Comparison of Multinomial Logit and Mixed Logit

Ludo Pfaff, 433964

Supervisor: Jeroen Vester

Second assessor: Kathrin Gruber

July 7, 2019

Bachelor Thesis

Programme BA & Quantitive Marketing

Abstract

This paper concentrates on the comparison of different discrete choice models using data concerning purchases of different brands of crackers in USA. Five logit models are proposed: multinomial logit (MNL), mixed logit (ML), multinomial logit which accounts for heterogeneity between the customers (MNLH), multinomial logit extended with a brand loyalty variable (MNLB), and multinomial logit with a combination of the two extensions to MNL mentioned above (MNLC). Using the Cracker data set from the Ecdat package in the R software, it was found that ML outperforms all the other models, and that the extensions applied to MNL help in improving MNL. Therefore, it was concluded that using a mixed logit model is the best way to model an unordered categorical dependent variable.

Contents

1 Introduction

How is a choice between different brands made by a consumer? Can the price or the way it is displayed influence the purchasing behaviour of the customer? And how do producers and sellers choose to manufacture or market these products? By using discrete choice models to summarise relevant data, conclusions can be drawn with respect to the actions of consumers, firms and households. With the help of these conclusions, an idea about the preferences of consumers can be constructed and, therefore, the supply can be adapted to the demand more effectively. That raises the question what discrete choice models are available? And which model can be used best in which situation?

Since the development of computer algorithms and the ability of working with huge sets of data, the growth of the application of discrete choice models has been rapid. The most frequently used branch of these models are the logit models. What started with the publication of Pierre-François Verhulst during the nineteenth century (Cramer, 2002), where he first named the logit function, is now at a point where multiple models, based on this logit function, are regularly used and still extended. While these logit models kept extending into more advanced applications, parameter estimation of these models became an issue. The solution for this barrier came with the development of simulation methods. With the possibility to estimate even the more advanced models of the logit family, the practice of discrete choice methods has become a lot easier.

In this paper, a multinomial logit model (MNL) is compared with the more advanced mixed logit model (ML). By reason of ML being an extension of MNL, ML has a few advantages, relative to MNL. Two main advantages are that ML does not exhibit the assumption of independence of irrelevant alternatives (IIA) in contrast to MNL, and that ML accounts for possible correlation over repeated choices made by each individual (Algers et al., 1998). IIA implies that the choice between two options only depends on the characteristics of these two options, and not on the characteristics of other possible options. Therefore, an individual's preference does not change after including or excluding any other option. While in many situations exhibiting the assumption of IIA does not cause any problems, it may be an unrealistic assumption in some cases (Franses and Paap, 2001). Furthermore, ML accounts for possible correlation over repeated choices made by each individual, while for MNL, repeated choices made by an individual are independent. To circumvent the latter shortcoming of MNL, two extensions are implemented in this paper. First, individual-specific parameters are included in MNL to represent the base preferences of every individual, which will be called unobserved heterogeneity in this paper. Secondly, brand loyalty is implemented in MNL. Brand loyalty is not an attribute of ML, but adding brand loyalty to a model gives an idea about the preferences of every individual, and it takes correlation between repeated choices into account for every individual. At last, a combination of the these two extensions is added to MNL to examine if this can improve MNL even more.

The first extension, i.e., including unobserved heterogeneity, has already been implemented before by, among others, Rossi and Allenby (1993), while brand loyalty has been included to a model by Guadagni and Little (1983). However, the models where these extensions were included, were not compared to ML, and a model which holds both extensions, has not even been implemented before.

It is hypothesised that using ML is a better way to model choices between different options, than using MNL. Besides, it is expected that adding the different extensions to MNL may help to improve MNL, and, maybe, even make MNL a better model than ML. Furthermore, MNL with its extensions is expected to be relatively easy to implement and to interpret, in comparison with the more complex ML. Therefore, if MNL with the extensions added, performs as well as ML, it can be suggested to use MNL with the extensions.

This paper will continue in Section 2 with a brief overview of relevant literature over the past years. Afterwards, in Section 3, an explanation of the used data will be presented. In Section 4 the used discrete choice models, extensions and performance measures will be thoroughly explained. In Section 5, the results of the different models will be displayed and discussed, and at last, in Section 6, the conclusion of this paper and recommendations for future work will be given.

2 Literature Review

In 1980, the first applications of ML appeared in the form of transportation research (Boyd and Mellman, 1980; Cardell and Dunbar, 1980). In both papers the automobile gasoline prices are analysed using ML, where Boyd and Mellman (1980) also look at the market shares of different models of cars. Since these two papers have been published, the amount of researches using ML started rising. Because the purpose of this paper is to compare MNL and ML, and to improve MNL with some extensions, this overview of relevant literature contains three parts: Multinomial Logit model, Mixed Logit model and comparison of the two models.

2.1 Multinomial Logit model

The multinomial logit model has been the basis for every generalised logit model that has been developed. Franses and Paap (2001) have given a clear overview of the representation and interpretation for multiple models for individual choice behaviour, including MNL and the comparable Conditional Logit model (CL). The application and interpretation of MNL and CL are similar, while the difference between these two models lies in the kind of variables used. For MNL, only individual-specific variables, like age or gender, are used. Therefore, these explanatory variables only differ per individual; not per category of choice or observation. In case of CL, the explanatory variables take different values across the different choice options and observations. An example of such a variable, is the price of a product experienced by an individual on a particular observation. Thus, these variables can differ per brand, person and observation, while for MNL the explanatory

variables only differ per person. While Franses and Paap (2001) made a distinction between MNL and CL, in this paper the term MNL is used for both MNL and CL. Furthermore, Franses and Paap (2001) used MNL to model the choice between four brands of crackers. The data set they used to model, is the same data set used in this paper. It is shown that MNL is a convenient and proper way to model these kind of discrete choice problems.

