^{\gamma}) \tag{7}$$

$$= \int \left(\prod_{t=1}^{T} L_{ij(i,t)t}(\alpha_i, \beta_i, \gamma_i) \right) f(\alpha_i) f(\beta_i) f(\gamma_i) d\alpha_i d\beta_i d\gamma_i \tag{8}$$

$$= \int \left(\prod_{t=1}^{T} L_{ij(i,t)t|c}(\alpha_i, \beta_i) L_{ic(i,t)t}(\gamma_i) \right) f(\alpha_i) f(\beta_i) f(\gamma_i) d\alpha_i d\beta_i d\gamma_i \tag{9}$$

where $j(i,t)$ is the alternative chosen by consumer i in period t and f denotes the standard normal probability distribution function. We maximize the log-likelihood $\sum_{i=1}^{N} \ln[P_i(\alpha, \beta, \gamma, \sigma^{\alpha}, \sigma^{\beta}, \sigma^{\gamma})]$ with respect to the coefficients $(\alpha, \beta, \gamma, \sigma^{\alpha}, \sigma^{\beta}, \sigma^{\gamma})$.

Marginal costs In the following, we omit the time subscript. Each retailer r set prices for all products in their assortment S_r. Specifically, they maximize their profits:

$$\Pi_{jr} = \sum_{j \in S_r} [p_j - mc_j^r] s_j(p) \tag{10}$$

where mc_j^r is retailer r's marginal cost of selling product j. The corresponding FOC is

$$s_{jt} + \sum_{m \in S_r} (p_{mt} - mc_m^r) \frac{\partial s_m}{\partial p_j} = 0 \tag{11}$$

For notational simplicity, we switch to matrix notation in the following. Let T denote the $j \times j$ retailer ownership matrix where element $T(j,k)$ is equal to 1 if products j and k are sold by the same retailer and 0 otherwise. Let Δ be a $j \times j$-matrix of first derivatives of all market shares with respect to all retail prices, i.e. element $\Delta(j,k)$ contains $\partial s_k / \partial p_j$. Stacking up the first-order conditions for all products and rearranging terms, we obtain the $j \times 1$-vector of marginal costs:

$$mc = p + (T * \Delta) s(p). \tag{12}$$

where mc is a $j \times 1$-vector of marginal costs, p is a $j \times 1$-vector of retail prices and $s(p)$ is a $j \times 1$-vector of market shares. We obtain $s(p)$ and the ownership matrix T from the data and the matrix Δ from the estimation.

Testing In order to test M different choice set specifications against each other, we recover for each specification m a $j \times 1$ vector of marginal costs mc^m and regress it on observed cost shifters. c is a $j \times l$-matrix of costshifters

where l is the number of different costshifters.

$$mc^m = c\delta + \nu \tag{13}$$

where ν is a $j \times 1$-vector of mean-zero i.i.d. errors and the $l \times 1$-vector of parameters δ is to be estimated.

The intuition is that each specification m generates a different vector of mc. We can then use external data on cost shifters to evaluate the "goodness" of each specifications marginal cost estimates. More specifically, we use a model selection test for non-nested models proposed by **?**. This test does not require one of the competing models to be the true model; instead, it indicates which model is closest to the true model.

The Vuong test statistic is computed as

$$V(1,2) = \frac{LR(\beta_{ML,1}, \beta_{ML,2})}{\sqrt{N}\omega_N} \longrightarrow N(0,1) \tag{14}$$

where $LR = L1_N(\beta_{ML,1}) - L2_N(\beta_{ML,2}) - \frac{K1-K2}{2} \cdot log(N)$. ω_N denotes the variance of LR. Under the null hypothesis that both models 1 and 2 fit equally well, the likelihood ratio statistic should equal zero. The asymptotic distribution of V is standard normal.

4 Data

We use household panel scanner data from the German market research company "Gesellschaft für Konsumforschung" (GfK). Each observation is one single purchase and includes characteristics of the purchased product (price, brand, private label, fat content, UHT, organic), of the shopping trip (date,

retail chain, total expenditure) and of the household (postcode of residence, income, education, age).

The German grocery market is characterized by an oligopolistic structure. According to a sector inquiry conducted by the German cartel office, the largest eight chains capture more than 90% of the market (**?**). In order to reduce the computational burden, we focus on only shopping trips to these eight retail chains in the following (see Table 1. Four chains are full-line retailers and the other four chains are discounters.

