Chapter 18

Discriminant Analysis

Review: Previous Techniques

- Regression models
  - Numeric response modeled by continuous and/or categorical predictors
- PCA
  - (Unknown) feature extraction and dimension reduction for correlated variables
- Cluster Analysis
  - Obtain (unknown) groups based on proximity of continuous predictors

Discriminant Analysis

- Data from known groups
- Want one or more functions based on explanatory variables to classify those groups
- Classification is based on group means
- Goal is to classify new observations (with unknown group) into one of the known groups

Terminology

- **Discriminant function** – effectively define dividing lines for groups
- **Prior probabilities** – tell relative sizes of known groups in the general population
- **Posterior probabilities** – probabilities that an observation with specific explanatory variable values would be from each of the known groups
- **Misclassification rate estimates** – estimated percentages of observations classified into the wrong group (methods include re-substitution, cross-validation, or using training and test sets)
- **Training set** – the set of observations the discrimination is based on
- **Test set** –observations with known group not included in training set used to estimate classification of new observations

A Few Types

- Assumed normal populations (LDA and QDA)
- Logistic discriminants
- Nonparametric or semi-parametric (e.g. kernel-based)
- Support vector machines (a more general methodology)
- Will focus on LDA and QDA in class

Linear Discriminant Analysis (LDA)

- Assume multivariate normal distribution for each group
- Assume same covariance for each group
- Separating surfaces will be linear(straight lines or planes or hyperplanes)

Quadratic Discriminant Analysis (QDA)

- Assume multivariate normal distribution for each group
- Covariances not assumed to be the same for each group
- Separating surfaces will be quadratic

proc discrim

- **class** statement for classification variable (really our response variable)
- **var** statement for our predictors, just like in other procedures
- Will give us classifications, misclassification rates, etc.
- Can use resubstitution, cross-validation, or test/training set misclassification estimates

Example: Skulls Data Set

Variables:

- **Length** – measurement of skull length
- **Width** – measurement of skull width
- **Height** – measurement of skull height
- **Faceheight** – measurement of face height
- **Facewidth** – measurement of face width
- **Type** – A (local graves) or B (battlefield)

Example: LDA with One Predictor

- Classify skulls based only on length using LDA
- Will look at the following:
  - the classifications
  - misclassification rates
  - the posterior probabilities (via the **out** option)
  - estimated densities (via the **outd** option) for the underlying populations

Example: Cross-Validation

- Leave-one-out crossvalidation is more realistic
- Removes influence of data point on its own classification
- Can be obtained with **crossvalidate** option

Exercise: Unequal Priors

- Prior probabilities are about the assumed proportion in the general population
- Specify with **priors** statement
- See docs for **priors** statement…
- Repeat previous analysis assuming proportional rather than equal priors
- Compare results

Example: Iris Data and Petal Width

- Now will have 3 groups (species)
- Will classify based only on petal width first
- Will look at LDA, testing for unequal covariances, and QDA
- Simplified version of **The DISCRIM Procedure>> Examples>> Univariate Density Estimates and Posterior Probabilities**

Example: Both Petal Measurements

- Will look at the example in the docs for this one
- **The DISCRIM Procedure>> Examples>> Bivariate Density Estimates and Posterior Probabilities**

Exercise: Skulls with All Predictors

With the skulls data set:

- Perform discriminant analysis using all 5 skull measurements
- What does the MANOVA tell us?
- Compare these classifications to those based only on length

stepdisc Procedure

- Stepwise selection for terms in discriminant analysis
- Can be used to determine predictors to use in **proc discrim**

Exercise: Selection for Skulls

- Use **stepdisc** to choose best skull measurements for discrimination
- Compare results with those based on **length**

Training and Testing Sets

- Will look at **The DISCRIM Procedure>> Examples>> Linear Discriminant Analysis of Remote-Sensing Data on Crops**
- Steps:
  - discrimination on a training data set
  - output training classification info via **outstat=**
  - use that data as a new input data set in **proc discrim** via **data=** and define a test set via **testdata=**
  - **testout=** to write the test classification information out like we did with **out=** before