Homework 6

Name: Zixin Ouyang

Exercise 1

Class Level Information					
group	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	49	49.0000	0.753846	0.753846
2	2	16	16.0000	0.246154	0.246154

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
15.535325	10	0.1137

Multivariate Statistics and Exact F Statistics					
S=1 M=1 N=29					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.68277761	6.97	4	60	0.0001
Pillai's Trace	0.31722239	6.97	4	60	0.0001
Hotelling-Lawley Trace	0.46460573	6.97	4	60	0.0001
Roy's Greatest Root	0.46460573	6.97	4	60	0.0001

Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into group			
From group	1	2	Total
1	46 93.88	3 6.12	49 100.00
2	9 56.25	7 43.75	16 100.00
Total	55 84.62	10 15.38	65 100.00
Priors	0.75385	0.24615	

Error Count Estimates for group			
	1	2	Total
Rate	0.0612	0.5625	0.1846
Priors	0.7538	0.2462	

1.
(a) Based on the chi-square test result, the null hypothesis of equal covariance matrices is not rejected at a 10% level of significance. Therefore, it is reasonable to assume a pooled covariance for the infant groups, indicating linear discriminant analysis will be fine.

The MANOVA tests for equal means of the four variables between the infant groups. All of the MANOVA tests are significant at a 5% level of significance. This tells us the linear discriminant function using the four variables have strong discriminating power between the infant groups.

(b) The overall cross validation error is 18.46%, which is successful. In fact, the group 1 was classified correctly by the model 93.88% of the time and was classified as group 2 6.12% of the time. Group 2 was classified correctly 43.75% of the time, with 56.25% of group 2 incorrectly classified as group 1.

The overall model predicted 55 observations in group 1 even though there were only 49 of these observations in the dataset. Furthermore, the model predicted 10 observations in group 2 even though there were 16 of these observations in the dataset. Although the model appeared to be more successful at predicting group 1 compared to group 2, it could partly be the result of this discrepancy in group rates between the model and the actual dataset.

Exercise 2

Stepwise Selection Summary										
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	factor68		0.2041	16.15	0.0002	0.79593517	0.0002	0.20406483	0.0002
2	2	bw		0.1374	9.88	0.0026	0.68657015	<.0001	0.31342985	<.0001

2. (a) From stepwise selection, only bw and factor68 are statistically significant for discriminating between the infant groups. Therefore, birthweight and factor68 two variables will be used in the discriminant analysis.

Class Level Information					
group	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	1	49	49.0000	0.753846	0.753846
2	2	16	16.0000	0.246154	0.246154

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
6.885791	3	0.0756

Multivariate Statistics and Exact F Statistics					
S=1 M=0 N=30					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.68657015	14.15	2	62	<.0001
Pillai's Trace	0.31342985	14.15	2	62	<.0001
Hotelling-Lawley Trace	0.45651541	14.15	2	62	<.0001
Roy's Greatest Root	0.45651541	14.15	2	62	<.0001

Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into group			
From group	**1**	**2**	**Total**
1	45 91.84	4 8.16	49 100.00
2	9 56.25	7 43.75	16 100.00
Total	54 83.08	11 16.92	65 100.00
Priors	0.75385	0.24615	

Error Count Estimates for group			
	1	**2**	**Total**
Rate	0.0816	0.5625	0.2000
Priors	0.7538	0.2462	

(b) The p-value for the test of homogeneity of within covariance is 0.0756. Thus we can conclude that the two groups have significantly different covariance and quadratic discriminant analysis should be used.

MANOVA test is strongly rejected at any reasonable level of significance, so the quadratic discriminant function based on the bw and factor68 should have some success discriminating between two infant groups. The MANOVA result is similar to that for the model in question 1.

(c) The overall cross validation error is 20%, which is worse than the error in question 1. Compared to the misclassification rate in question 1, the misclassification rate increased for group 1, while remained the same for group 2. Group 1 was classified correctly by the model 91.84% of the time and was classified as group 2 8.16% of the time. This is worse classification than question 1. Group 2 was classified correctly 43.75% of the time, with 56.25% of group 2 incorrectly classified as group 1, the same as the classification result from the model in question 1. Overall, results of this model after variable selection is slightly different from that of the model in question 1.

Exercise 3

Number of Observations Read	45
Number of Observations Used	45

Class Level Information					
group	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
1	1	35	35.0000	0.777778	0.777778
2	2	10	10.0000	0.222222	0.222222

Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	20
Number of Observations Used	20

Number of Observations and Percent Classified into group			
From group	**1**	**2**	**Total**
1	14 100.00	0 0.00	14 100.00
2	3 50.00	3 50.00	6 100.00
Total	17 85.00	3 15.00	20 100.00
Priors	0.77778	0.22222	

Error Count Estimates for group			
	1	2	Total
Rate	0.0000	0.5000	0.1111
Priors	0.7778	0.2222	

3. The results above from a quadratic discriminant analysis based on training and test set. Among 20 observations assigned to the test set, 3 observations are misclassified and the total error rate is 11.11%, which is less than the cross-validation error rate of 20% in Exercise 2. Based on the overall error rate estimates, this looks like a good model for discriminating between the two infant groups. However, the highest individual error rate estimate is for group 2, which is 50%. Therefore, this model performance is not good based on individual error rate estimates.