**Extended SKU choice modeling opportunities using a novel source of detailed big data from a large retailer**

**Abstract**

In this paper we introduce a recent “Complete Journey” dataset maintained by consulting company “84.51°” to a wide audience of economics and marketing scholars, discuss its advantages and disadvantages compared to the famous IRI marketing data set and provide an application of the new data source to the estimation of an extended microeconometric model of shredded cheese choice with multiple groups of predictors that could not be accounted for without such rich promotional and coupon metadata. We outline some ideas for future research and encourage empirical economists and marketing scientists to test their hypotheses and explore the “Complete Journey” dataset further.

Keywords: random coefficient logit, mixed logit, scanner data, demand estimation, big data, choice modeling

Acknowledgements. This paper is an output of a research project implemented within the framework of the Academic Fund Program at HSE University in 2019 — 2020 (grant №19-01-070) and within the framework of the Russian Academic Excellence Project «5-100».

**1. Introduction**

Estimating demand systems and analyzing consumer preferences is a central part of various economic studies as demand systems have widespread applications (Hoffmann and Bronnmann 2019). Scanner data has been one of the most sought after sources of data for empirical industrial organization (the examination of market power, the impact of innovations, the evaluation of new products in differentiated markets and the determination of the degree of horizontal or vertical market integration) and marketing (marketing strategies require information about how different product attributes and sociodemographic characteristics affect consumers’ purchase behavior).

Most of the literature on brand/SKU choice modeling using data on actual purchases has followed the logit modeling approach described in Guadagni and Little (1983) and applied by them to coffee market. Later, researchers used other product categories and estimation techniques. Jain, Vilcassim, and Chintagunta (1994) were the first to use a random parameters logit model to study demand for yougurt brands. Fader and Hardie (1996) proposed switching from brand-level to SKU-level choice modeling. Thunström (2010) assessed the strength of habit persistence in the cereal market, while Hoffman and Bronnmann (2019) applied a similar random coefficients logit model to model preference and response heterogeneity in the German carbonated soft drink market. The structure of most scanner datasets used in micro-level choice modeling remained basically the same: feature and/or display promotion binary variables, as well as price were the only marketing features used as predictors of choice. Some authors also accounted for brand loyalty and whether the previous purchase was made on promotion.

The use of microdata from scanners has been somewhat limited due to the lack of publicly available sources and reluctance of most retailers to disclose such data even for academic purposes. Not accidentally, since the mid-1990s a rigorous BLP model for demand estimation using aggregate rather than microdata has become dominant in empirical economics (Berry, Levinsohn, and Pakes 1995; Nevo 2000) with many applications to agricultural and health economics problems (E. Lopez and Lopez 2009; R. A. Lopez and Fantuzzi 2012) among other research areas. However, panel data on individual choice naturally remains the preferred type of data for demand estimation. Until recently the only publicly available source of microdata was the IRI Marketing Data Set first described by Bronnenberg, Kruger and Mela (2008). The data set initially contained 5 years of data, but was later expanded to include more than twice as much data (2001-2012). The bibliography associated with the dataset has been regularly updated by one of the above mentioned paper’s co-authors (Kruger 2017) and includes over 200 publications, theses, conference papers and dissertations which have used the data set. The main barrier to using this seminal dataset is a substantial handling fee of $1000, which prevents the growth of research based on such a valuable source especially for researchers from institutions with limited funding.

In this paper we describe advantages and disadvantages, as well as unique research opportunities associated with a new dataset made publicly available free of charge from customer data science company “84.51°”, a wholly owned subsidiary of The Kroger Co. – a major US retailer. We also present a preliminary version of our brand choice model using the unique features of the new dataset and some of the latest development in machine learning. The model extends the logit choice model being at the heart of many marketing science studies and allows answering the following key questions:

1. How effects of display and feature promotions on purchase probability are differentiated depending on the location of these promotions in the store and in the weekly mailer respectively?
2. How availability of a coupon for a product increases the probability of purchasing this product?

