Improving the conditional logit model in discrete choices of individuals

Econometrics and Operations Research
Business Analytics and Quantitative Marketing

Tijmen Schakel
433853

Supervisor: Anoek Castelein
Second assessor: Patrick J.F. Groenen

July 8, 2018

Abstract

In modeling discrete choices of individuals, many models are possible. The most well-known model is the conditional logit model, but this model has his limits in the assumptions it makes, namely homogeneity for all the individuals and so assumes the same model for all the individuals and the assumption of independence of irrelevant alternatives. To relief these assumptions, I first use a latent class model. This model construct different segments of individuals with different preferences, so that the not all the individuals have the same parameters anymore. Another model I use to relief the assumption of independence of irrelevant alternatives, is called the nested logit model. This model separate the individuals in different clusters of base-preference and so does not assume independence of irrelevant alternatives between different clusters. I also use past-choice, where I take the choice of the individual in the last visit to estimate the model. The model with latent class performs better than the conditional logit model, and including a past-choice variable improves the out-of-sample performance of the model. The nested logit model performs better than the conditional logit model, but does not improve the out-of-sample performance.

1 Introduction

In a lot of daily subjects, we are interested in how people choose their products and what can influence their choices. In many research, researches uses models to model the choices of individuals and want to predict their choices by a couple of variables, for example variables like income and age. One model that model the choice of individuals is called the conditional logit model. The conditional logit model is one of the most used models by researches for this kind of problems, but this model has also shortcomings.

One of the problems with this model is that the model assumes homogeneity in the data. This means that the model assumes that every individual have the same parameters and so are estimated by the same model for all the individuals. That is why, in this paper, I want to relief this assumption and observe the differences. Another issue with the conditional logit model is that it assumes independence of irrelevant alternatives, this means that, if an individual has to choose another brand if some brand does not exist anymore, the relative chances does not change. So to illustrate the problem of this, if there are three brands, A, B and C, where the chance to choose one of the three is for all the options $\frac{1}{3}$, but brand A does not exist anymore, the chance for the other two become now $\frac{1}{2}$. Assume that brand B is a brand that is similar to brand A, most of the people who would choose A will now choose B, so that the probabilities actually become $\frac{2}{3}$ for brand B, and $\frac{1}{3}$ for brand C. That is why I want to improve the conditional logit model in terms of performances and in the two main assumptions it makes described above. I want to relief these two assumptions, so that is why the research question is the following:

"How does changing the conditional logit model in terms of relaxing the assumptions of homogeneity in the data and the independence of irrelevant alternatives change the outcomes and conclusions in modeling individual choice?"

One of the models that does not make the assumption of homogeneity in the data, is called the latent class model. This model uses individual heterogeneity and uses classes to get homogeneous groups, and so constructs a model with different classes, such that not every individual is from the same group of individuals. Because of this, the assumption of independence of irrelevant alternatives and the assumption of homogeneity in the data are relaxed.

I also construct a nested logit model which does not take the independence of irrelevant alternatives into account. This model separate the model into two groups, so that individuals are separated by choice.

I also discuss a model where past-choice is included for every individual, because this say something about the loyalty towards a specific brand, and can be helpful in explaining choice of individuals. I also construct a prediction model, and calculate the out-of-sample performance, and compare this with the performances of the conditional logit model and the nested logit model.

The data I use is from Rome, Georgia about saltine crackers (Jain et al., 1994). This data uses

136 individuals with a total of 3292 visits at the supermarket, where individuals choose between 4 brands named Private, Sunshine, Keebler and Nabisco.

The main findings in my research are that past-choice improves modeling choice of the individuals. Also, it is indeed better to include individual heterogeneity and independence of irrelevant alternatives does not hold for all the four brands, so that relieving these assumptions improves modeling discrete choices of individuals. The out-of-sample performance is the best for the model with past-choice.

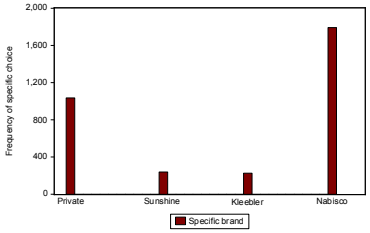
I organize the paper as follow. First I discuss the data, described in section 2, then I discuss the different models I use to compare with each other, described in section 3. In section 4, I show the results of the research and in section 5 the conclusion is presented and the paper ends with a discussion in section 6.

2 Data

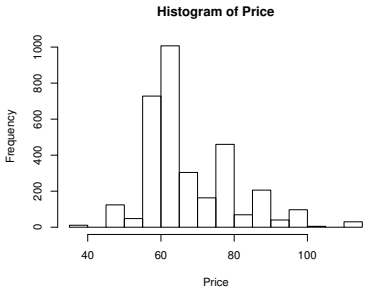
I use the data of 136 households and a total of 3292 visits in the Rome, Georgia, market, where I have the choice of the household for a brand of saltine crackers at a specific visit (Jain et al., 1994). The individuals in the data chooses between the following four brands of saltine crackers: Private, Sunshine, Keebler and Nabisco. Besides the choice, I also have given the price of the specific brand and if there was a display or a feature at that visit for each of the four brands, where display means a prominent place in the supermarket, and a feature means an advertisement for the brand. I want to use this information like display, feature and price to explain the choice of a specific household on a specific visit. The variables I use in modeling the discrete choice of choosing between the four brands of saltine crackers are:

1. $price_{it}$: Price at visit t for individual i in pennies
2. $display_{it}$: 0/1 equals 1 if there is a display at visit t for individual i
3. $feature_{it}$: 0/1 equals 1 if there is a feature at visit t for individual i
4. buy_{it}: 0/1 equals 1 if a product is bought at visit t for type j for individual i

For the visits $t = 1,...,T_i$ with T_i is equal to the last visit for person i and for the brands $j =$ Private, Sunshine, Keebler and Nabisco, and i = i,...,I with I is equal to 136. To see how the 136 households in the 3292 visits choose between the 4 brands, and to see the price of the chosen brand, a figure is made.



