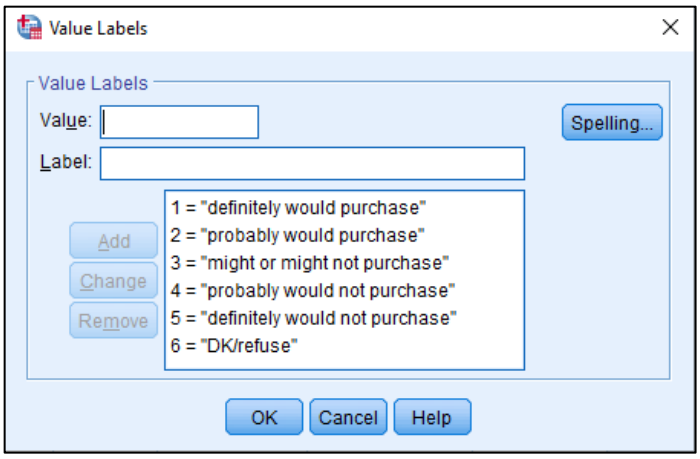
Kimberly-Clark Case

Discriminant Analysis and Logistic Regression

Andy Huang

- Recode the purchase likelihood (Likely_Purchase, Q3) into two groups by combining codes 2, 3, 4, and 5 into a single category. Run a two-group discriminant analysis with recoded Likely_Purchase as the dependent variable and responses to the message rating (Info_New_Different (Q6), Info_Approprate (Q7), Info_Believable (Q8), and Info_Understanding (Q9)) as the independent variables. Interpret the results.



The dependent variable, purchase likelihood, is measured on a Likert scale of 1 to 6, with 1 being "definitely would purchase" and 5 being "definitely would not purchase." In addition, 6 stands for Don't Know or Refuse to Answer.

Likely to Purhcase

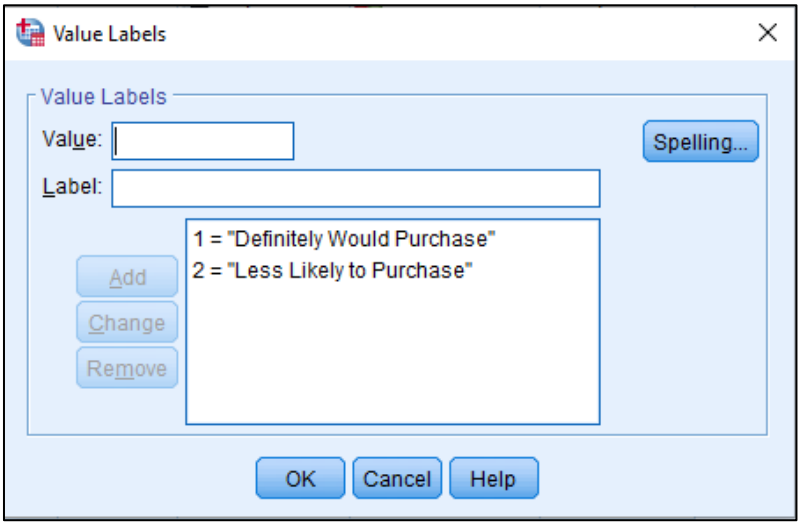
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	definitely would purchase	165	55.0	55.0	55.0
	probably would purchase	72	24.0	24.0	79.0
	might or might not purchase	33	11.0	11.0	90.0
	probably would not purchase	19	6.3	6.3	96.3
	definitely would not purchase	11	3.7	3.7	100.0
	Total	300	100.0	100.0	

By generating a Frequency Table for variable Likely_Purchase, we could see that there is no missing value within the purchase likelihood. In other words, no respondents indicated that they

do not know or refuse to answer their purchase likelihood. All our recorded responses are spread among 1 to 5 on the scale.

Among all 300 responses, 165 respondents indicated that they definitely would purchase this brand of diaper while only 11 respondents indicated that they definitely would not purchase this brand of diaper. In other words, 55% of the total respondents indicated that they definitely would purchase this brand of diaper while only 3.7 % of the total respondents indicated that they definitely would not purchase this brand of diaper. Moreover, 79% of the total respondents indicated that they probably or definitely would purchase this brand of diaper. On the other hand, only 10% (calculated by 6.3%+3.7%) of the total respondents show the unlikeliness of purchasing this brand of diaper by indicated that they probably or definitely would not purchase this brand of diaper. Furthermore, 11% of the total respondents indicated that they might or might not purchase this brand of diaper.

In general, the purchase likelihood tends to own an ascending pattern, with the majority of the respondents indicated that they definitely would purchase this brand of diaper.



