Spring 2019

BUAN 6337: Predictive Analytics using SAS

A Report of

Group Assignment 3 on Heinz Hunts Data Analysis

Submitted by Group 7

Vinay Singh 2021441554
Vaibhav Shrivastava 2021434681
Megan Malisani 2021440151
Pragati Mishra 2021434655
Ishan Jain 2021426222
Erhao Liang 2021435949

Under the Guidance of

Prof. Shervin Shahrokhi Tehrani

Data Overview: 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

A. **OBS:** The observation ID
B. **HOUSEHOLDID**: The Consumer ID
C. **HEINZ:** It is equal 1 if Heinz was purchased; otherwise it is zero
D. **HUNTS:** It is equal 1 if Hunts was purchased; otherwise it is zero
E. **PriceHeinz:** Price of Heinz per ounce
F. **PriceHunts:** Price of Hunts per ounce
G. **DisplayHeinz:** = 1 if Heinz had a store display, =0 if Heinz did not have a store display
H. **DisplayHunts:** = 1 if Hunts had a store display, =0 if Hunts did not have a store display
I. **FeatureHeinz:** = 1 if Heinz had a store feature, =0 if Heinz did not have a store feature
J. **FeatureHunts:** = 1 if Hunts had a store feature, =0 if Hunts did not have a store feature

1. Summary Statistics

There are 300 Household ID's with 2798 Observations in our dataset consisting of consumption of Heinz and Hunt Ketchup brands. The classification of variables is listed below:

Feature Relevance

- **Qualitative Variables:** HOUSEHOLDID | HEINZ | HUNTS | DISPLAYHEINZ | DISPLAYHUNTS FEATUREHEINZ |FEATUREHUNTS
- **Quantitative Variables**: PRICEHEINZ | PRICEHUNTS

Market Share

We also came to know that Heinz has 89% of the total Market share when compared to 10.97% of Hunt Market Share. Most of the Market Share is occupied by Heinz.

HEINZ	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	307	10.97	307	10.97
1	2491	89.03	2798	100.00

HUNTS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2491	89.03	2491	89.03
1	307	10.97	2798	100.00

Average Price Comparison

We can see that Heinz Ketchup is little bit expensive when compared to Hunts Ketchup while comparing the Average Price of the ketchups.

- Average Price of Heinz per ounce: $0.0348
- Average Price of Hunts per ounce: $0.0335

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PRICEHEINZ	2798	0.03483	0.00897	97.44767	0.00100	0.06100
PRICEHUNTS	2798	0.03356	0.00532	93.90400	0.00300	0.08700

Display & Feature Importance

- Heinz is featured more on Display when compared to Hunts. 16% of the total observation comprises of Heinz on a store display when compared to 3.54% of Hunts Sauce on Display.
- Like Store Display, Heinz is featured more in a Store. 12.47% of the total observation comprises of Heinz being featured in a store when compared to only 3.65% of Hunts Sauce featured in Store.

DISPLAYHEINZ	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2351	84.02	2351	84.02
1	447	15.98	2798	100.00

DISPLAYHUNTS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2699	96.46	2699	96.46
1	99	3.54	2798	100.00

FEATUREHEINZ	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2449	87.53	2449	87.53
1	349	12.47	2798	100.00

FEATUREHUNTS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2696	96.35	2696	96.35
1	102	3.65	2798	100.00

Store Display and Price

- The mean price of Heinz when it is not on Store Display is $0.0349 per ounce as compared with $0.0342 per ounce when it is on Store Display.
- We cannot see much difference between the average price of Heinz Ketchup when it is on Store Display, but we can say that the Price of Heinz Ketchup is less when it is on Store Display.
- The mean price of Hunts when it is not on Store Display is 0.0337 per ounce as compared with 0.0277 per ounce when it is on Store Display.
- We can clearly see that the average Price of the Hunts Ketchup is substantially less when it is on Store Display

Summary Statistics by Display

The MEANS Procedure

DISPLAYHEINZ=0

Analysis Variable : PRICEHEINZ	
N	Mean
2351	0.0349396

DISPLAYHEINZ=1

Analysis Variable : PRICEHEINZ	
N	Mean
447	0.0342386

Summary Statistics by Display

The MEANS Procedure

DISPLAYHUNTS=0

Analysis Variable : PRICEHUNTS	
N	Mean
2699	0.0337729

DISPLAYHUNTS=1

Analysis Variable : PRICEHUNTS	
N	Mean
99	0.0277879

Heinz: Feature Display and Price

- The mean price of Heinz when it is not on Feature Display is 0.0352 per ounce as compared with 0.0318 per ounce when it is on Feature Display.
- The mean price of Hunts when it is not on Feature Display is 0.0337 per ounce as compared with 0.0359 per ounce when it is on Feature Display.
- We cannot see much difference between the average price of Heinz & Hunts Ketchup when it is on Feature Display, but we can say that the Price of Heinz & Hunts Ketchup is less when it is on Feature Display.

Summary Statistics by Feature

The MEANS Procedure

FEATUREHEINZ=0

Analysis Variable : PRICEHEINZ	
N	Mean
2449	0.0352468

FEATUREHEINZ=1

Analysis Variable : PRICEHEINZ	
N	Mean
349	0.0318863

Summary Statistics by Feature

The MEANS Procedure

FEATUREHUNTS=0

Analysis Variable : PRICEHUNTS	
N	Mean
2696	0.0334714

FEATUREHUNTS=1

Analysis Variable : PRICEHUNTS	
N	Mean
102	0.0359314

<u>Correlation between Price of Heinz & Hunts</u>

- We can see that there is a negative correlation between the Price of Heinz and Hunts, but they are only 10% correlated. Hence, we can assume that these metrics are not correlated.

Pearson Correlation Coefficients, N = 2798 Prob > |r| under H0: Rho=0		
	PRICEHEINZ	PRICEHUNTS
PRICEHEINZ	1.00000	-0.10869 <.0001
PRICEHUNTS	-0.10869 <.0001	1.00000

2. Setting Aside Test Data

Please see SAS code as the end of the report.

