

# Statistics 579

## Applied Multivariate Methods

### Exam 2

1. The file `Magazines.sas7bdat` contains information from a subset of households from a suburban panel in a Midwestern U. S. market. Of the 141 subjects, each said that they subscribed to one and only one of the following 4 magazines: *Better Homes & Gardens*, *Reader's Digest*, *TV Guide*, and *Newsweek*. For each household, a variety of demographic information is available. The 4-level grouping variable is `MagazineSubscription`. Demographic variables to be used are `FamilySize`, `Income`, `Race`, `TVSets`, `NewspaperSubscriber`, `NoMaleHeadHousehold`, `NoFemaleHeadHousehold`, `Children`, `Age`, `Education`.
  - a) Using an appropriate variable selection procedure, determine which of the demographic variables are useful for discriminating among the 4 groups. Use  $\alpha=0.025$ .
  - b) Using the variables selected, classify the observations according to magazine subscription. Estimate the classification error rate. Comment on the usefulness of this model.
  - c) Plot the means of the groups on the canonical variables in the 2-dimensional canonical space, using different plotting symbols to identify the 4 groups. Characterize each of these 4 groups based on their average scores on the original measurements.
  
2. A sample of 111 students entering college was partitioned according to the college French course in which they enrolled. Thirty-three (33) enrolled in the beginning level, and seventy-eight (78) enrolled in the intermediate level. Thirteen (13) measures were obtained on each of the 111 students:
  - Five high school cumulative grade-point averages:
    - English (EGPA)
    - Mathematics (MGPA)
    - Social Science (SGPA)
    - Natural Science (NGPA)
    - French (FGPA)
  - The number of semesters of high school French (SHSF)

Four measures of academic aptitude:

ACT English (ACTE)

ACT Mathematics (ACTM)

ACT Social Studies (ACTS)

ACT Natural Sciences (ACTN)

Two scores on a French test:

ETS Aural Comprehension (ETSA)

ETS Grammar (ETSG)

The number of semesters since the last high school French course (SLHF)

Data are contained in the file French.txt.

- a) For the students enrolled in the beginning level of French, and separately for the students enrolled in the intermediate level of French, determine if the assumption of multivariate normality is reasonable (using all 13 variables).
- b) Using an appropriate variable selection procedure for discriminant analysis, determine which of the 13 variables are useful for discriminating between the 2 groups. Use  $\alpha=0.025$ .
- c) Using the selected variables, obtain the sample linear discriminant function, and estimate the classification error rate. Assume equal priors.
- d) Using the above linear discriminant function, classify 5 new students whose status is unknown that are contained in the file Frenchtest.txt.
- e) Plot all 116 values in a 2-dimensional canonical using different plotting symbols to identify whether the individual is known to be enrolled in beginning French, or is known to be enrolled in intermediate French, or is Unknown.
- f) Develop a logistic regression model that predicts enrollment based upon the subset of variables selected by an appropriate variable selection technique. Estimate the cross-classification error rate, and compare it with the error rate obtained from discriminant analysis.
- g) Using the above logistic regression model, classify the 5 new students whose status is unknown that are contained in the file Frenchtest.txt. Compare these results with the results obtained from discriminant analysis.