

**B9106: Applied Multivariate Statistics**

Fall 2020—Thursday/Friday 8:30AM-10:00AM

Professor Kamel Jedidi Office hours: TH/FR 1:30-2:30PM

518 Uris Hall e-mail: kj7@gsb.columbia.edu

Phone: (212) 854-3479 Classroom: Uris 332(TH), 303(FR)

TA: Isha Gupta ([ig2412@columbia.edu](mailto:ig2412@columbia.edu)) Overflow: Uris 306[[1]](#footnote-1)

Zoom link: <https://gsb-columbia-edu.zoom.us/j/96501069971?pwd=c1VaR24zRmhndEhrbGFwYnV2NkFRUT09>

Course Objectives

Multivariate Statistical techniques are important tools of analysis in all fields of management: Finance, Operations, Accounting, Marketing, and Management. In addition, they play key roles in the fundamental disciplines of the social science: Economics, Psychology, Sociology, ... etc.

This course is designed to provide students with a working knowledge of the basic concepts underlying the most important multivariate techniques, with an overview of actual applications in various fields, and with experience in actually using such techniques on a problem of their own choosing. The course addresses both the underlying mathematics and problems of applications. As such, a reasonable level of competence in both statistics and mathematics is needed.

Required Text:

Richard A. Johnson and Dean W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, (Sixth Edition).

Recommended Book:

Wilkinson, D. J., Multivariate Data Analysis with **R**—Downloadable for free from:

<http://www.staff.ncl.ac.uk/d.j.wilkinson/teaching/mas3314/notes.pdf>

Useful Books

* Chapman and McDonald Feit, R for Marketing Research and Analytics, Springer, 2015
* Bollen, Structural Equations with Latent Variables Wiley, 1989.
* Lattin, Carroll, and Green, Analyzing Multivariate Data, Duxbury, 2003.
* Hair, Anderson, Tatham, and Black, Multivariate Data Analysis, 1998.

Course Prerequisite

R-Programming data camp: <https://www.datacamp.com/courses/free-introduction-to-r> and a basic knowledge of statistics. The course assignments and course project use R for statistical computing. Before the start of the semester, students must download R from <http://www.r-project.org/> and RStudio, a powerful user interface for R, from <http://www.rstudio.com/>.

If you are not experienced with programming in R, we require that you complete by **Sept 15, 2020** an online interactive course on R through DataCamp.com. This is what you need to do to take the course:

1. Sign up for DataCamp using this link:

https://datacamp.com/groups/shared\_links/53dfb17d5ec259a52ad153e78a597ba7900f2d012330f0cee3ef724c0c1f82a0

You will need to use your UNI@columbia.edu email to sign-up.

1. Complete the "Introduction to R” course (about 4 hours).

Course Requirements:

1. Class participation 5%
2. Group project 25%
3. Assignments 20%
4. Midterm and Final Exams 50%

Group Project (25%)

The project requires you to work with a group of four students on a research problem of your choice. The task is to develop a series of research hypotheses based on theory or past empirical evidence and then apply some of the multivariate techniques covered in class on real data for testing.

Presentation

A typical presentation includes:

1. Research Questions/Hypotheses

2. Data

3. Data Analyses and Results

4. Limitations of the research and suggestions for future research.

A write-up of 10 pages (1.5 line spacing) needs to be submitted on the last day of class. Summarized tables and exhibits need to be appended to the write-up. They do not count towards the 10-page limit.

Assignments (20%)

These assignments involve statistical problems solving and data analysis exercises using R. Their purpose is to illustrate the material covered in class. You are expected to work on the problems both manually (on paper) and using R (**where applicable**). The assignments are graded on a scale ✓-, ✓, ✓+. Please see class schedule for the list of assigned problems (mostly from the required textbook). **These assignments are due by 11:59pm on Fridays in the week they are posted.**

Mid-Term and Final Exams (50%)

The exam purpose is to test student knowledge of the concepts and techniques covered in class. Both exams will be offered online, closed book, closed notes during the school’s exam periods. You are allowed, however, to bring two pages of notes to the exam. In addition, you are expected to bring your statistical tables as well as a calculator. The midterm will be based on all course material covered till the midterm date. The final will be based on all material covered after the midterm.

Class Participation (5%)

It is very important that you ask and answer questions during class. This will greatly help you and your classmates understand the material better. Another important aspect is your class attendance.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  | 30-Nov | 7-Dec |

**CLASS SCHEDULE**

**Week Date Topic**

1 9/10-11 **Course Introduction: Aspects of Multivariate Analysis**

**+ Review of Matrix Algebra and Random Vectors**

Read: Ch. 1 and 2

Complete the Introduction to R course (about 4 hours) before class. Link is on Canvas.

Do (on paper and with R): Problems 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 2.12

Recommended videos on Matrix Algebra: [3blue1brown’s Essence of Linear Algebra series](https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab) on YouTube provides an intuitive overview of linear algebra.