3. Estimating Linear Probability Model

- We coded 1 = purchase Heinz; 0 = purchase Hunts
- The overall probability Heinz is purchased is:

$$\mathrm{E}[\mathrm{Y}] = 1 \times \Pr(\mathrm{Y} = 1) + 0 \times \Pr(\mathrm{Y} = 0) = \Pr(\mathrm{Y} = 1) = .89$$

- The linear regression model we are estimating:

$$\mathrm{Y} = \beta_0 + \beta_1 I_{Price\ Hz} + \beta_2 I_{Price\ Ht} + \beta_3 I_{Display\ HZ} + \beta_4 I_{Display\ Ht} + \beta_5 I_{Feature\ Hz} + \beta_1 I_{Feature\ Ht} + \beta_1 I_{Display*Price\ Hz} + \beta_1 I_{Display*Price\ Ht} + \epsilon$$

Linear Regression Model

The GLM Procedure

Dependent Variable: HEINZ

Weight: Selected Selection Indicator

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	35.4825233	4.4353154	54.13	<.0001
Error	2230	182.7086334	0.0819321		
Corrected Total	2238	218.1911568			

R-Square	Coeff Var	Root MSE	HEINZ Mean
0.162621	32.14075	0.286238	0.890576

Explanation of Parameter Estimates

- We can see that the price of Heinz and price of Hunts are the biggest influence factor on the chance of purchasing Heinz. Other factors that are significant at the 5% level include Display of Hunts and Feature of Hunts which two have relatively small effect on the probability of purchasing Heinz.
  - If the price of Heinz increases by $1, the probability of buying Heinz will decrease by 8.753
  - If the price of Hunts Increase by $1, the Probability of buying Heinz will Increase by 13.723
  - If the Display of Hunts exists, the purchasing chance of Heinz will decrease by 0.885
  - If the Feature of Hunts exists, the purchasing chance of Heinz will decrease by 0.0714
  - The effects of other parameters are estimated by the model, but there is not evidence at the 5% level to suggest these relationships truly exist in reality.

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	0.73162358	B	0.04890339	14.96	<.0001
PRICEHEINZ	-8.75305389	B	0.71322218	-12.27	<.0001
PRICEHUNTS	13.72300526	B	1.19889018	11.45	<.0001
DISPLAYHEINZ 1	-0.03722429	B	0.07820205	-0.48	0.6341
DISPLAYHEINZ 0	0.00000000	B	.	.	.
DISPLAYHUNTS 1	-0.88504368	B	0.44125347	-2.01	0.0450
DISPLAYHUNTS 0	0.00000000	B	.	.	.
FEATUREHEINZ 1	0.00933878	B	0.01926489	0.48	0.6279
FEATUREHEINZ 0	0.00000000	B	.	.	.
FEATUREHUNTS 1	-0.07138969	B	0.03335318	-2.14	0.0324
FEATUREHUNTS 0	0.00000000	B	.	.	.
PRICEHEIN*DISPLAYHEI 1	2.67259974	B	2.24776215	1.19	0.2346
PRICEHEIN*DISPLAYHEI 0	0.00000000	B	.	.	.
PRICEHUNT*DISPLAYHUN 1	25.51679384	B	15.68820092	1.63	0.1040
PRICEHUNT*DISPLAYHUN 0	0.00000000	B	.	.	.

4. Selecting Best Linear Model

For the best model, AIC = -3495.79 and MSE_Test = 0.08.

The formula used to calculate AIC here is: $n \ln\left(\frac{SSE}{n}\right) + 2p$

In order to compare the AIC of this model with the logistic model in problem 5, AIC is calculated using a different formula in problem 7. Please see problem 7 for details.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	46.87112	5.85889	76.26	<.0001
Error	2230	171.32004	0.07683		
Corrected Total	2238	218.19116			

Root MSE	0.27717
Dependent Mean	0.89058
R-Square	0.2148
Adj R-Sq	0.2120
AIC	-3495.79257
AICC	-3495.69382
SBC	-5685.36850
ASE (Train)	0.07652
ASE (Test)	0.07678

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.619699	0.169664	15.44	<.0001
PRICEHEINZ	1	-62.788608	4.697441	-13.37	<.0001
PRICEHUNTS	1	-41.182584	4.852833	-8.49	<.0001
PRICEHEIN*PRICEHUNTS	1	1578.877673	135.092322	11.69	<.0001
DISPLAYHEINZ_1	1	-0.097958	0.075772	-1.29	0.1962
PRICEHEINZ*DISPLAYHEINZ_1	1	4.239363	2.181760	1.94	0.0521
DISPLAYHUNTS_1	1	0.318963	0.226326	1.41	0.1589
PRICEHEINZ*DISPLAYHUNTS_1	1	-10.688053	5.522733	-1.94	0.0531
FEATUREHUNTS_1	1	-0.077234	0.032276	-2.39	0.0168

5. Estimating Logistic Regression Model

The price of Heinz and price of Hunts are the biggest influence factors on the chance of purchasing Heinz. Other factors that are significant at the 5% level include Feature of Hunts and Feature of Heinz which have relatively small effects on the probability of purchasing Heinz. This is a highly price-sensitive situation. Price change of one product can highly affect the demand of other product.

