Homework 5

Group 7 (Lucky7)

Denzel Ignatius Arulmani Selvam | Jeffrey Chang
Ramakrishnan Girishankar | Chun-Li Hou
Tingyu Li | Kuan-Lun Yu

Question 1

Use PROC SURVEYSELECT to sample the original data into training and testing data sets. Use 75% for training and 25% for testing.

Code:

```
data Crackers;
  set ClsData.Crackers;
  run;

proc surveyselect data = Crackers out = Crackers_sampled outall samprate = 0.75 seed = 23;
  run;
```

Result:

VIEWTABLE: Work.Crackers_sampled

	Selected	OBS	PRIVATE	SUNSHINE	KEEBLER	NABISCO	PRICEPRIVATE	PRICESUNSHINE	PRICEKEEBLER	PRICENABISCO	FeatPrivate	FeatSun
1	1	1	0	0	0	1	0.709999979	0.99000001	1.089999914	0.99000001	0	0
2	1	2	0	1	0	0	0.779999912	0.49000001	1.089999914	1.089999914	0	0
3	1	3	0	0	0	1	0.779999912	1.029999971	1.089999914	0.889999986	0	0
4	1	4	0	0	0	1	0.639999986	1.089999914	1.089999914	1.190000057	0	0
5	1	5	0	0	0	1	0.839999914	0.889999986	1.089999914	1.190000057	0	0
6	1	6	0	1	0	0	0.779999912	1.089999914	1.089999914	1.289999962	0	0
7	1	7	0	0	0	1	0.779999912	1.089999914	1.190000057	1.289999962	0	0
8	1	8	0	0	0	1	0.779999912	1.089999914	1.209999919	1.089999914	0	0
9	1	9	0	0	0	1	0.779999912	0.789999962	1.209999919	1.089999914	0	0
10	1	10	0	0	0	1	0.959999979	1.089999914	1.129999995	1.089999914	0	0
11	1	11	0	0	0	1	0.859999955	1.089999914	1.209999919	0.99000001	0	0
12	1	12	0	0	0	1	0.859999955	0.889999986	1.209999919	0.99000001	0	0
13	0	13	0	0	0	1	0.959999979	1.089999914	1.089999914	1.289999962	0	0
14	0	14	0	0	0	1	0.789999962	1.089999914	1.089999914	1.289999962	0	0
15	1	15	0	0	0	1	0.959999979	1.289999962	1.039999962	1.289999962	0	0
16	1	16	0	0	0	1	0.689999998	0.789999962	0.99000001	0.689999998	0	0
17	0	17	0	1	0	0	0.649999976	0.689999998	1.049999952	0.889999986	0	1
18	1	18	0	1	0	0	0.689999998	0.789999962	1.25	1.059999824	0	1
19	1	19	0	1	0	0	0.689999998	0.789999962	1.25	1.059999824	0	0
20	0	20	0	1	0	0	0.689999998	0.789999962	1.25	0.99000001	0	0
21	0	21	0	1	0	0	0.569999933	0.970000029	1.149999976	1.089999914	1	0
22	1	22	0	1	0	0	0.649999976	0.789999962	1.349999905	1.089999914	0	0
23	1	23	0	0	1	0	0.599999964	0.970000029	0.99000001	1.25	0	0
24	1	24	0	1	0	0	0.730000019	1.089999914	1.349999905	1.289999962	0	0
25	0	25	0	1	0	0	0.759999931	1.089999914	1.349999905	1.149999976	0	0
26	1	26	0	0	0	1	0.589999914	1.049999952	1.25	0.99000001	0	0

Question 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.

The model we use to predict the choice probabilities is a multinominal logit model with alternative-specific characteristics. The general utility model to estimate this logit model is:

*V= price + displ*private + feat*private + Sunshine + displ*Sunshine + feat*Sunshine +Nabisco + displ*Nabisco + feat*Nabisco + Keebler + displ*Keebler + feat*Keebler*

Question 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.

Yes, the data formatted is needed to estimate the choice probabilities using the chosen multinominal model.

Code:

```
/*Question 3*/
data crackers_choice (keep=selected obs cracker choice price display feature);
array chosen[4] Sunshine Keebler Nabisco Private;
array prices[4] PriceSunshine PriceKeebler PriceNabisco PricePrivate;
array displays[4] DisplSunshine DisplKeebler DisplNabisco DisplPrivate;
array features[4] FeatSunshine FeatKeebler FeatNabisco FeatPrivate;
array allbrands[4] $ _temporary_ ('Sunshine' 'Keebler' 'Nabisco' 'Private');
set Crackers_sampled;
obs = _n_;
do i = 1 to 4;
Cracker = allbrands[i];
Price = prices[i];
Display = displays[i];
Feature = features[i];
Choice = chosen[i];
output;
end;
run;
```