**2. Data and Model**

R package *completejourney* provides access to retail shopping transactions for 2,469 households who are frequent shoppers at a retailer over one year and originates from the 84.51° “Complete Journey 2.0” source files and also includes useful metadata on products, coupons, campaigns, and promotions[[1]](#footnote-1). It contains all of each household’s purchases, not just those from a limited number of categories. For certain households, demographic information as well as direct marketing contact history are included.

While the IRI dataset and the “Complete Journey” are not identically structured, they are similar enough to allow a comparison. Taking into account that in the IRI dataset valuable media data are available only for two categories (salty snacks and beer) and panel microdata – for all 30 categories, but for a limited geography, the only major disadvantage of the “Complete Journey” dataset is the coverage of only a single retail chain over a relatively modest period of time. One more weakness is that SKU names are not provided making it impossible to enrich the data with product characteristics found in external sources. At the same time, some of the dataset’s clear advantages are listed below:

1. More recent data (2017 as opposed to 2001-2012 in the case of the IRI data).
2. Transaction-level data as opposed to weekly store data in the case of the IRI data (panel data is available in the IRI data set only for two BehaviorScan markets).
3. Larger number of product categories (304 as opposed to 30 in the IRI dataset) allows investigating patterns in some of the categories for which information is otherwise not available in the IRI dataset: beef, pharmacy, spices and extracts, imported wine, portable electric appliances to name a few. Cigarettes is another important product category for economic research with 388 SKUs contained in the “Complete Journey” database. In additional to researching underinvestigated niches, the large number of categories in the dataset is helpful for demonstrating the generalizability of empirical results.
4. Unique feature and display promotion information containing mailer page placement and in-store display placement.
5. Unique direct marketing information on coupons that were made available to each household and coupon redemptions that can be used to measure campaign efficacy and coupon usage on specific products.

We account for the following factors influencing SKU choice:

1. Price.
2. Feature promotion presence and location in the mailer.
3. Display promotion presence and location in the store.
4. Direct marketing: whether coupon for this product was available for a given household on a given purchase occasion. While most studies neglect direct marketing activity, coupon promotions are very popular among retailers as part of their loyalty programs. At the same time direct marketing activity often correlates with trade marketing (feature and display promotions). Thus ignoring coupon promotions when assessing the effects of feature and display promotion can lead to the omitted variable bias. It is important to take into account that if the campaign associated with the coupon was active on day t, but the coupon had been redeemed before that date by household j, then coupon was no longer available to household j on day t.
5. Demographic data containing missing values is used at the post-estimation phase to find pairwise associations between individual-level parameter estimates and household characteristics.

Factors 2, 3 and 4 have not been widely used in the literature. Factor 5 is rarely available as well due to the lack of demographic data. Additional factors such as brand loyalty can be used, but primarily for categories with longer history of purchases available (Hoffmann and Bronnmann 2019; Empen, Loy, and Weiss 2015).

The set of predictors used in our analysis is presented in Table 1. First levels of all factor variables were used as reference categories in the analysis.

Table 1. Predictors of SKU choice used in the random coefficient logit model

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Type | Values/Unit of measurement | Parameter type |
| *price\_per\_oz* | numeric | price per ounce accounting for retailer’s discount and not accounting for coupon discount, USD | random |
| *campaign\_type* | factor | 1 – none, 2 – coupon from Type A campaign, 3 – coupon from Type B campaign | fixed |
| *display\_location* | factor | Display location: 0 – no display, 1 – store front, 2 – store rear, 3 – front end cap, 4 – mid-aisle end cap, 5 – rear end cap, 6 – side aisle end cap, 7 – in-aisle, 9 – secondary location display, A – in-shelf) | fixed |
| *mailer\_location* | factor | Mailer location (0 – not on ad, A – interior page feature, C –interior page line item, D –front page feature, F – back page feature, H – wrap front feature, J – wrap interior coupon, L – wrap back feature, P – interior page coupon, X – free on interior page, Z – free on front page, back page, or wrap) | fixed |
| *manufacturer\_id* | factor | ID uniquely identifying each of the manufacturer: 69 (private label), 317, 435, 988 and 1126 (national brands) | random |
| *package\_size\_oz* | numeric | Package size, oz | random |

We apply our model to a subsample of shredded cheese purchases. Shredded cheese is a popular product subcategory (73.3% of households from the dataset purchased shredded cheese over the year) which is characterized by a wide variety of mailer and display locations, relative homogeneity of unobserved product characteristics, competition between several national brands and the grocery chain’s private label (manufacturer\_id=69). In addition, this category almost perfectly meets the mutual exclusivity assumption made in discrete choice models where buyers are assumed to purchase a single item on each purchase occasion. Indeed, according to the dataset 77% of market baskets contained a single SKU of shredded cheese.