(a) Number of purchases for the four brands for the 136 households and 3292 visits.



(b) Price of the chosen brand with the frequency.

From figure 1a, one can conclude that most of the individuals choose for Nabisco or Private. from figure 1b, one can observe that price is most of the times between 0.60 and 1 pound. Another important aspect to take into account is marketing. For companies, it is important that their display and feature is profitable and so they sell more products. So probably, the more display and features, the more saltine crackers the company sells. The number of display, features and the number of purchases for a specific saltine cracker brand is presented in the table:

Company	Number of displays	Number of features	Number of purchases
Private	325 (14.6%)	155 (22.0%)	1035 (31.4%)
Sunshine	424 (19.1%)	124 (17.6%)	239 (7.3%)
Keebler	350 (15.8%)	140 (19.9%)	226 (6.9%)
Nabisco	1120 (50.0%)	285 (40.0%)	1792 (54.4%)

Table 1: Number of displays, features and number of purchases for all the four brands with the percentage of the total number of displays, features and purchases between the brackets.

In the table, one can observe that especially Nabisco uses displays and features a lot. It is also the brand with the most purchases, so most probably, there is a positive effect of displays and features in the number of purchases of that same brand. The table also suggests that the number of purchases of Private is less influenced by displays and features, because the number of purchases are higher than Sunshine and Keebler, but it uses less displays and around the same features as those two brands.

3 Methodology

3.1 Conditional logit model

3.1.1 Model specification

To model the choice of the individuals, I use first a conditional logit model as explained in Theil (1969), where I use the probability of choosing brand j in visit t for individual i. The explanatory variables for $x_{it,j}$, are brand-specific for parameters β_1. The chance for choosing brand j for Private, Sunshine, Keebler and Nabisco is given by:

$$P[y_{it} = j|X_{it}] = \frac{\exp(\beta_{0,j} + x'_{j,it}\beta_1)}{\sum_{h=1}^{J} \exp(\beta_{0,h} + x'_{h,it}\beta_1)} \qquad for \qquad j = 1, ..., J \tag{1}$$

So $y_{it} = j$ is the choice for individual i in visit t. For identification restrictions, I set $\beta_{0,J}$ equal to zero, such that the $\exp(b_{0,J} + \beta_{1,J}x_{it})$ equals $\exp(\beta_{1,J}x_{it})$, and so this results in three intercept parameters for identification. In this research, I set Nabisco as the reference level, so Nabisco is the identification-brand and does not have an intercept parameter.

3.1.2 Parameter estimation

To estimate the parameters of the model, I use the maximum likelihood estimation. I use the BFGS method to maximize the log-likelihood function, described in Broyden (1970). The log likelihood of this model is:

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{j=1}^{J} I_{(y_{it}=j)} \log P[Y_{it} = j|X_{it}] \tag{2}$$

with $I_{(y_{it}=j)}$ is a dummy vector equals 1 if the choice for y_{it} is j and 0 otherwise, and i is individual i with t is $1...T_i$ with t is visit t for individual i and j is for brand-choice j. I maximize this function to estimate the parameters of the conditional logit model.

3.1.3 Standard errors

I use the maximum likelihood estimation to estimate the parameters. To calculate the standard errors, I need the hessian matrix which is given by:

$$H(\theta) = \frac{\partial^2 \log \mathcal{L}(\theta)}{\partial\theta\partial\theta'} \tag{3}$$

and the variance-covariance matrix is given by:

$$var(\theta) = \left(-E\left[\frac{\partial^2 \log \mathcal{L}(\theta)}{\partial\theta\partial\theta'} \right] \right)^{-1} \tag{4}$$

On the diagonal of this matrix, I have given the variance, so to calculate the standard-errors for the $\beta_{0,j}$ and $\beta_{1,j}$, the square roots of the diagonal gives the standard-errors.

3.2 Independence of irrelevant alternatives

An issue with the conditional logit model is that it assumes independence of irrelevant alternatives(IIA) (McFadden et al., 1977). We can see this in the log odds ratios. For the conditional logit model, the log odds ratios for choosing a choice of one choice only depends on the other choice. So the log odds ratio of for example Private in comparison with Keebler, only depends on these two brands and is equal to

$$(\beta_{0,h} - \beta_{0,j}) + (\beta_{1,h} - \beta_{1,j})x_{it} \tag{5}$$

With h is equal to the choice of Keebler in this case and j the choice of Private. So, if some brand does not exist anymore, the log odds ratio stays the same and the relative probability of choosing Nabisco or Keebler stays the same. This is not always the case, because, in this case, if someone chooses Keebler, the chance to choose Sunshine or Nabisco is most likely higher than the chance of choosing the Private brand. Sunshine and Nabisco are closer substitutes of Keebler in comparison with Private. In this example, the independence of irrelevant alternatives assumes that if Keebler does not exist anymore, the relative probability for choosing private in comparison with Sunshine and Nabisco is the same as before. For example if in the first case the probability of choosing a brand is for all the brands $\frac{1}{4}$, and Keebler does not exist anymore, the probabilities of the brands with the assumption of independence of irrelevant alternatives is still the same for all the brands, so in this case $\frac{1}{3}$, but most likely, the people who chooses Keebler chooses Sunshine or Nabisco in this case, so the real probability must be $\frac{1}{4}$ for Private and $\frac{3}{8}$ for both Sunshine and Nabisco.