Table 1: Market shares of retailers

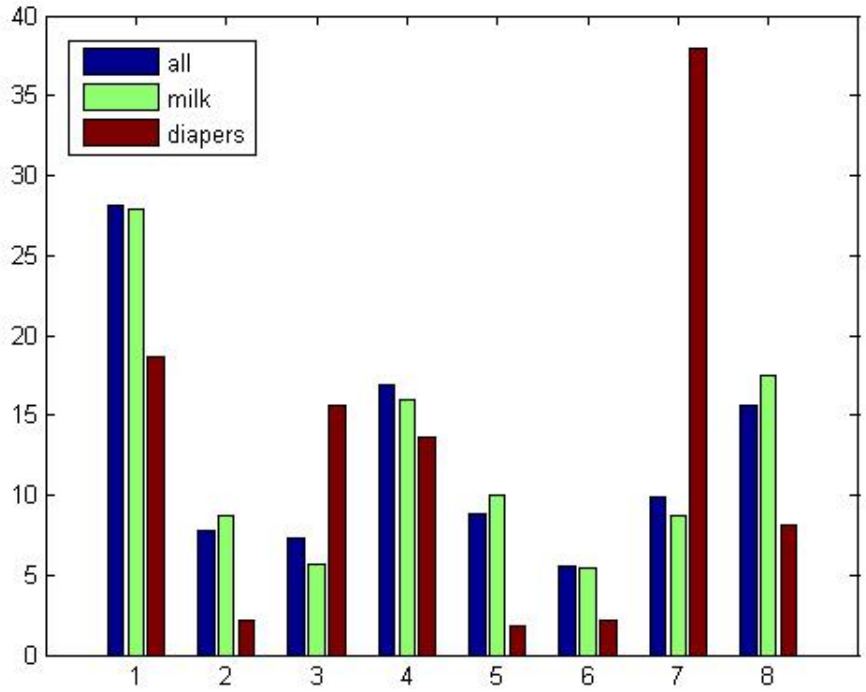
Retailer	Number of purchases	Market share
Retail Chain 1 (Discounter)	11,984	31.7
Retail Chain 2	2,478	6.6
Retail Chain 3	2,066	5.5
Retail Chain 4 (Discounter)	6,641	17.6
Retail Chain 5 (Discounter)	3,535	9.4
Retail Chain 6 (Discounter)	2,236	5.9
Retail Chain 7	2,447	6.5
Retail Chain 8	6,412	17.0
Total	37,799	100.0

Source: GfK

We choose milk for our analysis because milk purchases are highly representative of consumer shopping behavior. Graph 1 shows the share of retailers 1 to 8 for all purchases (blue), milk purchases (green) and diaper purchases (red). Retailer shares for milk purchases closely match those across all products, i.e. consumers do not seem to specifically target retailers for their milk purchases. (For other products, e.g. diapers which are a relatively expensive item, retailer shares differ strongly from the average.) Our data covers all milk purchases of 1261 households in 2010, with a total of 37,799

shopping trips. Due to computational reasons, we restrict our analysis to households living in the German state of North-Rhine Westphalia.[5]

Figure 1: Share (in %) of retailers 1-8 among all purchases, milk purchases and diaper purchases.



We define a product as a unique combination of retail chain, brand, a private label dummy, fat content, a UHT dummy and an organic dummy. We include only the 50 best-selling products in the market. Table 2 shows descriptive statistics of the milk characteristics for our sample.

The German milk market is largely dominated by private label products: More than 95% of all milk is sold under a private label, national brands

[5] The state of North-Rhine Westphalia corresponds to the Nielsen area 1.

capture only a small share of the market. Promotions are rarely offered for milk: Only about 3% of all milk sales have promotional prices. Milk is typically sold in cardboard cartons of 1 liter and it is almost always pasteurized, i.e. subjected to heating for a short time in order to increase its shelf-life.[6] Different pasteurization procedures yield produce either fresh milk or UHT milk, and they differ both in shelf life and taste. In Germany, fresh milk (64.3%) has a larger market share. Milk comes in two different fat levels: semi-skimmed milk and full fat milk come with a fat content of 1.5% and 3.5%, respectively. Both have roughly equal market shares. Organic milk is still a niche market with less than 3% market share.

Table 2: Product Characteristics

Stats	Price	Private Label	Organic	Fresh	Fat Content	Control Function	Accessability
mean	53.692	.952	.029	.643	2.315	-.0008	.168
sd	8.906	.213	.168	.479	1.021	4.015	.083
min	25	0	0	0	.1	-41.237	0
max	109	1	1	1	3.8	50.579	.406

Source: GfK

In our data, each observation corresponds to one purchase of milk. We observe the date of the purchase, the retail chain, the price, the brand and the number of units[7], the characteristics of the milk and the sociodemographic characteristics of the household.

[6] Heating milk for about 15 seconds up to 75 °C produces what is commonly known as fresh milk. Heating milk for 1-4 seconds up to 135-150 °C, i.e. ultra-high temperature processing (UHT), yields so-called UHT milk.

[7] We neglect the number of purchased units in our estimation. We are aware that this is an important variable if consumers stock milk. However, milk storability is limited, particularly for fresh milk which constitutes the majority of sales. Also, if consumer product preferences and price sensitivities are not linked to consumer storage preferences, then estimates will remain unbiased. As of now, a dynamic stockpiling model is beyond the scope of the paper.