Our data contains only 12 month of purchases, which is why the history is not long enough to reliably estimate brand loyalty for each household: 29% of households made only 1 or 2 purchases in this category over the year. At the same time, we account for brand preferences by including the set of brand dummies as predictors allowing their coefficients to vary across households. In accordance with Allender and Richards (2012) only the households purchasing one SKU per purchase were considered. If households bought more than one SKU, it could not be distinguished between households that were variety seekers and households that consisted of several members with diverging preferences. To obtain reasonable individual-level parameter estimates for each household we limit our sample to households who made at least 3 purchases.

The relationship between choice probability and the explanatory variables was estimated using the random coefficients logit model that overcomes the independence of irrelevant alternatives (IIA) property of the traditional multinomial logit model (Train 2009). The probability of household *i* choosing alternative *j* in choice set *t* is computed as:

**, (1)**

where  is a vector of explanatory variables, the coefficient vector  is known to the customer but not to the researcher. It varies over customers with density  , where  represents the parameters of this distribution. For example, if  is normally distributed in the population,  would represents means and covariances. The model was fitted by using maximum simulated likelihood (Train 2009) using Stata command *mixlogit* (Hole 2007). Individual-level parameters were computed using the method proposed by Revelt and Train ( 2000) and allow both enriching the retailer’s dataset with additional variables for segmentation and targeting as well as for inferring systematic relationships between preferences and household characteristics available for a large proportion of households in the sample.

**3. Results**

Parameters estimates of two models – with feature and display promotion effects without accounting for location (1) and with feature and display promotion effects indicators accounting for location (2) – are presented in Table 2. Not accounting for location leads to a somewhat misleading conclusion that mailer promotions tend to increase the probability of purchasing an SKU. However, further analysis has shown that it depends on location: more specifically only location A (interior page feature) is significantly better than no mailer promotion at all, while the effects of locations D (front page feature) and H (wrap front feature) are insignificantly different from zero. Surprisingly, there was no evidence that display or coupon promotions were effective in the shredded cheese product category. Other things equal, two of the national brands were less preferred than the private label used as the reference category, while two other national brands had an insignificantly different probability of being chosen compared to the baseline. Other things equal (price per ounce being one of them), customers prefer smaller packages. The results were robust to changing the minimum number of purchases threshold for the inclusion of a household in our subsample.

Table 2. Random parameter logit models:

parameter estimates (subsample: households with 2 or more purchases)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (1) |  | (2) |  |
|  | choice |  | choice |  |
| display | 0.0596 | (0.0493) |  |  |
| mailer | 0.124\* | (0.0604) |  |  |
| campaign\_type\_TypeA | 0.0855 | (0.216) | 0.0743 | (0.218) |
| campaign\_type\_TypeB | -0.261 | (0.422) | -0.280 | (0.430) |
| price\_per\_oz | -1.724\*\* | (0.662) | -1.524\* | (0.647) |
| manufacturer\_id\_317 | -0.274\*\* | (0.100) | -0.338\*\*\* | (0.102) |
| manufacturer\_id\_435 | -0.218 | (0.546) | -0.640 | (0.649) |
| manufacturer\_id\_988 | -1.656\*\* | (0.510) | -1.709\*\*\* | (0.474) |
| manufacturer\_id\_1126 | -0.303 | (0.184) | -0.273 | (0.187) |
| package\_size\_oz | -0.0468\*\*\* | (0.00805) | -0.0505\*\*\* | (0.00849) |
| display\_location\_1 |  |  | -0.307 | (0.623) |
| display\_location\_2 |  |  | 0.0555 | (0.0855) |
| display\_location\_3 |  |  | 0.240 | (0.926) |
| display\_location\_4 |  |  | 0.00670 | (0.415) |
| display\_location\_5 |  |  | 0.0638 | (0.770) |
| display\_location\_6 |  |  | 0.0793 | (0.142) |
| display\_location\_7 |  |  | 0.0887 | (0.0706) |
| display\_location\_9 |  |  | -0.0340 | (0.136) |
| display\_location\_A |  |  | -0.00716 | (0.232) |
| mailer\_location\_A[[2]](#footnote-2) |  |  | 0.165\* | (0.0786) |
| mailer\_location\_D |  |  | 0.146 | (0.0885) |
| mailer\_location\_H |  |  | -0.220 | (0.231) |
| Number of obs |  |  | 20,965 | 20,965 |
| LR chi2(21) |  |  | 1016.85 | 1028.11 |
| Log likelihood |  |  | -6627.4523 | -6620.5631 |
| Prob > chi2 |  |  | <0.0001 | <0.0001 |

