Homework 4

Jeffrey Chang

Question 1

Create a variable: LogPriceRatio = log (PriceHeinz/PriceHunts).
Also create interaction terms between display and feature for both brands (e.g., DisplHeinz * FeatureHeinz for Heinz, and the same for Hunts).

- Code:

```
*Question1;
data Heinz;
set data.heinzhunts(encoding=any);
run;

data Heinz2;
 set Heinz;
 LogPriceRatio = log (PriceHeinz/PriceHunts);
 IntHienz = DisplHeinz * FeatHeinz;
 IntHunts = DisplHunts * FeatHunts;
run;
```

- Output:

VIEWTABLE: Work.Heinz2

	HEINZ	HUNTS	PRICEHEINZ	PRICEHUNTS	FeatHeinz	FeatHunts	DisplHeinz	DisplHunts	LogPriceRatio	IntHienz	IntHunts
1	1	0	0.052000001	0.034000002	0	0	0	0	0.424883154	0	0
2	1	0	0.052000001	0.044	0	0	0	0	0.167054104	0	0
3	1	0	0.046	0.048	1	0	0	0	-0.04255961	0	0
4	1	0	0.052000001	0.034000002	0	0	0	0	0.424883154	0	0
5	1	0	0.046	0.048	1	0	0	0	-0.04255961	0	0
6	1	0	0.046999998	0.029999999	0	0	0	0	0.448950211	0	0
7	1	0	0.046	0.041000001	0	1	0	1	0.115069305	0	1
8	1	0	0.046	0.030999999	0	0	0	0	0.394654224	0	0
9	1	0	0.046999998	0.030999999	0	0	0	0	0.416160387	0	0
10	1	0	0.050000001	0.028000001	0	0	0	1	0.57981848	0	0
11	1	0	0.050000001	0.041000001	0	0	0	0	0.198450934	0	0
12	1	0	0.050000001	0.034000002	0	0	0	0	0.385662442	0	0
13	1	0	0.052999999	0.037	0	0	0	0	0.359373982	0	0
14	1	0	0.035999998	0.037	0	0	1	0	-0.02739903	0	0
15	1	0	0.052000001	0.037	0	0	0	0	0.340325825	0	0
16	1	0	0.030999999	0.034000002	0	0	0	0	-0.09237341	0	0
17	1	0	0.017999999	0.034000002	0	0	0	0	-0.63598888	0	0
18	1	0	0.041999999	0.035	1	0	0	0	0.182321533	0	0
19	1	0	0.050000001	0.030999999	0	0	0	0	0.478035853	0	0
20	1	0	0.043000001	0.034000002	0	0	0	0	0.234839556	0	0
21	1	0	0.039000001	0.037	0	0	0	0	0.052643759	0	0
22	1	0	0.028000001	0.035999998	0	0	0	0	-0.25131434	0	0
23	1	0	0.037	0.034000002	0	0	0	0	0.084557329	0	0
24	1	0	0.028000001	0.037	0	0	0	0	-0.27871337	0	0
25	1	0	0.037	0.035	0	0	0	0	0.055569851	0	0

Question 2

Randomly select 80% of the data set as the training sample, remaining 20% as test sample. Use the seed= option to set random seed to a value of 10.

- Code:

```
*Question2;
proc surveyselect data=Heinz2 out=Heinz2_sampled outall samprate=0.8 seed=10;
run;

data Heinz2_training Heinz2_test;
  set Heinz2_sampled;
  if selected then output Heinz2_training;
  else output Heinz2_test;
run;
```

- Output:

VIEWTABLE: Work.Heinz2_sampled

	Selected	HEINZ	HUNTS	PRICEHEINZ	PRICEHUNTS	FeatHeinz	FeatHunts	DisplHeinz	DisplHunts	LogPriceRatio	IntHienz	IntHunts
1	1	1	0	0.052000001	0.034000002	0	0	0	0	0.4248831544	0	0
2	1	1	0	0.052000001	0.044	0	0	0	0	0.1670541039	0	0
3	1	1	0	0.046	0.048	1	0	0	0	-0.042559614	0	0
4	0	1	0	0.052000001	0.034000002	0	0	0	0	0.4248831544	0	0
5	1	1	0	0.046	0.048	1	0	0	0	-0.042559614	0	0
6	1	1	0	0.046999998	0.029999999	0	0	0	0	0.4489502108	0	0
7	1	1	0	0.046	0.041000001	0	1	0	1	0.1150693054	0	1
8	0	1	0	0.046	0.030999999	0	0	0	0	0.3946542243	0	0
9	0	1	0	0.046999998	0.030999999	0	0	0	0	0.4161603869	0	0
10	1	1	0	0.050000001	0.028000001	0	0	0	1	0.5798184795	0	0
11	1	1	0	0.050000001	0.041000001	0	0	0	0	0.1984509343	0	0
12	1	1	0	0.050000001	0.034000002	0	0	0	0	0.385662442	0	0
13	1	1	0	0.052999999	0.037	0	0	0	0	0.359373982	0	0
14	1	1	0	0.035999998	0.037	0	0	1	0	-0.02739903	0	0
15	1	1	0	0.052000001	0.037	0	0	0	0	0.3403258252	0	0
16	1	1	0	0.030999999	0.034000002	0	0	0	0	-0.092373411	0	0
17	1	1	0	0.017999999	0.034000002	0	0	0	0	-0.635988881	0	0

VIEWTABLE: Work.Heinz2_training

	Selected	HEINZ	HUNTS	PRICEHEINZ	PRICEHUNTS	FeatHeinz	FeatHunts	DisplHeinz	DisplHunts	LogPriceRatio	IntHienz	IntHunts
1	1	1	0	0.052000001	0.034000002	0	0	0	0	0.4248831544	0	0
2	1	1	0	0.052000001	0.044	0	0	0	0	0.1670541039	0	0
3	1	1	0	0.046	0.048	1	0	0	0	-0.042559614	0	0
4	1	1	0	0.046	0.048	1	0	0	0	-0.042559614	0	0
5	1	1	0	0.046999998	0.029999999	0	0	0	0	0.4489502108	0	0
6	1	1	0	0.046	0.041000001	0	1	0	1	0.1150693054	0	1
7	1	1	0	0.050000001	0.028000001	0	0	0	1	0.5798184795	0	0
8	1	1	0	0.050000001	0.041000001	0	0	0	0	0.1984509343	0	0
9	1	1	0	0.050000001	0.034000002	0	0	0	0	0.385662442	0	0
10	1	1	0	0.052999999	0.037	0	0	0	0	0.359373982	0	0
11	1	1	0	0.035999998	0.037	0	0	1	0	-0.02739903	0	0
12	1	1	0	0.052000001	0.037	0	0	0	0	0.3403258252	0	0
13	1	1	0	0.030999999	0.034000002	0	0	0	0	-0.092373411	0	0
14	1	1	0	0.017999999	0.034000002	0	0	0	0	-0.635988881	0	0
15	1	1	0	0.050000001	0.030999999	0	0	0	0	0.4780358532	0	0

VIEWTABLE: Work.Heinz2_test

	Selected	HEINZ	HUNTS	PRICEHEINZ	PRICEHUNTS	FeatHeinz	FeatHunts	DisplHeinz	DisplHunts	LogPriceRatio	IntHienz	IntHunts
1	0	1	0	0.052000001	0.034000002	0	0	0	0	0.4248831544	0	0
2	0	1	0	0.046	0.030999999	0	0	0	0	0.3946542243	0	0
3	0	1	0	0.046999998	0.030999999	0	0	0	0	0.4161603869	0	0
4	0	1	0	0.041999999	0.035	1	0	0	0	0.182321533	0	0
5	0	1	0	0.043000001	0.034000002	0	0	0	0	0.2348395555	0	0
6	0	1	0	0.034000002	0.029999999	0	0	0	0	0.1251632351	0	0
7	0	1	0	0.034000002	0.028999999	0	0	0	0	0.1590647879	0	0
8	0	1	0	0.037	0.034000002	0	0	0	0	0.0845573292	0	0
9	0	1	0	0.039000001	0.030999999	0	0	1	0	0.2295744995	0	0
10	0	1	0	0.034000002	0.035	1	0	0	0	-0.028987478	0	0
11	0	1	0	0.030999999	0.044	0	0	0	0	-0.350202462	0	0
12	0	1	0	0.027000001	0.035	0	1	0	0	-0.259511158	0	0
13	0	1	0	0.037	0.028000001	0	0	0	1	0.2787133668	0	0
14	0	1	0	0.035999998	0.037	0	0	0	0	-0.02739903	0	0
15	0	1	0	0.034000002	0.041000001	0	0	0	0	-0.187211508	0	0

Question 3

Estimate a logit probability model for the probability that Heinz is purchased – using LogPriceRatio, DisplHeinz, FeatureHeinz, DisplHunts, FeatureHunts, and interaction terms (from part 1) as the explanatory variables.