Explanation of Parameter Estimates

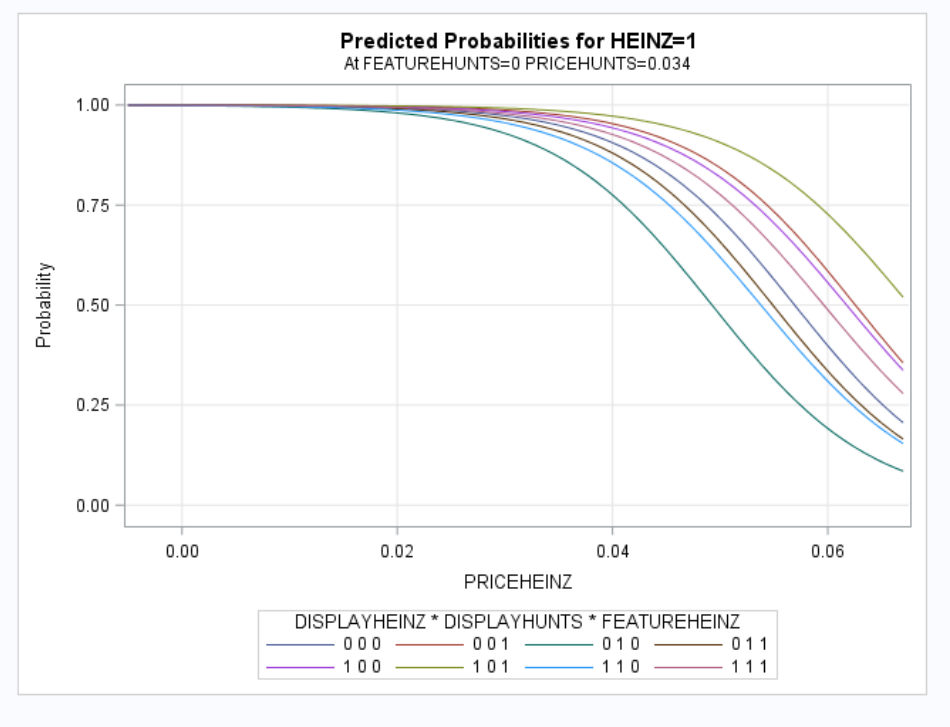
- If the price of Heinz increases by $1, the ratio of odds of buying Heinz will decrease by $e^{131.4}$
- If the price of Hunts Increase by $1, the ratio of odds of buying Heinz will increase by $e^{179.2}$
- If the Display of Hunts exists, the ratio of odds of buying Heinz will increase by $e^{1.2086}$
- If the Feature of Hunts exists, the ratio of odds of buying Heinz will increase by $e^{0.5184}$
- The effects of other parameters are estimated by the model, but there is no evidence at the 5% level to suggest these relationships truly exist in reality.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.1164	2.1335	0.2738	0.6008
PRICEHEINZ		1	-131.4	19.7468	44.2726	<.0001
PRICEHUNTS		1	179.2	69.9109	6.5728	0.0104
DISPLAYHEINZ	1	1	0.1730	0.7843	0.0486	0.8254
DISPLAYHUNTS	1	1	1.2086	1.9549	0.3822	0.5364
FEATUREHEINZ	1	1	0.3774	0.1900	3.9465	0.0470
FEATUREHUNTS	1	1	-0.5184	0.1916	7.3199	0.0068
PRICEHEIN*DISPLAYHEI	1	1	2.4293	19.7426	0.0151	0.9021
PRICEHUNT*DISPLAYHUN	1	1	-51.2824	69.9314	0.5378	0.4634

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
FEATUREHEINZ 1 vs 0	2.127	1.010	4.479
FEATUREHUNTS 1 vs 0	0.355	0.167	0.751



6. Estimating Probit Regression Model

Just like in the logistic regression model from problem 5, we can see that the price of Heinz and price of Hunts are the biggest influence factors on the chance of purchasing Heinz. Other significant factors include Feature of Hunts and Feature of Heinz which have relatively small effects on the probability of purchasing Heinz, which is also true in problem 5. A high price sensitivity is shown in this market segment. The demand of one brand can be highly affected by competitor's price.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4728	0.5346	0.7821	0.3765
PRICEHEINZ		1	-75.1116	10.1013	55.2916	<.0001
PRICEHUNTS		1	113.5	11.0610	105.3031	<.0001
DISPLAYHEINZ	1	1	0.0923	0.3935	0.0551	0.8145
DISPLAYHUNTS		1	0.5254	2.3146	0.0515	0.8204
FEATUREHEINZ	1	1	0.1710	0.0908	3.5455	0.0597
FEATUREHUNTS	1	1	-0.2855	0.1055	7.3294	0.0068
PRICEHEIN*DISPLAYHEI	1	1	1.0860	10.0924	0.0116	0.9143
PRICEHUNT*DISPLAYHUN		1	-29.3452	82.7456	0.1258	0.7229

7. Logit vs Linear Model

Linear and Logistic models can be compared using AIC and BIC. After ensuring that the same AIC formula is used for both models, we found that the logistic model outperformed the linear model. The logistic model had AIC=1140.50 while the linear model had AIC=1544.85.

The formula used for calculating AIC in PROC LOGISTIC is :

Akaike's information criterion:

$$AIC = -2 \text{ Log } L + 2p$$

Through use of this formula, SAS output for the PROC REG statement returns AIC=1140.50:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1548.294	1140.499
SC	1554.008	1191.923
-2 Log L	1546.294	1122.499

In order to compare the models, we must calculate AIC for the LPM model in the same way.

First, we need -2LogL:

$$-2logL = n + n\log(2\pi) + n\log\left(\frac{RSS}{n}\right)$$

$$-2logL = 2239 + 2239 * \log(2\pi) + 2239 * \log\left(\frac{171.32}{2239}\right) = 1526.85$$

Then, we calculate AIC:

$$AIC = -2LogL + 2p$$

$$AIC = 1526.85 + 2 * 9 = 1544.85$$

8. Expected Effect of Heinz Price Drop

The change in predicted probability that Heinz is purchased if the average price of Heinz dropped by 20%=89.30%

- P(Heinz) = exp(Beta X)/(1 + exp(Beta X))
- Beta(X) = 1.116 -131.4(PriceHeinz) + 179.2(PriceHunts) + 0.1730(DisplayHeinz) + 1.2086(DisplayHunts) + 0.3774(FeatureHeinz) -0.5184(FeatureHunts)
- Average price of Heinz = 0.03
- Average price of Heinz = 0.03
- Drops by 20% average price of Heinz = 0.03 –0.20 * 0.03 = 0.024
- Drops by 20% average price of Hunts = 0.03 –0.20 * 0.03 = 0.024
- Heinz does not use display and feature whereas Hunts uses display and feature. Thus, the equation is:
- Beta(X) = 1.116 -131.4*0.024 + 179.20.024 + 0.1730*0 + 1.2086*0 + 0.3774*1 -0.5184*1= 2.1222
- Exp(Beta X) = 8.347816
- P(Heinz) = 8.347816/9.347816 = 0.893023
- Thus, the predicted probability is 0.893023.