3.3 Latent class model

3.3.1 Model specification

To analyze for individual heterogeneity, I use the latent class model described in Greene and Hensher (2003). The idea of the model is, is that preferences of individuals is different, such that one construct Q clusters of individuals were the Q clusters are homogeneous. The latent class model is still a logit model were it is important to know the probability for choice j by individual i in choice situation t, for a given class q. This probability is given by:

$$\frac{\exp(\beta_{0,j,q} + x'_{j,it}\beta_q)}{\sum_{h=1}^{J} \exp(\beta_{0,h,q} + x'_{h,it}\beta_q)} = F(i,t,j|q) \tag{6}$$

The idea of the model is to separate the costumers into Q homogeneous groups, such that the assumption for homogeneity hold again in the Q clusters. To describe the probability for the choice made by a costumer I use

$$P_{it|q}(j) = Prob(y_{it} = j|class = q) \tag{7}$$

with the contribution of costumer i to the sample likelihood is given by:

$$P_{i|q} = \prod_{t=1}^{T_i} \prod_{j=1}^{J} P_{it|q}(j) \tag{8}$$

So that the assumption of independence choice of individual i in visit t is made. The assignment for the classes is unknown. Like stated in Greene and Hensher (2003), I denote the probability to be part of class q for individual i, such that

P[individual i ϵ q] = π_q and $\sum_{q=1}^{Q} \pi_q = 1$

3.3.2 Parameter estimation

I estimate the model with the EM-algorithm (Dempster et al., 1977). To find a (local) maximum of the log likelihood function, like done in section 3.1, I use the Expectation-Maximization algorithm, which is an iterative parameter estimation and has two steps. The estimates of the class probabilities for every person specific are $\hat{H}_{q|i}$, which is the E-step and given by

$$P[individual\ i\ \epsilon\ q \mid y_i] = \hat{H}_{q|i} = \frac{\hat{P}_{i|q}\hat{H}_{iq}}{\sum_{q=1}^{Q} \hat{P}_{i|q}\hat{H}_{iq}} \tag{9}$$

with $\hat{H}_{iq}$ is equal to P[individual i ϵ q], and for all the individuals the same. So if the probabilities for being in a specific class for the whole dataset is known, and also the probabilities of every individual for the choice made is known, one can get the posterior estimate of the class probabilities given by $\hat{H}_{q|i}$, and so the probability for every individual separated. If one know these class probabilities, one can get the likelihood for every individual. The likelihood for one individual is the expectation over all the classes. This is given by:

$$\log \mathcal{L}(L)\ P_i = \sum_{q=1}^{Q} \hat{H}_{q|i} \cdot P_{i|q} \tag{10}$$

such that the log likelihood for all the individuals is given by

$$\log \mathcal{L}(L) = \sum_{i=1}^{N} \left[\sum_{q=1}^{Q} \hat{H}_{q|i} \left(\log \pi_q + \sum_{t=1}^{T_i} \log P_{it|q} \right) \right] \tag{11}$$

The M step is to maximize the expected value, with respect to β_q and with respect to π_q. So I get

$$\hat{\pi}_q = \frac{\sum_{i=1}^{N} \hat{H}_{q|i}}{N} \tag{12}$$

For the new π_q, which is equal to the new $\hat{H}_{iq}$ in the next iteration. For the new β_q, I have to maximize the following with the maximum likelihood estimation:

$$\ln \mathcal{L}(L) = \sum_{i=1}^{N} \left[\sum_{q=1}^{Q} \hat{H}_{q|i} \left(\sum_{t=1}^{T_i} \log P_{it|q} \right) \right] \tag{13}$$

with respect to β_q. Comparing this with the model described in 3.1, it is almost the same model, the only difference now is the expected class probabilities for every individual separated, and I maximize over different classes. I get new $\hat{\beta}_q$ for every class, so I can estimate the new probabilities for the choice made by a costumer for every class, and with the new $\hat{\pi}_q$, I can estimate the new class probabilities for every person. The goal is to maximize the log likelihood, so I go on with this iterative process till the log likelihood for the new $\hat{\beta}_q$ and the new class probabilities is not an improvement comparing to the old log likelihood in the previous iteration. To achieve this, I set a stop condition ϵ, with ϵ is equal to 0.001. If the log likelihood for the next iteration is a difference of at least 0.001, I iterate again, and otherwise I stop and reached a (local) maximum. To make sure I reach the best maximum for the log likelihood, I use different random start values for $\hat{\beta}_q$ and $\hat{\pi}_q$, and repeat this and choose the maximum log likelihood. There is always a chance to reach a local maximum, but with repeating this process with different random start values, the chance of reaching a global maximum is getting bigger.

