Chapter 1

Introduction

1.1 Multivariate Analysis vs. Statistical Learning

Most statistical learning methods can be thought of as advanced multivariate methods.

1.1.1 What is multivariate analysis?

- Objective: Understand if and how response variables are related to one another and explanatory variables.

- Data are considered multivariate if 2 or more "response variables" are measured. The method employed depends upon whether there are 0 vs. 1 or more explanatory variables and whether they are continuous or categorical: PCA, cluster analysis, factor analysis, discriminant analysis, classification analysis,

MANOVA, canonical correlation

- Many multivariate methods are/were traditionally developed for designed experiments. But they can be applied to observational data and analysis is viewed as conditional upon the realized values of explanatory variables.

1.1.2 What is statistical learning?

- Objective: Predict "output" as a function of one or more "inputs". Can be used to model the relationship between input variables and output (careful here).

- Method employed depends on whether the output variable is observed, whether its categorical or continuous.

- Typically used in large observational studies. Because n is large, we let the data do the talking, i.e. do less statistical modeling.

- Many methods can be understood as advanced versions of multivariate methods or regression: PCA, cluster analysis, discriminant analysis, classification analysis, linear and logistic regression.

1.1.3 Examples

Example 1. *This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.*

Code

```
> # Iris data set
> data(iris) # may not need to run this line, but it won't hurt.
> help(iris)
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa

> plot(iris, pch = as.vector(iris$Species))
```

Multivariate Analysis

- What are the response and explanatory variables?
- Could this data be considered multivariate?
- Objective 1: Determine if species are different and, if they are, understand how (MANOVA).

  - Could you draw a line in one of the scatter plots that separates Setosa plants from the other 2 specie? This is called a *linear discriminant analysis*.

- Objective 2: Sample a plant, measure its leaves, and use the leaf measurements to *predict* whether or not its a Setosa or *classify* it as a Setosa or non-Setosa. Classical linear discriminant analysis or logistic regression can estimate parameters in

$$logit(\Pr(Setosa)) = \beta_0 + \beta_1 SL + \beta_2 SW + \beta_3 PL + \beta_4 PW$$

Statistical Learning

- We can use the line or logistic model above or more generally

$$\Pr(Setosa) \approx f(SL, SW, PL, PW)$$

- How would is n important here in determining what assumptions / restrictions should be put on f?

- Would this be considered supervised or unsupervised statistical learning?

Example 2. *The USArrests data set contains arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. A biplot (more later) groups similar observations in the data set together.*

Code

```
# USArrests
> help(USArrests)
> biplot(princomp(USArrests, cor = TRUE))
```

- Objective 1: Determine which states are similar to Oklahoma and which regions might cluster together. Why would this type of *cluster analysis* be called *unsupervised learning*? Are there observed inputs and outputs?

1.1.4 Key points

- In many instances, statistical learning methods build upon classical multivariate statistical methods by relaxing assumptions.

- The degree to which the assumptions can/should be relaxed depends upon

  1. Sample size

  2. Do we need to interpret f or just get the best prediction possible?

1.2 Matrices review

1.2.1 Notation

- A matrix $\mathbf{X}$ is a rectangular array of numbers. A data set of p measurements on n experimental units is an $n \times p$ (n rows and p columns) array, sometimes written $\mathbf{X}_{n\times p}$ or $\mathbf{X} \in \Re^{n\times p}$ to emphasize the dimension.

  $\mathbf{X}_{n\times p} =$

- What would a matrix of two exam scores for 3 students look like?

- What is x_{12} above?

- What does x_{ij} generically represent?

- An $n \times 1$ matrix $\mathbf{x}$ is a vector. By default, vectors are written as $n \times 1$ (column) vectors rather than $1 \times n$ (row) vectors. The jth column of a matrix is written

  $\mathbf{x}_j =$

- The data for the ith row of $\mathbf{X}$ is a $1 \times p$ row vector

  $x_i^T =$

- We can write x_i^T as a column vector

  $x_i =$

- An $n \times 1$ vector is generally bold and times new roman $\mathbf{a}$ while others, such as the p dimensional vector are written a.

- The vector of outputs, if observed, is denoted $\mathbf{y}$.

1.2.2 Operations

- The transpose of a vector $\mathbf{x}$ is written $\mathbf{x}^T$. The transpose of a matrix $\mathbf{X}$ is written $\mathbf{X}^T$. Transposes are found by permuting rows and columns. Write out a 3×2 matrix and compute the transpose.

- A matrix can be written in terms of $\mathbf{x}_1, ..., \mathbf{x}_p$ or in terms of $x_1, ..., x_n$.

- In general x_{ij} refers to the jth measurement for the ith experimental unit and may or may not refer to the element in the ith row and jth column. For example if $\mathbf{X}$ is a data matrix, consider $\mathbf{X}^T$.

- The multiplication of two matrices is performed $\mathbf{A}_{n\times k}\mathbf{B}_{k\times p} = \mathbf{C}_{n\times p}$ where $c_{ij} = \sum_k a_{ik} \times b_{kj}$. Observe the number of columns of $\mathbf{A}$ must be equal to the number of rows of $\mathbf{B}$ for matrix multiplication to be defined. Write out two matrices and multiply them.

- If $\mathbf{1}$ is a vector of 1's and $\mathbf{x}_j$ is data for the jth variable, what is $\frac{1}{n}\mathbf{1}^T\mathbf{x}_j$?

- Is $\mathbf{A}^T\mathbf{A}$ well defined?