**Homework 2: Things to do with Heinz Hunts data**



You should use the **Heinz Hunts Data.csv** dataset for this homework. The dataset **Heinz Hunts Data.csv** has data on grocery store purchases of Heinz and Hunts ketchup brands. Each observation denotes the consumer decision on one purchase occasion. In each purchase occasion, the shopper bought only one of these two brands. The description of columns is as follows:

if Heinz was purchased, =0 if Hunts was purchased

1. **OBS:** The observation ID
2. **HOUSEHOLDID**: The Consumer ID
3. **HEINZ:** It is equal 1 if Heinz was purchased; otherwise it is zero
4. **HUNTS:** It is equal 1 if Hunts was purchased; otherwise it is zero
5. **PriceHeinz:** Price of Heinz per ounce
6. **PriceHunts:** Price of Hunts per ounce
7. **DisplayHeinz:** = 1 if Heinz had a store display, =0 if Heinz did not have a store display
8. **DisplayHunts:** = 1 if Hunts had a store display, =0 if Hunts did not have a store display
9. **FeatureHeinz:** = 1 if Heinz had a store feature, =0 if Heinz did not have a store feature
10. **FeatureHunts:** = 1 if Hunts had a store feature, =0 if Hunts did not have a store feature

Please answer the following questions in a report document. Your submission should include your report and a SAS code. Please print your SAS code and attach it as an appendix to your report too. The SAS code should replicate your reported results completely. Please report your results based on **2 decimal places**.

**GOOD LUCK!**

1. Based on the previous lectures, provide a professional summary statistic of this dataset. Your summary should briefly explain the status of ketchup 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, … Give a picture of the market status based on the data. **Make sure that you must do research to understand the meaning of display and feature in store shopping.**
2. Randomly select 80% of the data set as the training sample, remaining 20% as test sample. Please set the seed=2.
3. Estimate a linear probability model for the probability that Heinz is purchased – using **PriceHeinz**, **PriceHunts,** **DisplayHeinz**, **DisplayHunts**, **FeatureHeinz**, and **FeatureHunts** as the explanatory variables. **Also, include interaction terms between display and feature for a particular brand** (e.g., **DisplayHeinz \* PriceHeinz,** **DisplayHunts \* PriceHunts, FeatureHeinz \* PriceHeinz**, and **FeatureHunts \* PriceHunts**). Provide a bullet point for each estimated parameter to explain its effect. Make sure you explain the interaction effects carefully.
4. Estimate a logit probability model for the probability that Heinz is purchased – using **PriceHeinz**, **PriceHunts,** **DisplayHeinz**, **DisplayHunts**, **FeatureHeinz**, and **FeatureHunts** as the explanatory variables. **Also, include interaction terms between display and feature for a particular brand** (e.g., **DisplayHeinz \* PriceHeinz,** **DisplayHunts \* PriceHunts, FeatureHeinz \* PriceHeinz**, and **FeatureHunts \* PriceHunts**). Provide a bullet point for each estimated parameter to explain its effect. Make sure you explain the interaction effects carefully.
5. Estimate a probit probability model for the probability that Heinz is purchased – using **PriceHeinz**, **PriceHunts,** **DisplayHeinz**, **DisplayHunts**, **FeatureHeinz**, and **FeatureHunts** as the explanatory variables. **Also, include interaction terms between display and feature for a particular brand** (e.g., **DisplayHeinz \* PriceHeinz,** **DisplayHunts \* PriceHunts, FeatureHeinz \* PriceHeinz**, and **FeatureHunts \* PriceHunts**). Provide a bullet point for each estimated parameter to explain its effect. Make sure you explain the interaction effects carefully.
6. Now, you should compare the models’ goodness-of-fit (on training dataset) in Q4 and Q5. Based on AIC criterion. Which one does explain the data better? Is there any significance difference between logit model and probit model in terms of explanatory power? If Yes or No Why? (Hint: Find difference of AIC of models and follow IIA property)
7. Based on the estimated logit model in Q4, and using the logit probability formula, calculate the change in predicted probability that Heinz is purchased () if the average price of Heinz dropped by 20% (i.e., changes from to , when Hunts charges the its average price, (i.e., ), Heinz does not use a feature or display, while Hunts uses a feature and a display (**Hint:** Recall that in the logit model: , where Y is the outcome variable, X are the predictor variables, and are the estimated model coefficients in Q5.)

1. Based on the estimated logit model in Q4, make predictions for the test data (which has been constructed in Q2.) You need to plot ROC curves for the test data based **all parameters in model Q4 (including the interaction terms** (e.g., **DisplayHeinz\*PriceHeinz,** **DisplayHunts \* PriceHunts, FeatureHeinz \* PriceHeinz**, and **FeatureHunts \* PriceHunts**), respectively. Based on the area under the ROC curves (AUC), which of the above explanatory variable is a more important factor in your prediction analysis? Does your model in Q4 provide a significant better prediction result relative to simple models which contain only one of the above explanatory variables?

Now, find the ROC curve of the estimated linear probability in Q3 by making predictions for the test data. Based on the area under the ROC curves (AUC), which of the above models, in Q3 and Q4, provide you a better predictive tool on the consumers’ ketchup purchase pattern?

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 Heinz or Hunts? Which variables have negative effects on the Sales Heinz or Hunts? As a brand manager, what will be your suggestion to increase the market share of these two brands in the market? And …

**Guideline**: For question 9, you should provide a professional, managerial one-page 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 9 will be marked very competitively. The best answer will receive the full mark. We will decrease marks according to your rank among groups.