3.3.3 Selecting the number of segments

I test for 2, 3, 4 and 5 classes and check which number of classes suits the best for this dataset. To check which number of classes suits the best, I use the Bayesian information criterion(BIC), described in Schwarz et al. (1978) which is given by

$$BIC = \frac{-2\dot{\log}(L) + k\dot{\log}(N)}{N} \tag{14}$$

with N is the number of individuals, the log(L) is the log likelihood of the model and k the number of parameters in the model. The literature suggest that the lowest value for the BIC is the model that suits the most.

3.3.4 Independence of irrelevant alternatives

Because of the different classes with different parameters, the latent class model does not assume independence of irrelevant alternatives anymore. The construction of the different groups, makes it possible to separate the individuals with different parameters for display, feature, price and the intercepts, which says something about their base-preference towards a specific brand, and so the assumption of independence of irrelevant alternatives does not hold as in the conditional logit model.

3.4 Nested logit model

Like stated in section 3.2, there is a problem with the conditional logit model in terms of independence of irrelevant alternatives. Most likely, there are roughly two types of people in our dataset, namely one that purchases Private brand over the Sunshine, Keebler and Nabisco, and one type that purchases Sunshine, Keebler and Nabisco over Private. So, if this is the case, independence of irrelevant alternatives do not hold anymore, so another model is necessary where I branch the 4 types in 2 branches, which is shown below how it works.

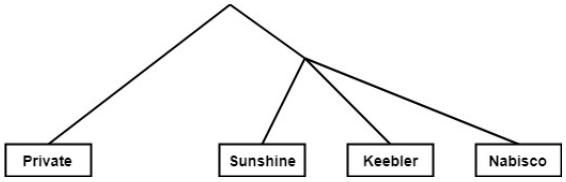


Figure 2: A nested logit model for the four saltine crackers brands

To make this model, I have to make a nested logit model, as stated in Amemiya (1981). I have two clusters, and I let S_{it} ϵ M, with M is the cluster choice, and I let C_1 ϵ {Private} for cluster one and for cluster two C_2 ϵ {Sunshine, Keebler, Nabisco}, so that the choice j in cluster m (Y_{it} = (j,m)) and the probability of choice j in cluster m is equal to Pr[y_{it} = (j,m)] = Pr[C_{it} = j|S_{it} = m]Pr[S_{it}=m] with

$$Pr[C_{it} = j|S_{it} = m] = \frac{\exp(\beta_{0,j} + x'_{mj,it}\beta_1)}{\sum_{h=1}^{J_m} \exp(\beta_{0,h} + x'_{mh,it}\beta_1)} \tag{15}$$

with $x_{mh,it}$ is the explanatory variables for individual i at time t for choice h in cluster m.
for $j = 1, ..., j_m$ with $\beta_{jm} = \beta^*_{jm}/\tau_m$ and

$$P[S_{it} = m] = \frac{\tau_m I_{m,it}}{\tau_l I_{h,it}} \tag{16}$$

And let

$$I_{it,m} = log\left(\sum_{h=1}^{J_m} exp(\beta_{0,j} + x'_{mh,it}\beta_1)\right) \tag{17}$$

Such that I can calculate the chance of Pr[y_{it} = (j,m)]. The reason why I choose to model the nested logit model stated above in this way, is because it is the most natural way to subtract the brands. Nabisco, Sunshine and Keebler are more likely substitutes of each other and Private is more likely on its own as a brand. If the variable τ is not significant different from 1, The odds-ratios for the nested-logit model results in the same odds-ratios for the conditional Logit model, so that in this case, independence of irrelevant alternatives still holds.

I use the two level nested logit model, with the error terms having a generalized extreme value CDF, and I construct τ_m as $\sqrt{(1 - cor(\epsilon_{i,mh}, \epsilon_{i,ml}}$, for the choice h for individual i in cluster m.

3.5 Conditional logit model with lagged choice variable

I have given the choices for the individuals on different times. I use this data to make a model of the choice for the next purchase, as done in Roy et al. (1996). I construct a model that does take the choice for every individual at time $t-1$ such that the model become:

$$P[y_{it} = j|X_{it}] = \frac{\exp(\beta_{0,j} + x'_{j,it}\beta_1 + \delta I[y_{i,t-1} = j])}{\sum_{h=1}^{J} \exp(\beta_{0,h} + x'_{h,it}\beta_1 + \delta I[y_{i,t-1} = h])} \qquad for \qquad j = J, ..., J \tag{18}$$

With $I[y_{i,t-1} = j]$ is one if the choice of individual i at $t-1$ is equal to j.

To construct this model, I have to manipulate the dataset. For the choice $t-1$, I use for every individual visit t with t is visit $1, ..., T-1$, and for the model, I use t is equal to visit $2, ..., T$, such that the new dataset is equal to 3156 visits instead of 3292.

Further, I use this parameter in estimating new latent class models and check if we can compare this with the other latent class models. Also I use this model, because I want to check the assumption of independent probabilities in choosing for different times. Most likely, this assumption is not valid.

To extend this model, I use also the following model:

$$P[y_{it} = j|X_{it}] = \frac{\exp(\beta_{0,j} + x'_{j,it}\beta_1 + \delta_j I[y_{i,t-1} = j])}{\sum_{h=1}^{J} \exp(\beta_{0,h} + x'_{h,it}\beta_1 + \delta_h I[y_{i,t-1} = h])} \qquad for \qquad j = J, ..., J \tag{19}$$

The difference is that in this model, I calculate the δ for every different brand independent. I use these two models for creating a new latent class model and compare this to the old one.

