**Exercise 1 Data Description**

1. **Average and dispersion in product characteristics**.

Below is the mean calculated for all product characteristics,

*Average*

|  |
| --- |
| **PPk\_Stk** **PBB\_Stk** **PFl\_Stk** **PHse\_Stk** **PGen\_Stk** **PImp\_Stk**  **PSS\_Tub PPk\_Tub**  0.5184362 0.5432103 1.0150201 0.4371477 0.3452819 0.7807785 0.8250895 1.0774094  **PFl\_Tub** **PHse\_Tub**  1.1893758 0.5686734  *Dispersion*  **PPk\_Stk PBB\_Stk** **PFl\_Stk PHse\_Stk PGen\_Stk PImp\_Stk PSS\_Tub**  0.15051740 0.12033186 0.04289519 0.11883123 0.03516605 0.11464607 0.06121159  **PPk\_Tub PFl\_Tub PHse\_Tub**  0.02972613 0.01405451 0.07245500  The red line in each of the graph distribution plots shows the mean. |
|  |
| |  | | --- | |  | |

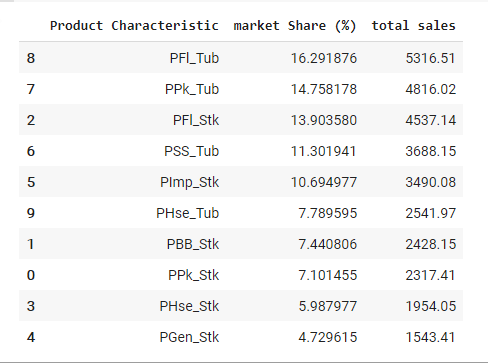
1. **Market share, and market share by product characteristics**

The market share refers to the portion of the market controlled by a particular product. To calculate the market share for each of the product based on its characteristics, we must first find the sum of the purchases made per product characteristic then calculate the total market share by summing up all the sum purchases for each product characteristic.

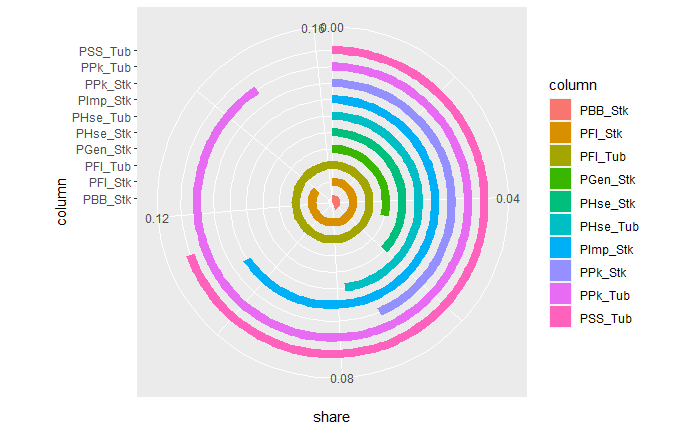
To get the market share for each, we must calculate and find out the proportion of each product characteristic relative to the total market share

I.e. market share by product characteristics **=** sum of the purchases made per product characteristic / the total market share

Results:

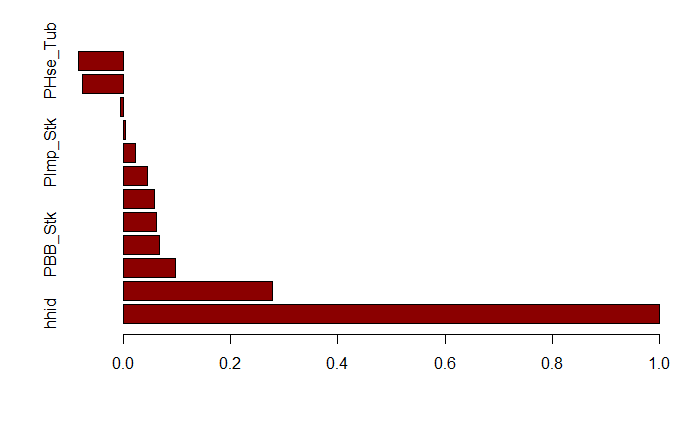


Visualization by bar and pie chart,



1. **Illustrate the mapping between observed attributes and choices**

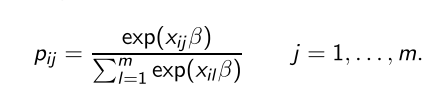
Below is a graph that reveals the correlation or rather relationship between each of the attributes and the final choice.



