Advanced Econometrics of Qualitative Data

Course by Thierry Kamionka at ENSAE ParisTech

Project paper: Choice models for brands of peanut butter and ketchup

Alexandre Combessie* Pierre Foulquie† Ismail Machraoui‡

January 29, 2016

Abstract

This project paper intends to explore the purchasing behaviors of a sample of American households regarding the choice of brands of two food products: ketchup and peanut butter. We study the econometric specification of three choice models, namely Multinomial Logit, Nested Multinomial Logit and Competing Risk. The first two models are implemented with the programming language R, with similar results. We confirm that pricing, promotion and advertising have statistically significant impacts on choices. Finally, we analyze the varying effects of household income and time since last purchase for each brand.

Introduction

The question of modeling household choices is central to both econometrics research and business life. We chose the subject of this project paper in order to apply our knowledge of econometrics of qualitative data to a practical case of choice between different brands of two popular food product categories in the USA: ketchup and peanut butter. This topic has special importance in the retail industry and relates to quantitative marketing.

*Email address: alexandre.combessie@ensae.fr
†Email address: pierre.foulquie@ensae.fr
‡Email address: ismail.machraoui@ensae.fr

It can be useful to estimate consumer demand, refine pricing, and optimize marketing campaigns in terms of socio-demographic targeting, couponing and advertising.

For this project, we leverage a dataset used by Ching, Erdem and Keane [1] in their article on a new "Price Consideration" model for brand choice. The dataset originates from the data analytics company Nielsen based on scanner data, i.e. the monitoring of purchase behaviors of a sample of American households through their use of an electronic scanner. We also leverage the vast body of literature on the subject, in particular [2, 3, 4, 5, 6, 7].

Our paper is organized as follows. In section 1, we examine the two datasets available in order to characterize their structure, the distribution of observations, brand choices and the duration of no-purchase spells. Afterwards in section 2, we describe the specifications used for the chosen econometric models: Multinomial Logit, Nested Multinomial Logit and Competing Risk models. Finally in section 3, we interpret the empirical results from the two models implemented and applied to our two datasets, i.e Multinomial Logit and Nested Multinomial Logit. These results will enable us to highlight which individual or product specific factors are most important for brand choices.

1 Descriptive analysis of the datasets

Two datasets are at our disposal in order to accomplish our work. Each dataset concerns only one category, namely ketchup and peanut butter, containing equally 4 brands. Both of them contain the same attributes for individuals and/or for brands, such as: household id, price of each 4 available brands, number of members of a household, store and market id, household income, education level, brand on feature in a flyer, brand on display and coupon availability.

The ketchup category includes data regarding 3189 households during 114 weeks, while the peanut butter category encompasses 7924 households during 51 weeks. It should be noted that households could choose whether to buy or not during those weeks. Moreover, it is important to mention that each household observation is his only weekly record. In other words, it is assumed that each household has been going to a single mall in a single store for its purchasing task.

In the following subsections, we will try to highlight some interesting characteristics of our datasets that could help us afterwards into implementing our choice models.

1.1 Purchase observations per household

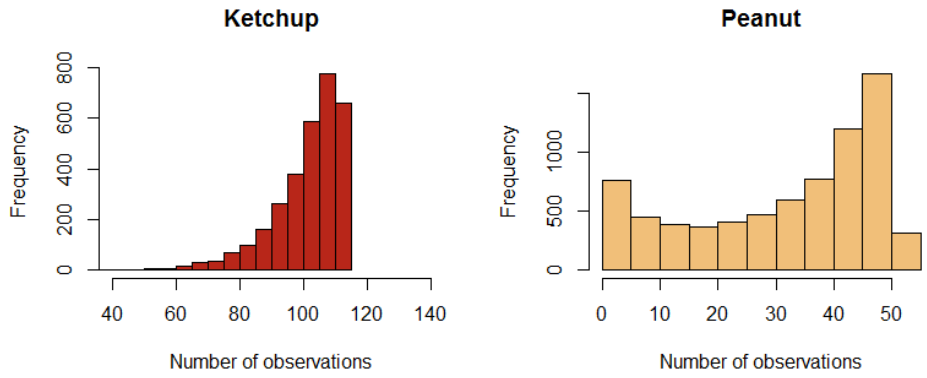

Figure 1: Distribution of number of observations per household

Regarding the ketchup category, we see that most households have been recorded between 80 and 120 times, while for the peanut butter category the distribution is relatively uniform between 0 and 50.

1.2 Brand market shares

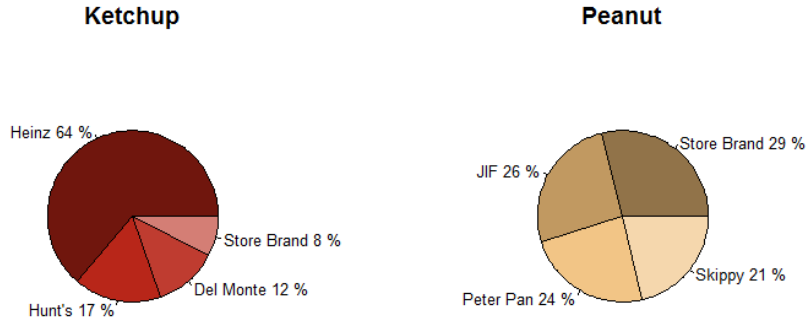

Figure 2: Share of purchases per brand

It is striking to realize that one ketchup brand, *Heinz*, stands out as the people's favourite. As for peanut butter, all brands seem to share the market equally: no brand is standing out from the rest.

1.3 No-purchase spells

Using the number of each observation week combined with the household decision of purchasing or not, we were able to determine the distribution of no-purchase spells for each household.