Table 3: German Price Indices of cost shifters (in €)

Statistics	Rawmilk	Diesel	Electricity	Labour
mean	30.97	100	99.992	102.081
std. dev.	2.558	3.35	.693	5.031
min	27.95	92.7	98.7	94.085
max	34.65	106.6	100.8	112.493

Source: German Federal Statistical Office

All retail chains do not have outlets close to each household. We define chain r's accessibility to household i as the number of r's outlets divided by the total number of retail outlets in a 10 km radius around household i's home.[8] Values of accessibility vary across households from 0, i.e. a chain not being in a households shopping radius at all, to 0.406. No retailer is a regional monopolist by being the only one to have outlets in the shopping radius of some households.

Finally, we add industry-wide data on cost shifters provided by the German Federal Statistical Office. We use input price indices for raw-milk, electricity, labor, paper packaging and diesel (see Table 3). A drawback of our industry-level cost shifter data is its coarseness. If available, firm-level cost shifter data would increase the power of our test.

Evidence of Choice Set Heterogeneity In the following, we present some empirical evidence that the standard assumption of homogeneous, full choice sets does not apply well to supermarket choice. More specifically, we show that (a) households typically visit only a subset of all retail chains, (b) all households do not visit the same subset of retail chains, (c) households do

[8] We know household addresses on a postcode level. Since Germany is decided into 28,683 post code areas, five-digit postcodes are a fairly precise measure of location.

Table 4: Number of different retailers visited at least once in 2010

Retailers visited	Frequency	Percentage	Cumulative
1	4,218	11.16	11.16
2	7,699	20.37	31.53
3	8,115	21.47	53.00
4	7,686	20.33	73.33
5	5,550	14.68	88.01
6	3,352	8.87	96.88
7	975	2.58	99.46
8	204	0.54	100.00
Total	37,799	100.00	

Source: GfK

not choose retailers at random and (d) transport costs play a role in retailer choice.

The vast majority of households visits only a subset of retail chains in 2010. The average household visits no more than three different retail chains, less than 5% of the households visit more than five different retail chains and only 0.54% of households visits all eight retail chains in our sample (see Table 4).[9] Furthermore, households' retailer choice varies strongly across households: 1261 households visit 208 different sets of retail chains. All in all, our data suggests that the standard assumption of homogeneous and full choice sets does not capture retailer choice well. In the following, we provide evidence on how retailer choice should be modeled instead.

Household retailer choice displays a large degree of persistence. We divide the year 2010 into the first six months and the last six months, and find that

[9] We do not observe every shopping trip of a consumer. However, we observe all shopping trips where consumers purchased at least one product from one of the six following categories: coffee, cheese, milk, yogurt, toilet paper, diapers. As these categories cover a large range of needs and are products that have to be repeatedly purchased, we believe we are able to capture a large share of shopping trips.

for 60.74% (32.64%) of the households the most often visited retailer (set of visited retailers) is the same in both half-years.

Retailer choice seems to be linked to travel costs. Retailers that are located farther from households are less likely to be visited. Graph 2 shows that the local accessibility of a retail chain is strongly linked to its market shares. Furthermore, retailer choice seems to be serially correlated: In 63.5% of all choice occasions, the household chooses the same retailer as on the previous choice occasion.

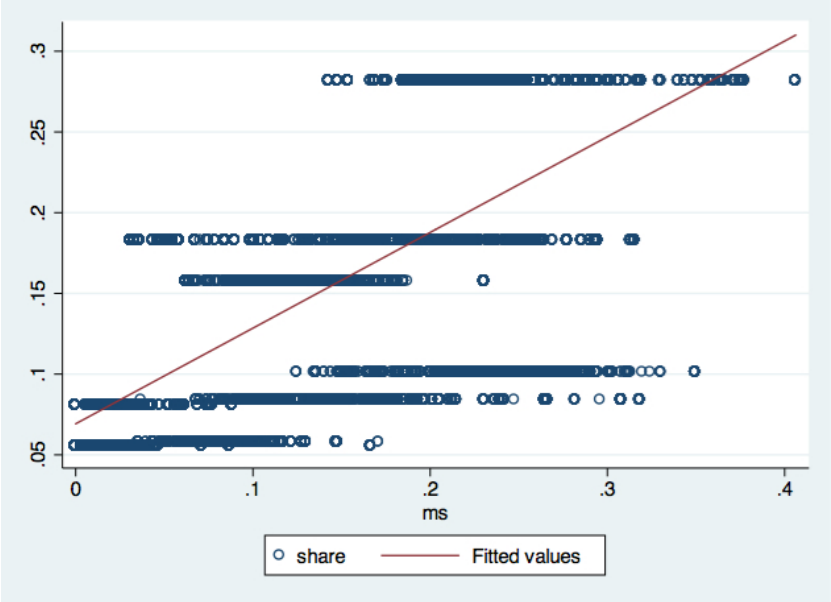


Figure 2: Local accessibility of retailers (for each postcode) on the x-axis and retail chain market shares on the y-axis. Fitted values are shown in the red line.

