Homework 5

Homework dataset (Crackers): This dataset contains store sales data of crackers at a supermarket that carries four brands of crackers. Each observation corresponds to one purchase occasion and provides data on the price, display and feature of each brand as well as which brand was chosen.

1. **OBS** : = Observation number
2. **Private, Keebler, Sunshine, Nabisco**: 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. **PricePrivate, PriceNabisco, PriceKeebler and PriceSunshine**: 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. **DisplSunshin**:= 1 if **Sunshin** had a store display, =0 if **Sunshin** 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 **Sunshin** had a store feature, =0 if **Sunshin** did not have a store feature
11. **FeatNabisco**: = 1 if **Nabisco** had a store feature, =0 if **Nabisco** did not have a store feature

Homework Questions:

1) Use PROC SURVEYSELECT to sample the original data into training and testing data sets. Use 75% for training and 25% for testing. Use the seed= option to set random seed to a value of 23.

2) The store manager would like to predict the choice probabilities for each brand of crackers depending on the price, display and promotion for all brands. What type of multinomial logit model would you estimate – a model with alternative-specific characteristics or with individual-specific characteristics? Write the general utility model to estimate this logit model.

3) Is the data formatted as needed to estimate the above multinomial logit model using PROC LOGISTIC? If not, how should the data be formatted? Reformat the data as necessary.
   (HINT: For estimating a multinomial logit model with individual-specific characteristic, 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 modify reformat SAS codes used during the lecture.)

4) Estimate the logit model on the training sample using PROC LOGISTIC and report the estimation results (model parameters, significance).

(**HINT:** You should use the training data to estimate the model. This is straightforward with PROC LOGISTIC**)**

5) Reproduce your results using multinomial discrete choice command 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. 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. To reproduce the results of PROC LOGISTIC, you will need to restrict the coefficients of the dummy variables that are omitted in PROC LOGISTIC to zero in PROC MDC)

6) Use PROC MDC to predict the choice probabilities for the test sample using the estimated model.

(**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.

Use this documentation for your help:
https://support.sas.com/documentation/cdl/en/etsug/67525/HTML/default/viewer.htm#etsug_mdc_sect003.htm)