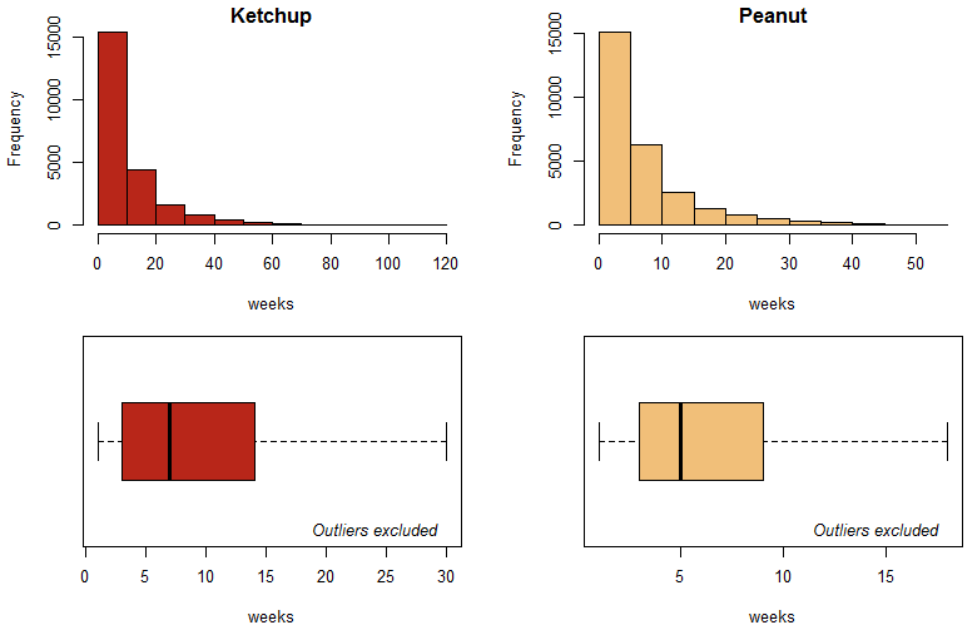

Figure 3: Distribution of no-purchase spells

For both ketchup and peanut butter, the spell distribution decreases exponentially. This clearly reflects the *inventory* effect: households have a regular consumption of these products during their meals, and they need to buy more once it is finished. From box plot figures, we notice that ketchup's no-purchase spells median is 7 weeks, to be compared with 5 weeks for peanut butter.

Since the data sets do not include this state of inventory, we can consider no-purchase spells as a good proxy for this missing data. It effectively represents the time gap since the last purchase. We will thus include it in our models, which means dropping the first observations for each household, which suffer from a left censoring bias[1]. We will see in the next section that this inventory effect is statistically significant.

[1]Following recommendations given in [1], we chose to drop the first 20 observations for ketchup and the first 10 observations for peanut, for each household.

2 Econometric model specifications

Three methods were chosen here as relevant econometric methods regarding our study. First, we set up a Multinomial Logit model to explain the purchase choice of households. Second, a Nested Multinomial Logit is assessed. The third model to be estimated is a Competing Risk model.

2.1 Multinomial Logit

This model is the most common model to deal with problems of discrete choices in econometrics. In our case, we aim to model a non-ordered qualitative variable, the choice of a household among several brands or the absence of purchase, as a function of features describing the household and brands.

Let's begin with some notations. As in [7], we write this equation to refer to the utility yielded to the household i when it chooses j at time t:

$$U_{ijt} = z'_{ijt}\beta + \epsilon_{ijt} \quad i = 1, ..., N \quad j = 1, ..., J \quad t = 1, ..., T \tag{1}$$

where U_{ijt} is the utility, z_{ijt} is the matrix of features, β is the vector of parameters and ϵ_{ijt} is the error term. The idea of the multinomial logit model is that household i will choose option j at time t if it provides the highest utility ($\Pr[U_{ijt} \geq U_{ikt}], \forall k \neq j$).

As shown in the previous part of this work, our dataset contains 2 sets of variables, individual and choice specific variables. The first one only changes across households while the latter vary across brands. The features' matrix can thus be separated in two parts: w_i and x_{ijt}, which respectively contain the individual and choice specific variables in such way that $z_{ijt} = [w_i, x_{ijt}]$. Accordingly, the vector of parameters β can be separated in two subvectors δ and γ respectively associated to w_i and x_{ijt} : $\beta = [\delta, \gamma]$.

The purchase process of households also depends on non-observable characteristics of brands and households tastes. These characteristics will be embodied by fixed effects in our model: α_{ij}. The last variable to be included in our model is the purchase gap defined in the previous part. It captures the inventory effect, that is to say the effect of the duration since the last purchase of the choice of the household. Given that we consider a logit model, the error terms are assumed to be independently and identically

distributed from a standard logistic distribution.

Hence, the probability that household i chooses option j at time t is given by :

$$\Pr[Y_{it} = j | z_{ijt}, \alpha_{ij}] = \frac{e^{z'_{ijt}\beta + \alpha_{ij}}}{\sum_{k=1}^{J} e^{z'_{ikt}\beta + \alpha_{ik}}}$$

$$\iff \Pr[Y_{it} = j | w_i, x_{ijt}, \alpha_{ij}] = \frac{e^{w'_i\delta + x'_{ijt}\gamma + \alpha_{ij}}}{\sum_{k=1}^{J} e^{w'_i\delta + x'_{ikt}\gamma + \alpha_{ik}}} \tag{2}$$

The model is estimated through the maximization of the likelihood associated to it.

The distinction between individual and choice specific features matters for the interpretation of the results of the model. As for the individual specific variables, the estimated parameters are used to compute odds ratios. These ratios are computed for each variable for each choices, except the one that is taken as a reference in our model. They are determined the following way :

$$ln\Big(\frac{\Pr[Y_i = j]}{\Pr[Y_i = k]}\Big) = w'_i(\beta_j - \beta_k) \tag{3}$$

Concerning choice specific variables, their interpretation is done through the study of marginal effects of a modality to the probability that a household formulates a particular choice. Given that these features take a different value depending on the choice of the household, marginal effects can be computed for every other choices. For one particular feature, it gives the marginal effect on the probability of formulating the initial choice of having the value of another choice for this feature. For a choice j, the marginal effect of choice specific features x_k is given by :

$$\frac{\delta \Pr[Y_i = j]}{\delta x_{ik}} = \beta \Pr[Y_i = j] \times \big(\mathbb{1}_{(j=k)} - Pr[Y_i = k]\big), \quad k = 1, ..., J \tag{4}$$

Alternatively, we can use elasticities of probabilities. The effect of having modality m from choice k on the probability to choose j is given by :

$$\frac{\delta \log \Pr[Y_i = j]}{\delta \log x_{ikm}} = \beta_m x_{ikm}\big(\mathbb{1}_{(j=k)} - Pr[Y_i = k]\big), \quad k = 1, ..., J \tag{5}$$

2.2 Nested Multinomial Logit

Nested Multinomial Logit models can be seen as a generalization of the previously described Multinomial Logit model. It allows to emphasize correlations between choices by grouping some of the alternatives.

