**Homework 3: Things to do with Crackers data**



You should use the **Crackers.csv** dataset for this homework. The dataset **Crackers.csv** has data on grocery store purchases of Crackers brands. Each observation denotes the consumer decision on one purchase occasion. There are three major brands in this market **Keebler**, **Sunshine**, and **Nabisco**. All other brands have been aggregated as the **Private** in this dataset. In each purchase occasion, the shopper bought only one of these brands. The description of columns is as follows:

* 1. **OBS:** = Observation number
  2. **Choice Columns:** = **Private**, **Keebler**, **Sunshine**, **Nabisco** are Indicator variables for which brand was chosen. Value of 1 indicates the brand that was chosen. Other 3 brands will be 0 in that observation.
  3. **Price Columns:** = **PricePrivate**, **PriceNabisco**, **PriceKeebler** and **PriceSunshine** contain prices that were offered by each brand for that purchase occasion.
  4. **DisplPrivate:** = 1 if Private had a store display, = 0 if Private did not have a store display
  5. **DisplKeebler:** = 1 if Keebler had a store display, = 0 if Keebler did not have a store display
  6. **DisplSunshine:** = 1 if Sunshine had a store display, = 0 if Sunshine did not have a store display
  7. **DisplNabisco:** = 1 if Nabisco had a store display, = 0 if Nabisco did not have a store display
  8. **FeatPrivate:** = 1 if Private had a store feature, = 0 if Private did not have a store feature
  9. **FeatKeebler:** = 1 if Keebler had a store feature, = 0 if Keebler did not have a store feature
  10. **FeatSunshin:** = 1 if Sunshine had a store feature, =0 if Sunshine did not have a store feature
  11. **FeatNabisco:** = 1 if Nabisco had a store feature, =0 if Nabisco did not have a store feature

**GOOD LUCK!** 

1. Based on the previous lectures, provide a professional summary statistic of this dataset. Your summary should briefly explain the status of crackers brands in the market, based on the above dataset? For example: which brand has the highest market share? Is there any difference between the average prices of these brands? Which brand does provide display or feature in store more frequently? And, …

1. Randomly select 80% of the data set as the training sample, remaining 20% as test sample. Please set the seed=2.
2. The brand manager would like to predict the choice probabilities for each brand of crackers depending on the price, display, feature for all brands in the market. Write the general utility model to estimate this logit model. A consumer receives a utility by purchasing each brand. You can assume that her utility is function of the of the chosen brand and **unobserved error term**. You should write the utility equation of each choice precisely.
3. Is the data formatted as needed to estimate the above multinomial logit model using PROC LOGISTIC or PROC MDC? If not, how should the data be formatted? Reformat the data as necessary (**HINT:** For estimating a multinomial logit model, each observation should correspond to the data for one purchase occasion. For estimating a multinomial logit model with alternative-specific characteristic, each observation should correspond to the data about the characteristics of one brand for each purchase occasion. So, there should be 4 observations corresponding to the 4 brands for each purchase occasion. ***You can check this*** [***link***](https://support.sas.com/documentation/onlinedoc/ets/132/mdc.pdf) ***at pages 992-993 to learn how to reformat the data set.***)
4. Estimate the logit model on the training sample using PROC LOGISTIC and report the estimation results. Please set the **Private** intercept equal to zero as the based. You need to report model parameters, significance. (HINT: You should use the training data to estimate the model. Remember to us the Q4 reformatted dataset.) Also, provide bullet points on the meaning of the estimations results in terms of consumers’ utility by purchasing a brand.
5. Reproduce your results in Q5 using PROC MDC (HINT: See the SAS code posted for the lecture for examples of replicating the results with PROC MDC. You can use the same dataset format. Refer to the SAS manual for more details about [PROC MDC](https://support.sas.com/documentation/onlinedoc/ets/132/mdc.pdf). ***You will need to use “type=clogit” to estimate a multinomial model, and “nchoice=4” to indicate there are four alternatives for each choice occasion***.