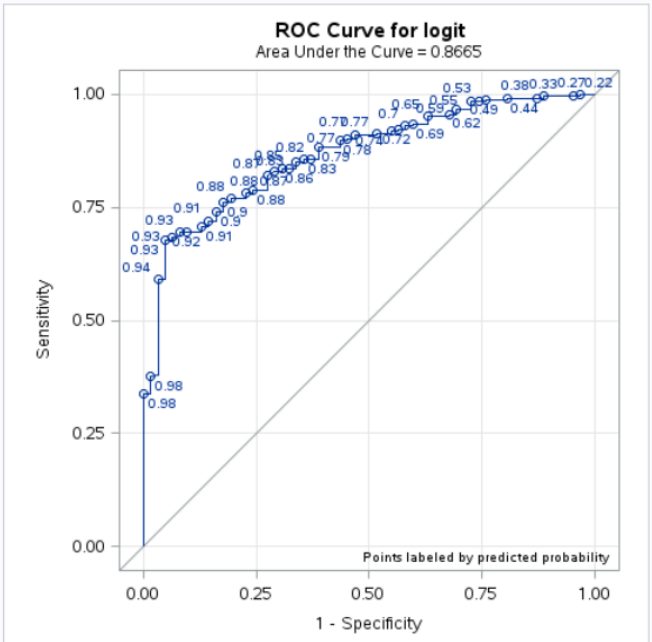
9. ROC Curves: Logistic vs Linear Model

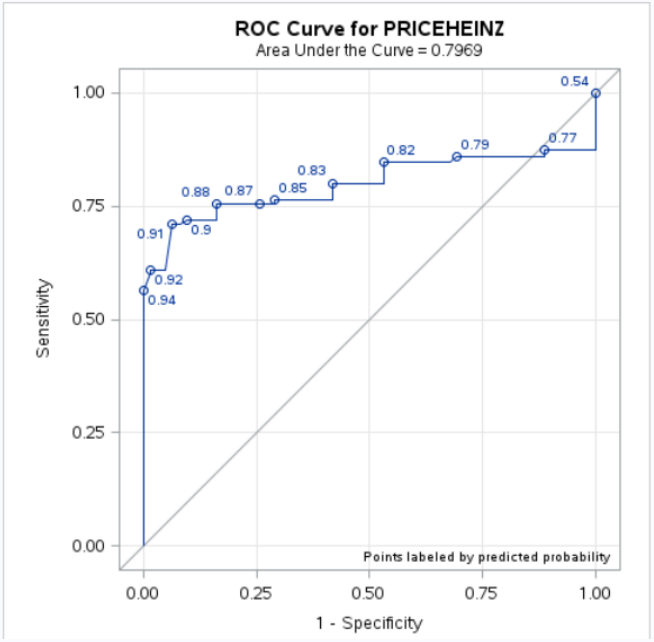
Predictions for the test data

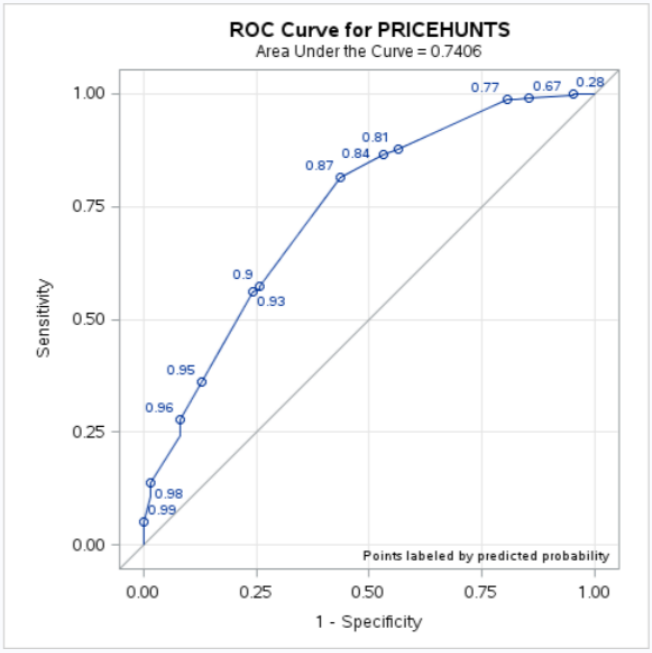
- The most important factor is Display_Heinz*Price_Heinz, which has the biggest AUC.
- This result is very different from what we get from linear model, in which the Display_Heinz*Price_Heinz term is not a significant factor.
- We can also conclude that factors including DisplayHeinz, DisplayHunts, FeatureHeinz, and FeatureHunts don't have much contribution for the logit model. They are very close to the straight line.