5 Estimation and Results

Identification Preference parameters for product characteristics are identified by variation in these characteristics (see Table 2). Store-fixed effects explain why consumers may choose a store which offers products at worse conditions than its competitor. The error term is individual-, time- and

alternative-specific. It rationalizes why, on two different shopping trips, a consumer may choose differently even when the conditions remain the exact same. The error term captures among others the momentary mood of the consumer, advertising exposure and end-of-aisle displays.

The error term could be correlated with the price. For example, marketing-instruments such as advertising can increase both the price and the demand of a product. We believe that this is not a major problem for the milk category because it is characterized by an extremely low rate of promotions (3% of all purchases are promotional), relatively uniform packaging in 1-liter cartons and limited advertising. Still, in order to tackle potential endogeneity, we follow the control function approach proposed by **?** (see appendix A.1).

Assumptions In our application, we make two assumptions about shopping behavior. Our first assumption is that the factors that make consumers choose a retail chain do not affect their product choice. In our application, a household's choice of the retail chain is conditional on chain fixed-effects and each retail chain's local accessibility; product choice is conditional on product characteristics. One may raise the concern that retailer choice is conditional on product characteristics as consumers may be drawn into stores by promotional prices. This is however negligible given that retailers barely offer promotions for milk.

Milk is also relatively cheap and constitutes only a small fraction of the total shopping basket. Graph 3 shows milk expenditure as a share of the total shopping trip expenditure, with the average share being less than ten per cent. Given this small share, it is unlikely that milk choice determines retailer choice. For this reason, we subsequently drop households that buy only or mainly milk on a given shopping trip. Specifically, we exclude shopping trips

for which milk expenditure makes up more than 20% of the total shopping trip expenditure.

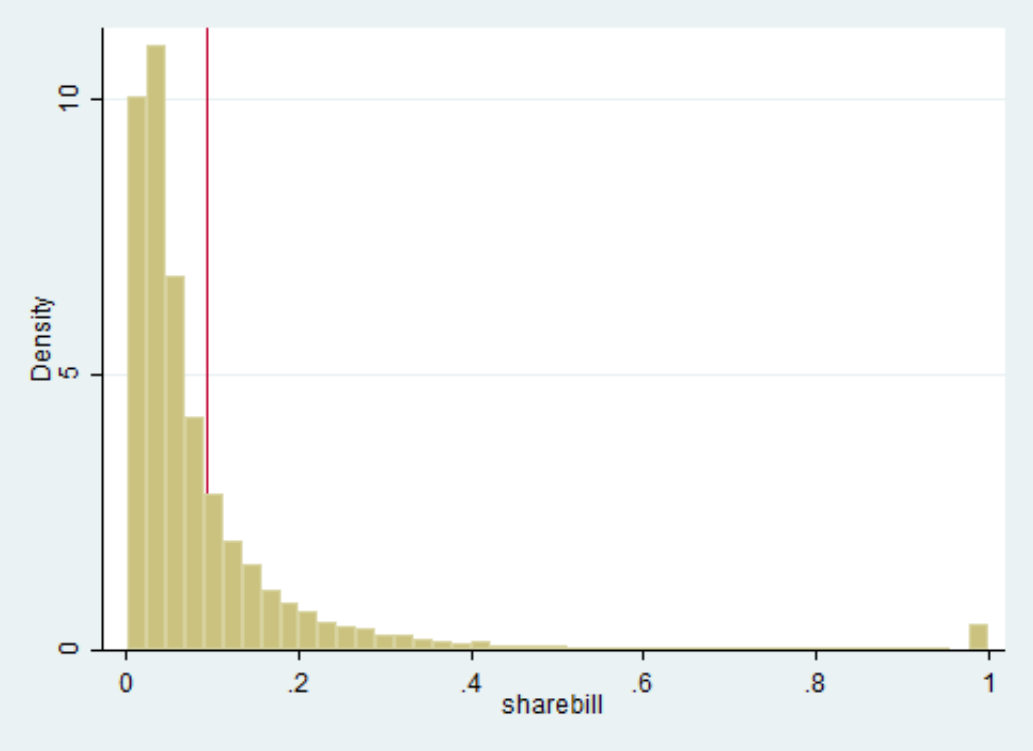


Figure 3: For every shopping trip on which milk was purchased, the graph shows milk expenditure as a share of total shopping trip expenditure.

