Sample Final

STA 825 Final Exam – Wednesday, March 18
SHOW ALL WORK

Name: Key

1. A study was conducted to evaluate the effects of different washing treatments on the quality of beef after storage. There were three washing treatments: normal, chlorine, and lactic acid. On each of four consecutive days, three samples of beef were prepared and each sample was randomly assigned to one of the three treatments such that each treatment was observed on one sample per day. The experiment had to be conducted over several days because it was not possible to prepare and treat more than three samples per day. After treatment, all samples were stored for 10 days. After storage, each of the samples was evaluated on six criteria: beefy aroma, bloody/serumy aroma, metallic aroma, grassy/barnyard aroma, sour aroma, and spoiled aroma.

   a. Why is this a multivariate rather than a univariate data situation?

Because we have multiple (6) responses as opposed to just one. Here we have

Beefy aroma
Bloody/Serumy "
Metallic "
grassy/barnyard "
Sour "
Spoiled "

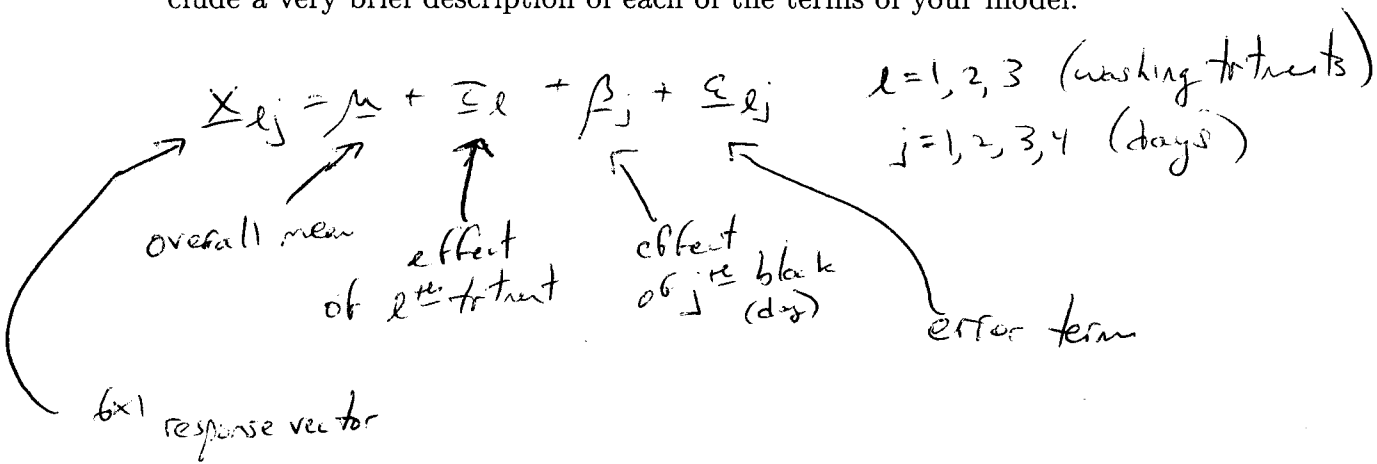
} all as response variables.

   b. What is the design here? (Hint: we only talked about three designs, and its one of those three.)

Randomized Complete Block
(Days as Blocks)

$X_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$ $i = 1, 2, 3$ (washing treatments)

c. Write down an appropriate MANOVA model for analyzing these data. Include a very brief description of each of the terms of your model.



d. In terms of the parameters of your model, write down the null hypothesis for testing that there is no difference among the three washing methods.

$H_0: \tau_1 = \tau_2 = \tau_3 = 0$

e. After testing the null hypothesis from (d), we will typically want to obtain simultaneous confidence intervals for specific comparisons among the treatment means on particular response variables. Give one reason that the Bonferroni method may be preferred in some cases over the Roy method for obtaining simultaneous confidence intervals. Give one reason that the Roy method may be preferred in some cases over the Bonferroni method.

Bonferroni intervals will often be shorter

Roy intervals allow data snooping (unplanned comparisons)

2. In the early 1900's, several investigators were interested in predicting behavioral and social outcomes among people based on physical characteristics. Macdonnell (1902) reports a correlation matrix for the following seven physical variables measured on 3000 British criminals: (1) head length, (2) head breadth, (3) face breadth, (4) left finger length, (5) left forearm length, (6) left foot length, and (7) height. Assume that all original variables were measured in centimeters.

   The attached SAS program and output (labelled "Final Exam Problem #2") performs a principal components analysis based Macdonnell's correlation matrix. Answer the following questions:

a. One of the goals of principal components analysis is to reduce the dimension of the original data. How would you choose the number of principal components to retain for subsequent analyses? In this example, how many principal components would you retain?

Methods: 1.) Scree plot
2.) Decide the % age variance you want to explain
3.) Use the # of eigenvalues > 0
4.) Subjective judgement

Here, I would choose 3 P.C.s so that I explained a sufficient % age of variance (85%)

b. Briefly interpret the first two principal components in this example. That is, what aspect of the original variables is captured by the first principal component? the second?

1st: Overall (total) size

2nd: Contrast between head size & shape and limb & body length.

c. Explain why it is not appropriate for this example to perform a principal components analysis on the covariance matrix rather than the correlation matrix.