We see that the PPk\_Stk has the most correlation value with respect to the rest of the observed attributes. This correlation reveals a weak positive linear relationship between PPk\_Stk and choice.. Attributes like PPk\_Tub and PSS\_Tub have small negative correlation with choice. Choice is this case is a categorical attribute.

**Exercise 2 First Model**

For this model, we shall use a conditional logit model from the survival package in r. The conditional logit model is an extension of the ordinary logistic regression that gives room for stratification plus matching as a method of limiting confounding. This model solves classification problems by maximizing the conditional likelihood. The equation of the conditional logit model is:



In this model we recorded our choice variable so that it could support the clogit model such that the resultant dataset set has a copy per outcome level. Additionally, we introduced the variable “tocc” -> target occupation for this copy and case, which validated whether the outcome is really the actual outcome for each observation.

**Below is the r-Implementation**

library(caret)

margarine$choicePrice$choice <- as.factor(margarine$choicePrice$choice)

margarine$choicePrice$choice2 = relevel(margarine$choicePrice$choice,ref="5")

resp <- levels(margarine$choicePrice$choice)

n <- nrow(margarine$choicePrice)

indx <- rep(1:n, length(resp))

logan2 <- data.frame(margarine$choicePrice[indx,],

id = indx,

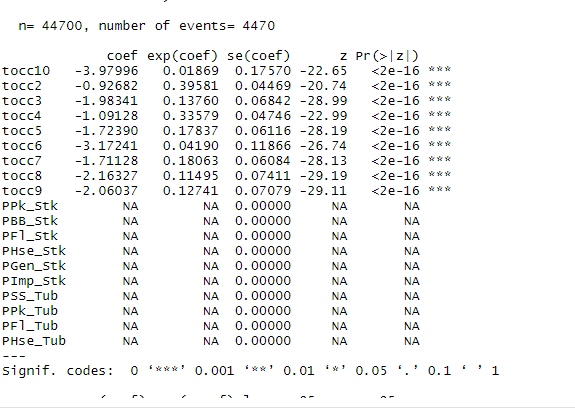
tocc = factor(rep(resp, each=n)))

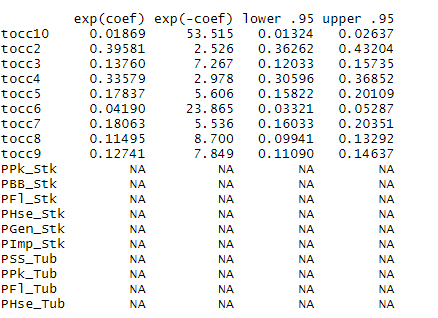
logan2$case <- (margarine$choicePrice$choice == logan2$tocc)

logan2 <- logan2[order(logan2$id),]  
res.clogit <- clogit(case ~ tocc + PPk\_Stk+PBB\_Stk+PFl\_Stk+PHse\_Stk+PGen\_Stk+PImp\_Stk+PSS\_Tub+PPk\_Tub+PFl\_Tub+PHse\_Tub + strata(id), logan2)

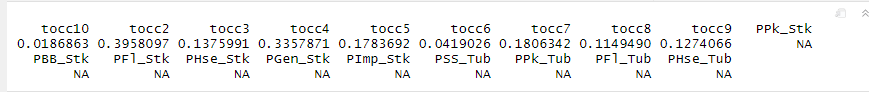
summ.clogit <- summary(res.clogit)

The conditional logit model (clogit) is declared as shown above. After fitting the model, we see the results as below:





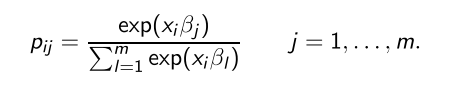
The coefficients were further converted into odds by raising them exponentially such that our new odds were:



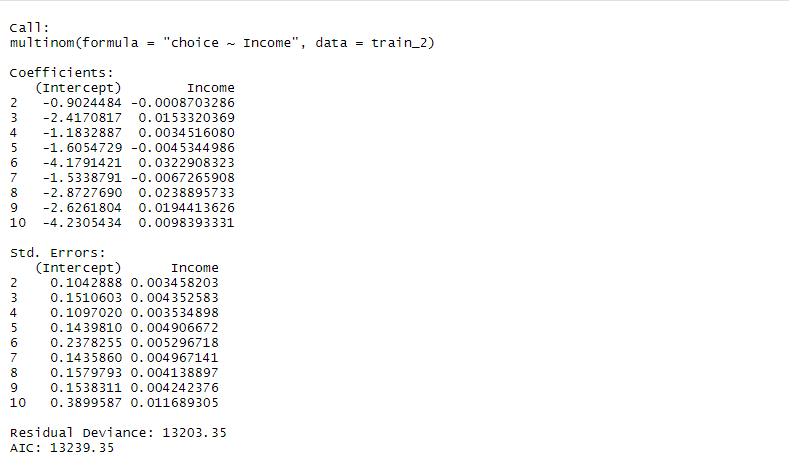
**Exercise 3 Second Model**

The model of choice in this exercise is the logistic regression just as in the previous question. Multinomial logistic regression to be precise with k = 10 (K denoting number of classes). The model - to its best degree - tries to capture any evident trend in terms of the relationship between the target and predictor variable. The model gives back a probability of the predictor of belonging to each of the ten classes through the SoftMax function. The model went through 256 iterations of training, both the “choice Price” and demography dataset were combined such that there was mapping between the two datasets through the hhid. The combined dataset now included the Income column of the demo’s dataset and all columns of the choice price.

The data was further split into the training set and testing set with a ratio of 4: 1. The test set had 894 records while the training set had 3,376 records. The formula is



**Co-efficients:** below is the set of coefficients and bias for each of the ten choices.



Each row represents the weights for each variable to calculate the probability of belonging to a particular class. Since it was a univariate classification problem, the set of co-efficients will not be as complex as that of the previous exercise. The positive and negative means more effects and less effects.

For example, for class 2 = z = -0.0008703286Income + -0.9024484.

**Exercise 4 Marginal Effects**

**Marginal Effects**

In this exercise, we will use sjPlot which is an R package containing the plot model function that plots the marginal effect of a given model. We used it to plot and visualize the marginal effects for both our models, i.e., both the conditional logit model and multinomial logit model and the results are as follows:

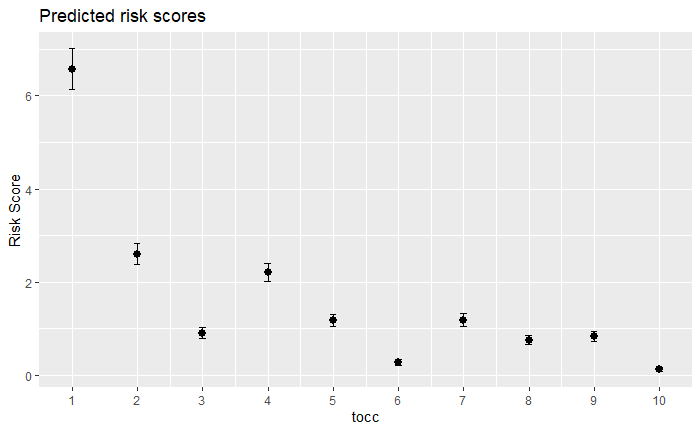
***Model 1***

*R-code Implementation*

plot\_model (res.clogit, type = "pred", terms = "tocc")

**NB:**

Model 1 was a conditional logit that was used to model the multinomial variable choice with respect to the choice’s prices. The data was converted to a long format that is supported by the clogit model. The results are as follows for the “target occupation for this copy and case” variable.