Our second assumption about shopping behavior is that if consumers consider a retailer, they consider its entire milk assortment. Milk is a category characterized by a relatively narrow assortment.[10] Also, the milk market is a mature market and, unlike other markets, it rarely sees innovations or product introductions. The typical consumer will therefore be familiar with the assortment. Alternatively, we could introduce an additional stage in

[10] Our raw data contains a total of 204 Universal Product Codes (UPC) of milk. For example, "Brand A, full-fat fresh milk" would be a different UPC than "Brand A, half-skimmed fresh milk". The number of milk UPCs is considerably lower than, e.g., for cheese and yogurt with 939 and 1178 UPCs, respectively.

which, after having chosen a retailer choice set, the consumer further selects a consideration set among the products of the considered retailers. These consideration sets could be modeled as a function of marketing instruments (**??**) or search costs (**?**). The focus of such an additional stage would be on *cognitive* availability which is beyond the scope of this paper and left for further research.

Lastly, we have to make assumptions on the supply side in order to recover marginal costs. We choose an oligopoly model of Bertrand-Nash pricing because it corresponds closely to actual German retail pricing: Prices are typically changed on Mondays and remain constant over the rest of the week. Also, we assume that retail chains know true consumer choice sets. Given retailers' large spending on market research[11], this assumption seems justified. Alternatively, we could model retailers to observe true choice sets with a measurement error.

Estimation We estimate demand using a simulated maximum likelihood estimator (see appendix A.2). We do not specify an outside option; instead, demand is estimated conditional on milk purchase. We do so because average milk consumption remains largely constant and unaffected by variations in milk prices (see appendix A.3).

[11] For example, more than 9.6 billion US dollars were spent on market research in the US in 2012 (**?**).

Table 5: Mixed Logit: Estimation Results

Table 6: Homogeneous full choice sets

Mean		
Retailer 2	-1.1110***	(0.0251)
Retailer 3	-1.1483***	(0.0363)
Retailer 4	0.1504***	(0.0213)
Retailer 5	-1.5625***	(0.0198)
Retailer 6	-1.8542***	(0.0302)
Retailer 7	-0.6304***	(0.0370)
Retailer 8	-0.6899***	(0.0182)
Local market share	4.0427***	(0.1228)
Fresh	0.4422***	(0.0119)
Fat content	2.7147***	(0.0937)
Private Label	-2.5760***	(0.1708)
Organic	5.6178***	(0.3591)
Control Function	0.1980***	(0.0104)
Price	-1.1523***	(0.0394)
Standard Deviation		
Price	0.5372***	(0.0218)
No. of households	1261	
No. of choice occasions	37799	
Standard errors in parentheses		
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$		

Table 7: Heterogeneous choice sets (current retailer)

Mean		
Retailer 2	-1.1966***	(0.0247)
Retailer 3	-0.8936***	(0.0354)
Retailer 4	-0.1606***	(0.0202)
Retailer 5	-1.1810***	(0.0192)
Retailer 6	-1.0522***	(0.0299)
Retailer 7	-0.7550***	(0.0337)
Retailer 8	-0.3869***	(0.0169)
Local market share	3.9851***	(0.1204)
Fresh	1.3219***	(0.0460)
Fat	1.0779***	(0.0379)
Private Label	-3.7089***	(0.2491)
Organic	6.5861***	(0.5210)
Control Function	0.2329***	(0.0152)
Price	-1.4989***	(0.0417)
Standard Deviation		
Price	0.7668***	(0.0308)
No. of households	1261	
No. of choice occasions	37799	
Standard errors in parentheses		
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$		

Table 8: Mixed Logit: Heterogeneous choice sets (past 3 months)

Mean		
Retailer 2	-1.3509***	(0.0096)
Retailer 3	-1.4917***	(0.0133)
Retailer 4	-0.4380***	(0.0093)
Retailer 5	-1.6239***	(0.0094)
Retailer 6	-1.4320***	(0.0070)
Retailer 7	-0.7343***	(0.0095)
Retailer 8	-0.7493***	(0.0093)
Local market share	40.9002***	(0.2361)
Fat	3.8610***	(0.0615)
Private Label	-3.8678***	(0.1059)
Organic	8.5289***	(0.2230)
Control Function	0.2713***	(0.0066)
Price	-1.4770***	(0.0188)
Standard Deviation		
Price	0.5637***	(0.0135)
No. of households	1261	
No. of choice occasions	37799	
Standard errors in parentheses		
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$		

Tables 6-7 present the results under the three different choice set specifications, respectively. In all specifications, all coefficients are significantly different from zero at the 1% level. The price coefficient is negative as expected. It is smaller (in absolute terms) for homogeneous than heterogeneous choice sets - the intuition is that, when choice sets are assumed to be homogeneous and consumer i does not react to a price change in product j, this is rationalized by consumer i having a low price sensitivity. When we allow for heterogeneous choice sets, a non-reaction can also be attributed to consumer i not including product j in her choice set. The standard deviation of the price coefficient is significantly different from zero, thus justifying an individual-specific price coefficient.