In PROC MDC, using the CLASS statement for a categorical variable with N levels will create N dummy variables, each for one level of the categorical variable. Use the restrict statement to set the coefficient for one of the dummy variables to zero – effectively omitting this dummy variable. You will need to do this for the main effects and any interaction effects that involve the variable used in the CLASS statement – refer to the SAS code for the lecture for an example. Also, you cannot insert the interaction effect

directly into PROC MDC. First define a new variable as ,then insert it into model.

1. Now use the model in Q5, and estimate the PROBIT model using PROC MDC. Do the estimation results look similar to Q6? Explain the differences.

1. In this Question, you need to estimate three new versions of the PROBIT model used in Q7.
2. Estimate the model where the error terms are normally distributed, independently and identically. This is equivalent that the covariance matrix is the identity matrix.
3. Estimate the model where the error terms are normally and independently distributed, but allow for existence of heteroscedasticity. This is equivalent that the covariance matrix is a diagonal matrix.
4. Estimate the model where the error terms are normally distributed, there is no existence of heteroscedasticity, and the errors are correlated. This is equivalent that the covariance matrix is a matrix such that (1) all diagonal elements are equal to 1, and (2) there are non-zero correlations among alternatives (i.e., the off-diagonal elements are non-zero).

(HINT: See the SAS code posted for the lecture about PROC MDC. Use the ***restrict statement*** to set the right parameters to zero to generate the above models.

Note that all the above four models (Q7 and three models in Q8) are nested model. Therefore, you can use the Loglikelihood to choose the best model that fits into data. ***Now, based on your answer, is there any evidence against IIA property in this data, which is existed by using logit model in Q6?***

1. Now re-consider your model in Q6, i.e., the logit model. First, find the predicted probability for each brand on your ***test dataset***. Nabisco brand manager believes if the predicted probability of busying Nabisco is greater than 80%, the consumer will buy it for sure. Use the predicted probabilities and 80% threshold to find Nabisco market share based on ***test dataset***.

(HINT: PROC LOGISTIC does not support the feature of estimating the model on one portion of the dataset and predicting for another portion. You can do this in PROC MDC by having one dataset with training and test data, and setting the choice data to be missing for the test sample (e.g., create a new variable that indicates the chosen alternative for each choice occasion and set it to missing by assigning a missing value – if x is the choice variable, then the “x =.” sets the value to missing. PROC MDC will estimate the model only for the observations with non-missing choice data. Predictions using the output statement are made using the estimated model for the entire dataset including the observations with missing choice data. ***You can check this*** [***link***](https://support.sas.com/documentation/onlinedoc/ets/132/mdc.pdf) ***at pages 994-996 to learn how to reformat the data set.***)

Now, the Nabisco brand manager has two strategies to increase its market share. I- ***reducing their price by 15%*** or II- ***having feature in store always and increase its price by 15%***. Now, based on ***test dataset*** and your model in Q6, find how far the Nabisco’s market share will increase by implementing one of these two strategies? **You must quantify the increase by each marketing strategy.** If we assume the implementation cost of both strategies are equal, which one will be more profitable? (Hint: You can follow the threshold 50% to find who buys Nabisco under the above strategies in ***test dataset***.)

1. Provide a paragraph which summarizes your analyses in this homework (or any other important trend you find in this data). For example: Which variables have positive effects on the Sales of these products? Which variables have negative effects on the Sales of these products? As a brand manager, what will be your suggestion to increase the market share of these brands in the market? And …

**Guideline**: For question 10, you should provide a professional, managerial paragraph to summarize your analyses. It must be short, informative, and inclusive. ***The last sentence must provide a clear strategy for this firm***. What can be done to increase the sales? Question 10 will be marked very competitively. The best answer will receive the full mark. We will decrease marks according to your rank among groups.