

SOCY 7703: Multivariate Statistics

Wen Fan

Fall 2024

Term: **Fall 2024**

Class location: **O'Neill 245**

Office location: **McGuinn 408**

Email: wen.fan@bc.edu

Instructor: **Dr. Wen Fan**

Class times: **MW (10:30-11:45 am)**

Office hours: **M (12:30-1 pm) or by appt**

Mailbox: **McGuinn 426**

1 Course Description

1.1 Overview

This course is designed for graduate students with a prior background in statistics at the level of *SOC7702: Introduction to Statistics and Data Analysis* or its equivalent. This means that at the time you enroll in this course, I expect you to have a high degree of competence with basic statistics, including measures of central tendency and dispersion, correlation and association, probability and hypothesis testing.

I have several objectives for this class.

- First, I will help you to achieve a deep understanding of ordinary least squares regression (OLS), including its mechanics, assumptions, consequences of violating assumptions, as well as diagnostic tools to evaluate model conformity to those assumptions.
- Second, I will introduce you to methods that relax certain assumptions of OLS—linear probability models and quantile regression models.
- Third, I will give you some of the tools you will need to conduct your analyses independently—issues related to missing data, complex survey design, and mediation analysis.
- Fourth, I will expose you to some of the challenges in conducting panel data analysis and the different approaches you might consider.

1.2 Software

We will use Stata as the primary software, but all of the models and statistics techniques discussed in this class can be estimated in other software such as R. We will spend two lectures in the beginning of the semester covering the basics to get you started up with Stata.

1.3 Course Communication

Please do not hesitate to email me (wen.fan@bc.edu) whenever you have questions. If you don't hear within 24 hours, try again. We can also meet during my office hour. If the current office hour does not work for you, I am happy to meet with you at another mutually convenient time.

After class, I may occasionally communicate with you by email. Please check your email regularly. If you prefer that I use an email address different from the one you have provided to the university, please send me an email indicating your preference.

All course materials will be posted at <http://www.wenfan.co/teaching/multivariate/>. I will announce the password during our first class.

1.4 Texts and Materials

Recommended:

Allison, Paul D. 1999. *Multiple Regression: A Primer*. Thousand Oaks, CA: Pine Forge Press.
Cameron, Colin, and Pravin Trivedi. 2022. *Microeconometrics using Stata, Second Edition*. College Station, TX: Stata Press. [CT]

Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage Publications.

Hamilton, Lawrence C. 2013. *Statistics with Stata, Version 12*. Boston, MA: Cengage Learning. (Useful for those new to Stata.)

Gandrud, Christopher. 2020. *Reproducible Research with R and RStudio, Third Edition*. Boca Raton, FL: CRC Press Book.

Long, J. Scott. 2008. *The Workflow of Data Analysis Using Stata*. College Station, TX: Stata Press. (If you plan to do a lot of data analysis in Stata, this book will save you tons of time while making your work replicable.)

2 Course Policies

2.1 Academic Integrity

I have no tolerance for cheating in any form. Such will earn the student an automatic zero. Academic misconduct includes (but is not limited to): cheating on exams; using material from the Internet without citing it; plagiarizing any part of work done by someone else; and submitting substantially similar work for two courses without consent. It is your responsibility to familiarize yourself with the university's policy on academic integrity: https://www.bc.edu/bc-web/academics/sites/university-catalog/policies-procedures.html#academic_integrity_policies. If you have any questions, always consult with me.

2.2 Withdrawals

If you decide to discontinue the course for any reason, please make an official withdrawal. If you fail to officially withdraw from a class which you are no longer attending, you may receive an F on your permanent transcript.
I will not give incompletes.

2.3 Absences

You are expected to come to class prepared to discuss the readings. If you should be absent for unavoidable reasons, you must check with me first.

2.4 Late Work

Up to one week late: 70% of grade awarded.
Two weeks late: 35% of grade awarded.
Later than that: Forget it.