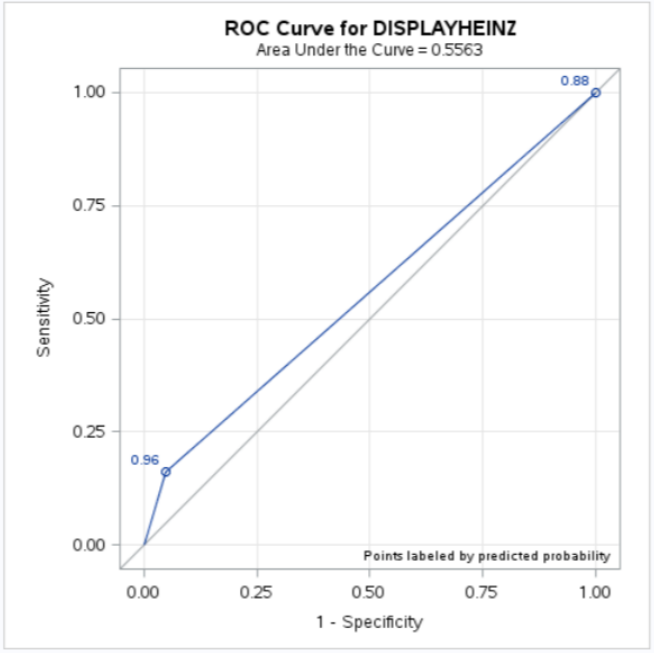
ROC Association Statistics

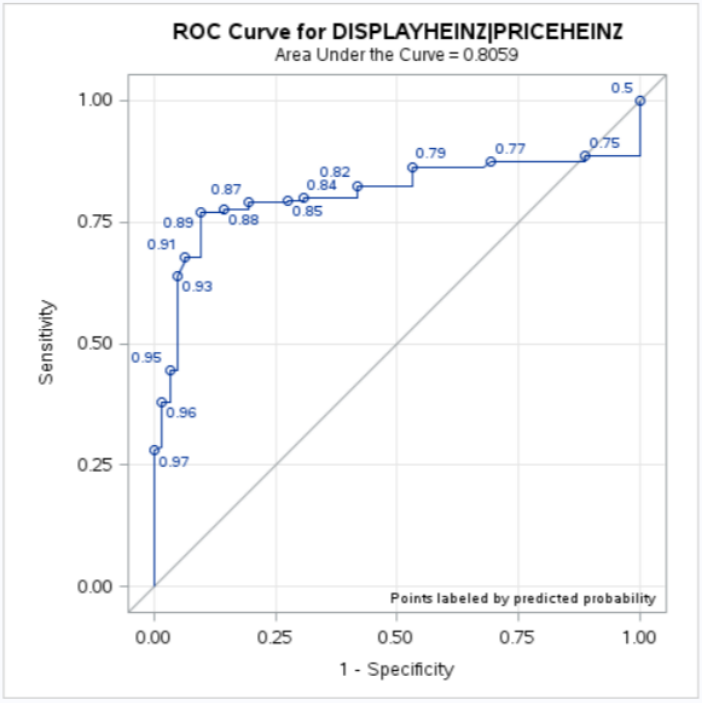
ROC Model	Mann-Whitney Area	Mann-Whitney Standard Error	Mann-Whitney 95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
logit	0.8665	0.0203	0.8268	0.9063	0.7330	0.7332	0.1448
all	0.8665	0.0203	0.8268	0.9063	0.7330	0.7332	0.1448
PRICEHEINZ	0.7969	0.0182	0.7612	0.8325	0.5937	0.5952	0.1173
PRICEHUNTS	0.7406	0.0348	0.6724	0.8087	0.4811	0.5406	0.0951
DISPLAYHEINZ	0.5563	0.0160	0.5249	0.5877	0.1126	0.5810	0.0222
DISPLAYHUNTS	0.5827	0.0256	0.5326	0.6328	0.1654	0.7845	0.0327
FEATUREHEINZ	0.5251	0.0190	0.4878	0.5623	0.0501	0.2634	0.00991
FEATUREHUNTS	0.5070	0.0123	0.4830	0.5310	0.0140	0.1856	0.00277
DISPLAYHEINZ|PRICEHEINZ	0.8059	0.0207	0.7654	0.8465	0.6119	0.6124	0.1209
DISPLAYHUNTS|PRICEHUNTS	0.7434	0.0350	0.6748	0.8120	0.4868	0.5384	0.0962

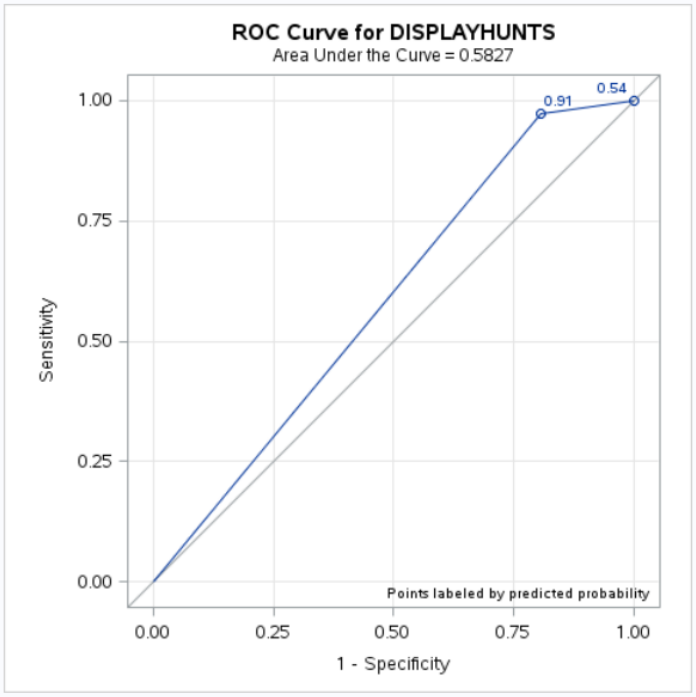

ROC Curve for logit
Area Under the Curve = 0.8665
Sensitivity
1 - Specificity
Points labeled by predicted probability

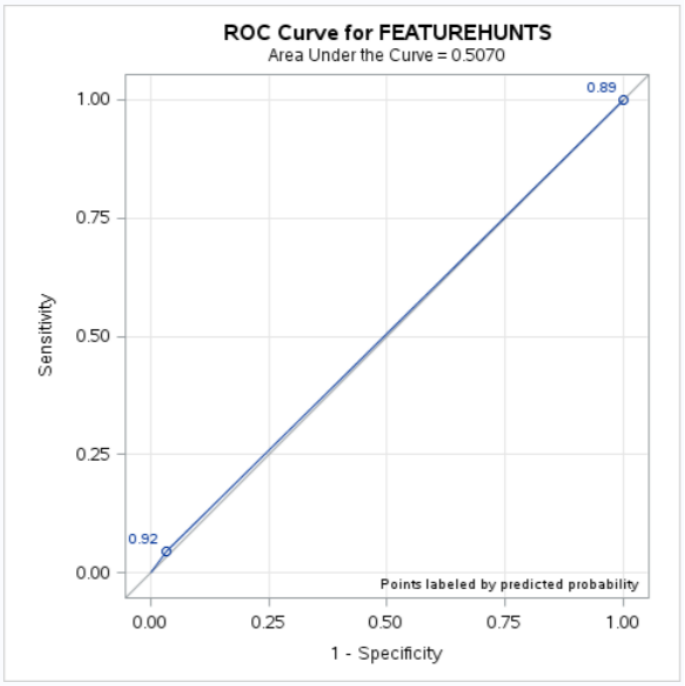

ROC Curve for PRICEHEINZ
Area Under the Curve = 0.7969
Sensitivity
1 - Specificity
Points labeled by predicted probability

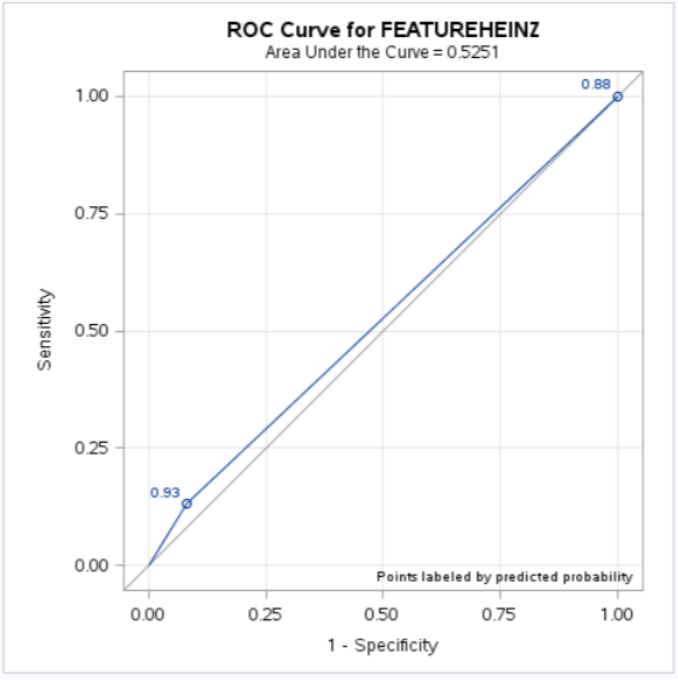

ROC Curve for PRICEHUNTS
Area Under the Curve = 0.7406
Sensitivity
1 - Specificity
Points labeled by predicted probability

