

SOCY 7704: Regression Models For Categorical Data*

Wen Fan

Spring 2024

Term: **Spring 2024**

Class location: **O'Neill 245**

Office location: **McGuinn 408**

Email: **wen.fan@bc.edu**

Instructor: **Dr. Wen Fan**

Class times: **Tu (3-5:30 pm)**

Office hours: **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 *SOC7703: Multivariate Statistics* 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, analysis of variance (ANOVA), and multivariate OLS regression. For those of you in the sociology department, this course provides a foundation for the "Advanced Quantitative Methods" area exam.

I have several objectives for this class.

- First, I will help you achieve a deeper and more nuanced 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 for analyzing categorical outcomes:
 - Dichotomous outcomes: binary logit and probit regression models
 - Multi-categorical (nominal) outcomes: multinomial logit regression models
 - Ordinal outcomes: ordered logit regression models
 - Count outcomes: Poisson, negative binomial, hurdle, zero-inflated, and zero-truncated regression models
 - Duration outcomes: survival analysis
- Third, I will give you some of the tools you will need to read the methodological literature independently—a basic familiarity with matrix algebra and maximum likelihood. Issues related to missing data and complex survey design will also be covered.

*This syllabus was heavily influenced by Eric Grodsky's 2009 syllabus for Advanced Social Statistics and Christopher Uggen's 2013 syllabus for Event History & Panel Methods, both at the University of Minnesota, as well as Natasha Sarkisian's and Sara Moorman's syllabi for this course at Boston College.

- Fourth, I will expose you to some of the challenges in conducting mediation analysis for categorical outcomes and the different approaches you might consider.
- Throughout the course, we will work through empirical pieces by some top social science researchers. As you develop your own research, it is useful to see how others have translated propositions into testable hypotheses, devised appropriate methodologies to test them, and presented the results to diverse audiences. This process will sharpen your written and oral research presentation skills.

1.2 Software

We will be using both Stata and R as the primary software tools. Familiarity with Stata is required for this course, and part of the first lecture will be dedicated to covering the basics of R to get you started.

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. I am also happy to meet with you at a mutually convenient time.

After class, I will 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, send me an email from that account indicating your preference.

All course materials will be posted at <http://www.wenfand.co/teaching/categorical/>. I will announce the password in our first class.

1.4 Texts and Materials

Required:

Long, J. Scott, and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata, 3rd Edition*. College Station, TX: Stata Press. [LF]

We will also read some research articles throughout the semester, but I've limited the number of required readings to just a few—emphasizing solid and accessible applications of the techniques we discuss, rather than the brilliant statistical innovations that brought us these techniques.

Recommended:

Agresti, Alan. 2012. *Categorical Data Analysis, 3rd Edition*. Hoboken, NJ: Wiley. (Especially useful if you are interested in the underlying mathematical details.)

Cameron, Colin, and Pravin Trivedi. 2022. *Microeconometrics Using Stata, Second Edition*. College Station, TX: Stata Press. (We will read several chapters from this well-written book.) [CT]

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

Gandrud, Christopher. 2020. *Reproducible Research with R and RStudio, 3rd Edition*. Boca Raton, FL: CRC Press Book.

Long, J. Scott. 2009. *The Workflow of Data Analysis Using Stata*. College Station, TX: Stata Press.

Meyer, David, and Michael Friendly. 2015. *Discrete Data Analysis with R*. Chapman and Hall/CRC.

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 (e.g., an A- becomes a C+).
Two weeks late: 35% of grade awarded (e.g., an A- becomes a D).
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. 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

Boston College is committed to providing reasonable accommodations and integrated access for students with disabilities to all available academic, social, and recreational programs and activities. Appropriate support and referral services are provided by the Disability Services Office, which serves students with hearing, visual, mobility, medical, and psychiatric disabilities. If you are a student with a documented disability seeking reasonable accommodations in this course, please contact Kathy Duggan, (617) 552-8093, dugganka@bc.edu, at the Connors Family Learning Center regarding learning disabilities and ADHD, or Paulette Durrett, (617) 552-3470, paulette.durrett@bc.edu, in the Disability Services Office regarding all other types of disabilities, including temporary disabilities. Advance notice and appropriate documentation are required for accommodations. If you are unsure whether or not Disability Services are appropriate for your needs, please make an appointment with Disability Services for a consultation.