Standard errors in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

Empirical Bayes estimates of household-level coefficients were obtained and were shown to be correlated with one another and with some of the demographic variables at the 5% significance level. For instance, income was positively correlated with the coefficient at price, indicating that higher income households are less price-sensitive. The larger the household and the more kids it has, the stronger the preference towards larger packages of shredded cheese. Preferences towards national brands are highly correlated. Those preferring large package size are also the least price sensitive (correlation between coefficients at *price\_per\_oz* and *package\_size\_oz* is 0.735). Utilities of brand 1126 and 435 are especially highly correlated (r=0.946) implying that they are viewed as substitutes. A purely empirical, but potentially useful for marketing purposes, observation is that households comprised of 1 adult with no kids, as well as of 2 adults with 2 kids have a higher preference to brands 1126 and 988 compared to households with one adult with kids.

**4. Conclusion and future research**

In this paper we present preliminary results associated with a random coefficient logit version of an SKU choice model utilizing some of the extended information from a new source of objective real-world household panel data from scanners. To our best knowledge, this is the first empirical study regarding promotions and heterogeneous preferences in the shredded cheese market in the US. While a standard approach where only the facts of feature and display promotions are captured without accounting for their location led us to a conclusion about the effectiveness of feature promotions, when we accounted for promotion location, it turned out that not all of them are equally effective. More specifically, two out of three locations of feature promotion in the chain’s weekly mailer have been shown to provide no gains compared to having no promotion at all, emphasizing the usefulness of accounting for promotion heterogeneity. Surprisingly, we have found no evidence of the effectiveness of any of the two types of coupon promotions, as well as display promotions for the shredded cheese product category in stimulating purchases. This result is likely to be category-specific, but it implies that effects of various types of promotions in different product categories on customer activity is worth investigating further.

Individual-level parameter estimates for each household were obtained by using the Empirical Bayes method and were shown to be potentially useful for both retailers and manufacturers for better understanding of customers’ revealed preferences, as well as for using as segmentation and targeting variables. Model parameters have been shown to actually correlate with one another, implying that demand models should allow for correlated parameters. Some parameters have been shown to be correlated with household characteristics as well. Such correlations can be accounted for in promotional planning by manufacturers and sellers.

Choice modeling for specific product categories allows making valuable conclusions about the sensitivity of customer choice to various promotions and product characteristics, brand preferences, as well as linking household characteristics and household-level parameter estimates. Manufacturers may be interested in purchasing reports based on such type of choice modeling to substitute or complement conjoint studies, while retailers can benefit from making more personalized offers to households based on their revealed preferences.

Retailers should aim to have their data organized at least as good as Kroger’s data is organized in the “Complete Journey” dataset to extract maximum insight about the effectiveness of various promotional activities. The main limitation of the “Complete Journey” dataset is that SKUs are anonymized in such a way that a researcher cannot enrich their models with product characteristics not contained in the “Complete Journey” dataset. While shredded cheese is a relatively homogeneous product, this limitation may be more important for other, more heterogeneous, products. In addition, the availability of SKUs in each store can only be inferred by looking up whether the SKU was purchased at least once in a given store on a given week. Such inference is more or less reliable only for product categories with a short purchase cycle.