Because of its ease of use and multi applicability, MNL has been applied in different levels. For instance, Guadagni and Little (1983) applied MNL to model panel data concerning purchases of different coffee brands and the loyalty of customers to the brands of coffee, in Kansas City in the late eighties. They showed that using MNL can give a good representation of the used panel data and, when it is extended by adding a variable representing brand loyalty, it can give an idea about the brand loyalty of the customers concerning different coffee brands.

In Section 4, a clear explanation of MNL will be given.

2.2 Mixed Logit model

As said earlier, ML is an extension of MNL which does not feature IIA, in contrast to MNL. One of the many papers where ML is discussed, is written by Brownstone and Train (1998). In this paper, it is explained how ML works and why it does not exhibit IIA. While ML resembles MNL, its main difference lies in one point. The difference is that the parameters are taken randomly from a distribution, while for MNL the parameters are estimated using data. Brownstone and Train (1998) used a survey data set on households' attitudes and preferences regarding alternative-fuel vehicles, e.g., a full electric car. It is shown that ML is a useful model to see what variables influence the choice of the different kinds of alternative-fuel vehicles.

Moreover, ML is frequently used for the computation of value-of-time measures (Hess et al., 2005). Such a research where ML is used to estimate the value of time for long-distance car travel, is written by Algers et al. (1998). The purpose of their paper is to estimate the values of the distribution parameters for the coefficients and to investigate if the value-of-time is affected by allowing varying coefficients in the population, i.e., including unobserved heterogeneity. Before Algers et al. (1998) explain their application of ML, they give six advantages of using ML in general, namely:

1. ML does not exhibit the IIA property;
2. ML accounts for possible correlation over repeated choices made by each individual;
3. ML can be derived from utility maximising behaviour;
4. ML can closely approximate multinomial probit[1];

[1]Multinomial probit is a discrete choice model which uses a normal distribution for the error terms, instead of a logistic distribution in case of MNL.

5. ML can, unlike a multinomial probit, represent situations where the coefficients follow other distributions than the normal distribution;

6. ML might have an advantage over the multinomial probit if the dimension of the mixing distribution is less than the number of alternatives, because the simulation is over fewer dimensions.

These points give a good summary of why ML is used frequently and in so many different cases. Algers et al. (1998) estimated the value-of-time with ML using several distributions for the coefficients, such as normal and log-normal. They found that the estimated value-of-time is very sensitive to how the model is specified. For instance, the value-of-time is less sensitive using a distribution for the coefficients, than when the coefficients are fixed. It is also concluded, that the model specifications fit the data significantly better, when allowing for unobserved heterogeneity in the population. Because these are two properties of ML, it is concluded that ML is indeed useful for computing value-of-time values.

ML will be explained thoroughly in Section 4.

2.3 Comparison

While it is proven that MNL can give weak forecasts because of its restrictive assumption of Independence of Irrelevant Alternatives (McFadden et al., 1977), it can be seen useful to compare ML with MNL. In this way, MNL can be seen as a baseline to show the performance of ML. In the paper written by Ye and Lord (2014), MNL is compared with Ordered Probit and ML models. The objective of this paper is to examine what influence the sample size has on these three models. The authors of this paper use a simulated data set of car crashes to compare the different models. The conclusion of Ye and Lord (2014) is that ML has a better interpretive power than MNL and that ML fits the data significantly better than MNL.

Furthermore, there are more papers where both MNL and ML are used. One of these papers has been written by Alfnes (2004), where MNL as well as ML is used to model the choice between several kinds of beef. He concludes that using ML is better to model the choice between different types of beef, than MNL, since ML allows the presence of correlated error terms. That is why, ML is better for identifying substitution patterns and predicting cannibalisation of consumer shares (Alfnes, 2004). In other words, ML identifies when a new option is a substitute of other options, or a new, unique option. Alfnes (2004) shows this distinction between the models, by simulating different markets where not every kind of beef is sold. The resulting market shares of the MNL market simulation are not in line with the actual market shares, while ML predicts the market shares more correctly. Therefore, it can be concluded, that MNL did not detect the substitution patterns, and hence, that using ML is superior to MNL in this case.

3 Data

The data set used to evaluate the implemented methods, comes from the Ecdat package in the R software. The set consists of scanner data of purchases of different brands of crackers in the United States of America. Four brands are taken into account, namely Sunshine, Kleebler, Nabisco and a private brand of the supermarket. The set contains 3 292 observations of 136 individuals. At every moment of purchase, the marketing mix variables of the four brands are observed, i.e., display, feature and price. Display and feature are both binary variables and represent whether there is a display and an advertisement for the relevant brand, respectively. The third variable, price, is an integer and gives the price of the crackers in US cents.

	Sunshine	Kleebler	Nabisco	Private
Total purchases	239	226	1792	1035
Total display	424	350	1120	325
Total feature	124	140	285	155
% in display	12.88	10.64	34.03	9.87
% in feature	3.77	4.25	8.66	4.71
Average price	95.7	112.6	107.9	68.07
Minimum price	49	88	0 (49)	38
Maximum price	129	139	169	115

Table 1: Descriptive Statistics

In Table 1, a summary of the data is given. It is shown that Nabisco is the most bought brand of crackers. The private brand of the supermarket is the second most bought brand, while Sunshine and Kleebler have been bought almost as much as each other. A reason that Nabisco is the most frequently bought brand of crackers could be that it is the most frequent brand, concerning display and feature, which make being in display and feature a significant reason to purchase a certain brand of crackers.