In a general case, we would consider j choice alternatives, $j = 1, ...J$, and k nests, $k = 1, ..., K$, denoted as $N_1, ..., N_K$. We start from the expression of utility of an household given by (1). It is in the definition of the probability to formulate choice j at time t that we include the affiliation of the household i to a nest N_k :

$$\Pr[Y_{it} = j] = \Pr[Y_i \in N_k] \Pr[Y_{it} = j | Y_i \in N_k] \tag{6}$$

where

$$\Pr[Y_i \in N_k] = \frac{e^{\tau_k IV_k}}{\sum_m e^{\tau_m IV_m}}$$

and

$$\Pr[Y_{it} = j | Y_i \in N_k] = \frac{e^{\frac{1}{\tau_k}} V_{jt}}{e^{IV_k}}$$

τ_k refers to dissimilarity parameters, measures of degrees of dissimilarity between the alternatives composing nest k. IV_k denotes the inclusive value of belonging to nest N_k:

$$IV_k = log \sum_{l \in N_k} e^{\frac{1}{\tau_k} V_{lt}} \quad \text{where} \quad V_{jt} = w_i' \delta + x_{ijt}' \gamma$$

The model is then estimated by a likelihood maximization procedure.

As for the multinomial logit model, we cannot interpret the coefficients in the model as such. For households in nest N_k, computing the elasticity of having modality m from choice l of the probability to choose j gives:

$$\frac{\delta \Pr[Y_{it} = j | Y_i \in N_k]}{\delta x_{ilm}} = \beta_m \Big(\mathbb{1}_{(Y_i \in N_k)} \Big(\mathbb{1}_{(j=l)} - \Pr[j|N_k] + \tau_k \big(\mathbb{1}_{(Y_i \in N_k)} - \Pr[N_k] \Pr[j] \big) \Big) \Big)$$

In our work, two obvious nests can be formed by separating observations regarding a binary purchasing choice. The first nest is constituted of observations of households where no purchase is recorded, while the second is formed of observations where households have purchased one of the four brands.

2.3 Competing Risk Model

Competing risk models are a case of duration models where different scenarios can apply at the end of a spell. Applying this kind of model in this work, we intend to take into account both the duration of the no-purchase spell and the choice made at the end of this spell. The competing risks are thus the four brands available to households. The difference of such a model with the usual duration model is the differentiation between the various functions composing the model, with respect to the "risks". For each brand, we define the associated hazard, survival and density functions.

For time $t = 1, ..., T$ and covariates Z, we denote the overall hazard rate as :

$$\lambda(t, Z) = \lim_{dt \to 0} \frac{\Pr[t \leq T < t + dt | T \geq t, Z]}{dt}$$

which is constituted of cause-specific hazard rates, namely the chance of ending the no-purchase spell by choosing brand $M = j$:

$$\lambda_j(t, Z) = \lim_{dt \to 0} \frac{\Pr[t \leq T < t + dt, M = j | T \geq t, Z]}{dt}$$

Given that households are assumed to either continue their spell or buy one of the four brands we consider, hazard rates are such that:

$$\lambda(t, Z) = \sum_{j=1}^{J} \lambda_j(t, Z)$$

Note that the covariate vector Z can be separated between brand and individual specific features as in before with logit models. Given that some of the features we dispose of are time-varying, we separate Z in two subvectors $X(t)$ and W, where $X(t)$ contains time-varying variables and W contains time-invariant variables. Hence, we have $Z = [X(t), W]$.

Similarly, we define the overall survival function as :

$$S(t, X(t), W) = e^{-\Lambda(t, X(t), W)}$$

where $\Lambda(t, X(t), W) = \int_0^t \lambda(u, X(u), W) \mathrm{d}u$ is the cumulative risk obtained when we integrate the overall hazard rate λ. $S(t, X(t), W)$ describes the probability to make no purchase at all during the period of observation.

The overall survival function can be differenciated with respect to the brand that is chosen by the household at the end of the spell of no purchase:

$$S_j(t, X(t), W) = e^{-\Lambda_j(t, X(t), W)}$$

where $\Lambda_j(t, X(t), W) = \int_0^t \lambda_j(u, X(u), W)\mathrm{d}u$ is the cumulative risk for brand j.

Finally, we define brand-specific densities, which stand for the unconditional hazard that an household purchases brand j at time t:

$$f_j(t, X(t), W) = \lambda_j(t, X(t), W)S(t, X(t), W)$$

As for hazard rate functions, we have:

$$f(t, X(t), W) = \sum_{j=1}^{J} f_j(t, X(t), W)$$

We choose to use a Cox regression to estimate this model. Thus, no assumption is made regarding the baseline hazard. The estimation is done through the maximization of its associated partial likelihood function, built from the densities f_j.

For the estimation, we use the Nelson-Aalen estimator, which demands the definition of two variables: n_{ij}, the number of households that are "at risk" to buy brand j just before t_{ij}, and d_{ij}, the effective number of households that buy brand j at time t_{ij}. The estimator is then given by :

$$\hat{\Lambda}_j(t) = \sum_{i:t_{ij}<t} \frac{d_{ij}}{n_{ij}}$$

3 Empirical model results

In this part, we present results of the implementation of the Multinomial Logit and its nested version. Unfortunately, we were not able to get results with the Competing Risk model in R. This model turned out to be too complex and computationally intensive with 5 possible choices, because of the size of the transition matrix.