Because the variance of the variables differs a great deal. Height will vary much more than finger length, say. Therefore, when using unstandardized variables for the P.C.A, height will dominate the 1st P.C. and the 1st P.C. will account for nearly all of the variance.

d. Suppose that in addition to the seven variables described above, an eighth variable, computed as head length minus head breadth, had been included in the analysis to capture head shape. What would be the variance of the last principal component in such an analysis and why?

The last eigenvalue, λ_8, would be 0 because head shape is a linear comb of other variables (head length and head bdth) so there is a structural relationship among the 8 variables

3. Annual financial data are available on firms. Four financial variables including x_1 = (cash flow)/(total debt), x_2 = (net income)/(total assets), x_3 = (current assets)/(current liabilities), and x_4 = (current assets)/(net sales), were collected for 21 firms that subsequently went bankrupt and 25 financially sound firms at about the same point in time. A discriminant analysis for these data is performed in the attached SAS program and output (labelled "Final Exam Problem #3"). The discriminant analysis is based only on x_2 and x_3. Answer the following questions.

   a. In the SAS program, a hypothesis test is performed using PROC GLM of the hypothesis that the mean vectors are the same in the two groups (financially sound and financially troubled firms). Why is this test performed prior to performing a discriminant analysis, and what does the result of the hypothesis test say about how the discriminant analysis will perform?

The test is performed because a discriminant function will only perform well if the groups are well separated (don't overlap too much). If the group means are not significantly diff't we should not expect a low error rate using our discrim. function. In this case, the means are significantly diff't ($p < .0001$) so we should expect that the discriminant function may perform well.

b. In this example, equal priors and costs of misclassification were assumed. Explain why it may be more appropriate to use different costs of misclassification in this analysis? Why might it be more appropriate to use different prior probabilities?

Diff't costs of misclassification may be appropriate because we'd rather tell a financially sound firm that it may be in trouble than tell a troubled firm that it is sound. The latter is more likely to be disastrous because it may result in inaction for a firm that should take steps to avoid bankruptcy.

Diff't priors may be appropriate because there are probably substantially more financially sound firms than troubled ones in the general population.

c. Write down the linear discriminant rule for classifying firms as financially sound or financially troubled.

$$\underline{b}'\underline{x} = \begin{pmatrix} -11.288 + 1.085 \\ 2.408 - 4.047 \end{pmatrix} x = -10.203x_2 - 1.639x_3$$

$$k = -(-2.108) - 5.219 = -3.111$$

So classify an observation $\underline{x} = (x_2, x_3)^T$ as financially troubled if

$$-10.203x_2 - 1.639x_3 + 3.111 > 0$$

d. There are two estimated error rates in the output. What error rate would you use and why, if you were describing how you expected the linear discriminant rule from (d) to perform in practice (when actually classifying firms of unknown status as either financial sound or troubled)?

I would use the cross-validation error rate 13.52% because resubstitution is biased downward. The error rate will be lower on the data from which the discriminant rule was formed than for new data.

4. The following table lists measurements on 5 nutritional variables for 12 breakfast cereals.

TABLE 12.9 BREAKFAST-CEREAL DATA

Cereal	x_1 Protein (gm)	x_2 Carbohydrates (gm)	x_3 Fat (gm)	x_4 Calories (per oz)	x_5 Vitamin A (% daily allowance)[a]
1. Life	6	19	1	110	0
2. Grape Nuts	3	23	0	100	25
3. Super Sugar Crisp	2	26	0	110	25
4. Special K	6	21	0	110	25
5. Rice Krispies	2	25	0	110	25
6. Raisin Bran	3	28	1	120	25
7. Product 19	2	24	0	110	100
8. Wheaties	3	23	1	110	25
9. Total	3	23	1	110	100
10. Puffed Rice	1	13	0	50	0
11. Sugar Corn Pops	1	26	0	110	25
12. Sugar Smacks	2	25	0	110	25

[a] 0 indicates less than 2%.

A cluster analysis for these 12 cereal brands is performed in the attached SAS program and output (labelled "Final Exam Problem #4"). Answer the following questions.

a. Identify the type of clustering algorithm that was used in the SAS program. (Don't describe the algorithm, just tell me the type of algorithm it is — more than one adjective is necessary).

Agglomerative Hierarchical clustering using average linkage.

b. What is the cluster structure for 5 clusters given by the results of PROC CLUSTER in this example. That is, assuming there are 5 clusters, tell me which cereals belong together.

Cluster 1 = Life, & Special K

Cluster 2 = Grape Nuts, Super Sugar Crisp, Rice Krispies, Sugar Smacks, Sugar Corn Pops

Cluster 3 = Raisin Bran, Wheaties

Cluster 4 = Product 19, Total

Cluster 5 = Puffed Rice

c. Based on the pseudo-T^2 statistics printed in the SAS output, what is the appropriate number of clusters for this example.

5 because $T^2 = 14.3$ when we go from 5 to 4 clusters indicating that the clusters merged at that step were relatively far apart.

d. Attached to the output is a page of star plots for the 12 cereals. Explain why just a dot appears for Puffed Rice.

Because Puffed Rice has the smallest value for all 5 variables.