2 9/17-18 **Sample Geometry and Random Sampling**

Read: Ch. 3

Do on paper and with R: 2.19, 2.20, 2.21, 2.25, 2.26, 2.27, 2.30, 2.32, 2.34

3 9/24-25 **Multivariate Normal Distribution**

Read: Ch. 4

Do on paper and with R:1.6, 3.6, 3.10, 3.11, 3.14,3.15

4 10/01-02 **Multivariate Normal Distribution**

Read: Ch. 5 and 6 (Sections 6.1-6.3; 6.7-6.8)

Do on paper and with R: 4.2, 4.3, 4.4, 4.5, 4.8, 4.14, 4.15, 4.16

5 10/08-09 **Regression**

Read Ch. 7

Do on paper and with R: 4.18, 4.19, 4.21, 4.23, 4.26, 5.1, 5.3, 5.5, 6.5, 6.6, 6.11, 6.26, 6.27

6 10/15-16 **Analysis of Variance**

Read: Ch. 6 (Sect 6.4 pp. 314-320 and Section 6.6 pp. 331-334)

Do on paper and with R: 7.1, 7.2, 7.8, 7.14, 7.17, 7.19

**Mid-Term Exam (Oct 20-23, 2020 exam period)**

7 10/29-30 **Multinomial Logit Choice Model**

**Cluster Analysis**

Read: Ch. 12 and MNL chapter TBD

Guest speaker: Elliot Shin Oblander will discuss causality

Do on paper and with R: Problems 1-3 on page 4-6 of syllabus

8 11/05-06 **Principal Components Analysis/Factor Analysis**

Read: Ch. 9

Do with R: MNL exercise on page 6 of syllabus

9 11/12-13 **Structural Equations Models**

Readings: TBD

Do on paper and with R: 8.6, 8.7, 8.10, 9.1, 9.2, 9.9, 9.17, 9.19

10 11/19-20 **Structural Equations Models**

Read: Ch. 12

11 11/26-27 **Natural Language Processing**

Read: TBD

12 12/03-04 **Course Review and Presentations**

**Final Exam (Dec 15-18 exam period)**

## Analysis of Variance Assignment (See Week 7)

**Problem 1:**

Consider the following experimental design data consisting of one treatment (at four levels) and one covariate:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Case | Y | W1 | W2 | W3 | Covariate  X |
| Level  1 | a  b  c | 2  6  5 | 0  0  0 | 0  0  0 | 0  0  0 | 1  3  2 |
| Level  2 | d  e  f  g | 4  7  9  8 | 1  1  1  1 | 0  0  0  0 | 0  0  0  0 | 1  2  4  4 |
| Level  3 | h  i  j | 6  8  10 | 0  0  0 | 1  1  1 | 0  0  0 | 2  4  8 |
| Level  4 | k  l | 12  14 | 0  0 | 0  0 | 1  1 | 3  10 |

The columns W1, W2, and W3 refer to the dummy-variable coding of the four-level treatment variable. Assume that you wish to test whether the four treatment-level mean responses are significantly different, ignoring the covariate:

1. Write the single-factor ANOVA model for the present problem.
2. Perform an ANOVA analysis for the four-level treatment variable; use an alpha level of 0.05.
3. Perform a regression analysis in which W1, W2, and W3 are dummy-variable regression and compare your results with those of part b.
4. Change the coding of the treatment variable to: level 1 ⇒ 2, 2, 2; level 2 ⇒ 3, 2, 2; level 3 ⇒ 2, 3, 2; level 4 ⇒ 2, 2, 3, and repeat the regression run. Compare your results with those of part c.

**Problem 2:**

Five teaching assistants for the recitation section of a large basic statistics course were rated by their students with respect to overall ability. The ratings on the five-points scale had the following frequencies:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Teaching assistant | | | |  |  |
| Scale value | A | B | C | D | E | Total |
| 1 (highest)  2  3  4  5 (worst) | 20  10  4  0  0 | 12  17  6  1  0 | 14  18  9  2  0 | 10  24  8  4  0 | 16  30  14  4  1 | 72  99  41  11  1 |
| Mean  ∑ | 34  1.53  16.4706 | 36  1.89  21.556 | 43  1.98  30.9767 | 46  2.13  33.2174 | 65  2.14  53.7538 | 224  −  − |

We shall assume that inferences are to be made only to the five instructors; the fixed-effects model should be used.

* 1. Complete the analysis of variance for the hypothesis of equal teaching assistant means.
  2. Use the Bonferroni methods to determine which instructors are different.
  3. What are some other contrasts of the sampler means that are “significant” at the 0.05 level?

Multinomial Logit Model Exercise (See Week 8)

Both the dataset and the R program are posted on Canvas.

The data file contains choices of transportation modes for travelers. A set of independent variables also accompanies these choices. The variables are described below. Your task in this assignment is to:

1. Estimate a variety of multinomial logit models that predict the choice of mode using both alternative-specific variables and individual-specific variables.
2. Your write-up should include:
   * Description of the model that you have selected
   * Description of the computer program
   * An explanation of the coefficients
   * An explanation of the goodness of fit of the model
   * Some suggestions for model improvement.

**Data**

840 observations, 4 choices, 210 respondents (4 rows per respondent)

* Mode = 0/1 for four alternatives: 1=Air, 2=Train, 3=Bus, 4=Car,
* Time = terminal waiting time,
* Invc = Invehicle cost for all stages,
* Invt = Invehicle time for all stages,
* Hinc = Household income in thousands,
* Psize = Travelling party size.

1. Uris 332 has capacity for 14 students, Uris 303 (17 students), and Uris 306 (9 students). [↑](#footnote-ref-1)