2.5 Classroom Conduct

Please turn off cell phones or other electronic devices that may disrupt class. If you know you need to leave class early, it is less distracting for me if you let me know before class starts, and then choose a seat close to the exit. Other disruptive behaviors in class include (but are not limited to): using your laptop for purposes other than note taking, engaging in personal conversations, arriving late, making rude and sarcastic comments.

2.6 Disability Statement

I am committed to supporting the learning of all students in my class. If you have already registered with Connors Family Learning Center (learning disabilities and ADHD) or Disability Services (all other disabilities including temporary disabilities) and have your letter of accommodations, please meet with me early in the course to discuss, plan, and implement your accommodations in the course. If you have or think you have a disability but are not registered, the Connors Family Learning Center and Disability Services websites provide information on the registration process.

3 Assessments and Grades

3.1 One Publishable Paper

The main requirement is **one empirical research paper** that applies some of the methods covered in class to a substantive problem in social science. The paper should include problem formulation, linkage to substantive literature, analysis of suitable data, and discussion of findings. Purely methodological papers will not be accepted. Each of you is starting in a different place, so I expect great variation in the sorts of projects that you produce. If you are not sure which research question(s) to pursue, come talk with me as soon as possible—we can together come up with a project idea that suits your research interests and the goal of this course.

The dependent variable for your project must be continuous. Your independent variables (of which there should be at least 3) must include at least one continuous measure and one categorical measure. If your data include a sample weight variable, you should weight all your analysis. If there are multiple sample weights (as is often the case), consult with me if you do not know which one to use.

I highly recommend you submit a two-page proposal that outlines the problem, data, and empirical analyses. The proposal is optional, but I promise feedback on proposals that come in by **October 22**. This proposal should consist of the following elements: (1) introduction to the problem, (2) hypotheses to be tested, and (3) data and methods to use.

In our last two classes (**December 4 and December 9**), every student will give a presentation, showing (preliminary) results. These are not a big part of your grade, but in my opinion are truly essential to your professional success.

The final paper should be no more than 15 pages (double-spaced, excluding references, tables, and figures). It must come with an Appendix which includes all codes used in the analysis of this project. The paper is due on **December 15**.

3.2 Assignments

There will be **three homework assignments**. To obtain credit for homework, you must hand it in on time. Homework assignments are meant to achieve three goals: (1) to provide practice with Stata, (2) to provide practice with the statistical concepts discussed in class, and (3) to provide a chance to demonstrate your mastery of material and highlight areas where more

work is needed. You may work in a group consisting of up to three persons in size, but all write-ups must be done independently. All collaborators should be appropriately cited in your write up.

3.3 Assignment Presentation

Students will be divided into three groups and each group will give **one presentation** of one homework assignment. More details will be discussed at a later point in class.

3.4 Readings

Readings are assigned on the course syllabus. All readings are assumed to be completed *before* each lecture.

3.5 Participation and Collaboration

Build class camaraderie: prepare, participate, help others. Seminars are constructed in interaction. I will try to provide an environment in which everyone feels comfortable participating, but responsibility for the seminar is borne collectively. This means that you must come to class prepared to discuss the readings (preferably with written comments and at least one question of your own) and having considered how the course materials will affect your work. A rough guideline: speak at least once every week.

3.6 Grading

The term paper determines 50 percent of the final grade. Homework assignments, assignment presentation, and class participation determine 30, 10, and 10 percent, respectively.

Letter grade assignment:

A	93%+
A-	90-92.99%
B+	87-89.99%
B	83-86.99%
B-	80-82.99%
C+	77-79.99%
C	73-76.99%
C-	70-72.99%
D	60-69.99%
F	<60%

4 Data

You are welcome to work with any dataset you wish. Some commonly-used social science datasets are listed below. If you wish to use a dataset that is not publicly available, it is especially important that you get started quickly in order to have time to request and receive the data. Some of the datasets listed below are available through **ICPSR**, rather than directly from the study website.