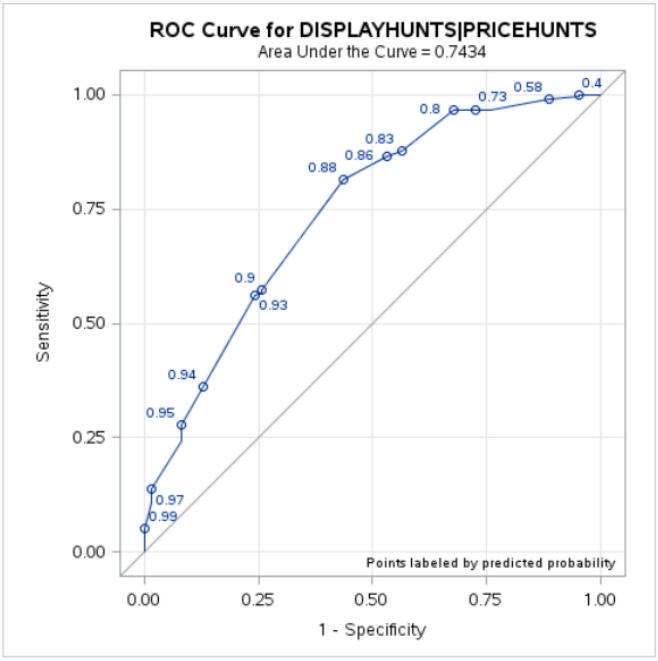

ROC Curve for DISPLAYHEINZ
Area Under the Curve = 0.5563
Sensitivity
1 - Specificity
Points labeled by predicted probability

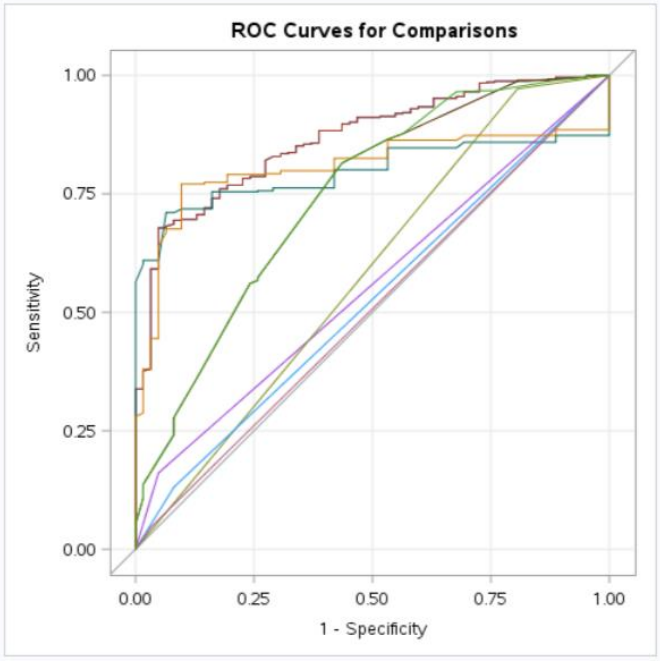

ROC Curve for DISPLAYHEINZ|PRICEHEINZ
Area Under the Curve = 0.8059
Sensitivity
1 - Specificity
Points labeled by predicted probability
0.00
0.25
0.50
0.75
1.00
0.5
0.75
0.77
0.79
0.82
0.84
0.85
0.87
0.88
0.89
0.91
0.93
0.95
0.96
0.97


ROC Curve for DISPLAYHUNTS
Area Under the Curve = 0.5827
Sensitivity
1 - Specificity
Points labeled by predicted probability
0.00
0.25
0.50
0.75
1.00
0.54
0.91


ROC Curve for FEATUREHUNTS
Area Under the Curve = 0.5070
Sensitivity
1 - Specificity
Points labeled by predicted probability
0.00
0.25
0.50
0.75
1.00
0.89
0.92

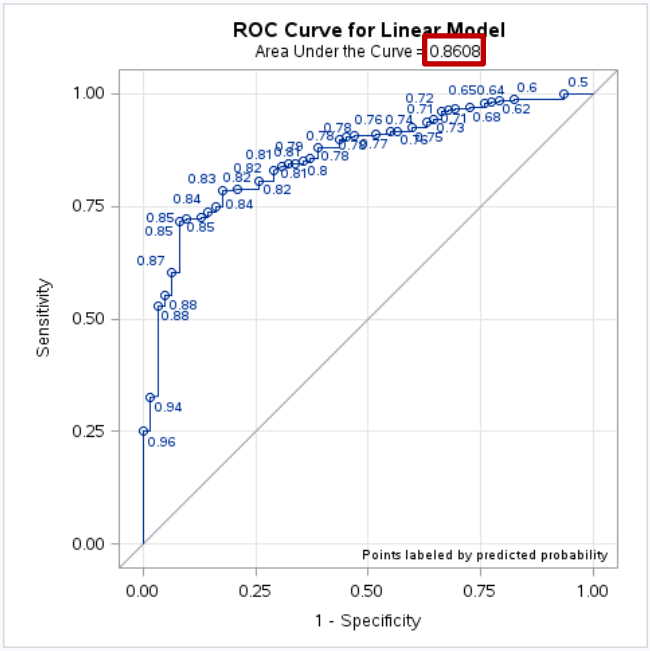

ROC Curve for FEATUREHEINZ
Area Under the Curve = 0.5251
Sensitivity
1 - Specificity
Points labeled by predicted probability
0.00
0.25
0.50
0.75
1.00
0.88
0.93