The question specifically asking us to recode the purchase likelihood (Likely_Purchase) into two groups by combining codes 2, 3, 4, and 5 into a single category. Therefore, by assigning 1 to *group 1* and grouping 2,3,4,5 to *group 2*, we will generate two groups for later analysis.

27	Info_New_Different	Numeric	1	0	Information is N...	{1, extremel...	6	8	Right	Scale	Input
28	Info_Appropriate	Numeric	1	0	Information is A...	{1, very appr...	6	8	Right	Scale	Input
29	Info_Believable	Numeric	1	0	Information is B...	{1, extremel...	6	8	Right	Scale	Input
30	Info_Understanding	Numeric	1	0	Information is U...	{1, very eas...	6	8	Right	Scale	Input

Discriminant analysis can be used for two main objectives – to explain the differences between groups in a multivariate manner or to be used as a procedure to classify observations with known attribute values while group membership remains unknown. One assumption of discriminant analysis is that the dependent variable should be measured in nominally scaled while the independents variable should be measured in metrically scaled. Therefore, in this case, it is necessary to change our independent variables – Info_New_Different, Info_Appropriate, Info_Believable, and Info_Understanding – to metrically scaled, as the above image presents. The recoded purchase likelihood – Likely_Purchase_Recoded – is already set to be measured in Scale, which is what it should be.

In addition, the groups to be analyzed need to be predetermined. In this case, the groups to be analyzed will be the two groups that we created previously for purchase likelihood, denoted by 1 and 2.

Discriminant

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		299	99.7
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	1	.3
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	1	.3
Total		300	100.0

The output after running the two-group discriminant analysis indicates that the number of valid observations become 299 – there is one missing discriminating variable. In other words, 299 responses will be used in the outputs. Nevertheless, there is not big a difference between 300 and 299, therefore, the slightly difference in the number of valid observations could be safely ignored. The analysis results would not be severely affected.

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.081[a]	100.0	100.0	.274

a. First 1 canonical discriminant functions were used in the analysis.

We have only recoded purchase likelihood into two 2 groups. Consequently, there is only one discriminant function.

The eigenvalue represents the variance explained in the observations. In other words, eigenvalue is considered to access the overall difference between the two groups. The higher eigenvalue yields for higher discriminant function. In other words, the larger the eigenvalue, the more of the variance in the dependent variable is explained by the function. In our case, the eigenvalue is 0.081, which signaled a low discriminant function.

The eigenvalue has a disadvantage of not being standardized to values between 0 and 1, other metrics based on the eigenvalue have been established for quality assessment, including the canonical correlation coefficient.

The canonical correlation is the measure of association between the discriminant function and the dependent variable, recoded purchase likelihood. When there are only two groups, the canonical correlation is an extremely useful measure in the table.

In addition, the square of canonical correlation coefficient is the percentage of variance explained in the dependent variable. In our case, the value of canonical correlation coefficient is 0.274, which means only approximately 7.5% of the variance within the recoded purchase likelihood are explained (calculated by 0.274 x 0.274). Specifically, 7.5% of the variance between “definitely would purchase this brand of diaper” and “less likely to purchase this brand of diaper” are explained. Lastly, the fact that the canonical correlation coefficient is only 0.274 signaled that it is ta low discriminant function, same with what eigenvalue indicates.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.925	23.089	4	.000

In a sufficiently large sample size, even small differences are likely to have statistical significance. Therefore, even though it shows statistically significant in the above table, we

should still pay close attention to other criteria, such as the canonical correlation coefficients, Wilk's lambda, or the absolute values of the mean differences between the groups.

Furthermore, chi-square statistic is also presented in the above table. Chi-square statistic can help us gain insights by testing the hypothesis that the means of the function is equal across the two groups, which are "definitely would purchase this brand of diaper" and "less likely to purchase this brand of diaper" in our case.

Wilks' Lambda is another critical indicator, it is a measure of how well each function separates cases into groups. Wilks' Lambda is an inverse measure of goodness, the smaller values of Wilks' lambda indicate greater discriminatory ability of the function. In other words, the lower values imply a better discriminant power of the discriminant function while the higher values imply a worse discriminant power of the discriminant function. As shown above, The Wilks' Lambda shows 0.925, which implies a low discriminant power of the discriminant function.

Standardized Canonical Discriminant Function Coefficients

	Function 1
Information is New & Different	.347
Information is Appropriate	.471
Information is Believable	.578
Information is Understandable	-.176

The influence of the independent variables can be accessed from standardized discriminant coefficients. The standardized discriminant function coefficients table offers insights of the relative importance of the independent variables in predicting the recoded purchase likelihood.