British Household Panel Survey: <https://www.understandingsociety.ac.uk/about/british-househol>

Current Population Survey: <http://www.census.gov/cps>

Early Childhood Longitudinal Study: <http://nces.ed.gov/ecls/index.asp>

Future of Families & Child Wellbeing Study: <https://ffcws.princeton.edu>

General Social Survey: <http://www3.norc.org/gss+website>
 German Socio-Economic Panel Study: <http://www.diw.de/en/soep>
 Health and Retirement Study: <http://hrsonline.isr.umich.edu>
 High School & Beyond: <http://nces.ed.gov/surveys/hsb>
 International Social Survey Programme: <http://www.issp.org/index.php>
 National Longitudinal Study of Adolescent Health: <http://www.cpc.unc.edu/projects/addhealth>
 National Longitudinal Surveys: <http://www.bls.gov/nls>
 National Social Life, Health, and Aging Project: <http://www.norc.org/Research/Projects/Pages/national-social-life-health-and-aging-project.aspx>
 Panel Study of Income Dynamics: <https://psidonline.isr.umich.edu/default.aspx>
 Survey of Income and Program Participation: <http://www.census.gov/sipp>
 Wisconsin Longitudinal Study: <http://www.ssc.wisc.edu/wlsresearch>

5 Readings

August 26: Welcome and introduction to Stata (I)

- **Lecture:** introduction; course logistics; requirements; basic Stata commands
- **Readings:** [ICPSR User's Manual](#)

August 28: Introduction to Stata (II)

- **Lecture:** basic Stata commands

September 4: Review of Descriptive Statistics, Correlation, and ANOVA

- **Lecture:** mean; standard deviation; correlation; summarize; corr; anova

September 9: OLS: estimation

- **Lecture:** `reg`; `estimates store`; `estimates table`; `outreg2`; `lm`
- **Readings:** CT 3.3-3.4

September 11: OLS: assumptions (I)

- **Lecture:** mechanics and assumptions of OLS; LINE

September 16: OLS: assumptions (II)

- **Lecture:** mechanics and assumptions of OLS; LINE

September 18: OLS: model fit

- **Lecture:** standardized coefficients; hypothesis testing; heteroskedasticity; test; predict; rvfplot; multicollinearity
 - **Readings:** CT 3.5.1-3.5.4
-

September 23 and 25: No Class This Week

October 2: OLS: diagnostics

- **Lecture:** outliers; influence; vif; dfbeta
-

October 7 and 9: Transformations and interaction terms (I)

- **Lecture:** transformations; interaction (categorical by categorical)
 - **Readings:** [Decomposing, Probing, and Plotting Interactions in Stata](#)
-

October 16: Interaction terms (II)

- **Lecture:** interaction (categorical by continuous; continuous by continuous)
-

October 21: OLS models better presented

- **Lecture:** prediction; plots

October 23: OLS summary

- **Lecture:** review; assignment 1 presentation
-

October 28: Linear probability models

- **Lecture:** linear discriminant model; predict_ldm

October 30: Quantile regression

- **Lecture:** modeling mean versus quantile; qreg; bsqreg
-

November 4 and 6: Missing, missing, missing

- **Lecture:** missing data; listwise deletion; MAR; multiple imputation
 - **Readings:** Acock, Alan C. 2012. "What to do about missing values." In Harris Cooper (Ed.), *APA Handbook of Research Methods in Psychology, Vol. 3. Data Analysis and Research Publication* (pp. 27-50). Washington, DC: American Psychological Association.
-

November 11 and 13: To weight or not?

- **Lecture:** weight; svyset; svy:; assignment 2 presentation
 - **Readings:** [Survey data analysis in Stata](#)
-

November 18 and 20: Mediation

- **Lecture:** mediation; Baron and Kenny approach; decomposition; total/direct/indirect effects
-

November 25 and 27: No Class This Week

December 2: Panel Analysis

- **Lecture:** fixed effects; random effects; lagged dependent variables; assignment 3 presentation

December 4 and 9: Student Presentations