3.1 Multinomial Logit

	Brand choice per category	
	Ketchup	**Peanut**
Brand 1: (intercept)	-3.229^{***}	-3.864^{***}
Brand 2: (intercept)	-4.802^{***}	-3.768^{***}
Brand 3: (intercept)	-5.615^{***}	-4.295^{***}
Brand 4: (intercept)	-5.090^{***}	-3.800^{***}
Price	-1.276^{***}	-0.499^{***}
Product on feature	2.118^{***}	1.932^{***}
Product on display	1.057^{***}	1.543^{***}
Coupon availability	1.639^{***}	1.807^{***}
Brand 1: Household income	0.016^{***}	-0.001
Brand 2: Household income	0.014^{***}	0.020^{***}
Brand 3: Household income	0.038^{***}	0.022^{***}
Brand 4: Household income	0.011^{*}	0.034^{***}
Brand 1: Household size	0.326^{***}	0.306^{***}
Brand 2: Household size	0.378^{***}	0.269^{***}
Brand 3: Household size	0.345^{***}	0.223^{***}
Brand 4: Household size	0.338^{***}	0.232^{***}
Brand 1: No-purchase gap	0.065^{***}	0.034^{*}
Brand 2: No-purchase gap	0.102^{**}	0.071^{***}
Brand 3: No-purchase gap	0.022	0.015
Brand 4: No-purchase gap	0.131^{***}	0.083^{***}
Log Likelihood	$-76{,}376.510$	$-82{,}672.410$
LR Test (df = 20)	$18{,}629.510^{***}$	$9{,}377.721^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 1: Key results from the Multinomial Logit models[2]

Our model specification seems appropriate, as we can reject the hypothesis of joint nullity of coefficients at 99% level.

[2]See complete results in appendix.

Taken separately, almost all of our coefficients are statistically significant at 95% level, except for one brand-specific no-purchase gaps. We can thus interpret our coefficients in terms of sign and side-by-side comparison. Both models are equally likely with respect to the order of log-likelihood (or R^2 and AIC criterion). Note that we transformed income from an ordered qualitative factor ranging from 1 to 14 into a numeric variable. We found that it facilitates the interpretation without oversimplifying the results.

To begin with, all signs are behaving according to our natural intuition, for both ketchup and peanut:

- Price coefficients (elasticities) are negative;
- Advertising and promotion have a positive impact on purchases, through features in a flyer, in-store displays or coupons;
- Larger and wealthier families buy more products, from all brands;
- No-purchase gaps have a positive impact on purchases, which confirms the inventory effect, as households need regular "refills" of their favorite products.

Besides, interesting insights can be derived from the comparison of coefficients. For instance, consumers seem to more sensitive to the price of ketchup than the price of peanut butter. It may indicate that peanut butter is more of a "base product" than ketchup. Second, feature is more efficient than display in terms of advertising, for both ketchup and peanut butter.

Third, all brands are not equal with respect to income: it seems that richer households prefer Brand 3 (Del Monte) for ketchup, contrary to Brand 1 (Store Brand) for peanut butter. For ketchup too, the Store Brand (Brand 4) is less preferred by richer households than other brands. That would confirm the hypothesis that consumers are ready to pay a premium for non-generic brands.

Last but not least, ketchup seems to appeal more to larger families than peanut butter: for all brands, household size coefficients are larger in the ketchup case. Lastly, the inventory effect is more significant for ketchup than for peanut butter, for all brands except Brand 2. It would mean that ketchup requires more regular refills than peanut butter. However, in the case of Brand 3 (Del Monte for ketchup and Peter Pan for peanut butter) this inventory effect disappears. A possible explanation would be that these brands are not purchased regularly, but as occasional trials of new brands.

3.2 Nested Multinomial Logit

	Brand choice per category	
	Ketchup	**Peanut**
Brand 1: (intercept)	-3.148^{***}	-5.086^{***}
Brand 2: (intercept)	-5.366^{***}	-5.107^{***}
Brand 3: (intercept)	-6.489^{***}	-6.089^{***}
Brand 4: (intercept)	-5.775^{***}	-5.047^{***}
Price	-1.643^{***}	-0.506^{***}
Product on feature	2.524^{***}	2.821^{***}
Product on display	1.190^{***}	2.404^{***}
Coupon availability	1.894^{***}	2.424^{***}
Brand 1: Household income	0.016^{***}	-0.016^{**}
Brand 2: Household income	0.015^{**}	0.022^{***}
Brand 3: Household income	0.046^{***}	0.024^{***}
Brand 4: Household income	0.007	0.045^{***}
Brand 1: Household size	0.331^{***}	0.334^{***}
Brand 2: Household size	0.382^{***}	0.273^{***}
Brand 3: Household size	0.342^{***}	0.218^{***}
Brand 4: Household size	0.328^{***}	0.213^{***}
Brand 1: No-purchase gap	0.061^{**}	0.022
Brand 2: No-purchase gap	0.112^{**}	0.088^{***}
Brand 3: No-purchase gap	-0.001	0.006
Brand 4: No-purchase gap	0.139^{**}	0.088^{***}
Inclusive value (2nd Nest)	1.504^{***}	1.915^{***}
Log Likelihood	$-76{,}189.760$	$-82{,}425.270$
LR Test (df = 21)	$19{,}003.020^{***}$	$9{,}872.008^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2: Key results from the Nested Multinomial Logit models[3]

This model is relatively similar to the previous one: the same coefficients are available, with an additional one to account for the second nest, which encompasses actual purchases. As before, our specification seems correct, rejecting the joint nullity of coefficients at 99% level and the nullity of each coefficient taken separately at 95% level, except one for the no-purchase gap. We can interpret our coefficients as before in terms of sign and side-by-side comparison. The nest-specific coefficient is statistically significant at 99%, which confirms the interest of adding the nest structure to the Multinomial Logit specification. However, the likelihood of the model is only slightly higher than before, when studying the log-likelihood, the R^2 or the AIC criterion.