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. Do not waste your time on a paper that will only be used to meet course requirements. The seminar project should advance your own career and research agenda. If you are uncertain about which research question(s) to pursue, please come talk with me as soon as possible. Together, we can develop a project idea that aligns with your research interests and the objectives of this course.

The dependent variable for your project must be categorical (binary, nominal, ordinal, count, or duration). Your independent variables (of which there should be at least 5) 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 **March 28**. This proposal should consist of the following elements: (1) introduction to the problem, (2) hypotheses to be tested, (3) data and methods to use, and (4) expected results.

In our last class (**April 30**), 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 20 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 **May 8**.

3.2 Assignments

There will be **three homework assignments**. To obtain credit for homework, you must hand it in on time and be prepared to present solutions orally in class. Homework assignments are meant to achieve three goals: (1) to provide practice with Stata/R, (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 Readings and Case Studies

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

Case studies are empirical pieces to help you understand how to apply the methods we learn in class. While most are optional readings and will not be discussed in class, there are two exceptions. These two papers, scheduled for detailed examination on **February 6** and **March 19**, are required readings. For these sessions, the class will be divided into pairs to assume the roles of “reviewer” and “author.” Reviewers will critique the paper’s contribution, focusing particularly on methodologies or method-related substantive issues. Authors will respond to these critiques, defending their positions and, if the critiques are valid, suggesting the types of analyses they would perform to address the issues. Regardless of your assigned

role, please ensure that your write-up and in-class presentation are concise and to the point. We will discuss more details about this activity later in the course.

3.4 Software/AI Tool Presentation

Each student (or pairs of students) is expected to give a brief presentation (10 minutes) on an Stata/R package or some AI tools. Possible options are listed below:

- knitr (R)
- ggeffects (R)
- sjPlot (R)
- dyndoc or markstat (Stata)
- Strategies for integrating ChatGPT and Stata
- Strategies for integrating ChatGPT and R

Please keep in mind that these are just suggestions. If there is a package or AI tool you are interested in exploring and introducing to your peers, please let me know!

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, case studies, software presentation, and class participation determine 30, 10, 5, and 5 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.iser.essex.ac.uk/bhps>
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>
Los Angeles Family and Neighborhood Survey: <http://lasurvey.rand.org>
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 Class Schedule Overview

The following schedule is tentative. Please keep yourself updated by checking email regularly.

	Week	Lecture	Due
I	1/23	Introduction (course, R)	
	1/30	OLS: assumptions, estimation, and model fit	
	2/6	OLS: diagnostics and interaction	Case study 1 (2/6)
II	2/13	Missing data and survey design	
	2/27	Maximum likelihood and testing methods	
III	3/12	Binary logit and probit	Assignment 1
	3/19	Binary logit and probit	Case study 1 (3/19)
	3/26	Multinomial logit	Proposal (optional) (3/28)
	4/2	Ordered logit	Assignment 2
	4/9	Count models	
IV	4/23	Mediation	Assignment 3
	4/30	Presentations	Project (5/8)

6 Readings

January 23: Welcome, math review, and R introduction

- **Lecture:** introduction; course logistics; requirements; math review

- **Readings:**
 - [Stata cheat sheets](#)
 - [Stata to R cheat sheets](#)
-

January 30: OLS models: assumptions, estimation, and model fit

- **Lecture:** mechanics and assumptions of OLS; LINE; standardized coefficients; hypothesis testing; heteroskedasticity; multicollinearity
 - **Readings:** CT 3.4-3.5.12
-

February 6: OLS models: diagnostics, interaction terms, and presentation

- **Lecture:** outliers; influential; interaction; plots
 - **Readings:**
 - LF Chapter 4 (skip 4.5; mentally replace `logit` with `reg` while reading)
 - [Decomposing, Probing, and Plotting Interactions in Stata](#)
 - **Case study (required):** Donato, Katharine M, Gabriela León-Pérez, Kenneth A Wallston, and Sunil Kripalani. 2018. "Something old, something new: When gender matters in the relationship between social support and health." *Journal of Health and Social Behavior* 59(3): 352-370.
-

February 13: Missing data and survey weights

- **Lecture:** missing data; listwise deletion; MAR; multiple imputation; weight
 - **Readings:**
 - LF 3.1.6; LF 3.1.7
 - 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.
 - [Survey data analysis in Stata](#)
-