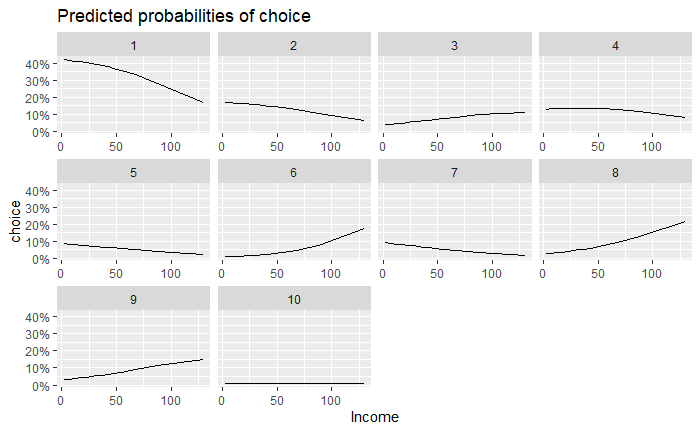
*A predicted risk score was also evaluated for each of the tocc levels.* It’s important to note that the price variables were rejected by the conditional model as it was a stable and had no transitions within the outcomes of a single entry.

***Model 2***

*R-code Implementation*

plot\_model (multinom\_model\_2, type = "pred", terms = "Income")

Below are visualization plots that shows the marginal effects between The Income variable and each of the choices.



For choices 1, we see that as Income increases us starting to witness that the probability of the choice recedes just as seen in classes 2, slightly in 4, 5, and 7.

In contrast, for choices 3, 6 (in a non-linear manner), 8, 9 tends to see an Increase in probability as the Income increases. In the choice 10, there is no marginal effect between the Income and the choice.

**Exercise 5 IIA**

pR2(clogit.model1)

## fitting null model for pseudo-r2  
## # weights: 20 (9 variable)  
## initial value 10292.555366   
## iter 10 value 8386.905805  
## final value 8285.856670   
## converged

## llh llhNull G2 McFadden r2ML   
## -7059.4930049 -8285.8566698 2452.7273299 0.1480069 0.4223046   
## r2CU   
## 0.4329301

pR2(mlogit.model2)

## fitting null model for pseudo-r2  
## # weights: 20 (9 variable)  
## initial value 10292.555366   
## iter 10 value 8386.905805  
## final value 8285.856670   
## converged

## llh llhNull G2 McFadden r2ML   
## -8.285857e+03 -8.285857e+03 5.863891e-06 3.538494e-10 1.311832e-09   
## r2CU   
## 1.344839e-09

In the model we considered mixed logit setting so as to realize the effect of price and family income. Thus, the optimize the likelihood of the mixed logit is as shown below.

## Concordance= 0.777 (se = 0.004 )  
## Likelihood ratio test= 4013 on 9 df, p=<2e-16  
## Wald test = 3481 on 9 df, p=<2e-16  
## Score (logrank) test = 5189 on 9 df,

The probabilities do not exhibit the well-known independence from irrelevant alternatives property (IIA), and different substitution patterns are obtained by appropriate specification of f. The mixed logit model recognizes the role of such information and handles it in two ways (both leading to the same model only when the random effects model has a non-zero mean).

Likelihood-ratio test assesses the goodness of fit of two competing statistical models based on the ratio of their likelihoods, specifically one found by maximization over the entire parameter space and another found after imposing some constraint. Therefore, the optimized likelihood of the mixed logit is ## Likelihood ratio test= 4013. on 9 df, p=<2e-16

This is an implementation of the Hausman and Mcfadden’s consistency test for multinomial logit models. If the independence of irrelevant alternatives applies, the probability ratio of every two alternatives depends only on the characteristics of these alternatives. Consequently, the results obtained on the estimation with all the alternatives or only on a subset of them are consistent, but more efficient in the first case. On the contrary, only the results obtained from the estimation on a relevant subset are consistent.