When looking at the prices of the four brands, it can be seen that Kleebler is on average the most expensive brand, while the private brand has the lowest average price. The relatively low price of the private brand could be a reason that this brand is the second-most bought brand. Furthermore, it stands out that the minimum price of Nabisco is zero. When looking at the data, it can be seen that it occurs three times that Nabisco is free. This can be due to an error in the data set, or to a deal where one can get Nabisco crackers for free. When the price of zero cents for Nabisco is not taken into account, the minimum price is 49 cents.

What is further noticeable, is that Sunshine and Kleebler are bought almost the same amount of times, while Kleebler is significantly more expensive and less frequent in display, than Sunshine. However, Kleebler has an advertisement, i.e., is in feature, more frequently than Sunshine. This

could mean that being in feature has a greater impact on the choice between Sunshine and Kleebler, than the price or being in display. Another explanation could be that the customers who buy Kleebler, are more loyal to their brand of crackers, than the customers who buy Sunshine. Due to such a brand loyalty, the customers for Kleebler are less sensitive to changes in price, display or feature, than the customers for Sunshine.

At last, when looking at the data, it stands out that every individual mostly buys one certain brand of crackers. This possibly implies that every customer has a specific base preference regarding brands of crackers, and that every customer is loyal to "their" brand of crackers. This is in line with the extensions for MNL which are introduced in Section 1, i.e., unobserved heterogeneity and brand loyalty. Because it can be seen that these phenomena are present in the data set, it gives reason to think that extending MNL with unobserved heterogeneity and brand loyalty, might indeed improve the goodness-of-fit to this data set.

4 Methodology

In this section, the characteristics and application of both MNL and ML will be thoroughly explained. Afterwards, the applied extensions for MNL will be illustrated and the performance measures will be introduced.

Before the different methods are implemented, the data set is slightly adjusted. Instead of using the price of every brand as an explanatory variable, the natural logarithm of the price is taken, because the parameter that represents the influence of the log of the price, is equal to the elasticity of the price (Franses and Paap, 2001). Consequently, the influence of the prices is easier to interpret. To prevent any errors to occur due to taking the logarithm of zero, the data points where the price of Nabisco equals zero, are taken out of the data set.

Short explanations of the R programs used for the methods in this paper, are given in the Appendix.

4.1 Multinomial Logit model (MNL)

MNL can be used to model an unordered categorical dependent variable. This means that the dependent variable for observation t, Y_t, can take discrete values 1,2, ..., J, where J equals the number of possible options. Every value of Y_t represents one of the J choices in an unordered way, that is, there is no logical order of the different categories.

Moreover, the probability that category j is chosen at observation t given explanatory variables w_t, equals

$$\Pr[Y_t = j|w_t] = \frac{\exp(\beta_{0,j} + \gamma w_{j,t})}{\sum_{l=1}^{J} \exp(\beta_{0,l} + \gamma w_{l,t})}, \quad \text{for } j = 1, \ldots, J,\ t = 1, \ldots, T, \tag{1}$$

where $w_t = (w_{1,t}, ..., w_{J,t})$ represents the explanatory variables at observation t. Hence, in case of this research, the explanatory variables differ between categories and observations. Furthermore, $w_{j,t}$ is a column vector with three elements, which represent the display, feature and the log of price of category j at observation t. The explanatory variables have an equal impact γ on the probabilities for every category j, where γ is a row vector with three parameters that represent the influence of the display, feature and log of price on the probability of buying one of the J categories, independently. Besides the parameter γ, there is an intercept parameter $\beta_{0,j}$ for $\forall j$. Therefore, when MNL is used in this paper, the individuals are not considered, because the parameters do not differ per individual, but only per category. For other models used in this paper, which will be explained later, the difference between individuals are considered, in contrast to MNL. For identification of the intercept parameters, $\beta_{0,J}$ is set to zero, such that the parameters can be uniquely estimated. Without setting one of the intercept parameters to zero, it is possible that a unique estimation of the parameters is not found.

To find the best fitting parameters for MNL, Maximum Likelihood estimation (MLE) is used. The idea of MLE is to make the desired probability distribution "most likely", such that the model fits the data as good as possible. This is accomplished by seeking the value of the parameters that optimises the likelihood function $\mathcal{L}(\theta)$ (Myung, 2003).
The likelihood function for MNL is as follows:

$$\mathcal{L}(\theta) = \prod_{t=1}^{T}\prod_{j=1}^{J} \Pr(Y_t = j|w_t)^{\mathrm{I}[y_t=j]} = \prod_{t=1}^{T}\prod_{j=1}^{J} \left(\frac{\exp(\sum_{j=1}^{J}(\beta_{0,j} + \gamma w_{j,t})}{\sum_{l=1}^{J}\exp(\beta_{0,l} + \gamma w_{l,t})}\right)^{\mathrm{I}[y_t=j]}, \tag{2}$$

where θ stands for the parameters in the likelihood function, i.e., γ and $\beta_{0,j}$ for $\forall j$, and $\mathrm{I}[y_t = j]$ is the indicator function for y_t. So, $\mathrm{I}[y_t = j]$ equals one if category j is chosen on observation t, and zero otherwise. To find the optimum of the likelihood function, it is more convenient to optimise the log-likelihood function $\ell(\theta)$, that is, the logarithm of the likelihood function. The log-likelihood function for MNL is