[3]See complete results in appendix.

Note that we performed the same factor-to-numeric transformation on the household income variable as before.

The signs of statistically significant coefficients are all the same, which confirms our previous observations: negative price elasticities, positive impact of advertising and promotion, increased effects for larger and wealthier families, and positive inventory effect.

The comparison of coefficients leads to similar insights, with higher intensity. The price sensitivity differential between ketchup and peanut butter is still present, but even more important. The intensity of advertising and promotion is also more intense, and the efficiency gap between feature and display is higher.

As for income, richer households still prefer Brand 3 (Del Monte) for ketchup, and despise Brand 1 (Store Brand) for peanut butter. This time the coefficient differentials are more marked. Richer households clearly purchase less of Store Brand products: the coefficient for ketchup is not significant, and that of peanut butter is negative.

In the case of household size, the coefficients are very similar, there are no additional comments to be made. The same applies to the inventory effect (no-purchase gap coefficients). It still has a larger effect for ketchup, which disappears for Brand 3.

Conclusion

Using panel data on purchases, pricing, advertising, promotion, and socio-demographic characteristics, we were able to model the choice between several brands. In order to do so, we implemented usual econometric techniques to model discrete choice: a Multinomial Logit and its nested version.

The Multinomial Logit model provides intuitive results regarding pricing, advertising, households' income and inventory effects. Assessments drawn from it are slightly varying between product categories and brands. The Nested Multinomial Logit, which separates observations regarding the act of purchase, confirms insights from the previous model.

In addition of detailing the econometric specification of several classic models in the context of brand choices, our study confirms the relevance of these models to better understand consumers' behaviors. These models are easy to interpret and can be of interest to food producers to optimize their marketing strategy, especially pricing, socio-demographic positioning, advertising, and promotion.

To go further, i.e. to be able to predict changes in consumers' behaviors over time, newer econometric models would be needed, such as the Price Consideration model [1] and Dynamic Discrete Choice models [8]. Although it can improve the model fit, in particular for inter-purchase spells, these models bring new computational challenges [9, 10]. Recent research is done to apply Statistical Learning models to this field [11, 12] which are less interpretable but improve the model fit, seen as a predictive quality.

A Appendix: complete model summaries

A.1 Multinomial Logit

	Dependent variable:	
	Brand choice	
	Ketchup	**Peanut**
Brand 1: (intercept)	-3.229^{***}	-3.864^{***}
	(0.079)	(0.075)
Brand 2: (intercept)	-4.802^{***}	-3.768^{***}
	(0.103)	(0.092)
Brand 3: (intercept)	-5.615^{***}	-4.295^{***}
	(0.134)	(0.101)
Brand 4: (intercept)	-5.090^{***}	-3.800^{***}
	(0.105)	(0.088)
Price	-1.276^{***}	-0.499^{***}
	(0.057)	(0.036)
Product on feature	2.118^{***}	1.932^{***}
	(0.020)	(0.025)
Product on display	1.057^{***}	1.543^{***}
	(0.027)	(0.039)
Coupon availability	1.639^{***}	1.807^{***}
	(0.103)	(0.151)
Brand 1: Household income	0.016^{***}	-0.001
	(0.003)	(0.004)
Brand 2: Household income	0.014^{***}	0.020^{***}
	(0.005)	(0.005)
Brand 3: Household income	0.038^{***}	0.022^{***}
	(0.008)	(0.005)
Brand 4: Household income	0.011^{*}	0.034^{***}
	(0.006)	(0.004)
Brand 1: Household size	0.326^{***}	0.306^{***}
	(0.007)	(0.009)
Brand 2: Household size	0.378^{***}	0.269^{***}
	(0.012)	(0.010)
Brand 3: Household size	0.345^{***}	0.223^{***}
	(0.018)	(0.011)
Brand 4: Household size	0.338^{***}	0.232^{***}
	(0.015)	(0.009)
Brand 1: No-purchase gap	0.065^{***}	0.034^{*}
	(0.023)	(0.019)
Brand 2: No-purchase gap	0.102^{**}	0.071^{***}
	(0.040)	(0.017)
Brand 3: No-purchase gap	0.022	0.015
	(0.062)	(0.022)
Brand 4: No-purchase gap	0.131^{***}	0.083^{***}
	(0.047)	(0.016)
Observations	252,657	169,356
McFadden R^2	0.109	0.054
Log Likelihood	$-76{,}376.510$	$-82{,}672.410$
LR Test (df = 20)	$18{,}629.510^{***}$	$9{,}377.721^{***}$
AIC	152,793	165,385
Note:	*p<0.1; **p<0.05; ***p<0.01	

A.2 Nested Multinomial Logit

	Dependent variable:	
	Brand choice	
	Ketchup	**Peanut**
Brand 1: (intercept)	-3.148^{***}	-5.086^{***}
	(0.081)	(0.131)
Brand 2: (intercept)	-5.366^{***}	-5.107^{***}
	(0.133)	(0.152)
Brand 3: (intercept)	-6.489^{***}	-6.089^{***}
	(0.187)	(0.175)
Brand 4: (intercept)	-5.775^{***}	-5.047^{***}
	(0.146)	(0.140)
Price	-1.643^{***}	-0.506^{***}
	(0.057)	(0.044)
Product on feature	2.524^{***}	2.821^{***}
	(0.029)	(0.049)
Product on display	1.190^{***}	2.404^{***}
	(0.031)	(0.059)
Coupon availability	1.894^{***}	2.424^{***}
	(0.111)	(0.200)
Brand 1: Household incomee	0.016^{***}	-0.016^{**}
	(0.003)	(0.007)
Brand 2: Household income	0.015^{**}	0.022^{***}
	(0.007)	(0.007)
Brand 3: Household income	0.046^{***}	0.024^{***}
	(0.012)	(0.008)
Brand 4: Household income	0.007	0.045^{***}
	(0.010)	(0.007)
Brand 1: Household size	0.331^{***}	0.334^{***}
	(0.009)	(0.015)
Brand 2: Household size	0.382^{***}	0.273^{***}
	(0.018)	(0.018)
Brand 3: Household size	0.342^{***}	0.218^{***}
	(0.026)	(0.020)
Brand 4: Household size	0.328^{***}	0.213^{***}
	(0.022)	(0.016)
Brand 1: No-purchase gap	0.061^{**}	0.022
	(0.027)	(0.035)
Brand 2: No-purchase gap	0.112^{**}	0.088^{***}
	(0.052)	(0.034)
Brand 3: No-purchase gap	-0.001	0.006
	(0.088)	(0.039)
Brand 4: No-purchase gap	0.139^{**}	0.088^{***}
	(0.066)	(0.029)
Inclusive value (2nd Nest)	1.504^{***}	1.915^{***}
	(0.032)	(0.057)
Observations	252,657	169,356
R^2	0.111	0.057
Log Likelihood	$-76,189.760$	$-82,425.270$
LR Test (df = 21)	$19,003.020^{***}$	$9,872.008^{***}$
AIC	152,421	164,892
Note:	*p<0.1; **p<0.05; ***p<0.01	