Our future efforts in SKU-level choice modeling will be concentrated around building more accurate machine learning models of product choice enhanced by using model-agnostic techniques for model interpretation. We expected such models to be good at overcoming some limitations of traditional models originating from the random utility theory, especially at working with high dimensional data and accounting for cross-price and promotion effects more flexibly.

The “Complete Journey” dataset allows verifying some influential models related to brand choice, such as The Dirichlet (aka NBD-Dirichlet) model describing the purchase incidence and brand choice of consumer products. While SKU/brand choice models are of special interest to manufacturers, for retailers it is more important to know the effects of their direct marketing campaigns on customer value. A possible direction for future research using the same dataset is the identification of causal effects of promotions on customer value outcomes (e.g. inferring whether households who received coupon promotions became more active buyers overall). To control for possible non-random assignment of households to campaigns, it will be important to match households from the treatment and the control groups carefully using their pre-treatment behavior and demographic characteristics.

**5. References**

Allender, William J, and Timothy J Richards. 2012. “Brand Loyalty and Price Promotion Strategies: An Empirical Analysis.” *Journal of Retailing* 88 (3): 323–42.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. “Automobile Prices in Market Equilibrium.” *Econometrica* 63 (4): 841–90.

Bronnenberg, Bart J, Michael W Kruger, and Carl F Mela. 2008. “Database Paper—The IRI Marketing Data Set.” *Marketing Science* 27 (4): 745–48.

Empen, Janine, Jens-Peter Loy, and Christoph Weiss. 2015. “Price Promotions and Brand Loyalty.” *European Journal of Marketing*.

Fader, Peter S, and Bruce G S Hardie. 1996. “Modeling Consumer Choice among SKUs.” *Journal of Marketing Research* 33 (4): 442–52.

Guadagni, Peter M, and John D C Little. 1983. “A Logit Model of Brand Choice Calibrated on Scanner Data.” *Marketing Science* 2 (3): 203–38.

Hoffmann, Julia, and Julia Bronnmann. 2019. “Bottle Size Matters: Heterogeneity in the German Carbonated Soft Drink Market.” *Agribusiness* 35 (4): 556–73.

Hole, Arne Risa. 2007. “Fitting Mixed Logit Models by Using Maximum Simulated Likelihood.” *The Stata Journal* 7 (3): 388–401.

Jain, Dipak C, Naufel J Vilcassim, and Pradeep K Chintagunta. 1994. “A Random-Coefficients Logit Brand-Choice Model Applied to Panel Data.” *Journal of Business & Economic Statistics* 12 (3): 317–28.

Kruger, Michael W. 2017. “Research Use of the IRI Marketing Data Set: Bibliography.” *Available at SSRN 2342688*.

Lopez, Elena, and Rigoberto A Lopez. 2009. “Demand for Differentiated Milk Products: Implications for Price Competition.” *Agribusiness: An International Journal* 25 (4): 453–65.

Lopez, Rigoberto A, and Kristen L Fantuzzi. 2012. “Demand for Carbonated Soft Drinks: Implications for Obesity Policy.” *Applied Economics* 44 (22): 2859–65.

Nevo, Aviv. 2000. “A Practitioner’s Guide to Estimation of Random-Coefficients Logit Models of Demand.” *Journal of Economics and Management Strategy* 9 (4): 513–48.

Revelt, David, and Kenneth Train. 2000. “Customer-Specific Taste Parameters and Mixed Logit: Households’ Choice of Electricity Supplier.”

Thunström, Linda. 2010. “Preference Heterogeneity and Habit Persistence: The Case of Breakfast Cereal Consumption.” *Journal of Agricultural Economics* 61 (1): 76–96.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge university press.

1. <https://CRAN.R-project.org/package=completejourney> [↑](#footnote-ref-1)
2. While all possible display locations were used in the studied product category, only mailer locations A, D and H were used. [↑](#footnote-ref-2)