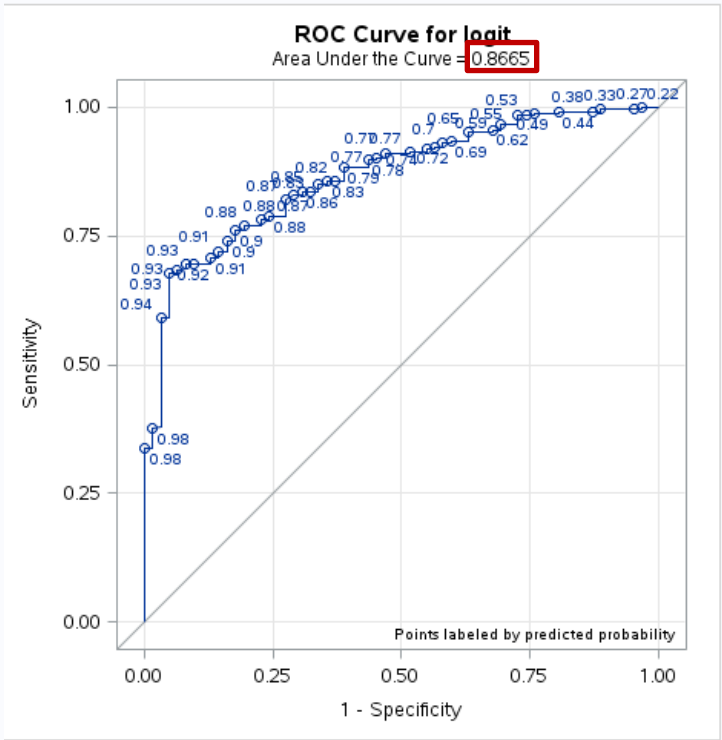




Model selection under ROC curve

- According to the AUC under Roc curve, there is only a slight difference (0.0057) between linear model (0.8608) and Logit model (0.8665).
- Based on the result Logit model (from question 5) has a better performance on the consumers' ketchup purchase pattern





ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Linear Model	0.8608	0.0217	0.8183	0.9033	0.7216	0.7218	0.1426

10. Conclusions

- Heinz is the dominant brand with 90% of the market share even though Hunts is a bit less expensive. Heinz is also more likely to be on display or featured in store. For both brands, prices tend to be a bit lower when featured or on display.
- Regardless of whether a logistic, LPM or probit model is used, price of Heinz and Hunts are shown to be the most important factors affecting whether or not Heinz is purchased. Other factors shown to be significant below the 5% level are Feature of Heinz and Feature of Hunts.
- Through AIC and ROC curve comparisons, the best model for understanding ketchup buying patterns is shown to be the logistic regression model.

