

# Lecture 1

## Overview

DSA 8070 Multivariate Analysis  
August 18-20, 2021

Whitney Huang  
Clemson University



Notes

---

---

---

---

---

---

---

## Agenda

### 1 Introduction

### 2 Objectives of Multivariate Analysis

### 3 Useful Tools for Multivariate Analysis



Notes

---

---

---

---

---

---

---

## Introduction

- In many observational or experimental studies, observations are collected simultaneously on **more than one variable** on each unit

```
> head(Boston)
  crim zn indus chas nox rm age dis rad tax ptratio black lstat medv
1 0.08632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98 24.0
2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14 21.6
3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03 34.7
4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94 33.4
5 0.00905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33 36.2
6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21 28.7
```

- **Multivariate analysis** is the collection of statistical methods that can be used to analyze these multiple measurements  
⇒ some are extensions of familiar methods (**t-test**, **ANOVA**, **linear regression**,...) while others are unique to multivariate analysis
- Idea is to exploit potential **"correlations"** among the multiple measurements to improve inference (see an example in the next slide)



Notes

---

---

---

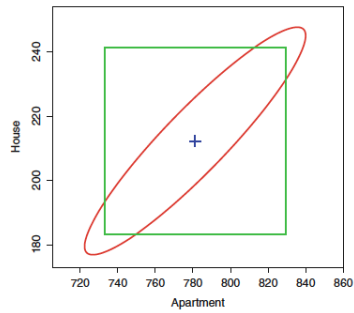
---

---

---

---

Using Multivariate Methods Could Lead to Sharper Inference



Source: Fig. 1.1 of Applied Multivariate Statistics with R by Zelterman

CLEMSON  
UNIVERSITY

Introduction

Objectives of  
Multivariate  
Analysis

Useful Tools for  
Multivariate  
Analysis

1.4

Notes

---

---

---

---

---

---

---

Objectives of Multivariate Analysis

CLEMSON  
UNIVERSITY

Introduction

Objectives of  
Multivariate  
Analysis

Useful Tools for  
Multivariate  
Analysis

1.5

Notes

---

---

---

---

---

---

---

Dimensionality Reduction or Structural Simplification

- **Goal:** to reduce the “dimensionality” by considering a small number of (linear) combinations of a large number of measurements without losing important information
- **Examples:**
  - A single index of patient reaction to radiotherapy can be constructed from measurements on several response variables
  - Wildlife ecologists can construct a few indices of habitat preference from measurements of dozens of features of nesting sites selected by a certain bird species
- **Techniques:**
  - **Principal Component Analysis** (Week 9)
  - **Factor Analysis** (Week 10)
  - **Multidimensional Scaling** (Week 14)

CLEMSON  
UNIVERSITY

Introduction

Objectives of  
Multivariate  
Analysis

Useful Tools for  
Multivariate  
Analysis

1.6

Notes

---

---

---

---

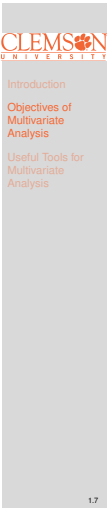
---

---

---

Grouping or Classification

- **Goal:** to **identify** groups of “similar” units or to **classify** units into previously defined groups
- **Examples:**
  - The US IRS uses data collected from tax returns (income, amount withheld, deductions, ...) to **classify** taxpayers into two groups: those who will be audited and those who will not
  - Using the concentration of elements (copper, silver, tin, antimony) in the lead alloy used in bullets, the FBI **identifies** ‘similar’ bullets that may be used to infer whether bullets were produced from the same batch of lead
- **Techniques:**
  - **Classification Analysis** (Week 12)
  - **Cluster Analysis** (Week 13)



Notes

---

---

---

---

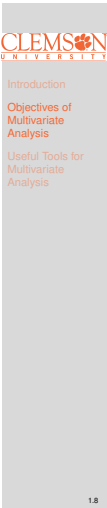
---

---

---

Investigation of the Dependence among Variables and Prediction

- **Goal:** to estimate the relationship among variables and, if the variables are associated, to predict the value of some of them given information on the others
- **Examples:**
  - The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance
  - The association between test scores, and several college performance variables were used to develop predictors of success in college
- **Techniques:**
  - **Multivariate Regression** (Week 7)
  - **Canonical Correlation Analysis** (Week 11)



Notes

---

---

---

---

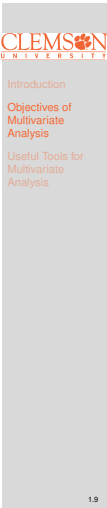
---

---

---

Hypothesis Testing

- **Goal:** to test if differences in sets of response mean vectors for two or more groups large enough to be distinguished from sampling variation
- **Examples:**
  - A transportation company wants to know if means for gasoline mileage, repair costs, downtime due to repairs differ for different truck models
  - An insurance company wants to know if changing case management practices leads to changes in mean length of hospital stay, mean infection rates, and mean costs
- **Techniques:**
  - **Hotelling's  $T^2$  and MAVONA** (Week 5 and Week 6)



Notes

---

---

---

---

---

---

---

# Useful Concepts/Tools for Multivariate Analysis

CLEMSON  
UNIVERSITY

Introduction  
Objectives of  
Multivariate  
Analysis  
Useful Tools for  
Multivariate  
Analysis

1.10

Notes

---

---

---

---

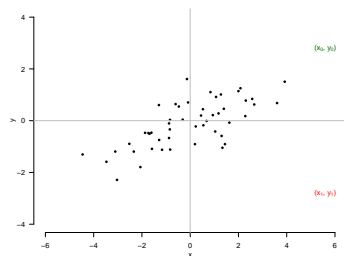
---

---

---

## Statistical Distance

Multivariate methods rely on “distances” between data points: **clustering** (group units that are “close”); **classification** (allocate each unit to the “closest” group)



**Question:** which one ( $(x_0, y_0)$  or  $(x_1, y_1)$ ) is closer the center of the observations?

CLEMSON  
UNIVERSITY

Introduction  
Objectives of  
Multivariate  
Analysis  
Useful Tools for  
Multivariate  
Analysis

1.11

Notes

---

---

---

---

---

---

---

## Matrix Algebra

The study of multivariate methods is greatly facilitated by the use of matrix algebra

- Many operations performed on multivariate data are presented using vector/matrix notation, e.g.,  $\mathbf{X}_{n \times p}$  (Data matrix);  $\hat{\boldsymbol{\mu}}_{p \times 1}$  (estimated mean vector);  $\hat{\boldsymbol{\Sigma}}_{p \times p}$  (estimated covariance matrix)
- The computation of **eigenvalues** and **eigenvectors** (i.e., the **spectral decomposition**) plays an important role in multivariate analysis
- We will use R to perform the needed matrix operations

CLEMSON  
UNIVERSITY

Introduction  
Objectives of  
Multivariate  
Analysis  
Useful Tools for  
Multivariate  
Analysis

1.12

Notes

---

---

---

---

---

---

---

Multivariate Normal Distribution

- We will (almost always) assume the joint distribution of  $X = (X_1, X_2, \dots, X_p)^T$  follows a multivariate normal distribution with the probability density function:

$$f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\Sigma)^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$

- The multivariate normal assumption is often appropriate:
  - Variables can sometimes be assumed to be multivariate normal (perhaps after transformation)
  - **Central limit theorem** tells us that distribution of many **multivariate sample statistics** is approximately normal, regardless of the form of the population distribution



Notes

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---