- Code:

```
*Question3;
proc logistic data=Heinz2_sampled;
  logit: model Heinz (event='1') = LogPriceRatio DisplHeinz FeatHeinz DisplHunts FeatHunts IntHienz IntHunts / clodds=wald orpvalue;
  *'clodds=wald orpvalue' generates p-values for Odds Ratios';
  weight selected; /*only training sample is used for estimation, since selected = 0 for test sample */
  title 'Logit';
run;
```

- Output:

Model Information		
Data Set	WORK.HEINZ2_SAMPLED	
Response Variable	HEINZ	
Number of Response Levels	2	
Weight Variable	Selected	Selection Indicator
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	2798
Number of Observations Used	2239
Sum of Weights Read	2239
Sum of Weights Used	2239

Response Profile			
Ordered Value	**HEINZ**	**Total Frequency**	**Total Weight**
1	0	252	252.0000
2	1	1987	1987.0000

Probability modeled is HEINZ='1'.

Note: 559 observations having nonpositive frequencies or weights were excluded since they do not contribute to the analysis.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

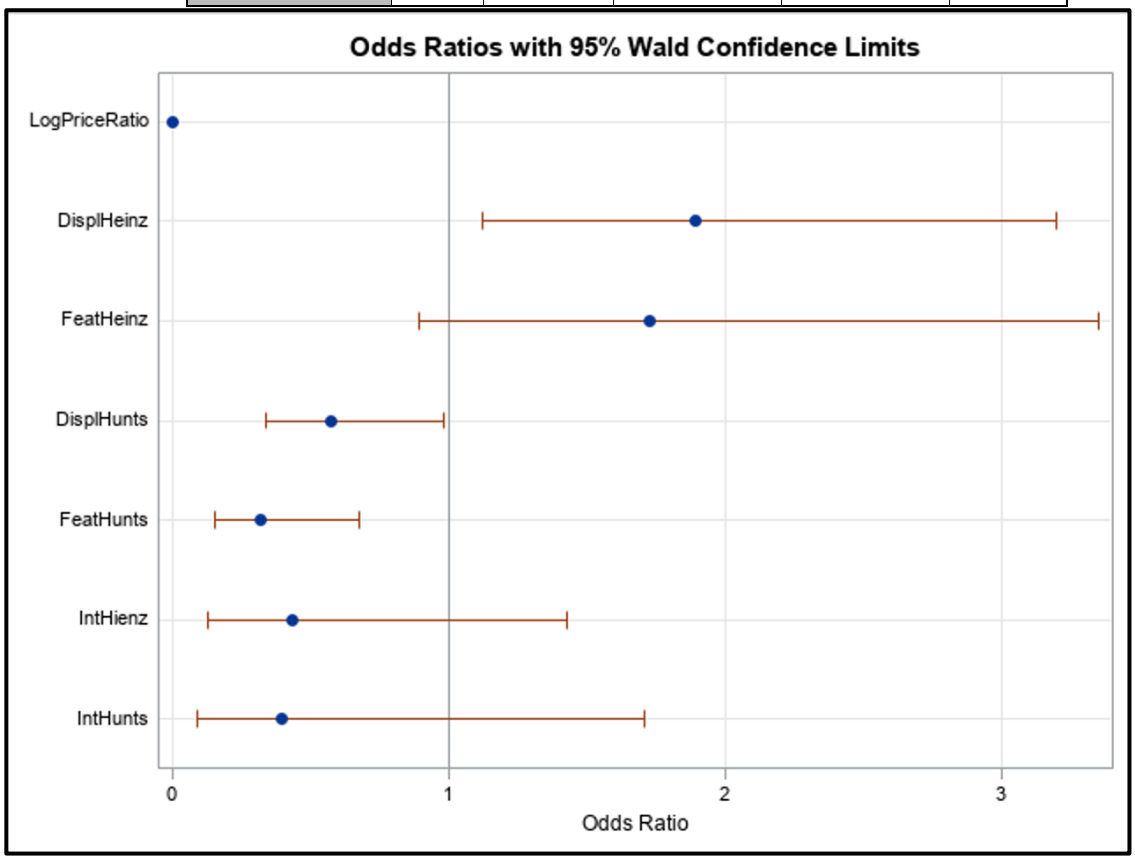
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1577.424	1112.699
SC	1583.138	1158.409
-2 Log L	1575.424	1096.699