February 20: Case study 1

- **Lecture:** case study
-

February 27: Maximum likelihood estimation and testing methods

- **Lecture:** maximum likelihood; Wald test; LR test; score test
 - **Readings:** LF 3.1 (read until 3.1.4); LF 3.2; CT 11.3-11.5.2
-

March 12: Logit and probit: setup, estimation, and interpretation of coefficients

- **Lecture:** logit; probit; odds; latent variable; odds ratios; predicted probabilities; marginal effects; MER vs. MEM vs. AME
 - **Readings:**
 - LF 5.1-5.2; LF 5.4.1; LF 6.1-6.3 (skip 6.1.2, 6.2.6, 6.3.3; read 6.4 and 6.5 if you have time)
 - Norton, Edward C., Bryan E. Dowd, and Matthew L. Maciejewski. 2018. "Odds ratios—Current best practice and use." *JAMA* 320(1): 84-85.
-

March 19: Logit and probit: hypothesis testing, model fit, unobserved heterogeneity and presentation

- **Lecture:** Wald test; LR test; internal and external assessment of model fit (R^2 -type; AIC/BIC); perfect prediction; heterogeneous choice models
 - **Readings:**
 - LF 3.3; LF 5.2.3; LF 5.3; LF 5.5; LF 6.6 (skip 6.6.6)
 - Williams, Richard. 2010. "Fitting heterogeneous choice models with oglm." *Stata Journal* 10(4): 540-567.
 - Mustillo, Sarah A., Omar A. Lizardo, and Rory M. McVeigh. 2018. "Editors' comment: A few guidelines for quantitative submissions." *American Sociological Review* 83(6): 1281-1283.
 - **Case study (required):** Mittleman, Joel. 2022. "Intersecting the academic gender gap: The education of lesbian, gay, and bisexual America." *American Sociological Review* 87(2): 303-335.
-

March 26: Multinomial logit model

- **Lecture:** nominal outcomes; independence of irrelevant alternatives; predicted probabilities; marginal effects
 - **Readings:** LF 8.1-8.8 (skip 8.8.1); LF 8.10-8.11
 - **Case study (optional):** Scarborough, William J., Joanna R. Pepin, Danny L. Lambouths, III, Ronald Kwon, and Ronaldo Monasterio. 2021. "The intersection of racial and gender attitudes, 1977 through 2018." *American Sociological Review* 86(5): 823-855.
-

April 2: Ordered logit model

- **Lecture:** ordinal outcomes; parallel regression assumption; predicted probabilities; marginal effects
 - **Readings:** LF 7.1-7.12 (skip 7.5 and 7.12.1); LF 7.14-7.15
 - **Case study (optional):** Doan, Long, and Matthew K. Grace. 2022. "Factors affecting public opinion on the denial of healthcare to transgender persons." *American Sociological Review* 87(2): 275-302.
-

April 9: Count models

- **Lecture:** Poisson distribution; exposure time; negative binomial distribution; zero-truncated count models; zero-inflated count models; hurdle regression
 - **Readings:** LF 9.1-9.7
 - **Case study (optional):** Correll, Shelley J., Katherine R. Weisshaar, Alison T. Wynn, and JoAnne Delfino Wehner. 2020. "Inside the black box of organizational life: The gendered language of performance assessment." *American Sociological Review* 85(6): 1022-1050.
-

April 23: Duration models and mediation

- **Lecture:** censor; discrete-time event history analysis; survival analysis; Baron and Kenny; decomposition; total/direct/indirect effects; Sobel-Goodman tests; bootstrap; KHB method
 - **Readings:**
 - Bloome, Deirdre, and Shannon Ang. 2020. "Marriage and union formation in the United States: Recent trends across racial groups and economic backgrounds." *Demography* 57: 1753-1786.
 - Breen, Richard, Kristian Bernt Karlson, and Anders Holm. 2018. "Interpreting and understanding logits, probits, and other nonlinear probability models." *Annual Review of Sociology* 44: 39-54.
 - (Optional) Kohler, Ulrich, Kristian Bernt Karlson, and Anders Holm. 2011. "Comparing coefficients of nested nonlinear probability models." *The Stata Journal* 11(3): 420-438.
-

April 30: Student presentations