$$\begin{aligned}
\ell(\theta) = \log \mathcal{L}(\theta) &= \log\left(\prod_{t=1}^{T}\prod_{j=1}^{J} \left(\frac{\exp(\sum_{j=1}^{J}(\beta_{0,j} + \gamma w_{j,t})}{\sum_{l=1}^{J}\exp(\beta_{0,l} + \gamma w_{l,t})}\right)^{\mathrm{I}[y_t=j]}\right) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J} \mathrm{I}[y_t = j] \log\left(\frac{\exp(\beta_{0,j} + \gamma w_{j,t})}{\sum_{l=1}^{J}\exp(\beta_{0,l} + \gamma w_{l,t})}\right) \\
&= \sum_{t=1}^{T}\sum_{j=1}^{J} \mathrm{I}[y_t = j](\beta_{0,j} + \gamma w_{j,t} - \log(\sum_{l=1}^{J}\exp(\beta_{0,l} + \gamma w_{l,t}))).
\end{aligned} \tag{3}$$

When the log-likelihood function $\ell(\theta)$ is optimised, the resulting parameters θ make the model fit the data in the best way possible.

4.2 Mixed Logit model (ML)

ML can also be used to model an unordered categorical dependent variable. The main idea of ML is that the parameters of the model are drawn randomly from a distribution. The most frequently used distributions for the parameters in ML are the normal, log-normal, triangular and uniform distribution (Hensher and Greene, 2003). ML avert limitations of MNL by allowing for unrestricted substitution patterns, random taste variation, and correlation in unobserved factors over time (Train, 2009). The probability that category j is chosen at observation t in ML, equals

$$\Pr[Y_t = j|w_t] = \int L(\beta) f(\beta) d\beta, \quad \text{for } j = 1, \ldots, J, \ t = 1, \ldots, T, \tag{4}$$

where β represents the parameters of ML, $f(\beta)$ represents the density of β, which is called the density of the mixing distribution, and $L(\beta)$=$\frac{\exp(\beta' w_{j,t})}{\sum_{l=1}^{J} \exp(\beta' w_{l,t})}$. Notice, if $f(\beta)$ is equal to one, ML is similar to MNL. The parameters and variables are similar to the ones which are used for MNL. So, there are intercept parameters for every category j, and parameters that represent the influence of the marketing mix variables, i.e., display, feature and the log of price. Furthermore, just like for MNL, the intercept parameter for category J is set to zero for identification.

The mixing distribution can, for instance, be a normal distribution. Then, the probability equals

$$\Pr[Y_t = j|w_t] = \int L(\beta) \phi(\beta|\mu, \sigma) d\beta, \quad \text{for } j = 1, \ldots, J, \ t = 1, \ldots, T, \tag{5}$$

where ϕ is the density of the normal distribution, and μ and σ are the mean and standard deviation of the normal distribution, respectively. Both μ and σ need to be estimated to give β its best fitting values. Consequently, Equation 4 can be written as

$$\Pr[Y_t = j|w_t] = \int L(\beta) f(\beta|\theta) d\beta, \quad \text{for } j = 1, \ldots, J, \ t = 1, \ldots, T, \tag{6}$$

where θ represents the parameters of the relevant density for β.

Considering that taking an integral of a function can be hard, the log-likelihood function of ML is approximated by a simulation. The simulated log-likelihood function (SLL) looks as follows

$$SLL = \sum_{t=1}^{T} \sum_{j=1}^{J} \mathrm{I}[y_t = j] \log \hat{\Pr}[Y_t = j|w_t], \tag{7}$$

where $\mathrm{I}[y_t = j]$ is one if y_t equals j, and zero otherwise, and $\hat{\Pr}[Y_t = j|w_t]$ is the simulated probability that Y_t equals j. This simulated probability $\hat{\Pr}[Y_t = j|w_t]$ can be determined by

$$\hat{\Pr}[Y_t = j|\beta, w_t] = \frac{1}{R} \sum_{r=1}^{R} L(\beta_r) = \frac{1}{R} \sum_{r=1}^{R} \frac{\exp(\beta_r' w_{j,t})}{\sum_{l=1}^{J} \exp(\beta_r' w_{l,t})}, \quad \text{for } j = 1, \ldots, J, \ t = 1, \ldots, T, \tag{8}$$

where β_r is the r^{th} draw from the mixing function $f(\beta|\theta)$, and R is the total number of draws, i.e., simulations.

One of the advantages of ML is that it can accommodate respondent preference correlation across repeated choice observations (Rose et al., 2009). In other words, if an individual is faced with the same choice repeatedly, such as a choice between different brands of crackers, ML can incorporate the correlation between the choices for an individual. To incorporate this respondent preference correlation, a few adjustments need to be made with regard to the function L. Instead of looking at one observation t, now the combination of choices of every individual is considered separately. Thus, for every individual $i \in \{1, \ldots, N\}$, the probabilities that the chosen categories were chosen, are multiplied for every observation $t_i \in \{1, \ldots, T_i\}$, with t_i being an observation for individual i and T_i being the total number of observations for individual i. The simulated probability that a certain sequence of choices has been made by individual i, looks now as follows