Result:

VIEWTABLE: Work.Crackers_choice

	Selected	OBS	Cracker	Price	Display	Feature	Choice
1	1	1	Sunshine	0.99000001	0	0	0
2	1	1	Keebler	1.089999914	0	0	0
3	1	1	Nabisco	0.99000001	0	0	1
4	1	1	Private	0.709999979	0	0	0
5	1	2	Sunshine	0.49000001	1	0	1
6	1	2	Keebler	1.089999914	0	0	0
7	1	2	Nabisco	1.089999914	0	0	0
8	1	2	Private	0.779999912	0	0	0
9	1	3	Sunshine	1.029999971	0	0	0
10	1	3	Keebler	1.089999914	0	0	0
11	1	3	Nabisco	0.889999986	0	0	1
12	1	3	Private	0.779999912	0	0	0
13	1	4	Sunshine	1.089999914	0	0	0
14	1	4	Keebler	1.089999914	0	0	0
15	1	4	Nabisco	1.190000057	0	0	1
16	1	4	Private	0.639999986	0	0	0
17	1	5	Sunshine	0.889999986	0	0	0
18	1	5	Keebler	1.089999914	0	0	0
19	1	5	Nabisco	1.190000057	0	0	1
20	1	5	Private	0.839999914	0	0	0
21	1	6	Sunshine	1.089999914	0	0	1
22	1	6	Keebler	1.089999914	0	0	0
23	1	6	Nabisco	1.289999962	1	0	0
24	1	6	Private	0.779999912	0	0	0
25	1	7	Sunshine	1.089999914	0	0	0
26	1	7	Keebler	1.190000057	0	0	0
27	1	7	Nabisco	[illegible]	1	0	1

Question 4

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

Code:

```
/*Question 4*/
data Crackers_training Crackers_test;
set crackers_choice;
if selected then output crackers_training;
else output crackers_test;
run;

proc logistic data = Crackers_training;
strata obs;
class cracker (ref = 'Private') / param = ref;
model choice (event = '1') = cracker price display feature cracker*display cracker*feature / clodds = wald orpvalue;
run;
```

Result:

The SAS System

The LOGISTIC Procedure

Conditional Analysis

Model Information	
Data Set	WORK.CRACKERS_TRAINING
Response Variable	Choice
Number of Response Levels	2
Number of Strata	2469
Model	binary logit
Optimization Technique	Newton-Raphson ridge

Number of Observations Read	9876
Number of Observations Used	9876
Number of Observations Informative	9876

Response Profile		
Ordered Value	Choice	Total Frequency
1	0	7407
2	1	2469

Probability modeled is Choice=1.

Class Level Information				
Class	Value	Design Variables		
Cracker	Keebler	1	0	0
	Nabisco	0	1	0
	Private	0	0	0
	Sunshine	0	0	1

Strata Summary				
Response Pattern	Choice		Number of Strata	Frequency
	0	1		
1	3	1	2469	9876

Newton-Raphson Ridge Optimization

Without Parameter Scaling

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

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
AIC	6845.522	5089.535
SC	6845.522	5175.909
-2 Log L	6845.522	5065.535

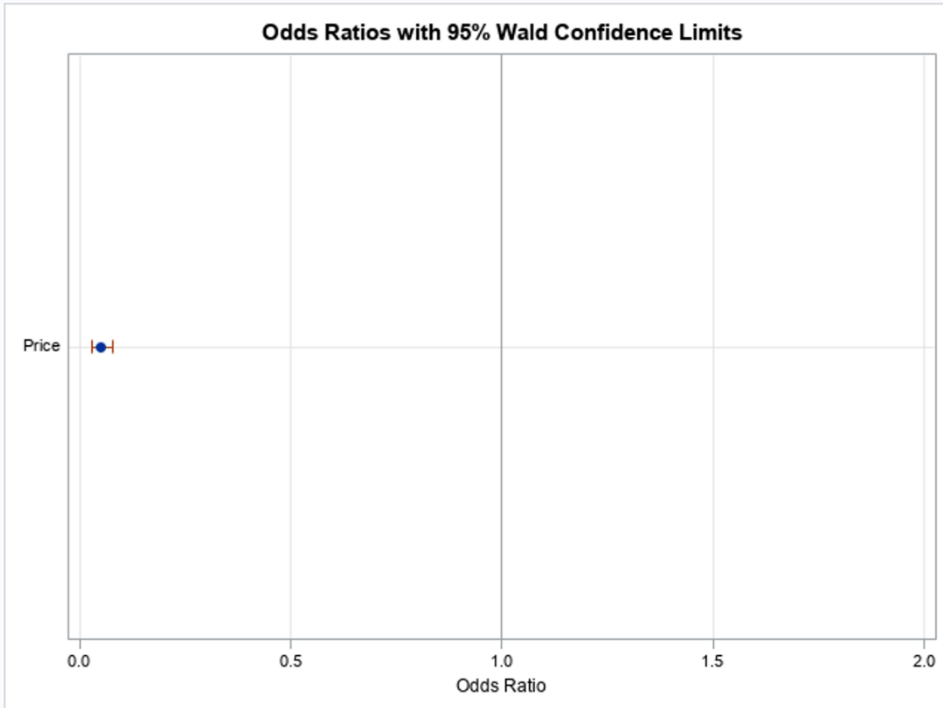
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1779.9869	12	<.0001
Score	1696.6539	12	<.0001
Wald	1214.3634	12	<.0001