The higher the absolute value of a standardized coefficient, the greater the discriminatory power of the associated variable. In our case, in terms of absolute values, *Information is New and Different* owns a value of 0.347, *Information is Appropriate* owns a value of 0.471, *Information is Believable* owns a value of 0.578, and *Information is Understandable* owns a value of 0.176.

Among all four independent variables, *Information is Believable* has the strongest discriminatory power while *Information is Understandable* has the least discriminatory power.

Information is Appropriate has slightly greater discriminatory power than *Information is New and Different* but has slightly less discriminatory power than *Information is Believable.*

Structure Matrix

	Function 1
Information is Believable	.825
Information is Appropriate	.692
Information is New & Different	.637
Information is Understandable	.135

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

As mentioned above, since we have only created two groups for purchase likelihood, we have only one discriminant function. Nevertheless, Structure Matrix can be useful to provide us insights by identifying the largest absolute correlations within our discriminant function.

Information is Believable owns the largest absolute size of correlation in the discriminant function while *Information is Understandable* owns the smallest absolute size of correlation in the discriminant function. In other words, *Information is Believable* shows the strongest correlation with our discriminant function.

Classification Statistics

Classification Processing Summary

Processed		300
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	1
Used in Output		299

Prior Probabilities for Groups

Recoded Purchase Likelihood	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Definitely Would Purchase	.500	165	165.000
Less Likely to Purchase	.500	134	134.000
Total	1.000	299	299.000

Prior Probabilities are used in classification, and it affect researcher's decision regarding group membership. When having different group sizes, the group sizes observed in the sample have been used to determine the prior probabilities of membership in the groups – "definitely would purchase this brand of diaper" and "less likely to purchase this brand of diaper." As shown in the Prior Probabilities table, the groups sizes are unequal. There are 165 respondents indicated that they would definitely purchase this brand of diaper while there are also 134 respondents shown a tendency of less likely to purchase this brand of diaper.

Classification Results[a,c]

			Predicted Group Membership		
		Recoded Purchase Likelihood	Definitely Would Purchase	Less Likely to Purchase	Total
Original	Count	Definitely Would Purchase	110	55	165
		Less Likely to Purchase	60	74	134
	%	Definitely Would Purchase	66.7	33.3	100.0
		Less Likely to Purchase	44.8	55.2	100.0
Cross-validated[b]	Count	Definitely Would Purchase	109	56	165
		Less Likely to Purchase	61	73	134
	%	Definitely Would Purchase	66.1	33.9	100.0
		Less Likely to Purchase	45.5	54.5	100.0

a. 61.5% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 60.9% of cross-validated grouped cases correctly classified.

From Classification Results table, we are able to access how well the discriminant function works. In addition, we could also observe if the discriminant function works equally well for "definitely would purchase this brand of diaper" and "less likely to purchase this brand of diaper" of the recoded purchase likelihood.

Our discriminant function correctly classifies 66.7% of the responses for "definitely would purchase this brand of diaper" while correctly classifies 55.2% of the responses for "less likely to purchase this brand of diaper."

In terms of misclassified, the discriminant function misclassifies 33.3% of the responses for "definitely would purchase this brand of diaper" while misclassifies 44.8% of the responses for "less likely to purchase this brand of diaper."

Comparing to that of "less likely to purchase this brand of diaper," the correctly classified proportion for "definitely would purchase this brand of diaper" is approximately 10% more, 11.5% specifically.

Comparing to that of "definitely would purchase this brand of diaper," the misclassified proportion for "less likely to purchase this brand of diaper" is approximately 10% more, 11.5% specifically.

Lastly, overall, there are 61.5% of the responses are classified correctly. Note that if we had a total of 300 valid originally grouped responses, there would be 60.95% of the responses been classified correctly, calculated by *(66.7+55.2)/2=60.95*. However, since there are only 299 responses been used in the output, the overall percentage of correctly classified responses

became 61.5%, which is slightly greater than 60.95%, but no significant difference between them. Therefore, it could be safely ignored.

Conclusion

The eigenvalue, canonical correlation, and Wilks' Lambda all indicate a low discriminant function for our case. In terms of independent variables, *Information is Believable* has the strongest discriminatory power and the strongest correlation with the discriminant function. *Information is Understandable* has the least discriminatory power and the weakest correlation with discriminant function.

As for the classification, the groups sizes within the recoded purchase likelihood are unequal. There are 165 respondents indicated that they would definitely purchase this brand of diaper while there are 134 respondents shown a tendency of less likely to purchase this brand of diaper. Our discriminant function correctly classifies 66.7% of the responses for “definitely would purchase this brand of diaper.” On the other hand, 55.2% of the responses for “less likely to purchase this brand of diaper” are classified correctly. Overall, there are 61.5% of the responses are classified correctly.