The rest of the estimated coefficients is in large parts expected. Consumers prefer (ceteris paribus) national brands over private labels, fresh milk over UHT milk, full-fat over skimmed, and organic over conventional milk. The larger a retailer's market share, i.e. the more accessible it is, the more likely are consumers to buy from it.

Testing We use a formal test to select the choice set specification that is closest to the "true" model. This test runs in the vein of the so-called menu approach: The researcher compares the fit of multiple models from a finite set of models, e.g. to infer industry conduct (**??**). Since this work is preliminary, we test only three different choice set scenarios against each other:[12]

(A) Consumers consider all products at all retailers.

(B) Consumers consider all products at all retailers they visited in the past three months.

(C) Consumers consider all products at only the retailer they are currently shopping in.

Specification A corresponds to the standard assumption of homogeneous, full choice sets. Specification C is the other extreme: Consumers have singular choice sets where each choice set consists of only one retailer. Lastly, we test for specification B in which choice sets are shaped by purchase history. More specifically, we test for the relevant time horizon of the purchase history that determines the retailer choice set.

For each choice occasion and household in scenario B, we keep all the products from the retailer that were visited in the past three months and drop

[12] We are currently working on testing alternative scenarios for a later version of this paper, such as consumers considering those retailers they visited in the past month, or consumers considering those retailers that are within a 10 km radius around their residence.

all other retailers' products. For scenario C we keep for each choice occasion only the products of the retailer at which the household made a purchase, so if a household goes to retailer r and purchases product j, then we assume all of the products at retailer r were also in the household's choice set.

Table 9: Marginal Costs

Variable	Mean	Std. Dev.	Min.	Max.	N
price	59.99	15.483	43.891	107.47	600
Model A: Homogeneous Full Choice Sets					
cost	59.695	15.491	43.881	107.34	600
margin	0.295	0.405	-0.352	3.368	600
margin (in %)	0.515	0.741	-0.704	6.762	600
Model B: Heterogeneous Choice Sets: Retailers from past 3 months					
cost (in euro cent)	59.522	15.416	43.787	107.16	600
margin (in euro cent)	0.468	0.574	0.069	4.577	600
margin (in %)	0.792	0.945	0.074	7.252	600
Model C: Heterogeneous Choice Sets: Current Retailer					
cost (in euro cent)	59.673	15.443	43.813	107.25	600
margin (in euro cent)	0.318	0.363	0.048	2.783	600
margin (in %)	0.54	0.609	0.053	4.408	600

Using the demand estimates and solving equation 12, we recover marginal costs for all three scenarios (see Table 9). Next, we regress them on observed cost shifters. We perform a model selection test for non-nested models à la **?**. Intuitively, the best-performing scenario is the one for which the estimated corresponding marginal costs are best explained by the observed cost shifters.

We compute the Vuong test statistics (see appendix A.4) and find that the best-performing model is the model in which consumers consider only the products of the retailer they currently shop at: Model C significantly outperforms models A and B. Models A and B do not have a significantly different fit.

6 Conclusion

In this paper, we develop a novel test to infer information on typically unobserved choice sets. The main advantage of our approach is that, compared to previous approaches, it has relatively limited data requirements. Our approach requires only increasingly available micro-sales data and widely available cost shifter data. Unlike previous approaches, it does not require any data on stated choice sets.

In our application, we find that the standard model of full and homogeneous choice sets is outperformed by a model in which consumers consider only products of the retailer at which they are currently shopping. If choice set heterogeneity is ignored, demand estimates will be biased. Crucially, this bias will carry over to the supply side and affect all following policy evaluations. In this situation, our test can help to select an appropriate choice set specification and consequently reduce the estimation bias.

A Appendix

A.1 Control Function Approach

The key idea behind the control function approach is that the price is exogenous conditional on the unobserved product-specific portion of the utility. In the first step, we regress the potentially endogenous price variable on a number of instruments as well as on exogenous variables of the demand equation:

$$p_{jt} = \delta J_{jt} + \gamma W_{jt} + \lambda_r + \eta_{jt} \tag{15}$$

where J_{jt} and W_{jt} are vectors of product characteristics and cost shifters, respectively. λ is a retailer dummy and η_{jt} is an iid error-term. The error-term contains unobserved product characteristics that are neither captured by the observed product characteristics nor the cost shifters. In the second step, the residual retained from (10) is plugged into the utility function for ξ_{jt}:

$$U_{ijt} = \alpha_i p_{jt} + x_{jt}\beta_i + w_t\gamma_i + \tau\hat{\eta}_{jt} + \bar{\varepsilon}_{ijt}. \tag{16}$$

For the control function approach, we regress milk retail prices on a number of cost shifters and exogenous characteristics, namely the (German) raw-milk price index, a diesel price index, an electricity price index, the fat content, by which retailer the product is sold, whether it is sold under a private label, whether it is fresh milk, whether it is sourced organically; the fat content and retailer dummies. Table 10 displays the results from the OLS regression. The results are as expected. The retail price increases in the cost of the input factors raw milk, diesel and electricity and decreases in its fat content. Being

from a national brand or organically sourced increase the price. Most retailers sell at a higher price than the baseline retailer 1 which is a hard-discounter.