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Cracker	3	874.6607	<.0001
Price	1	156.0034	<.0001
Display	1	2.8245	0.0928
Feature	1	0.2076	0.6487
Display*Cracker	3	10.1306	0.0175
Feature*Cracker	3	3.1261	0.3726

Note: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Analysis of Conditional Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Cracker	Keebler	1	-0.2664	0.1409	3.5763	0.0586
Cracker	Nabisco	1	1.6731	0.1178	201.8435	<.0001
Cracker	Sunshine	1	-0.8451	0.1193	50.2187	<.0001
Price		1	-3.0324	0.2428	156.0034	<.0001
Display		1	-0.2838	0.1689	2.8245	0.0928
Feature		1	0.1090	0.2392	0.2076	0.6487
Display*Cracker	Keebler	1	0.5730	0.2870	3.9851	0.0459
Display*Cracker	Nabisco	1	0.3503	0.1885	3.4553	0.0631
Display*Cracker	Sunshine	1	0.7705	0.2524	9.3181	0.0023
Feature*Cracker	Keebler	1	0.4633	0.3796	1.4898	0.2223
Feature*Cracker	Nabisco	1	0.4943	0.2895	2.9149	0.0878
Feature*Cracker	Sunshine	1	0.4358	0.3702	1.3861	0.2391

Odds Ratio Estimates and Wald Confidence Intervals					
Effect	Unit	Estimate	95% Confidence Limits		p-Value
Price	1.0000	0.048	0.030	0.078	<.0001



Interpretations:

1. If the cracker is Keebler, then its utility will be 0.266 less than the utility of the private cracker. This estimate is not statistically significant at the 5% level.

2. If the cracker is Nabisco, then its utility will be 1.6731 more than the utility of the private cracker. This estimate is statistically significant at the 5% level.

3. If the cracker is Sunshine, then its utility will be 0.8451 less than the utility of the private cracker. This estimate is statistically significant at the 5% level.

4. If the price of the cracker increases by 1, then its utility will decrease by 3.0324. This estimate is statistically significant at the 5% level.

5. If the private cracker is on display, then its utility will decrease by 0.2838. This estimate is not statistically significant at the 5% level.

6. If the private cracker is featured, then its utility will increase by 0.1090. This estimate is not statistically significant at the 5% level.

7. If the Keebler cracker is on display, then its utility will increase by 0.573. This estimate is statistically significant at the 5% level.

8. If the Nabisco cracker is on display, then its utility will increase by 0.3503. This estimate is not statistically significant at the 5% level.

9. If the Sunshine cracker is on display, then its utility will increase by 0.7705. This estimate is statistically significant at the 5% level.

10. If the Keebler cracker is featured, then its utility will increase by 0.4633. This estimate is not statistically significant at the 5% level.

11. If the Nabisco cracker is featured, then its utility will increase by 0.4943. This estimate is not statistically significant at the 5% level.

12. If the Sunshine cracker is featured, then its utility will increase by 0.4358. This estimate is not statistically significant at the 5% level.

Question 5

Reproduce your results using multinomial discrete choice command PROC MDC.

Code:

```
/*Question 5*/
proc mdc data = Crackers_training;
id obs;
class cracker;
model choice = cracker price display feature cracker*display cracker*feature / type = clogit nchoice = 4;
restrict crackerprivate = 0;
restrict crackerprivatedisplay = 0;
restrict crackerprivatefeature = 0;
run;
```

Result:

The SAS System

The MDC Procedure

Conditional Logit Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|	Parameter Label
CRACKERSunshine	1	-0.8451	0.1193	-7.09	<.0001	
CRACKERKeebler	1	-0.2664	0.1409	-1.89	0.0586	
CRACKERNabisco	1	1.6731	0.1178	14.21	<.0001	
CRACKERPrivate	0	0	0			
Price	1	-3.0324	0.2428	-12.49	<.0001	
Display	1	-0.2838	0.1689	-1.68	0.0928	
Feature	1	0.1090	0.2392	0.46	0.6487	
CRACKERSunshineDISPLAY	1	0.7705	0.2524	3.05	0.0023	
CRACKERKeeblerDISPLAY	1	0.5730	0.2870	2.00	0.0459	
CRACKERNabiscoDISPLAY	1	0.3503	0.1885	1.86	0.0631	
CRACKERPrivateDISPLAY	0	0	0			
CRACKERSunshineFEATURE	1	0.4358	0.3702	1.18	0.2391	
CRACKERKeeblerFEATURE	1	0.4633	0.3796	1.22	0.2223	
CRACKERNabiscoFEATURE	1	0.4943	0.2895	1.71	0.0878	
CRACKERPrivateFEATURE	0	0	0			
Restrict1	1	-2.47E-12	5.3131	-0.00	1.0000*	Equal BC CRACKERPrivateFEATURE
Restrict2	1	-5.22E-12	7.6969	-0.00	1.0000*	Equal BC CRACKERPrivateDISPLAY
Restrict3	1	-5.78E-11	12.5405	-0.00	1.0000*	Linear EC [1]

* Probability computed using beta distribution.

Comparing the above estimates with the estimates for question 4, we can see that the model in question 4 has been recreated using PROC MDC.

Question 6

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

Code:

```
/*Question 6*/
data cracker_6;
set crackers_choice;
if selected = 0 then choice =.;
run;

proc mdc data = cracker_6;
id obs;
class cracker;
model choice = cracker price display feature cracker*display cracker*feature / type = mprobit nchoice = 4;
restrict crackerprivate = 0;
restrict crackerprivatedisplay = 0;
restrict crackerprivatefeature = 0;
output out = cracker_prediction pred = p;
run;
```

Result:

	p	Selected	Subject	Cracker	Price	Display	Feature	Choice	CRACKERSunshine	CRACKERKeebler	CRACKERNabisco	CRACKER
1	0.0449315549	1	1	Sunshine	0.99000001	0	0	0	1	0	0	0
2	0.0697477964	1	1	Keebler	1.089999914	0	0	0	0	1	0	0
3	0.6124911072	1	1	Nabisco	0.99000001	0	0	1	0	0	1	0
4	0.2781531457	1	1	Private	0.709999979	0	0	0	0	0	0	1
5	0.337644395	1	2	Sunshine	0.49000001	1	0	1	1	0	0	0
6	0.0260635535	1	2	Keebler	1.089999914	0	0	0	0	1	0	0
7	0.4131644089	1	2	Nabisco	1.089999914	0	0	0	0	0	1	0
8	0.2278721097	1	2	Private	0.779999912	0	0	0	0	0	0	1
9	0.0307093762	1	3	Sunshine	1.029999971	0	0	0	1	0	0	0
10	0.0611558964	1	3	Keebler	1.089999914	0	0	0	0	1	0	0
11	0.7075380448	1	3	Nabisco	0.889999986	0	0	1	0	0	1	0
12	0.2023540821	1	3	Private	0.779999912	0	0	0	0	0	0	1
13	0.0415814859	1	4	Sunshine	1.089999914	0	0	0	1	0	0	0
14	0.1010776309	1	4	Keebler	1.089999914	0	0	0	0	1	0	0
15	0.4467995507	1	4	Nabisco	1.190000057	0	0	1	0	0	1	0
16	0.4100664797	1	4	Private	0.639999986	0	0	0	0	0	0	1
17	0.1035765481	1	5	Sunshine	0.889999986	0	0	0	1	0	0	0
18	0.1016564751	1	5	Keebler	1.089999914	0	0	0	0	1	0	0
19	0.5057065472	1	5	Nabisco	1.190000057	0	0	1	0	0	1	0
20	0.2800852811	1	5	Private	0.839999914	0	0	0	0	0	0	1
21	0.0548579547	1	6	Sunshine	1.089999914	0	0	1	1	0	0	0
22	0.1362909672	1	6	Keebler	1.089999914	0	0	0	0	1	0	0
23	0.4535968797	1	6	Nabisco	1.289999962	1	0	0	0	0	1	0
24	0.3689616573	1	6	Private	0.779999912	0	0	0	0	0	0	1
25	0.0632917038	1	7	Sunshine	1.089999914	0	0	0	1	0	0	0
26	0.094433883	1	7	Keebler	1.190000057	0	0	0	0	1	0	0
27	0.4642983551	1	7	Nabisco	1.289999962	1	0	1	0	0	1	0
28	0.3737044606	1	7	Private	0.779999912	0	0	0	0	0	0	1
29	0.0439772449	1	8	Sunshine	1.089999914	0	0	0	1	0	0	0

The probability for each choice has been prediction for all observations in the training and test datasets.