Columbia Business School

B9106: Applied Multivariate Statistics
Fall 2022

Professor Kamel Jedidi
518 Uris Hall
Phone: (212) 854-3479
TA: TBA

Office hours: M/W 4:30-5:30PM
e-mail: kj7@gsb.columbia.edu

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 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/.

We will make substantial use of the R programming language in this course. If you are not experienced with programming in R, we expect that you devote time to catching up on R within the first week of class. The study guide (posted on Canvas) we have provided recommends several online courses on DataCamp along with instructions for signing up for free access, as well as alternative resources for learning and reviewing R depending on your background. Please review these resources ASAP and work through as much as you need to feel comfortable.

Course Requirements:

1.	Class participation	5%
2.	Group project	27%
3.	Assignments	18%
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.

Project schedule

1. Proposal due on September 15 (research questions and data source)
2. Schedule a 30min meeting with professor to discuss project in the week of September 20-24
3. Submit preliminary results of descriptive analyses of the data on October 13
4. Submit intermediate data analyses results on November 10
5. Submit final data analyses results on December 1
6. Project presentations on December 6 and 8

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 15 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 15-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 $\checkmark^-$, $\checkmark$, $\checkmark^+$. Please see class schedule for the list of assigned problems (mostly from the required textbook). **These assignments are already posted on Canvas and are due by 11:59pm (ET) on Wednesday in the week they are due.**

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%)

Your active participation in the class benefits everyone involved: it helps you to stay engaged, get your questions answered, and gauge your understanding of the class material; it helps your classmates who most likely have similar questions; and it helps me assess how good of a job I am doing with pacing (e.g., whether I need to slow down and/or revisit a previous topic).
That said, I recognize that everyone has different comfort levels with speaking in class. Accordingly, there are many different ways to "participate" that count towards your participation grade. For example, asking questions on ED Discussions (available on Canvas), answering other students' questions, posting follow-ups, or posting notes (e.g., interesting articles/websites you find related to class or other online resources for learning the course material) also count for course participation. I place particularly high weight on answering other students' questions and posting notes to encourage deeper engagement with the course material.

CLASS SCHEDULE

Week	Date	Topic
1		**Course Introduction: Aspects of Multivariate Analysis** • Read: Ch. 1 and Chapter 2 • If you are not experienced with programming in R, complete the Introduction to R course before class—see study guide posted on Canvas • Recommended videos on Matrix Algebra: 3blue1brown's Essence of Linear Algebra series on YouTube provides an intuitive overview of linear algebra.
2		**Matrix Algebra and Random Vectors** • 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 • Group Project proposal due (research questions and data sources)—One page write-up
3		**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 • Schedule a group meeting with professor this week to discuss project
4		**Multivariate Normal Distribution** • Read: Ch. 4 • Do on paper and with R:1.6, 3.6, 3.10, 3.11, 3.14,3.15
5		**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
6		**Regression** • Read Ch. 7 or Morrison Chapters 1 and 2 (posted on Canvas). Read selectively based on what we discussed in class. • 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 • Preliminary results of descriptive data analysis—Two-page write-up **Regression Assignment due** • Do on paper and with R: 7.1, 7.2, 7.8, 7.14, 7.17, 7.19

----------Mid-Term Exam Period (October 19-22)----------

7 **Analysis of Variance and Causality**

- Read: Ch. 6 (Sect 6.4 pp. 314-320 and Section 6.6 pp. 331-334)
- Guest speaker: Elliot Shin Oblander will discuss causality on 10/25

8 **Multinomial Logit Choice Model**

- Read: MNL chapter posted on Canvas
- This link is a useful book on discrete choice models
- Do on paper and with R: Problems 1-3 on page 4-6 of syllabus

9 **Principal Components Analysis/Factor Analysis**

Read: Ch. 8

- Do with R: MNL exercise on page 6 of syllabus
- Intermediate data analysis results—Two-page write-up

10 **Structural Equations Models (SEM)**

- Read: Ch. 9
- Do on paper and with R: 8.6, 8.7, 8.10, 9.1, 9.2, 9.9, 9.17, 9.19

11 **Structural Equations Models (SEM)**

- Read: SEM chapter posted on Canvas

12 **Cluster Analysis and Natural Language Processing**

- Read: Ch. 12
- Final data analysis results—Two-page write-up

13 **Course Review and Presentations**

----------Final Exam Period (TBD)----------

Analysis of Variance Assignment (See Week 8)

Problem 1:

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

	Case	Y	W_1	W_2	W_3	Covariate X
Level 1	a	2	0	0	0	1
	b	6	0	0	0	3
	c	5	0	0	0	2
Level 2	d	4	1	0	0	1
	e	7	1	0	0	2
	f	9	1	0	0	4
	g	8	1	0	0	4
Level 3	h	6	0	1	0	2
	i	8	0	1	0	4
	j	10	0	1	0	8
Level 4	k	12	0	0	1	3
	l	14	0	0	1	10

The columns W_1, W_2, and W_3 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:

a. Write the single-factor ANOVA model for the present problem.
b. Perform an ANOVA analysis for the four-level treatment variable; use an alpha level of 0.05.
c. Perform a regression analysis in which W_1, W_2, and W_3 are dummy-variable regression and compare your results with those of part b.
d. Change the coding of the treatment variable to: level 1 $\Rightarrow$ 2, 2, 2; level 2 $\Rightarrow$ 3, 2, 2; level 3 $\Rightarrow$ 2, 3, 2; level 4 $\Rightarrow$ 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)	20	12	14	10	16	72
2	10	17	18	24	30	99
3	4	6	9	8	14	41

4	0	1	2	4	4	11
5 (worst)	0	0	0	0	1	1
N_j	34	36	43	46	65	224
Mean	1.53	1.89	1.98	2.13	2.14	–
$\Sigma(x_{ij} - \overline{x_j})^2$	16.4706	21.556	30.9767	33.2174	53.7538	–

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

a) Complete the analysis of variance for the hypothesis of equal teaching assistant means.
b) Use the Bonferroni methods to determine which instructors are different.
c) What are some other contrasts of the sampler means that are "significant" at the 0.05 level?

Multinomial Logit Model Exercise (See Week 9)

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.