3.6 Model comparison

3.6.1 Conditional logit model versus Nested logit model

To test for if this model is better than the conditional logit model, I use a LR-test given by:

$$-2(\theta_1 - \theta_0) \sim \chi^2(df) \tag{20}$$

With df is the degrees of freedom and where θ_1 is the log likelihood for the nested logit model and θ_0 is the log likelihood for the conditional logit model. In my case, the difference between the parameters in the conditional logit model and the nested logit model is one, so that the degrees of freedom is equal to one.

3.7 Prediction out of sample

I have data of 3292 visits and a total of 136 individuals. I want to use this data to predict other visits and what they buy in their next visits. To check this, I split the data in two sets, where I

use the first part to set up the model and estimate the parameters in the usual way, and use this to predict the second part, and calculate the hit-rate which says something about the out-of-sample performance of the model. I calculate the hit-rate for the model with past-choice behaviour, the nested logit model and the conditional logit model and compare the differences in out-of-sample performance. I use the first 80 households to set-up the model and use this to predict the choice of the next 56 households.

3.8 Marginal effects

To check for the change in the probability of a specific brand due to a change in a explanatory variable, I use the marginal effects. In the dataset, I have continuous variables, namely price, and I have dummy variables, namely display, feature and past-choice For both, I use a different approach to calculate the marginal effects described below.

3.8.1 Continuous variables

For the continuous variables, the marginal effect is given by:

$$P[y_{it} = j|X_{it}] \cdot (1 - P[y_{it} = j|X_{it}]) \cdot \beta_1 \tag{21}$$

For the betas which are continuous, so in my case, price. With the marginal effects, I can say something about the change in probability to choose a product, due to a change in price.

3.8.2 Dummy variables

For the dummy variables, the marginal effects are quite different and given by:

$$P[y_{it} = j|X_{it} = 1] - (1 - P[y_{it} = j|X_{it} = 0]) \tag{22}$$

With this equation, I can again say something about the change in probability to choose a product, due to setting a feature, display, or due to past-choice.

4 Results

In the first place, I compared the results from the conditional logit model and the latent class model for 2, 3, 4 and 5 segments. I compare in terms of how the individuals are separated and compare the BIC to see which model fits the best. After this, I extend the model with a past-choice variable to explain the separation of the individuals better and I connect this with the results of the latent class model. In the table on the next page, the estimated parameters for the five different models with the different number of segments are presented for the conditional logit model and the latent class models.

	E1	SE1	E2	S2	E3	SE3	E4	S4	E5	SE5
Private(intercept)	-1.79***	0.10								
Sunshine(intercept)	-2.46***	0.08								
Keebler(intercept)	-1.96***	0.07								
Price	-0.03***	0.00								
Display	0.09**	0.06								
Feature	0.50***	0.10								
keebler(intercept)	-1.28***	0.18	-2.09***	0.082						
private(intercept)	0.56***	0.16	-4.73***	0.22						
sunshine(intercept)	-1.20***	0.15	-2.88***	0.11						
Price	-0.03***	0.00	-0.03***	0.00						
Display	0.41***	0.13	-0.02	0.10						
Feature	0.85***	0.19	0.47***	0.16						
Fraction	0.34		0.66							
keebler(intercept)	-3.94***	0.20	-0.25***	0.10	-1.48***	0.32				
private(intercept)	-5.00***	0.40	-2.40***	0.20	1.05***	0.21				
sunshine(intercept)	-4.53***	0.26	-1.09***	0.12	-1.27***	0.24				
Price	-0.03***	0.01	-0.05***	0.00	-0.04***	0.00				
Display	0.32***	0.21	0.31***	0.11	0.00	0.22				
Feature	1.01***	0.30	0.55***	0.16	-0.13	0.39				
Fraction	0.48		0.28		0.24					
keebler(intercept)	-1.27***	0.15	0.96***	0.17	-1.50***	0.33	-3.94***	0.20		
private(intercept)	-1.90***	0.24	-6.13***	0.70	1.13***	0.22	-5.02***	0.40		
sunshine(intercept)	-1.46***	0.16	-0.71***	0.21	-1.37***	0.25	-4.55***	0.26		
Price	-0.04***	0.01	-0.07***	0.01	-0.04***	0.01	-0.03***	0.01		
Display	0.50***	0.14	0.07	0.22	-0.14	0.23	0.34***	0.21		
Feature	0.84***	0.19	-0.12	0.32	0.18	0.39	1.00***	0.31		
Fraction	0.19		0.10		0.24		0.48			
keebler(intercept)	-1.96***	0.63	-1.22***	0.16	0.97***	0.18	-3.95***	0.20	-1.37***	0.38
private(intercept)	-1.51***	0.42	-1.91***	0.25	-6.18***	0.71	-5.01***	0.40	2.40***	0.33
sunshine(intercept)	-2.12***	0.37	-1.46***	0.161	-0.70***	0.21	-4.55***	0.27	-1.24***	0.39
Price	-0.12***	0.01	-0.04***	0.01	-0.07***	0.01	-0.03***	0.01	0.08***	0.01
Display	0.58***	0.31	0.52***	0.14	0.052	0.22	0.34***	0.21	-0.20	0.32
Feature	-0.44	0.51	0.86***	0.20	-0.11	0.32	0.99***	0.31	0.41	0.54
Fraction	0.16		0.17		0.10		0.48		0.09	

Table 2: Conditional logit model with Nabisco as reference and the parameters estimated by the maximum likelihood estimation for 1, 2, 3, 4 and 5 classes. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

	1	2	3	4	5
BIC	2.05	1.39	1.15	1.08	1.07
Log-Likelihood	-3347.71	-2241.59	-1817.34	-1675.69	-1618.33

Table 3: BIC and the log-likelihood for the 5 different models, where a lower BIC means a better fitted model.