B Appendix: R scripts implemented

B.1 Data loading and preparation

```
lib<-c("plm","mlogit", "mnlogit", "mstate","pbapply","stargazer")
sapply(lib, require, character.only = TRUE, quietly=T)

features<-c("hhold_id","brand","i_obs","m_obs","hhold_income","hhold_members",
            "hhold_male_educ","hhold_female_educ","erim_market","store_id","n_wk", "ind_feature",
            "ind_display","Coupon_val","Coupon_av","P.1","P.2","P.3","P.4","ft.1","ft.2","ft.3","ft.4",
            "dp.1","dp.2","dp.3","dp.4","cp_av.1","cp_av.2","cp_av.3","cp_av.4")
totalclasses<-c(rep("factor",2),rep("numeric",2),"factor","numeric",
                rep("factor",3),"numeric","numeric",rep("factor",2),
                rep("numeric",6),rep("factor",8),rep("numeric",4))
df_ketchup <- read.table("./ketchup_est_m.txt",col.names = features, colClasses = totalclasses)
df_peanut <- read.table("./pbut_est_m.txt", col.names = features, colClasses = totalclasses)

length_ketchup<-nrow(df_ketchup)
P.0<-rep(0.0,length_ketchup)
ft.0<-rep(0.0,length_ketchup)
dp.0<-rep(0.0,length_ketchup)
cp_av.0<-rep(0.0,length_ketchup)
df_ketchup<-cbind(df_ketchup,P.0,ft.0,dp.0,cp_av.0)

# [Ketchup] gaps computing
gap_ketchup<-append(df_ketchup[-1,11],0)-df_ketchup[,11]
gap_ketchup[gap_ketchup<0]=0
df_ketchup <- cbind(df_ketchup,gap_ketchup)
colnames(df_ketchup)[ncol(df_ketchup)]='gaps'

mlogit.ketchup<-mlogit.data(df_ketchup, choice='brand', shape='wide',
                            alt.levels=c(0,1,2,3,4),varying=16:35,sep=".",id="hhold_id")
mlogit.ketchup[,"alt"]<-as.factor(mlogit.ketchup[,"alt"])
mlogit.ketchup[,"ft"]<-as.factor(mlogit.ketchup[,"ft"])
mlogit.ketchup[,"dp"]<-as.factor(mlogit.ketchup[,"dp"])

length_peanut<-nrow(df_peanut)
P.0<-rep(0.0,length_peanut)
ft.0<-rep(0.0,length_peanut)
dp.0<-rep(0.0,length_peanut)
cp_av.0<-rep(0.0,length_peanut)
df_peanut<-cbind(df_peanut,P.0,ft.0,dp.0,cp_av.0)
```

```
# [Peanut] gaps computing
gap_peanut<-append(df_peanut[-1,11],0)-df_peanut[,11]
gap_peanut[gap_peanut<0]=0
df_peanut <- cbind(df_peanut,gap_peanut)
colnames(df_peanut)[ncol(df_peanut)]='gaps'

mlogit.peanut=mlogit.data(df_peanut, choice='brand', shape='wide',varying=16:35,sep=".",
                          alt.levels=c(0,1,2,3,4),id="hhold_id")
mlogit.peanut[,"alt"]<-as.factor(mlogit.peanut[,"alt"])
mlogit.peanut[,"ft"]<-as.factor(mlogit.peanut[,"ft"])
mlogit.peanut[,"dp"]<-as.factor(mlogit.peanut[,"dp"])

mlogit.ketchup.reduced=lapply(households.ids.ketchup,erase,mlogit.ketchup,5)

erase <- function(hhold_id,data,n){
  total = data[data$hhold_id==hhold_id,]
  if(total$m_obs >=10){
    total = total[-(1:5),]}
  return(total)}

indexestodelete<-function(x,data,n){
  minIndex=min(which(data$hhold_id==x))
  maxIndex=max(which(data$hhold_id==x))
  mobs = data[minIndex,]$m_obs
  if(mobs >=n){
    offset= 5 * n -1
    return(seq(minIndex,minIndex +offset) )
  }else{
    return(seq(minIndex,maxIndex))}}

indexes.delete.ketchup=pblapply(households.ids.ketchup,indexestodelete,mlogit.ketchup,20)
mlogit.ketchup.reduced = mlogit.ketchup[-unlist(indexes.delete.ketchup),]
indexes.delete.peanut=pblapply(households.ids.peanut,indexestodelete,mlogit.peanut,10)
mlogit.peanut.reduced = mlogit.peanut[-unlist(indexes.delete.peanut),]
```

