Chapter 17

Cluster Analysis

Review: Previous Techniques

- Regression models (linear, logistic, generalized linear, ANOVA)
    - Predicting response based on continuous and/or categorical predictors
- PCA
    - Feature extraction and dimension reduction for correlated variables

Cluster Analysis: General Idea

- Want to identify groups (or clusters)
- Groups we want to identify are unknown
- Have continuous variables (not necessary to be continuous in general…)
- Want to group observations based on how similar (or dissimilar) the continuous variable values are
- Want well-separated groups

Terminology

- **hierarchical clustering**: nested grouping of points
- **distance** or **dissimilarity**: measure of how different points are (we will use Euclidean distance)
- **linkage**: method for measuring difference (e.g. distance) between clusters
- **dendrogram**: tree plot showing cluster hierarchy and distances

Linkages

- Define distances between clusters of points
- Few examples:
  - **single**: closest points in two different clusters
  - **complete**: most distant points in the two clusters
  - **average**: average of all distances between points in one cluster and points in the other cluster
- Others given in **The Cluster Procedure >>Details>>Clustering Methods** docs

Dendrograms

- Show distances between points and clusters
- Show the clustering history
- All points start in their own clusters
- Merge closest points or cluster of points
- Eventually all points in one big cluster
- Help us visually see well-separated clusters giving ideas of when to stop merging

proc cluster

- SAS procedure for hierarchical clustering
- Need **method** option to set linkage
- Distance based on variables in **var** statement
- **id** variable used in cluster history and tree labels
- **copy** statement includes additional variables in the output data set

proc tree

- Useful for creating dendrograms
- Also useful for just creating output data sets with cluster information in them

Initial Example: Iris Data

- Use complete linkage and guess number of clusters from dendrogram
- Obtain 3 clusters and compare with species
- Try single linkage & average linkage
- Guess number of clusters based on single and average linkages
- How well do the single and average linkage clusters match species?

Number of Clusters

Some diagnostics:

- Cubic clustering criterion (the **ccc** option)
- Pseudo t^2 and F statistics (the **pseudo** option)
- Higher ccc and pseudo F values indicate better clustering
- Lower pseudo t^2 values indicate better clustering

Example: Iris Diagnostics

- Look at plots of the ccc and pseudo F and t^2 statistics using the average linkage
- How many clusters would we choose based on each?

US Air Quality

Contains Cities and the following:

- SO_2 content in air ($\frac{mg}{m^3}$)
- Average temperature (F)
- Manufacturing companies employing 20 or more
- Population in thousands
- Average wind speed (mph)
- Average precipitation (inches)
- Average number of days with precipitation

Example: Pollution by City Groups

We will do the following:

- Identify and remove extreme cities
- Use complete linkage clustering using variables which could be predictors of SO_2 level
- Do means analysis by cluster
- Pick the two largest principal components and see where the clustered values fall
- Visualize SO_2 level by cluster
- Perform ANOVA of SO_2 level on cluster