So, to choose the best model, I choose the model with the lowest BIC and in this case, the model with 5 groups suits the best for this dataset. Because of the improvement of the BIC, relieving the assumption of homogeneity in the data and so use different parameters for the individuals improves the model and so improves modeling individual choice.

In the first place, one can observe that in the conditional logit model price has a negative effect on choice, and feature and display has a positive effect on the purchases of crackers, like expected.

In the table, I also observe that if I split the model in two groups, I mainly see differences in display and feature, where there is one group of 34 % which is more influenced by display and feature.

If I split the individuals into three groups, I observe that there is now a difference in price, and still in display and feature. There is one group wherefore feature is important and one group that seems like there is no effect in choice behaviour because of display and feature. It looks like that this group is brand loyal towards a specific brand, in comparison with the other two groups.

When I split the individuals in four groups, I observe that there is one group wherefore both display and feature is important, one group wherefore feature is mainly important and two groups wherefore the effect of feature and display both are not significantly different from zero. For one group, price is more important than the other three groups. In this model, the individuals are split on all the three explanatory variables.

Like shown above, split the model in five groups suits the most. What I observe in this model, is that there is one group that is most influenced by price, and not by display and feature, there is one group that is influenced by display, one group influenced most by feature and one group influenced by both display and feature. Also there is now one group that is not significant influenced by display and feature, and even influenced positive by price. This group seems loyal towards a specific brand.

4.1 Including past-choice behaviour

4.1.1 Conditional logit model with past-choice behaviour

	Estimate	Standard error
keebler(intercept)	-1.13***	0.09
private(intercept)	-1.69***	0.13
sunshine(intercept)	-1.77***	0.10
Price	-0.04***	0.00
Display	0.17***	0.08
Feature	0.74***	0.12
Past-Choice	2.06***	0.05
Log-likelihood	-2100.63	

Table 4: Output with past-choice behaviour. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

If I extend the model with a parameter for past choice behaviour, I observe that this parameter is highly significant for choice behaviour. The other three variables are still significant as before.

4.1.2 2-class model with Past-Choice

	Estimate1	Standard error 1	Estimate2	Standard error 2
keebler(intercept)	-1.24***	0.10	-1.05***	0.19
private(intercept)	-3.92***	0.24	0.09	0.18
sunshine(intercept)	-2.13***	0.13	-1.09***	0.16
Price	-0.04***	0.00	-0.04***	0.00
Display	-0.01	0.11	0.46***	0.14
Feature	0.70***	0.17	0.91***	0.21
Past-Choice	1.74***	0.07	1.10***	0.09
Fraction	0.66		0.34	
Log-likelihood	-1784.95			

Table 5: Latent class model for 2 classes for the model with past-choice. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

If I split the model into two groups, I observe that one group is more likely to choose Private as base preference, because of the positive intercept. This group consist of mainly the same people as in the latent class model for two segments. What I observe that the group who chooses private as base preference, has a lower value for past-choice than the other group, and a higher value for display. To show the result for past-choice for all the brands specific, I also construct a model with past-choice for all the four brands.

4.1.3 2-class model with separate Past-Choice

	Estimate1	Standard error 1	Estimate2	Standard error 2
keebler(intercept)	-1.11***	0.20	-1.67***	0.20
private(intercept)	-0.53***	0.24	-3.84***	0.29
sunshine(intercept)	-1.19***	0.18	-2.33***	0.21
Price	-0.04***	0.00	-0.04***	0.00
Display	0.55***	0.13	-0.11	0.12
Feature	0.73***	0.20	0.85***	0.18
Past-Choice(Private)	1.69***	0.18	0.14	0.46
Past-Choice(Sunshine)	0.36	0.34	1.99***	0.24
Past-Choice(Keebler)	1.01***	0.47	2.56***	0.24
Past-Choice(Nabisco)	0.60***	0.22	1.56***	0.19
Fraction	0.35		0.65	
Log-likelihood	-1769.77			

Table 6: Latent class model for 2 classes for the model with Past-Choice for every brand. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

What I see in the model above, are the same two groups as seen in the latent class model with two classes. I also see that there is one group that has a higher value for past-choice for Private and the other group has a higher value for past-choice for Nabisco,Sunshine and Keebler. Also, I see that the group that has the highest value for Private past-choice has a higher value for display than the other group. This is the same result as shown in the latent class model with two classes. Combining these results, one can observe that the group that mainly chooses Private is more sensitive for display than the other group.

I also observe that the assumption made in the latent class model of independent probabilities of choice for time t for individual i is a strong assumption made in the latent class model, because people are influenced by past-choice.

4.2 Main Results of the conditional logit models and the latent class models

The main results found in the models are that individuals can be separated mainly by means of display and feature in two classes. For more than two segments, individuals are separated by price, display and feature. Also the individuals who chooses mainly Private, are also the people that are influenced more by display. Past-choice is mainly important for the individuals who chooses Sunshine, Keebler and Nabisco and less important for individuals who choose Private.

In terms of performances, one can observe that relieving the assumption of homogeneity in the data, the model suits better. This means that relieving this assumption improves the model and

so improves modeling individual choice.