B.2 Descriptive analysis

```
brandshare_ketchup<-sort(table(df_ketchup$brand)[2:5],decreasing = T)
brandshare_peanut<-sort(table(df_peanut$brand)[2:5],decreasing = T)
pct_ketchup <- round(brandshare_ketchup/sum(brandshare_ketchup)*100,0)
pct_peanut <- round(brandshare_peanut/sum(brandshare_peanut)*100,0)

lbls_ketchup <- c("Heinz","Hunt's","Del Monte","Store Brand")
```

```
lbls_peanut <- c("Store Brand","JIF","Peter Pan","Skippy")
lbls_ketchup <- paste(lbls_ketchup, pct_ketchup,"%",sep=" ") # add percents to labels
lbls_peanut <- paste(lbls_peanut, pct_peanut,"%",sep=" ") # add percents to labels

par(mfrow = c(1, 2),mar=c(-2,-1,-1,-1))
pie(brandshare_ketchup,labels = lbls_ketchup,col=c("#70160D","#b82619","#BF3C30","#D47E75"),
    main="Ketchup",cex=0.8,pty='m')
pie(brandshare_peanut,labels = lbls_peanut,col=c("#917349","#C19961","#F2C586","#F5D7AD"),
    main="Peanut",cex=0.8,pty='m')

hist(as.numeric(df_ketchup[df_ketchup$brand!=0,2])-1,main="Ketchup",
        xlab="brands",breaks=c(0,1,2,3,4),cex=0.8,col="#b82619")
hist(as.numeric(df_peanut[df_peanut$brand!=0,2])-1,main="Peanut",
        xlab="brands",breaks=c(0,1,2,3,4),cex=0.8,col="#f1bf79")

# Gap distribution
par(mfrow = c(1, 2))
hist(gap_ketchup[gap_ketchup!=0],5,xlim = c(0,5))
hist(gap_peanut[gap_peanut != 0],100,xlim = c(0,6))

households.ids.ketchup <- unique(df_ketchup[,1])
spells.ketchup = aggregate(brand~hhold_id +i_obs+n_wk,data=mlogit.ketchup,
                           function(x) return(which(x==T)-1))
b=lapply(households.ids.ketchup,function(x)
        spells.ketchup[spells.ketchup$hhold_id==x &spells.ketchup$brand!=0,3] )
nopurchase_durations.ketchup=lapply(b,function(x) return(x - append(0,x[1:length(x)-1])))
hist(unlist(nopurchase_durations.ketchup),100,xlab='weeks',
        main='[Ketchup] Overall distribution of no purchase durations')

mean.per.hhold=lapply(nopurchase_durations.ketchup,mean)
hist(unlist(mean.per.hhold),40,xlab='weeks',main = '[Ketchup] Mean duration
            no purchase per household')
households.ids.peanut <- unique(df_peanut[,1])
spells.peanut = aggregate(brand~hhold_id +i_obs+n_wk,data=mlogit.peanut,
                           function(x) return(which(x==T)-1))
b=pblapply(households.ids.peanut,function(x)
            spells.peanut[spells.peanut$hhold_id==x & spells.peanut$brand!=0,3] )
nopurchase_durations.peanut=pblapply(b,function(x) return(x - append(0,x[1:length(x)-1])))

par(mfrow = c(2, 2),mar=c(4.2,4.2,1,2))
hist(unlist(nopurchase_durations.ketchup),100,xlab='weeks',
        main='Ketchup',col="#b82619",cex=0.8,breaks=10)
hist(unlist(nopurchase_durations.peanut),100,xlab = 'weeks',
        main='Peanut',col="#f1bf79",cex=0.8,breaks=10)
```

```
boxplotketchup=boxplot(unlist(nopurchase_durations.ketchup),outline=F,
        horizontal = T,col="#b82619",cex=0.8,xlab='weeks')
text(x=24,y=0.55,labels="Outliers excluded",font=3)
boxplotpeanut=boxplot(unlist(nopurchase_durations.peanut),outline=F,
        horizontal = T,col="#f1bf79",cex=0.8,xlab='weeks')
text(x=14.5,y=0.55,labels="Outliers excluded",font=3)


mean.per.hhold.peanut=lapply(nopurchase_durations.peanut,mean)
hist(unlist(mean.per.hhold.peanut),40,xlab='weeks',
        main = '[Peanut] Mean duration no purchase per household')

sumPurchases <- function(x,mlogit_data){
  filtered =mlogit_data[mlogit_data$hhold_id==x,]
  return(sum(filtered$P))}

purchases_cost = lapply(households.ids.ketchup,sumPurchases,
        mlogit.ketchup[mlogit.ketchup$brand==T,])
purchases_cost.ketchup=purchases_cost
hist(unlist(purchases_cost.ketchup),100,main="[Ketchup] Total Purchases per household")

purchases_cost.peanut = lapply(households.ids.peanut,
        sumPurchases,mlogit.peanut[mlogit.peanut$brand==T,])
hist(unlist(purchases_cost.peanut),100,main="[Peanut] Total Purchases per household")

#Unique on m_obs and household id
per.obs.ketchup = unique(mlogit.ketchup[,c(1,4)])
priceagainstobs.ketchup = cbind(per.obs.ketchup,purchases_cost.ketchup)
weeklymean.price.ketchup = priceagainstobs.ketchup[,3]/(priceagainstobs.ketchup[,2])
hist(weeklymean.price.ketchup)

per.obs.peanut = unique(mlogit.peanut[,c(1,4)])
priceagainstobs.peanut = cbind(per.obs.peanut,purchases_cost.peanut)
weeklymean.price.peanut = priceagainstobs.peanut[,3]/(priceagainstobs.peanut[,2])
hist(weeklymean.price.peanut)

weeks.ketchup = seq(1,114)
lapply(weeks.ketchup,)

purchasesperweek.ketchup= aggregate(brand~alt+n_wk+ind_feature,
data=mlogit.ketchup[mlogit.ketchup$brand==T&mlogit.ketchup$alt!=0
        &mlogit.ketchup$erim_market==1,],length)
plot(purchasesperweek.ketchup[,2],purchasesperweek.ketchup[,3],
    col=c("black","red","blue","green",'yellow')[purchasesperweek.ketchup[,1]],
```

```
        ylim=c(0,300),type='b')
colours=c("black","red","blue","green",'yellow')
featured = unique(purchasesperweek.ketchup$ft)
whichlines=which(purchasesperweek.ketchup$alt==1)
plot(weeks.ketchup,purchasesperweek.ketchup[whichlines,4],
        col=c("black","red")[purchasesperweek.ketchup$ft],ylim=c(0,300),type='l')
plot(weeks.ketchup,purchasesperweek.ketchup[purchasesperweek.ketchup$alt==1
    &purchasesperweek.ketchup$ft==1 ,3],ylim=c(0,300),type='l')

for (i in 2:4){
  lines(weeks.ketchup,purchasesperweek.ketchup[purchasesperweek.ketchup$alt==i,3],
    col=colours[i])}
par(mfrow=c(1,2))
obsPerhhold.ketchup= unique(mlogit.ketchup[,c(1,4)])
obsPerhhold.peanut= unique(mlogit.peanut[,c(1,4)])
hist(obsPerhhold.ketchup[,2],xlim=c(40,140),
    xlab="Number of observations",main="Ketchup",col="#b82619",cex=0.8)
hist(obsPerhhold.peanut[,2],xlab="Number of observations",main="Peanut",col="#f1bf79",cex=0.8)
```

