

## STAT 448 HW #6

Tianqi Wu

2018/ 11/15

### Problem 1.(a)

Here is the code that produces the stepwise selection:

```
proc stepdisc data=wine method=stepwise sls=0.05 sle=0.05;  
class alcohol;  
run;
```

From the result, 9 variables are significant at 5% level. The variables are nonflavanoid\_phenols, hue, malic\_acid, magnesium, alcalinity\_ash, od280\_od315, proline, ash and proanthocyanins.

The STEPDISC Procedure

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	nonflavanoid_phenols		0.7278	233.93	<.0001	0.27222451	<.0001	0.36388775	<.0001
2	2	hue		0.6235	144.08	<.0001	0.10249051	<.0001	0.62136638	<.0001
3	3	malic_acid		0.4006	57.80	<.0001	0.06143622	<.0001	0.73590105	<.0001
4	4	magnesium		0.1532	15.55	<.0001	0.05202633	<.0001	0.75251993	<.0001
5	5	alcalinity_ash		0.2131	23.15	<.0001	0.04094029	<.0001	0.78878774	<.0001
6	6	od280_od315		0.1172	11.29	<.0001	0.03614114	<.0001	0.79933202	<.0001
7	7	proline		0.1037	9.78	<.0001	0.03239310	<.0001	0.80706733	<.0001
8	8	ash		0.0552	4.91	0.0085	0.03060568	<.0001	0.81176624	<.0001
9	9	proanthocyanins		0.0374	3.24	0.0415	0.02946112	<.0001	0.81553790	<.0001

### Problem 1.(b)

Here is the code that produces the discriminant analysis:

```
proc discrim data=wine manova pool=test crossvalidate;  
class alcohol;  
var malic_acid--proline;  
run;
```

From the Test of Homogeneity of Within Covariance Matrices, the p-value(<0.0001) rejects the null and we conclude that QDA is more appropriate. From the MANOVA tests, all of the tests have p-value(<0.0001). Therefore, we reject the null and conclude that group means are different. Hence, it is very likely to discriminate between alcohol types based on these variables.

#### The DISCRIM Procedure Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
597.189174	156	<.0001

#### The DISCRIM Procedure

Multivariate Statistics and F Approximations					
S=2 M=4.5 N=81					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02832411	67.54	24	328	<.0001
Pillai's Trace	1.63745462	62.10	24	330	<.0001
Hotelling-Lawley Trace	10.79988562	73.42	24	280.09	<.0001
Roy's Greatest Root	7.77768507	106.94	12	165	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

### Problem 1.(c)

Here is the code that produces the discriminant analysis:

```
proc discrim data=wine manova pool=test crossvalidate;  
class alcohol;  
var malic_acid--proline;  
run;
```

From the cross-validation error results, there is 2.54% of observations classified into the wrong group. Hence, the discrimination matches the groups pretty well. All of the observations from group 3 are classified correctly. 2 out of 59 observations from group 1 are misclassified as group 2. 3 out of 71 observations from group 2 are misclassified as group 1.

**The DISCRIM Procedure**  
**Classification Summary for Calibration Data: WORK.WINE**  
**Cross-validation Summary using Quadratic Discriminant Function**

Number of Observations and Percent Classified into alcohol				
From alcohol	1	2	3	Total
1	57 96.61	2 3.39	0 0.00	59 100.00
2	3 4.23	68 95.77	0 0.00	71 100.00
3	0 0.00	0 0.00	48 100.00	48 100.00
Total	60 33.71	70 39.33	48 26.97	178 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for alcohol				
	1	2	3	Total
Rate	0.0339	0.0423	0.0000	0.0254
Priors	0.3333	0.3333	0.3333	

Gus Theofanis works as a statistician for State Farm and he shared some valuable information during the Q&A. First, he talked about some statistical methods important to his job. The methods mostly are linear and logistic regression, A/B testing and simple multivariate anaova tests. He believes that it is useful to understand the model validation and theories behind. Also, some soft skills are important for his job. The business partner would appreciate if he could relate his statistical knowledge back to the business perspective. Instead of being a pure math person, it is suggested to make more conversation with others and share the understanding. There are some challenges he facing in his role. He has to understand the need for different groups of people clearly and be ready to make changes accordingly. A lot of back-and-forth may be needed to finish the task.

He left the company and returned because he thinks the work at State Farm is more statistical and there is no culture issue here. It is also better for him to develop leadership. He enjoys the job that he could try different models and use the knowledge to help the business customers. He views himself more as a consultant to figure out the best way to address the client questions.

However, he does feel like doing the similar work over and over again sometimes. If he were not in this career, he would be a musician. In order to have a positive work/life balance in this career, he suggests to think less about work and dissociate from work when off duty.

He will look for self-motivated new members when hiring. The new member needs to show interest to the business and has a solid technical background. Master or above may be more desirable. He also gave some career advice that one should not be afraid of making mistake but learn form the mistake and avoid them next time.