To zoom further in on the independence of irrelevant alternatives, I construct a nested logit model, and to measure the performances, I show the out-of-sample performances of the different models.

4.3 Nested logit model

	Estimate	Standard error
keebler(intercept)	-1.94***	0.07
private(intercept)	-2.64***	0.31
sunshine(intercept)	-2.38***	0.08
Price	-0.03***	0.00
Display	0.11***	0.05
Feature	0.52***	0.09
Tau	1.37***	0.14
Log-likelihood	-3343.43	

Table 7: Output for the nested logit model for the first 80 households. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

Above the output of the nested logit model is shown. To be sure that this is an improvement over the conditional logit model, I use a LR test. The LR-test gives a value of 8.56, which is bigger than $\chi^2(1)$ on a 10% ,5% and 1% level, so the nested logit model is an improvement over the conditional logit model.

4.4 Out-of-sample performances

The hit-rates for the last 56 individuals for the conditional logit model, the conditional logit model with past-choice and the nested logit model are below. The model for the first 80 households to set up the model can be found in A.1.

4.4.1 Conditional logit model

Below is the prediction table used with the parameters of the first 80 households to predict the choice for the next 56 households and 1315 observations, for the conditional logit model.

	Private	Sunshine	Keebler	Nabisco
Prediction	194	0	0	1121
Good predicted	109	0	0	609
Real choice	454	109	74	678
Hit-rate	54.6 %			

Table 8: Hit-rate for the last 56 households for the conditional logit model.

If I use the first 80 households to predict the choice of the next 56 households, I get a hit-rate of 54.6 %, so that means that the model predicted for the next 56 households with the first 80 households 54.6% correct.

4.4.2 Conditional logit model with past-choice

Below is the prediction table used with the parameters of the first 80 households to predict the choice for the next 56 households and 1259 observations for a conditional logit model with past-choice.

	Private	Sunshine	Keebler	Nabisco
predicted	437	75	60	687
Good predicted	341	28	31	556
Real choice	437	103	71	648
Hit-rate	75.9 %			

Table 9: Hit-rate for the last 56 households for the conditional logit model with past-choice.

If I use the first 80 households to predict the choice of the next 56 households, I get a hit-rate of 75.9 %, so that means that the model predicts 75.9 % good for the next 56 households with the first 80 households. In section A.2, the results of the same model, but without a past-choice variable are shown. If I compare the hit-rates, I observe that the model without a past-choice variable has a hit-rate of 54.6%, so the model with a past-choice has a better out-of-sample performance in comparison with the conditional logit model.

4.4.3 Nested logit model

Below is the prediction table used with the parameters of the first 80 households to predict the choice for the next 56 households and 1315 observations, for the nested logit model.

	Private	Sunshine	Keebler	Nabisco
Predicted	223.00	0.00	0.00	1092.00
Good Predicted	120.00	0.00	0.00	598.00
Real Choice	454.00	109.00	74.00	678.00
Hit-rate	54.6 %			

Table 10: Hit-rate for the last 56 households for the nested logit model.

If I use the model with the first 80 households to predict the choice of the next 56, I see no improvement compared with the conditional logit model. So the nested logit model for this data is not an improvement in terms of out-of-sample performance.

4.5 Marginal effects

Below are the marginal effects for the conditional logit model and the conditional logit model with past-choice.

4.5.1 Conditional logit model

	Private	Sunshine	Keebler	Nabisco
Price	-0.63	-0.20	-0.20	-0.73
Feature	0.11	0.04	0.04	0.11
Display	0.02	-0.00	-0.00	-0.02

Table 11: Marginal effects for the conditional logit model for price, feature and display.

If I look at the marginal effects, I observe that price has the biggest effect on choice for Nabisco and Private. For feature, if there is a feature, the probability of buying Private increases with 0.11.

4.5.2 Conditional logit model with past-choice

	Private	Sunshine	Keebler	Nabisco
Price	-0.40	-0.18	-0.18	-0.46
past-choice	0.37	0.25	0.26	0.41
Feature	0.09	0.04	0.04	0.09
Display	0.02	-0.00	-0.00	-0.01

Table 12: Marginal effects for the conditional logit model with past-choice, for price, past-choice, feature and display.

In the table above, one can observe that price is most important for Private and Nabisco. The same result I can observe for past-choice. Past-choice seems again like an important parameter for the model. Feature is for all the 4 brands important and again, display is only important for Private.

4.6 Main results out-of-sample performance and marginal effects

For the out-of-sample perfomances, it is shown that including past-choice improves the out-of-sample performance. A nested logit model is an improvement over the conditional logit model, and because of this relieving the assumption of independence of irrelevant alternatives is a good way to improve the model, but in terms of out-of-sample performance, there is no trade-off in using a nested logit model.

For the marginal effects, I observed that past-choice has a high marginal effect.This means that past-choice is important for the choice in visit t for individual i, and so including past-choice improves the model in both out-of-sample performances, but also in understanding how people choose.

5 Conclusion

In this paper, I researched how to describe a model of discrete choices of individuals in its best way. In the first place, I have used a conditional logit model, but this model is really restrictive, because it assumes independence of irrelevant alternatives and also describes every person with the same parameters, and so assumes homogeneity in the data. In general I saw that relieving the assumptions and so use different models, improves modeling the discrete choice of individuals. Also, using a past-choice improves the model in understanding the model and out-of-sample performance.