Table 10: Control Function: Results from OLS Regression

	price	
rawmilk_ger	0.217***	(0.0101)
diesel_price_index	0.0177*	(0.00856)
electricity_price_index	0.528***	(0.00762)
privlab	-16.16***	(0.110)
fresh	-0.348***	(0.0452)
fat	2.933***	(0.0211)
organic	34.11***	(0.129)
ret_2	0.298***	(0.0893)
ret_3	0.0934	(0.0963)
ret_4	0.536***	(0.0629)
ret_5	-0.232**	(0.0773)
ret_6	0.00378	(0.0927)
ret_7	1.385***	(0.0895)
ret_8	0.0814	(0.0683)
N	37799	
adj. R^2	0.995	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

From the first-stage control function approach, we recover the fitted errors $\bar{\eta}_{jt}$ and plug them into the indirect utility **??**:

$$V_{ijrt} = \alpha_i p_{jrt} + x_{jrt}\beta_i + \xi_{jrt} + \tau\hat{\eta}_{jt} + \bar{\varepsilon}_{ijrt} \tag{17}$$

A.2 Simulated Maximum Likelihood

One complication of the mixed logit model is that there is no analytic solution to the integral in equation (9), i.e. the probabilities. We approximate (9) via

simulation based on given values of $(\alpha, \beta, \sigma^\alpha, \sigma^\beta)$:

$$SP_i(\alpha, \beta, \sigma^\alpha, \sigma^\beta) = \frac{1}{R} \sum_{r=1}^{R} \left(\prod_{t=1}^{T} L_{ij(i,t)t}(\alpha^r, \beta^r) \right) \tag{18}$$

where R is the number of simulations and α^r and β^r are the r^{th} draws of the distributions $f(\alpha_i|\alpha, \sigma^\alpha)$ and $f(\beta_i|\beta, \sigma^\beta)$. We use Halton draws for faster convergence. The simulated analogue to the log-likelihood function of the entire sample is:

$$SLL(\alpha, \beta, \sigma^\alpha, \sigma^\beta) = \sum_{i=1}^{N} \ln[SP_i(\alpha, \beta, \sigma^\alpha, \sigma^\beta)] \tag{19}$$

We maximize equation (19) using the simulated maximum likelihood with respect to the coefficients $\alpha, \beta, \sigma^\alpha$ and σ^β. We use the estimated model parameters to predict probabilities $\widehat{L_{ijt}}$ and compute the average choice probability $\widehat{s_{jt}}$:

$$\begin{aligned} \widehat{s_{jt}} &= \frac{1}{C} \sum_{i=1}^{N_j} \widehat{P_{ijt}} \\ &= \frac{1}{C} \sum_{i=1}^{N_j} \int \widehat{L_{ijt}}(\alpha_i, \beta_i) f(\alpha_i) f(\beta_i) d\alpha_i d\beta_i \end{aligned} \tag{20}$$

where, again, C denotes the total number of choice occasions. The own-price elasticity of product j is then computed as

$$SE_{jjt} = -\frac{p_{jt}}{\widehat{s_{jt}}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{R} \sum_{r=1}^{R} \alpha^r \widehat{L_{ijt}}(\alpha^r, \beta^r)(1 - \widehat{L_{ijt}}(\alpha^r, \beta^r)) \right) \right] \tag{21}$$

and the cross-price elasticity of product j with respect to p_k

$$SE_{jkt} = \frac{p_{kt}}{\widehat{s_{jt}}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{R} \sum_{r=1}^{R} \alpha^r \widehat{L_{ijt}}(\alpha^r, \beta^r) \widehat{L_{ikt}}(\alpha^r, \beta^r) \right) \right] \tag{22}$$

