Columbia Business School

B9106: Applied Multivariate Statistics
Fall 2023—Tuesday/Thursday 6:07-7:30PM, Geffen 390

Professor Kamel Jedidi
983 Kravis Hall
Phone: (212) 854-3479
Office hours: by request
e-mail: kj7@gsb.columbia.edu

TA: Sanjana Rosario (sr3551@gsb.columbia.edu)
Office hours: by request

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).

Useful Resources:

- Wilkinson, D. J., Multivariate Data Analysis with **R**—Downloadable for free from:
- https://darrenjw.github.io/work/teaching/mas8381/notes14.pdf
- R-Course: http://stat.wharton.upenn.edu/~buja/STAT-470-503-770/CHAPTERS/
- Matrix Algebra for Engineers: https://www.coursera.org/learn/matrix-algebra-engineers
- Other resources (Python, R, Algebra, Probability Theory): https://github.com/columbiamarketing/msmk/wiki/Summer-preparation

Course Prerequisite

R-Programming 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 two weeks of class. The study guide (posted here) 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	30%
3.	Assignments	15%
4.	Midterm and Final Exams	50%

- For all group work, students will evaluate each other's performance. These evaluations will be considered in assigning final grades.
- Assignments will be due on the posted due dates and times (no exceptions). If you and/or your group has a valid reason for not being able to turn in the assignment on time, let the teaching team know in advance.
- Electronics (e.g., laptop, smartphone, tablet) are not allowed in class.
- Class attendance and on time arrival/departure are required. If you arrive to class after 10 minutes, you will be considered absent. You should notify the instructor for class absence/late arrival/early deparure prior to class.

Group Project (30%)

The project requires you to form a group of about 5 students to work 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. Several datasets are posted on Canvas for possible group projects.

Project schedule

1. Proposal due on September 19 (research questions and data sources)
2. Schedule a 30min meeting with professor and TA to discuss project in the week of 09/19
3. Intermediate project report due on October 28. In this report, summarize your group progress on project so far and describe preliminary findings and remaining work to be done.
4. Project presentations on December 5 and 7.
5. Final Report due on December 7. This report consists of about 20 PPT slides (with an appendix) that will contain a summary of your methods, findings, insights, and recommendations. Use speaker notes to add any details to the information you provided on the slides.

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.

As indicated above, you will grade your group members on their effort and cooperation for this project. Individual project grades will be adjusted up or down depending on the evaluations. We will also adjust the project grade based on our perception of your effort on behalf of the group. It is to your benefit to be involved when your group meets with us to discuss the project and to make us aware of your contributions to the group project.

Assignments (15%)

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 Thursday in the week they are due.** Please refrain from searching for solutions to these problems online. This is a violation of the CBS Honor Code.

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 in-class, closed book, closed notes during the school's exam periods. You are allowed to bring four 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	9/5&7	**Course Introduction: Aspects of Multivariate Analysis** • Read: Ch. 1 and Chapter 2 • Do on paper (and check your answers with R where possible): Problems 1.2, 1.4, 1.17, 2.3, 2.4, 2.12 • If you are not experienced with programming in R, complete the Introduction to R course before class—see this study guide (posted here) for details.
2	9/12&14	**Matrix Algebra and Random Vectors** • Do on paper (and check your answers with R where possible): 2.20, 2.22, 2.24, 2.27
3	9/19&21	**Sample Geometry and Random Sampling** • Read: Ch. 3 • Do on paper (and check your answers with R where possible): 2.30, 2.32, 2.34, 2.41 • Group Project proposal due (research questions and data sources)—One page write-up. • Schedule meeting with professor and TA to discuss your group project this week.
4	9/26&28	**Multivariate Normal Distribution** • Read: Ch. 4 • Do on paper (and check your answers with R where possible): 3.6, 3.10, 3.11, 3.14, 3.15 • Project progress report due (preliminary findings and remaining work)—Three-page write-up.
5	10/03&05	**Multivariate Normal Distribution** • Read: Ch. 5 and 6 (Sections 6.1-6.3; 6.7-6.8) • Do on paper (and check your answers with R where possible): 4.2, 4.3, 4.4, 4.5, 4.16
6	10/10&12	**Regression** • Read Ch. 7 or Morrison Chapters 1 and 2 (posted on Canvas). Read selectively based on class discussion. • Do on paper (and check your answers with R): 4.18, 4.19, 4.23, 4.26, 5.1, 5.3, 5.5, 6.5, 6.6, 6.26, 6.27
	10/16	**Regression Assignment due** • Do on paper and check answers with R: 7.1, 7.2, 7.8, 7.14, 7.17, 7.19

----------Mid-Term Exam Period (October 16-20, 2023)----------

7	10/24&26	**Causality (Guest Speaker TBA)** Read: Panel Data Analysis (posted on Canvas)
8	10/31&11/02	**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 1-2 on page 4-6 of syllabus
9	11/09	**Principal Components Analysis/Factor Analysis** Read: Ch. 8 • Do with R: MNL exercise on page 6 of syllabus
10	11/14&16	**Structural Equations Models (SEM)** • Read: Ch. 9 • Do on paper and with R: 8.6, 8.7, 8.10
11	11/21	**Structural Equations Models (SEM)** • Read: SEM chapter posted on Canvas
12	11/28-11/30	**Cluster Analysis and Natural Language Processing** • Read: Ch. 12 • Do on paper and with R: 9.1, 9.2, 9.9, 9.10
13	12/05&07	**Course Review and Presentations**

----------Final Exam Period (December 11-15, 2023)----------

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.