In this paper, I have mainly focused on two parts. The first part was relaxing the two assumptions the conditional logit model makes, and the second part was to compare the out-of-sample performance and investigate if relieving the assumptions and using a new variable for past-choice improves this.

For the latent class model I split the group of individuals into different groups where groups are split by preferences. I split the groups in 2,3,4 and 5 groups and with the bayesian information criterion concluded that 5 segments suits the best in this dataset. For two groups, individuals are split mainly because of display and feature, but in more segments, price is also an important variable to separate the individuals. To extend this model, I have used a past-choice variable and this variable is the most important for the individuals who chooses Keebler, Sunshine or Nabisco. Individuals who chooses Private, are more sensitive for displays, and has a lower value for past-choice. Relieving the assumption of homogeneity in the data improves the model in terms of the BIC, and also in terms of understanding the choices of different individuals.

To zoom further in on the assumption of independence of irrelevant alternatives, I constructed a nested logit model. This model is better than the conditional logit model, so relieving the independence of irrelevant alternatives improves modeling discrete choice of individuals, but does not improve the out-of-sample performance. The model with past-choice is the best model to use for out-of-sample performance.

6 Discussion

An important point to take into account, is that with the latent class model I used the assumption that the choice of the people are independent for choice t, which is a strong assumption as shown with the model of past-choice. In further research, it is important to look at this, with taking brand-loyalty into account. A parameter for brand-loyalty could be used to describe the choice of an individual better than I did now, and use this in constructing a better latent class model.

Another important point with the latent class model, is that there is a probability to be in a local maximum with maximizing the log likelihood. To limit the chance of ending in a local maximum, I used different start-values for $\beta_{0,j}$ and β_1, but it still can be the case that there is a better optimum than I have found right now.

For further research, it is important to look at the past-choice model. Due to time-constraints, it was not possible to construct a latent class model for 3,4 and 5-groups with past-choice, because the probability to reach a local-maximum was too high, but it can be for further research interesting to take a look at it and compare, like done with the two groups, with the latent class model for 3,4 and 5 groups. It could explain better than done now how the groups are constructed by means of past-choice and brand-preferences.

References

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 19(4):1483–1536.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Greene, W. H. and Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8):681–698.

Jain, D. C., Vilcassim, N. J., and Chintagunta, P. K. (1994). A random-coefficients logit brand-choice model applied to panel data. *Journal of Business & Economic Statistics*, 12(3):317–328.

McFadden, D., Tye, W. B., and Train, K. (1977). *An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model*. Institute of Transportation Studies, University of California.

Roy, R., Chintagunta, P. K., and Haldar, S. (1996). A framework for investigating habits,âĂIJthe hand of the past,âĂİ and heterogeneity in dynamic brand choice. *Marketing science*, 15(3):280–299.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Theil, H. (1969). A multinomial extension of the linear logit model. *International Economic Review*, 10(3):251–259.

A Appendix

A.1 Models for first 80 households

Below is the output for 80 households and with 1977 observations for the conditional logit model.

	Estimate	Standard error
Private(intercept)	-1.88***	0.09
Sunshine(intercept)	-1.90***	0.13
Keebler(intercept)	-2.59***	0.11
Price	-0.03***	0.00
Display	0.11**	0.08
Feature	0.52***	0.12
Log-Likelihood	-1987.56	

Table 13: Model for the conditional logit model for the first 80 households. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

Below is the model of the first 80 households with 1897 observations for the conditional logit model with past-choice

	Estimate	Standard error
keebler(intercept)	-1.03***	0.11
private(intercept)	-1.75***	0.17
sunshine(intercept)	-1.84***	0.13
Price	-0.04***	0.00
Display	0.23***	0.11
Feature	0.78***	0.16
Past-Choice	2.14***	0.06
Log-likelihood	-1211.37	

Table 14: Output for the first 80 households for the conditional logit model with past-choice. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

Below is the output for 80 households and with 1977 observations for the nested logit model.

	Estimate	Standard error
Private(intercept)	-1.86***	0.09
Sunshine(intercept)	-2.77***	0.42
Keebler(intercept)	-2.53***	0.11
Price	-0.03***	0.00
Display	0.10**	0.07
Feature	0.52***	0.11
Tau	1.35***	0.17
Log-Likelihood	-1985.02	

Table 15: Model for the nested logit model for the first 80 households. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

A.2 Conditional logit model without past-choice

Below is the model of the first 80 households with 1897 observations for a conditional logit model without past-choice, but with the same dataset as the model with past-choice

	Estimate	Standard error
Private(intercept)	-1.87***	0.09
Sunshine(intercept)	-1.88***	0.13
Keebler(intercept)	-2.58***	0.11
Price	-0.03***	0.00
Display	0.12***	0.08
Feature	0.52***	0.12
Log-Likelihood	-1909.77	

Table 16: Output for the Conditional logit model for the first 80 households without Past-Choice, but with the same dataset. * significant at 0.1 level, ** significant at 0.05 level and *** significant at 0.01 level.

Below is the prediction table used with the parameters of the first 80 households to predict the choice for the next 56 households and 1259 observations for a conditional logit model without past-choice, but with the same dataset as the model with past-choice.

	Private	Sunshine	Keebler	Nabisco
predicted	194.00	0.00	0.00	1065.00
Good predicted	109.00	0.00	0.00	579.00
Real choice	437.00	103.00	71.00	648.00
Hit-rate	54.6 %			

Table 17: Hit-Rate for the conditional logit model for the last 56 households without past-choice, but with the same dataset as the model with past-choice.