B.3 Econometric modelling

```
#Multinomial Logit
mlogit.ketchup.reduced[,"hhold_income"]<-as.numeric(mlogit.ketchup.reduced[,"hhold_income"])
mlogit.peanut.reduced[,"hhold_income"]<-as.numeric(mlogit.peanut.reduced[,"hhold_income"])
test_ketchup_reduced2 <- mlogit(brand ~  P + ft + dp +cp_av  |
                                 hhold_income + hhold_members+gaps,
                               data=mlogit.ketchup.reduced,reflevel = "0")
summary(test_ketchup_reduced2)
test_peanut_reduced2 <- mlogit(brand ~  P + ft + dp +cp_av  |
                                  hhold_income + hhold_members+gaps,
                                data=mlogit.peanut.reduced,reflevel = "0")
summary(test_peanut_reduced2)
stargazer(test_ketchup_reduced2,test_peanut_reduced2,report = "vc*")
stargazer(test_ketchup_reduced2,test_peanut_reduced2)

#Nested Multinomial Logit
test_ketchup_reduced_nested <- mlogit(brand ~  P + ft + dp +cp_av  |
                                  hhold_income + hhold_members+gaps,
                                data=mlogit.ketchup.reduced,
                                nests = list(nopurchase = c("0"),
                                purchase = c("1","2","3","4")),
                                un.nest.el = TRUE)
```

```
summary(test_ketchup_reduced_nested)
test_peanut_reduced_nested <- mlogit(brand ~  P + ft + dp +cp_av  |
                                       hhold_income + hhold_members+gaps,
                                     data=mlogit.peanut.reduced,
                                     nests = list(nopurchase = c("0"),
                                     purchase = c("1","2","3","4")),
                                     un.nest.el = TRUE)
summary(test_peanut_reduced_nested)
stargazer(test_ketchup_reduced_nested,test_peanut_reduced_nested,report = "vc*")
stargazer(test_ketchup_reduced_nested,test_peanut_reduced_nested)
```

References

[1] Michael Keane Andrew Ching, Tülin Erdem. The price consideration model of brand choice. *Journal of Applied Econometrics*, 24(3):393–420, 2009.

[2] William H Greene. *Econometric analysis*, volume 5. Prentice hall, 2003.

[3] Thierry Kamionka. *Econométrie des Variables Qualitatives sur Données de Panel*. 2015.

[4] Henri Theil. A multinomial extension of the linear logit model. *International Economic Review*, pages 251–259, 1969.

[5] Baohong Sun, Scott A Neslin, and Kannan Srinivasan. Measuring the impact of promotions on brand switching when consumers are forward looking. *Journal of Marketing Research*, 40(4):389–405, 2003.

[6] Bo E Honoré and Ekaterini Kyriazidou. Panel data discrete choice models with lagged dependent variables. *Econometrica*, pages 839–874, 2000.

[7] Daniel McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

[8] Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.

[9] Stefano Ermon, Yexiang Xue, Russell Toth, Bistra Dilkina, Richard Bernstein, Theodoros Damoulas, Patrick Clark, Steve DeGloria, Andrew Mude, Christopher Barrett, et al. Learning large-scale dynamic discrete choice models of spatio-temporal preferences with application to migratory pastoralism in east africa. In *Meeting Abstract*, 2014.

[10] Andrew T Ching, Tülin Erdem, and Michael P Keane. A simple method to estimate the roles of learning, inventories and category consideration in consumer choice. *Journal of choice modelling*, 13:60–72, 2014.

[11] Zan Huang, Huimin Zhao, and Dan Zhu. Two new prediction-driven approaches to discrete choice prediction. *ACM Transactions on Management Information Systems (TMIS)*, 3(2):9, 2012.

[12] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. Demand estimation with machine learning and model combination. Technical report, National Bureau of Economic Research, 2015.

[13] Yves Croissant et al. Estimation of multinomial logit models in r: The mlogit packages.

[14] Kenneth Train and Yves Croissant. Kenneth train's exercises using the mlogit package for r. *R*, 25:0–2.

[15] Philip A Viton. Discrete-choice logit models with r, 2012.

[16] Changzheng Liu and David L Greene. Consumer vehicle choice model documentation. Technical report, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 2012.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[18] Yves Croissant. *mlogit: multinomial logit model*, 2013. R package version 0.2-4.

[19] J. Bruin. R data analysis examples: Multinomial logistic regression, 2011.

[20] Liesbeth C. de Wreede, Marta Fiocco, and Hein Putter. *mstate: An R Package for the Analysis of Competing Risks and Multi-State Models*, 2011.

[21] Liesbeth C. de Wreede, Marta Fiocco, and Hein Putter. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, 99:261–274, 2010.

[22] Arthur Allignol and Aurelien Latouche. Cran task view: Survival analysis. 2014.

[23] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Harvard University, Cambridge, USA, 2015. R package version 5.2.