$$\begin{aligned} \hat{\Pr}[Y_i = \{y_{i,t_i}\}_{t_i=1}^{T_i} | \{w_{t_i}\}_{t_i=1}^{T_i}] &= \frac{1}{R} \sum_{r=1}^{R} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \Pr[Y_{t_i} = j | \beta_r, w_{j,t_i}]^{\mathrm{I}(y_{t_i}=j)} \\ &= \frac{1}{R} \sum_{r=1}^{R} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \left(\frac{\exp(\beta_r' w_{j,t_i})}{\sum_{l=1}^{J} \exp(\beta_r' w_{l,t_i})} \right)^{\mathrm{I}(y_{t_i}=j)}, \quad \text{for } i = 1, \ldots, N, \end{aligned} \tag{9}$$

where $Y_i = \{y_{i,t}\}_{t_i=1}^{T_i}$ represents the sequence of choices by individual i and $\{w_{t_i}\}_{t_i=1}^{T_i}$ represents the set of explanatory variables for every observation of individual i. Due to these adjustments in the approach of determining the simulated probabilities, the simulated log-likelihood function is also computed slightly differently. The SLL equals now

$$SLL = \sum_{i=1}^{N} \log(\hat{\Pr}[Y_i = \{y_{i,t_i}\}_{t_i=1}^{T_i} | \{w_{t_i}\}_{t_i=1}^{T_i}]). \tag{10}$$

4.3 Extensions to MNL

It is shown and explained in Section 2 and 4, that ML has a few advantages over MNL. In this section, three extensions will be illustrated, which are applied to MNL to possibly improve it as a model. First, an extension concerning unobserved heterogeneity will be explained. Secondly, the explanation and implementation of brand loyalty will be shown. At last, the combination of these two extensions will be given.

4.3.1 Unobserved Heterogeneity (MNLH)

It is reasonable to assume that individuals have different base preferences. The heterogeneity between the preferences of individuals can be taken along in a model using individual-specific variables, like age or gender. However, this kind of variable is not always available. In that case,

to take the unobserved heterogeneity along in the model, individual-specific intercept parameters can be used, i.e., $\beta_{0,i,j}$ for $\forall i, j$ (Franses and Paap, 2001). $\beta_{0,i,j}$ is an intercept parameter, which represents the base preference of an individual i towards a category j. To apply this extension to MNL, not much has to be changed to MNL. Instead of using intercept parameters that only alter over the different categories, the intercept parameters now also change over the different individuals. When this adaption is applied, the probability that individual i chooses category j in observation t_i, looks like

$$\Pr[Y_{i,t_i} = j | w_{t_i}] = \frac{\exp(\beta_{0,i,j} + \gamma w_{j,t_i})}{\sum_{l=1}^{J} \exp(\beta_{0,i,l} + \gamma w_{l,t_i})}, \quad \text{for } i = 1, \ldots, N,\ j = 1, \ldots, J,\ t_i = 1, \ldots, T_i, \quad (11)$$

where t_i is an observation for individual i, T_i is the total amount of observations for individual i, and $\beta_{0,i,J}$ is set to zero for identification for every i. So, now the probability is different compared to the "normal" MNL, the likelihood function $\mathcal{L}$ and log-likelihood function ℓ change, as well. The likelihood function looks as follows

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \Pr(Y_{i,t_i} = j | w_{t_i})^{\mathrm{I}[y_{i,t_i}=j]} = \prod_{i=1}^{N} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \left(\frac{\exp(\beta_{0,i,j} + \gamma w_{j,t_i})}{\sum_{l=1}^{J} \exp(\beta_{0,i,l} + \gamma w_{l,t_i})} \right)^{\mathrm{I}[y_{i,t_i}=j]}, \quad (12)$$

where θ represents the parameters for this model, i.e., $\beta_{0,i,j}$ and γ for $\forall i, j$. Due to this transformation of the likelihood function, the log-likelihood function equals

$$\begin{aligned}
\ell(\theta) = \log \mathcal{L}(\theta) &= \log \left(\prod_{i=1}^{N} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \left(\frac{\exp(\beta_{0,i,j} + \gamma w_{j,t_i})}{\sum_{l=1}^{J} \exp(\beta_{0,i,l} + \gamma w_{l,t_i})} \right)^{\mathrm{I}[y_{i,t_i}=j]} \right) \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \mathrm{I}[y_{i,t_i} = j] \log \left(\frac{\exp(\beta_{0,i,j} + \gamma w_{j,t_i})}{\sum_{l=1}^{J} \exp(\beta_{0,i,l} + \gamma w_{l,t_i})} \right) \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \mathrm{I}[y_{i,t_i} = j] (\beta_{0,i,j} + \gamma w_{j,t_i} - \log(\sum_{l=1}^{J} \exp(\beta_{0,i,l} + \gamma w_{l,t_i}))).
\end{aligned} \quad (13)$$

Moreover, the individual-specific intercept parameters can be determined in two ways; estimating the intercept parameter for every individual i and category j, or drawing the intercept parameters from a distribution and estimating the parameters for this distribution (Franses and Paap, 2001). The first way to determine the intercept parameters can cause some problems if some individuals do not consider one or more categories (Franses and Paap, 2001). If that is the case, not every intercept parameter can be estimated, and, consequently, the latter way has to be used to determine the individual-specific intercept parameters. In this paper, the individual-specific parameters are drawn from a normal distribution, and the parameters of the normal distribution are estimated using the data.