Appendix

```
/* Importing Data-set */

LIBNAME HW3 'H:\My SAS Files\HW3';

PROC IMPORT OUT= HW3.Heinz_Hunt
            DATAFILE= "H:\My SAS Files\HW3\Heinz_Hunts_Data.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

DATA Storedata;
SET HW3.Heinz_Hunt;
RUN;

/* 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, …  */

/* Qualitative Variables : HOUSEHOLDID | HEINZ | HUNTS | DISPLAYHEINZ | DISPLAYHUNTS
|FEATUREHEINZ |FEATUREHUNTS
   Quantative Variables  : PRICEHEINZ | PRICEHUNTS
*/

/* We can see that there are 300 Household ID information with 2798 Observations in our
dataset consisting of consumptiopn of Heinz and Hunt Ketchup brands */
proc sql;
 select count(distinct HOUSEHOLDID) as Number_of_Household
 from
 Storedata;
 title 'Number of Household';
quit;
run;

/*We also came to know that Heinz as 89% of the total Market share when compared to
10.97% of Hunt Market Share. Most of the Market Share is occupied by the Heinz Ketchup
*/
proc freq data= Storedata;
table HEINZ HUNTS;
 title 'Summary Statistics of categorical variables';
run;

/* Average Price of Heinz per ounce : 0.0348
   Average Price of Heinz per ounce : 0.0335

 We can see that the Heinz Ketchup is little bit expensive when compared to Hunts Ketchup
while comparing the Average Price of the Ketchups*/
proc means data= Storedata mean;
var PRICEHEINZ PRICEHUNTS;
 title 'Average of Quantative Variables';
run;
```

```
/* We can see that */

proc freq data= Storedata;
table DISPLAYHEINZ DISPLAYHUNTS FEATUREHEINZ FEATUREHUNTS;
 title 'Summary Statistics of categorical variables';
run;

proc sort data = Storedata;
by DISPLAYHEINZ;
run;

/* Summary Statistics using PROC means: by gives us the information based on group */
proc means data= Storedata n mean;
var PRICEHEINZ;
 title 'Summary Statistics by Display';
 by DISPLAYHEINZ;
run;

proc sort data = Storedata;
by DISPLAYHUNTS;
run;

/* Summary Statistics using PROC means: by gives us the information based on group */
proc means data= Storedata n mean;
var PRICEHUNTS;
 title 'Summary Statistics by Display';
 by DISPLAYHUNTS;
run;

proc sort data = Storedata;
by FEATUREHEINZ;
run;

/* Summary Statistics using PROC means: by gives us the information based on group */
proc means data= Storedata n mean;
var PRICEHEINZ;
 title 'Summary Statistics by Feature';
 by FEATUREHEINZ;
run;

proc sort data = Storedata;
by FEATUREHUNTS;
run;

/* Summary Statistics using PROC means: by gives us the information based on group */
proc means data= Storedata n mean;
var PRICEHUNTS;
 title 'Summary Statistics by Feature';
 by FEATUREHUNTS;
run;

proc corr data=Storedata;
 var PRICEHEINZ PRICEHUNTS;
 run;

 /* Ques 2 : 2.Randomly select 80% of the data set as the training sample, remaining 20%
as test sample. Please set the seed=2. */

proc surveyselect data=Storedata out=store_sampled outall samprate=0.8 seed=2;
```

```
run;

data store_training store_test;
 set store_sampled;
 if selected then output store_training;
 else output store_test;
run;


/* Ques 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 * FeatureHeinz and
DisplayHunts * FeatureHunts).
Provide a bullet point for each estimated parameter to explain its effect.*/

/* Linear probability model using linear regression */
proc glm data=store_sampled ;
 class HEINZ(ref = '0') DISPLAYHEINZ(ref='0') DISPLAYHUNTS(ref='0') FEATUREHEINZ(ref='0')
FEATUREHUNTS(ref='0');
 model HEINZ = PRICEHEINZ PRICEHUNTS DISPLAYHEINZ DISPLAYHUNTS FEATUREHEINZ FEATUREHUNTS
DISPLAYHEINZ|PRICEHEINZ DISPLAYHUNTS|PRICEHUNTS/solution;
 weight selected; /*only training sample is used for estimation, since selected=0 for
test sample */
 TITLE 'Linear Regression Model';
run;

quit;

/*Ques 4: By using Proc Glmselect, find the best models according to Forward
algorithm.
(1) - you must use the AIC criterion for selection procedure, (
2) adding all interaction effects at the second order, and
(3) in a hierarchal order, when the main effects should be included first. Meanwhile,
find and record its AIC and the MSE_test of the above best selected model, based on the
above procedure.

*/

/*ASE in train vs. test data */
/* Forward selection with significant level of coefficients as criteria */
proc glmselect data=store_training testdata=store_test  plots=all;
 class DISPLAYHEINZ(split) DISPLAYHUNTS(split) FEATUREHEINZ(split) FEATUREHUNTS(split);
 model HEINZ = PRICEHEINZ|PRICEHUNTS|DISPLAYHEINZ|DISPLAYHUNTS| FEATUREHEINZ
|FEATUREHUNTS @2
  /selection=forward(select=aic) hierarchy=single showpvalues ;
 performance buildsscp=incremental;
 TITLE 'Forward Regression Model';
run;

/*Ques 5: 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 * FeatureHeinz and
DisplayHunts * FeatureHunts).
```

```
Provide a bullet point for each estimated parameter to explain its effect.  */

proc logistic data=store_sampled plots=all ;
 class HEINZ(ref = '0') DISPLAYHEINZ(ref='0') DISPLAYHUNTS(ref='0') FEATUREHEINZ(ref='0')
FEATUREHUNTS(ref='0');
 logit: model HEINZ(event = '1') = PRICEHEINZ PRICEHUNTS DISPLAYHEINZ DISPLAYHUNTS
FEATUREHEINZ FEATUREHUNTS DISPLAYHEINZ|PRICEHEINZ DISPLAYHUNTS|PRICEHUNTS;
 weight selected; /*only training sample is used for estimation, since selected=0 for
test sample */
 TITLE 'Logistic Regression Model';
run;
quit;

/*Ques 6 : 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 * FeatureHeinz and DisplayHunts * FeatureHunts). Compare your results
with Q5.  */

proc logistic data=store_training outmodel=Probitmodel;
class HEINZ(ref = '0') DISPLAYHEINZ(ref='0') DISPLAYHUNTS(ref='0') FEATUREHEINZ(ref='0')
FEATUREHUNTS(ref='0');
probit: model HEINZ(event = '1') = PRICEHEINZ PRICEHUNTS DISPLAYHEINZ DISPLAYHUNTS
FEATUREHEINZ FEATUREHUNTS DISPLAYHEINZ|PRICEHEINZ DISPLAYHUNTS|PRICEHUNTS;
weight selected; /*only training sample is used for estimation, since selected=0 for test
sample */
TITLE 'Probalistic Regression Model';
run;

/* 8 Based on the estimated logit model in Q5, and using the logit probability formula,
calculate the change in predicted probability that Heinz is purchased (Y=1) if the
average price of Heinz dropped by 20%
(i.e., changes from Average(PriceHeinz) to (1-2/10)×Average(PriceHeinz),
when Hunts charges the its average price, (i.e.,  Average(PriceHunts)), Heinz does not
use a feature or display,
while Hunts uses a feature and a display (Hint: Recall that in the logit model:
Pr(Y=1)=e^βX/(1+e^βX ) ,
 where Y is the outcome variable, X are the predictor variables, and β are the estimated
model coefficients in Q5.) */

/*9. Based on the estimated logit model in Q5, 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 Q5 (including the interaction
terms (e.g., DisplayHeinz * PriceHeinz and DisplayHunts * PriceHunts)),
PriceHeinz, PriceHunts, DisplayHeinz, DisplayHunts, FeatureHeinz, and
FeatureHunts, 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 Q5 provide a significant better
prediction result relative to simple models which contain only one of the
above explanatory variables?
Noe, 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 Q5, provide you a better predictive tool
on the consumersâ€™ ketchup purchase pattern?*/
ods rtf file='result';
proc logistic data=store_sampled plots=roc(id=prob);
      class HEINZ(ref='0') DISPLAYHEINZ(ref='0') DISPLAYHUNTS(ref='0')
```

```
            FEATUREHEINZ(ref='0') FEATUREHUNTS(ref='0');
      logit: model HEINZ(event='1')=PRICEHEINZ PRICEHUNTS DISPLAYHEINZ DISPLAYHUNTS
            FEATUREHEINZ FEATUREHUNTS DISPLAYHEINZ|PRICEHEINZ DISPLAYHUNTS|PRICEHUNTS;
      roc 'all' PRICEHEINZ PRICEHUNTS DISPLAYHEINZ DISPLAYHUNTS FEATUREHEINZ
            FEATUREHUNTS DISPLAYHEINZ|PRICEHEINZ DISPLAYHUNTS|PRICEHUNTS;
      roc 'PRICEHEINZ' PRICEHEINZ;
      roc 'PRICEHUNTS' PRICEHUNTS;
      roc 'DISPLAYHEINZ' DISPLAYHEINZ;
      roc 'DISPLAYHUNTS' DISPLAYHUNTS;
      roc 'FEATUREHEINZ' FEATUREHEINZ;
      roc 'FEATUREHUNTS' FEATUREHUNTS;
      roc 'DISPLAYHEINZ|PRICEHEINZ' DISPLAYHEINZ|PRICEHEINZ;
      roc 'DISPLAYHUNTS|PRICEHUNTS' DISPLAYHUNTS|PRICEHUNTS;
      where selected=0;/*only training sample is used for estimation, since selected=0
for test sample */
run;
ods rtf close;

proc glm data=store_sampled;
      class HEINZ(ref='0') DISPLAYHEINZ(ref='0') DISPLAYHUNTS(ref='0')
            FEATUREHEINZ(ref='0') FEATUREHUNTS(ref='0');
      model HEINZ=PRICEHEINZ PRICEHUNTS DISPLAYHEINZ DISPLAYHUNTS FEATUREHEINZ
            FEATUREHUNTS DISPLAYHEINZ|PRICEHEINZ DISPLAYHUNTS|PRICEHUNTS;
      weight selected;
    output out=Heinz_Lin_predict p=linear_predictions;
      /* predictions are made for all observations - training and test */
run;
quit;

ods graphics on;

proc logistic data=Heinz_Lin_predict plots=roc(id=prob);

      logit: model HEINZ(event='1')=PRICEHEINZ PRICEHUNTS DISPLAYHEINZ DISPLAYHUNTS
            FEATUREHEINZ FEATUREHUNTS DISPLAYHEINZ|PRICEHEINZ
DISPLAYHUNTS|PRICEHUNTS/nofit;
      roc 'Linear Model'pred=linear_predictions;
      where selected=0;
      /*only training sample is used for estimation, since selected=0 for test sample */

run;
```