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	478.7251	7	<.0001
Score	399.8103	7	<.0001
Wald	261.9204	7	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.2142	0.1560	424.4032	<.0001
LogPriceRatio	1	-6.0112	0.4234	201.5199	<.0001
DisplHeinz	1	0.6390	0.2671	5.7238	0.0167
FeatHeinz	1	0.5460	0.3384	2.6032	0.1066
DisplHunts	1	-0.5529	0.2730	4.1019	0.0428
FeatHunts	1	-1.1403	0.3819	8.9140	0.0028
IntHienz	1	-0.8363	0.6093	1.8843	0.1699
IntHunts	1	-0.9322	0.7490	1.5490	0.2133

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	87.1	Somers' D	0.741
Percent Discordant	12.9	Gamma	0.741
Percent Tied	0.0	Tau-a	0.148
Pairs	500724	c	0.871

Odds Ratio Estimates and Wald Confidence Intervals					
Effect	**Unit**	**Estimate**	**95% Confidence Limits**		**p-Value**
LogPriceRatio	1.0000	0.002	0.001	0.006	<.0001
DisplHeinz	1.0000	1.895	1.122	3.198	0.0167
FeatHeinz	1.0000	1.726	0.889	3.351	0.1066
DisplHunts	1.0000	0.575	0.337	0.982	0.0428
FeatHunts	1.0000	0.320	0.151	0.676	0.0028
IntHienz	1.0000	0.433	0.131	1.430	0.1699
IntHunts	1.0000	0.394	0.091	1.709	0.2133



Question 4

Interpret all parameter estimates. What promotional methods (feature/display) are effective for Hunts? For Heinz? How would you interpret the results for the interaction effects? Interpret all your results in in terms of odds ratio. Refer to lecture notes for examples.

Interpretation:

1. A one unit increase in "LogPriceRatio" decreases the odds of purchasing Heinz by 0.002 times or by (0.002-1)*100 = -99.8%". This is statistically significant at 5%.

2. A one unit increase in " DisplHeinz " or "having Heinz on display increases the odds of purchasing Coke by 1.895 times or by (1.895-1)*100 = 89.5%". This is statistically significant at 5%.
3. A one unit increase in “FeatHeinz” or "having Heinz on feature increases the odds of purchasing Heinz by 1.726 times or by (1.726-1)*100 = 72.6%“. This is not statistically significant at 5%.
4. A one unit increase in “DisplHunts” or "having Hunts on display decreases the odds of purchasing Heinz by 0.575 times or by (0.575-1)*100 = -42.5%“. This is statistically significant at 5%.
5. A one unit increase in “FeatHunts” or "having Hunts on feature decreases the odds of purchasing Heinz by 0.32 times or by (0.32-1)*100 = -68%“. This is statistically significant at 5%.
6. A one unit increase in “IntHeinz” or "having Heinz on display and feature decreases the odds of purchasing Heinz by (0.433-1) *100 = 56.7%“. This is not statistically significant at 5%.
7. A one unit increase in “IntHunts” or "having Hunts on feature decreases the odds of purchasing Heinz by (0.394-1) *100 = -60.6%“. This is not statistically significant at 5%.
8. According to the (2) and (3), both featuring and displaying are effective for Heinz and Hunts.
9. Having Heinz on display and feature together have counterproductive effect. That is, we expect a 162.1% (89.5%+72.6%) increases of the odds of purchasing Heinz. However, it only increases by (1.895+1.726+0.433-3) *100 = 105.4%.
10. Having Hunts on display and feature together have synergistic effect. That is, we expect a 110.5% (42.5%+68%) decreases of the odds of purchasing Heinz. However, it decreases by (0.575+0.32+0.394-3) *100 = -171.1%.