4.3.2 Brand Loyalty (MNLB)

Another way to possibly make MNL a better model, is to include brand loyalty of individuals in the model. Brand loyalty implies that, when looking at the purchasing behaviour of an individual, a certain brand is bought most of the times, or even bought all the time. To implement such behaviour in a model, the previous purchases of the individual have to be taken into account. An easy way to do this, is to include a lagged choice variable ($\mathrm{I}[y_{i,t-1} = j]$), that equals one if category j has been bought in the previous period, and zero otherwise (Franses and Paap, 2001). After adding such a variable to MNL, the probability that category j is chosen by individual i at observation t_i equals

$$\Pr[Y_{i,t_i} = j | w_{t_i}, y_{i,t_i-1}] = \frac{\exp(\beta_{0,j} + \gamma w_{j,t_i} + \delta \mathrm{I}[y_{i,t_i-1} = j])}{\sum_{l=1}^{J} \exp(\beta_{0,l} + \gamma w_{l,t_i} + \delta \mathrm{I}[y_{i,t_i-1} = l])}, \quad \begin{array}{l} \text{for } i = 1, \ldots, N, \\ \qquad j = 1, \ldots, J, \\ \qquad t_i = 1, \ldots, T_i, \end{array} \tag{14}$$

where δ represents the influence of the previous purchase being category j.

Furthermore, Guadagni and Little (1983) have included brand loyalty and size loyalty, of different coffee brands, to their logit model in a more advanced way. Instead of using a lagged choice variable, they introduce a variable being an exponentially weighted average of past purchase decisions. This variable, that embodies the brand loyalty towards category j for individual i, is defined as

$$b_{i,j,t_i} = \delta b_{i,j,t_i-1} + (1 - \delta)\mathrm{I}[y_{i,t_i-1} = j], \quad \text{for } i = 1, \ldots, N,\ j = 1, \ldots, J,\ t_i = 1, \ldots, T_i, \tag{15}$$

with $0 < \delta < 1$. For $t_i = 1$, b_{i,j,t_i} is set to δ if category j was chosen, and $(1 - \delta)/(J - 1)$ otherwise, for every individual i. This initialisation is chosen in order to insure that the sum of the variables $b_{i,1,j}$ over the categories j always equal one for every individual (Guadagni and Little, 1983). After including this variable to MNL, the probability that individual i chooses category j at observation t_i is equal to

$$\Pr[Y_{i,t_i} = j | w_{t_i}, b_{i,t_i}] = \frac{\exp(\beta_{0,j} + \gamma w_{j,t_i} + \beta_{1,j} b_{i,j,t_i})}{\sum_{l=1}^{J} \exp(\beta_{0,l} + \gamma w_{l,t_i} + \beta_{1,l} b_{i,l,t_i})}, \ \begin{array}{l} \text{for } i = 1, \ldots, N,\ j = 1, \ldots, J, \\ \qquad t_i = 1, \ldots, T_i, \end{array} \tag{16}$$

where $\beta_{1,j}$ and δ, just as the parameters γ and $\beta_{0,j}$ for $\forall j$, are estimated using the data. The associated log-likelihood function is as follows

$$
\begin{aligned}
\ell(\theta) = \log \mathcal{L}(\theta) &= \log \left(\prod_{i=1}^{N} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} (\Pr[Y_{i,t_i} = j | w_{t_i}])^{\mathrm{I}[y_{i,t_i}=j]} \right) \\
&= \log \left(\prod_{i=1}^{N} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \left(\frac{\exp(\beta_{0,j} + \gamma w_{j,t_i} + \beta_{1,j} b_{i,j,t_i})}{\sum_{l=1}^{J} \exp(\beta_{0,l} + \gamma w_{l,t_i} + \beta_{1,l} b_{i,l,t_i})} \right)^{\mathrm{I}[y_{i,t_i}=j])} \right) \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \mathrm{I}[y_{i,t_i} = j] \log \left(\frac{\exp(\beta_{0,j} + \gamma w_{j,t_i} + \beta_{1,j} b_{i,j,t_i})}{\sum_{l=1}^{J} \exp(\beta_{0,l} + \gamma w_{l,t_i} + \beta_{1,l} b_{i,l,t_i})} \right) \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \mathrm{I}[y_{i,t_i} = j](\beta_{0,j} + \gamma w_{j,t_i} + \beta_{1,j} b_{i,j,t_i} - \log(\sum_{l=1}^{J} \exp(\beta_{0,l} + \gamma w_{l,t_i} + \beta_{1,l} b_{i,l,t_i}))).
\end{aligned}
\tag{17}
$$

4.3.3 Combination (MNLC)

In addition to the individual explanations of the both extensions, the combination of the two extensions is applied. The log-likelihood function for MNL, which includes unobserved heterogeneity, as well as brand loyalty, is equal to

$$
\ell(\theta) = \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \mathrm{I}[y_{i,t_i} = j](\beta_{0,i,j} + \gamma w_{j,t_i} + \beta_{1,j} b_{i,j,t_i} - \log(\sum_{l=1}^{J} \exp(\beta_{0,i,l} + \gamma w_{l,t_i} + \beta_{1,l} b_{i,l,t_i}))) \tag{18}
$$

4.4 Performance measures

To measure the performances of the different models, the predictive power of the models is examined. This is done by splitting up the data set into a training and a test data set, where the training data set is used to estimate the parameters, and the test data set is used to examine whether the models fit the data well.

To measure the predictive power of the models, the accuracy and the F1-score are calculated for every model. The accuracy equals the number of correctly predicted observations divided by the total number of observations in the test data set. This is an easily applicable and understandable way of measuring the predictive power of a model. On the other hand, the F1-score is intuitively not as understandable as the accuracy. The F1-score is determined using the following equation

$$
F1 = \frac{2 \cdot (recall \cdot precision)}{recall + precision}, \tag{19}
$$

where recall is calculated by dividing the number of correctly predicted observations for a category by the total number of observations of the relevant category in the data set, and precision is determined by dividing the number of correctly predicted observations for a category by the total

number of times that that category is predicted. The F1-score is a more useful performance measure than the accuracy, due to the fact that the F1-score also takes the incorrect and missing predictions in account, while the accuracy only considers the correct predictions. The model with the highest accuracy and F1-score is considered the best model.