A.3 Variation in Milk Consumption and Milk Prices

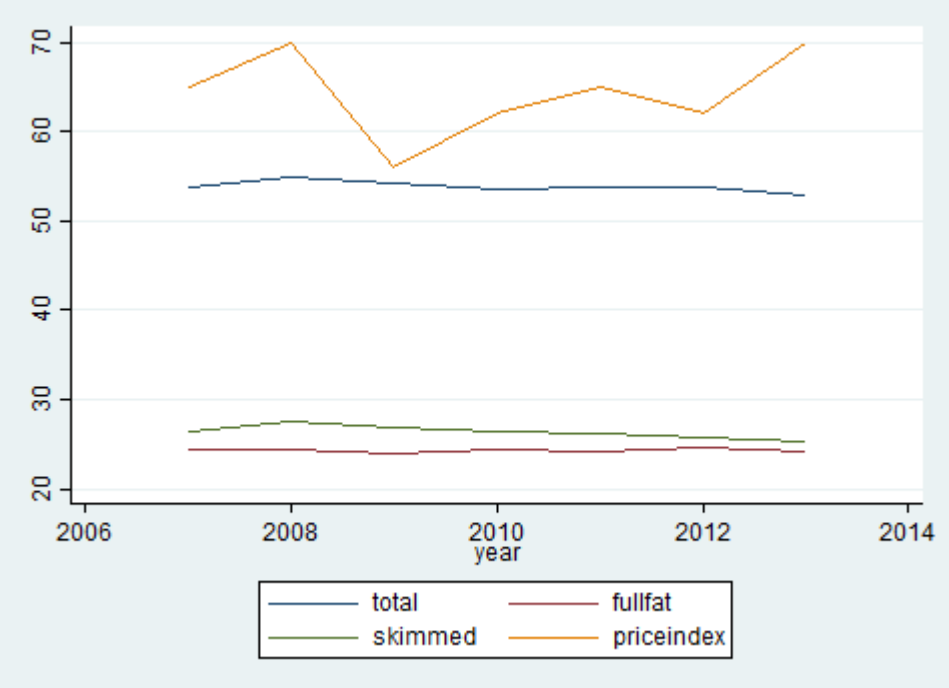


Figure 4: Variation in milk consumption and milk prices. Data Source: German Federal Ministry of Food and Agriculture.

A.4 Vuong test

We test the three models A,B and C against each other.

Table 11: Vuong test statistic

Model	Vuong test statistic	Comparison
$V(B,C)$	0.424	B $\prec$ C
$V(B,A)$	2.807	B $\prec$ A
$V(C,A)$	2.137	C $\succ$A

(A) Homogeneous choice sets. (B) Heterogeneous, past 3 months. (C) Heterogeneous, current retailer.

A.5 Elasticities

The computation of own-price elasticities is similar to those under choice set homogeneity. However, a change in product j's price will only affect a consumer's choice probability if it is included in their choice set to begin with. Otherwise the impact will be zero for that consumer. Hence, the elasticities are weighted by the probability of choosing the respective choice set.

$$\begin{aligned} E_{jjt} &= \frac{\partial s_{jt}}{\partial p_{jt}} \frac{p_{jt}}{s_{jt}} = -\frac{p_{jt}}{s_{jt}} \cdot P_{ict} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial P_{ijt|c}}{\partial p_{jt}} \right) \\ &= -\frac{p_{jt}}{s_{jt}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\int \alpha_i L_{ijt}(1 - L_{ijt}) f(\alpha_i) f(\beta_i) f(\gamma_i) d\alpha_i d\beta_i d\gamma_i \right) \right) \end{aligned} \tag{23}$$

where L_{ijt} is conditional on α_i,β_i and γ_i for all j. Cross-price elasticities are very similarly calculated:

$$\begin{aligned} E_{jkt} &= \frac{\partial s_{jt}}{\partial p_{kt}} \frac{p_{kt}}{s_{jt}} = \frac{p_{kt}}{s_{jt}} \cdot P_{ict} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial P_{ijt|c}}{\partial p_{kt}} \right) \\ &= \frac{p_{kt}}{s_{jt}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\int \alpha_i L_{ijt} L_{ikt} f(\alpha_i) f(\beta_i) f(\gamma_i) d\alpha_i d\beta_i d\gamma_i \right) \right) \end{aligned} \tag{24}$$

Cross-price elasticities under choice set heterogeneity depart from the standard elasticities in that they are only positive if a) both products are included in the same choice set, i.e. $\partial s_{jt}/\partial p_{kt} > 0$ and b) the respective choice set is considered by the consumer, i.e. $P_{ict} > 0$.[13]

Table 12: Mixed Logit: 3mo

Mean		
R2	-1.4815	(0.0134)
R3	-1.6168	(0.0140)
R4	-0.5617	(0.0122)
R5	-1.6425	(0.0154)
R6	-1.5970	(0.0133)
R7	-0.9451	(0.0117)
R8	-0.8376	(0.0133)
Local market share	34.3265	(0.3812)
fresh	0.4534	(0.0122)
fat	0.5580	(0.1307)
privlab	-0.6629	(0.2374)
organic	2.3734	(0.4990)
CF	0.0601	(0.0147)
price	0.1812	(0.1228)
price	-0.0958	(0.0129)
No. of households	1261	
No. of choice occasions	37799	

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

[13] This only applies if we consider marginal price changes. If prices change dramatically, this may influence choice set probabilities.