Question 5

Based on the estimated model, and using the logit probability formula, calculate the change in predicted probability that Heinz is purchased if LogPriceRatio changes from 0.3 to 0.4 and Heinz does not use a feature or display, while Hunts uses a feature and a display. Recall that in the logit model, where Y is the outcome variable, X are the predictor variables, and are the estimated model coefficients.

$$\Pr(Y=\text{Heinz}) = \frac{1}{1+e^{-(3.2142-6.0112*0.3-0.5529-1.1403-0.9322)}} \approx 22.9\%$$

$$\Pr(Y=\text{Heinz}) = \frac{1}{1+e^{-(3.2142-6.0112*0.4-0.5529-1.1403-0.9322)}} \approx 14\%$$

The change in predicted probability is: 22.9% - 14% = 8.9%

Question 6

The estimated model is to be used for targeting customers for Hunts coupons to build loyalty for the brand. Coupons are to be sent to customers who are likely to buy Hunts, and not to customers who are likely to buy Heinz. In other words, the coupons should be sent to customers whose predicted probability of buying Heinz is below a certain threshold level that needs to be determined based on the costs of misclassifications (incorrectly sending / not sending a coupon) The following information about the costs of incorrect classification is available: The cost of incorrectly sending a coupon to a customer who would have bought Heinz is $1 per customer, and the cost of incorrectly failing to send a coupon to a customer who would have bought Hunts is $0.25 per customer.

Based on these costs, what is the optimal threshold probability level that should be used with the estimated model to decide which consumers should receive coupons.

- Code:

```
*Question6 0.0523178601;
proc logistic data=Heinz2_training outmodel=Logitmodel;
  logit: model Heinz (event='1') = LogPriceRatio DisplHeinz FeatHeinz DisplHunts FeatHunts IntHienz IntHunts;
  weight selected;
  title 'Step 1';
run;

proc logistic inmodel=Logitmodel;
  score data=Heinz2_test outroc=Heinz2_logit_roc;
  title 'Step 2';
run;

data roc;
set Heinz2_logit_roc;
TotalCost = _FALPOS_ * 0.25 + _FALNEG_ * 1;
run;
```

- Output:

VIEWTABLE: Work.Roc

	PROB	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_	TotalCost
239	0.6498214605	484	20	35	20	0.9603174603	0.6363636364	28.75
240	0.6431269703	484	19	36	20	0.9603174603	0.6545454545	29
241	0.6430702444	485	19	36	19	0.9623015873	0.6545454545	28
242	0.6260784832	487	19	36	17	0.9662698413	0.6545454545	26
243	0.6150270565	488	19	36	16	0.9682539683	0.6545454545	25
244	0.6003181605	488	18	37	16	0.9682539683	0.6727272727	25.25
245	0.5983065398	489	18	37	15	0.9702380952	0.6727272727	24.25
246	0.5843339853	494	18	37	10	0.9801587302	0.6727272727	19.25
247	0.5787721434	494	17	38	10	0.9801587302	0.6909090909	19.5
248	0.5682618297	495	17	38	9	0.9821428571	0.6909090909	18.5
249	0.5617629685	495	16	39	9	0.9821428571	0.7090909091	18.75
250	0.5433872177	496	16	39	8	0.9841269841	0.7090909091	17.75
251	0.5206109594	496	15	40	8	0.9841269841	0.7272727273	18
252	0.4849286673	497	15	40	7	0.9861111111	0.7272727273	17
253	0.4747537048	497	14	41	7	0.9861111111	0.7454545455	17.25
254	0.4654698702	497	13	42	7	0.9861111111	0.7636363636	17.5
255	0.4635400704	497	12	43	7	0.9861111111	0.7818181818	17.75
256	0.4326018181	498	12	43	6	0.9880952381	0.7818181818	16.75
257	0.4199664373	499	12	43	5	0.9900793651	0.7818181818	15.75
258	0.3995013471	499	9	46	5	0.9900793651	0.8363636364	16.5
259	0.3281155956	500	9	46	4	0.9920634921	0.8363636364	15.5
260	0.3048869182	501	9	46	3	0.994047619	0.8363636364	14.5
261	0.2948316685	501	8	47	3	0.994047619	0.8545454545	14.75
262	0.2856930113	501	7	48	3	0.994047619	0.8727272727	15
263	0.2852050241	502	7	48	2	0.996031746	0.8727272727	14
264	0.2767914591	502	5	50	2	0.996031746	0.9090909091	14.5
265	0.1891391038	502	4	51	2	0.996031746	0.9272727273	14.75
266	0.1331606491	502	3	52	2	0.996031746	0.9454545455	15
267	0.0993811011	502	2	53	2	0.996031746	0.9636363636	15.25
268	0.0916343341	502	1	54	2	0.996031746	0.9818181818	15.5
269	0.0523178601	504	1	54	0	1	0.9818181818	13.5
270	3.3658996E-6	504	0	55	0	1	1	13.75

Interpretation:

The optimal threshold probability level that should be used is 0.0523178601, since the total cost is 13.5(lowest).