After considering the predictive power of every model, it should be possible to draw a suitable conclusion concerning the performance of the different models.

5 Results

All of the five models described in Section 4 are implemented using a data set concerning the purchases of different brands of crackers in USA. First, the data set is split randomly over all observations into a training data set containing 70 percent of the data, and a test data set containing the remaining 30 percent of the initial data set. In Table 2, the amount of purchases of every category in the training and test data set is shown. Using the training data set, the parameters for every model are estimated. With the estimated parameters, the performance of every model is measured with the help of the test data set, i.e., it is examined how well the models fit the test data with the estimated parameters from the training data. The methods that help to measure the performance of the different models are explained in Section 4.4.

	Sunshine	Kleebler	Nabisco	Private	Total
Training	168	163	1277	741	2349
Test	71	63	512	294	940

Table 2: Total purchases per category

For ML, a normal distribution is used as the mixing distribution for density function $f(\theta)$, just as for the distribution for the individual-specific parameters of MNLH and MNLC. Furthermore, after considering the accuracy and F1-score for different values for the number of simulations for ML, R, it is found that the accuracy and F1-score keep rising until R=500. For higher values than 500 for R, the accuracy and F1-score stagnate, while it takes more time to run the program if R increases. That is why, 500 is considered the optimum number of simulations for ML.

After implementing the several models, the predictive power of the five models is examined. In Table 3 and 4, the total number of predictions of purchases per category and the total number of correct predictions for every model is shown, respectively. It stands out that the models where unobserved heterogeneity is included (ML, MNLH and MNLC), predict indeed more diversely; especially ML and MNLH. While MNL almost only predicts that the brands Nabisco and Private are bought, which are the most frequently present brands in the test data (see Table 2), MNLB nearly only predicts purchases of Nabisco.

Model	Sunshine	Kleebler	Nabisco	Private
MNL	8	0	775	157
ML	6	76	815	38
MNLH	233	24	678	5
MNLB	15	0	925	0
MNLC	0	44	858	38

Table 3: Total predictions per category for test data set

According to the results in Table 4, predicting diversely is not always the best way, seeing that MNLB predicts the most purchases correctly, while this model predicts in the least diverse way. Besides, it can be seen in Table 4 that MNLH predicts the most purchases incorrectly of all the five models. ML and MNLC have almost the same amount of correctly predicted purchases, and MNL follows closely.

Model	Sunshine	Kleebler	Nabisco	Private	Total
MNL	0	0	420	43	463
ML	3	13	452	18	486
MNLH	13	0	363	3	379
MNLB	0	0	507	0	507
MNLC	0	6	477	10	493

Table 4: Correct predictions per category for test data set

Furthermore, with the help of the results in Table 2, 3 and 4, the accuracy and F1-score of the models can be determined. The resulting accuracy, recall, precision and F1-score of every model is shown in Table 5.

Model	Accuracy	Recall	Precision	F1-score
MNL	0.493	0.242	0.204	0.221
ML	0.517	**0.298**	**0.411**	**0.346**
MNLH	0.403	0.226	0.298	0.257
MNLB	**0.539**	0.248	0.137	0.176
MNLC	0.524	0.265	0.239	0.251

Table 5: Predictive power

In line with the number of correct predictions, shown in Table 4, the accuracy of MNLB is the highest, compared to the other models. ML and MNLC follow closely concerning their accuracy, and MNLH has the lowest accuracy.

As mentioned in Section 4.4, the F1-score is more useful than accuracy for comparing the predictive power of different models. That is a result of not only analysing the correctly predicted purchases, but also the recall and precision, which also consider the missing and incorrect predictions, re-

spectively. Inspecting the F1-scores in Table 5, something remarkable stands out. In contrast to the conclusions about the accuracy of the different models, MNLB has the lowest F1-score of all models. Besides, ML has the highest F1-score with at least eight percentage points difference, when compared to the F1-scores of the other models. MNLH and MNLC perform almost equally well, regarding their F1-scores.

The reason that MNLB performs badly, concerning its F1-score, is that the precision is remarkably lower than the precision of the other models. This is in line with the findings about the results in Table 3 and 4, that is, that MNLB almost only predicts the purchase of Nabisco. Predicting only purchases of Nabisco leads to many correctly predicted purchases, but many incorrectly predictions, too. Consequently, the precision of MNLB becomes relatively low, making the F1-score also low.

Furthermore, the recall and, especially, the precision of ML are substantially high, compared to the other models. Moreover, it is remarkable that MNLH has the lowest recall, but, not considering ML, the highest precision, which results in the second highest F1-score of the five models, while its accuracy is by far the lowest. Moreover, it is found that both MNLH and MNLC perform significantly better than MNL, while ML outperforms MNLH and MNLC significantly.

In summary, MNLB predicts the most purchases correctly, which results in the highest accuracy, however, considering the F1 score, ML performs significantly better than every other model, while MNLB performs the worst. Besides, MNLH and MNLC have significantly higher F1-scores than MNL.

6 Conclusion

This paper focuses on comparing different forms of discrete choice models, applied on scanner data concerning the purchases of different brands of crackers in USA. Five logit models are implemented, namely, multinomial logit (MNL), mixed logit (ML), multinomial logit which accounts for heterogeneity between the customers (MNLH), multinomial logit extended with a brand loyalty variable (MNLB), and multinomial logit with a combination of the two extensions to MNL (MNLC). The models were compared based on their predictive power, where ML outperformed the other models significantly, when the F1-score is considered. To refer back to the hypotheses stated in Section 1, the expectations regarding the relative performance between the different models were proven to be true; ML is indeed a better way to model an unordered categorical dependent variable than MNL, and some of the extensions to MNL were proven to have a significantly positive impact on the predictive performance, but not enough to outperform ML. It can also be concluded, that MNL with the extensions was relatively easy to implement compared to ML. So, if one is not experienced with implementing models, and wants to model an unordered categorical dependent variable, it can be suggested to use MNLH or MNLC, instead of the better performing ML, due to the improved performance in comparison with MNL.

Finally, it can be concluded that ML is the best performing model, and the applied extensions to MNL are helpful to improve MNL.

6.1 Recommendations for future research

In this paper, a data set concerning the purchases of four brands of crackers, is used to compare different discrete choice models. Because of the small amount of variables, and the limited number of different bought brands of crackers by the customers (that is, Nabisco and the private brand of the supermarket were bought approximately 90 percent of the time), it can be questioned if this data set is really useful for the comparison of different discrete choice models. For instance, if a model predicts only purchases of Nabisco, like MNLB, it seems to outperform other models, which predict more diversely. So, it is recommended to do this research using a different kind of data set, where the choices are more diverse, to examine if the same conclusions are drawn as in this paper. Besides, a greater amount of different variables could help to improve the different models, and maybe make the conclusion about the performance of the models more reliable.

Furthermore, for ML, MNLH and MNLC a distribution for (some of) the parameters is required. In this paper, for the three mentioned models, only the normal distribution is used, while other distributions, like the log-normal or triangular distribution, could also be used. Using a different distribution for the parameters can give significantly different results (Hensher and Greene, 2003). It is possible, that using other distributions than the normal distribution for the parameters, the performance of ML, MNLH or MNLC improves.

At last, to evaluate the performance of the five models, the predictive power is considered. To measure the predictive power of the models, the data set is randomly split into a training and test data set. With the training data set, the parameters of the models are estimated, and it is tested how every model fits the test data. In this paper, the data set is split only once, while a k-fold cross validation could be used. The main idea of k-fold cross validation is that, instead of splitting the data set into a train and test part once, it is split k times. Therefore, it gives more realistic estimates than other methods to evaluate predictive power.

References

Alfnes, F. (2004). Stated preferences for imported and hormone-treated beef: application of a mixed logit model. *European Review of Agricultural Economics*, 31(1):19–37.

Algers, S., Bergström, P., Dahlberg, M., and Lindqvist Dillén, J. (1998). Mixed logit estimation of the value of travel time. *Working Paper Series*, (15).

Boyd, J. H. and Mellman, R. E. (1980). The effect of fuel economy standards on the us automotive market: an hedonic demand analysis. *Transportation Research Part A: General*, 14(5-6):367–378.

Brownstone, D. and Train, K. (1998). Forecasting new product penetration with flexible substitution patterns. *Journal of econometrics*, 89(1-2):109–129.

Cardell, N. S. and Dunbar, F. C. (1980). Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General*, 14(5-6):423–434.

Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute Working Paper*, (2002-119/4).

Franses, P. H. and Paap, R. (2001). *Quantitative models in marketing research*. Cambridge University Press.

Guadagni, P. M. and Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238.

Hensher, D. A. and Greene, W. H. (2003). The mixed logit model: the state of practice. *Transportation*, 30(2):133–176.

Hess, S., Bierlaire, M., and Polak, J. W. (2005). Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice*, 39(2-3):221–236.

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.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100.

Rose, J. M., Hess, S., Bliemer, M. C., and Daly, A. (2009). The impact of varying the number of repeated choice observations on the mixed multinomial logit model. *European Transport Conference*, pages 5–7.

Rossi, P. E. and Allenby, G. M. (1993). A bayesian approach to estimating household parameters. *Journal of Marketing Research*, 30(2):171–182.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Ye, F. and Lord, D. (2014). Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. *Analytic methods in accident research*, 1:72–85.

A Appendix

In this section, brief explanations of the R software programs used for this paper are given.

- MNL-logl.R: Computes the log-likelihood for the Multinomial Logit model.
- Mixed Logit logl.R: Computes the simulated log-likelihood for the Mixed Logit model.
- MNL_hetero_distr.R: Computes the log-likelihood for the Multinomial Logit model extended with unobserved heterogeneity.
- MNL_brandloyal.R: Computes the log-likelihood for the Multinomial Logit model extended with a brand loyalty variable.
- MNL_combi.R: Computes the log-likelihood for the Multinomial Logit model extended with unobserved heterogeneity and a brand loyalty variable.
- forecast_MNL.R: Predicts the purchases using the Multinomial Logit model.
- ML_forecast.R: Predicts the purchases using the Mixed Logit model.
- Het_forecast.R: Predicts the purchases using the Multinomial Logit model extended with unobserved heterogeneity.
- brand_forecast.R: Predicts the purchases using the Multinomial Logit model extended with a brand loyalty variable.
- combi_forecast.R: Predicts the purchases using the Multinomial Logit model extended with unobserved heterogeneity and a brand loyalty variable.
- brandloy_var.R:Creates brand loyalty variables.