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 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 bivariate OLS regression. If you are sketchy on any of this material (we’ll have a math review for you to determine) you should reconsider taking this class. For those of you in the Sociology Department, the course can also provide a foundation for the "Advanced Quantitative Methods" area examination.

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, as well as consequences of violating those assumptions. I will also cover diagnostic tools to evaluate model conformity to those assumptions. (Weeks 2-4)

• Second, 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. (Week 5)

• Third, I will introduce you to methods for analyzing categorical outcomes:

  – Dichotomous outcomes: binary logit and probit regression (Weeks 6-7)
  – Multi-categorical (nominal) outcomes: multinomial regression (Weeks 8)
  – Ordinal: ordered logit and probit regression (Week 9)
  – Count: Poisson, negative binomial, and zero-inflated logit regression (Weeks 10-11)

---

*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 you will face in engaging in causal inference using non-experimental data and some of the solutions you might consider if you are using quantitative methods. Issues related to mediation analysis, missing data, and complex survey design will also be covered. (Week 12-14)

Throughout the course, we will work through empirical pieces by some top sociological 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

Almost all of the models discussed in this class can be estimated in Stata. While Stata includes commands for estimating these models, we will also use a set of ado files that researchers have written to make it easier to interpret these models. If you are not familiar with Stata, don’t worry. I will show you everything you need in our second lecture. Other software will be introduced (R, SAS, and SPSS) if necessary.

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 12 hours during the week, try again. You can also visit me 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.

I will communicate with you by email to let you know about the availability of lecture notes, assignments and other material. 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.

1.4 Texts and Materials

Required:

We will also read some challenging 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:
  Cameron, Colin, and Pravin Trivedi. 2009. *Microeconometrics using Stata*. College Station, TX: Stata Press. (We’ll read several chapters from this extremely well-written book.) [CT]
  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, this book will save you a lot of time and make your work replicable.)
1.5 Course Feedback

It is my first time teaching this course; any feedback is greatly appreciated. In particular, I will conduct a course evaluation in the end of each Thursday class. I will hand over a review form to you consisting of 3-4 short questions, such as whether the articles you read for that week was helpful or not, whether a particular technique we covered was easy to understand or not, anything you might need further explanation, etc. This review is conducted for my own improvement and for the benefit of future students; it is anonymous and won’t be used against you:)

2 Course Policies

2.1 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.

2.2 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 before I consider allowing any make-up work.

2.3 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!

2.4 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.

2.5 Cheating and Other Forms of Dishonesty

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: www.bc.edu/offices/stserv/academic/integrity.html. If you have any questions, always consult with me.

3 Assessments and Grades

3.1 One Publishable Paper

The main requirement is an empirical research paper that applies some of the methods covered in class to a substantive problem in social science. The paper should include problem formu-
lation, linkage to substantive literature, analysis of suitable data, and discussion of findings. Purely methodological papers will not be accepted. Every student 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 (a lesson I learned the hard way...). The seminar project should advance your own career and research agenda.

The dependent variable for your project must be categorical (nominal, ordinal, or count). You independent variables (of which there should be at least 5) must include at least one interval/ratio 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.

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 20. This proposal should consist of the following elements: (1) introduction to the problem, (2) hypothesis to be tested, (3) methods to use, and (4) expected results.

The final paper should be no more than 25 pages (excluding references, tables, and figures). It must come with an Appendix which includes all Stata code (or codes from whatever software you use) used in the analysis of this project. The paper is due on December 12.

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 results: (1) provide practice with Stata, (2) provide practice with the statistical concepts discussed in class, and (3) 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; any Stata code from a given source should also be cited as you would a journal article.

3.3 Readings

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

3.4 Participation and Collaboration

Build class camaraderie: prepare, participate, help others. Seminars are constructed in interaction. I will provide a brief setup at each meeting and 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 to have considered how the course materials will affect your work. A rough guideline: speak at least once during each course meeting.

3.5 Grading

The term paper determines 60 percent of the final grade (detailed rubrics will be distributed in late October or early November). Homework and class participation determine 30 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 <59.99%

4 Class Schedule Overview

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

<table>
<thead>
<tr>
<th>Week</th>
<th>Lecture</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>9/1</td>
<td>Introduction (course, Stata)</td>
</tr>
<tr>
<td></td>
<td>9/8</td>
<td>OLS: mechanics and assumptions</td>
</tr>
<tr>
<td></td>
<td>9/15</td>
<td>OLS: diagnostics</td>
</tr>
<tr>
<td></td>
<td>9/22</td>
<td>Interaction terms</td>
</tr>
<tr>
<td>II</td>
<td>9/29</td>
<td>Matrix algebra and maximum likelihood</td>
</tr>
<tr>
<td></td>
<td>10/6</td>
<td>Binary logit and probit (I)</td>
</tr>
<tr>
<td></td>
<td>10/13</td>
<td>Binary logit and probit (II)</td>
</tr>
<tr>
<td></td>
<td>10/20</td>
<td>Multinomial logit</td>
</tr>
<tr>
<td></td>
<td>10/27</td>
<td>Ordered logit</td>
</tr>
<tr>
<td></td>
<td>11/3</td>
<td>Poisson and negative binomial</td>
</tr>
<tr>
<td></td>
<td>11/10</td>
<td>Zero-inflated count models</td>
</tr>
<tr>
<td>III</td>
<td>11/17</td>
<td>Missing data and survey design</td>
</tr>
<tr>
<td></td>
<td>11/24</td>
<td>Project</td>
</tr>
<tr>
<td></td>
<td>12/1</td>
<td>Causal inference and mediation</td>
</tr>
<tr>
<td></td>
<td>12/8</td>
<td>Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assignment 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proposal (optional)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assignment 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assignment 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Project (12/12)</td>
</tr>
</tbody>
</table>

5 Readings

September 1: Welcome and math review

- **Lecture**: introduction; course logic; requirements; math review
- **Readings**: ICPSR User’s Manual

September 3: Introduction to Stata

- **Lecture**: basic Stata commands
- **Readings**: LF Chapter 2; LF 3.1.5
September 8: OLS assumptions
- **Lecture**: mechanics and assumptions of OLS; identification; LINE
- **Readings**: Allison Chapters 3, 5, and 6

September 10: OLS in Stata
- **Lecture**: `reg`, `test`, `estimates store`, `estimates table`, `outreg2`
- **Readings**: Allison Chapter 4; CT Chapters 3.3-3.4

September 15: OLS diagnostics (I)
- **Lecture**: outliers, influential, heteroskedasticity, `predict`, `rvfplot`, `dfbeta`
- **Readings**: CT 3.5.1-3.5.3

September 17: OLS diagnostics (II) and interaction
- **Lecture**: multicollinearity, interaction, `vif`
- **Readings**: Allison Chapters 7-8

September 22: OLS models better presented
- **Lecture**: prediction, plots
- **Readings**: LF Chapter 4 (skip 4.5; mentally replace `logit` with `reg` while reading)

September 24: OLS summary
- **Lecture**: review
- **Readings**: Allison Chapter 9

September 29: Maximum likelihood estimation
- **Lecture**: maximum likelihood, BLUE
- **Readings**: LF 3.1 (read until 3.1.4)

October 1: Testing methods
- **Lecture**: Wald test, LR test
- **Readings**: LF 3.2, CT 12.2-12.4
October 6: Linear probability model
- Lecture: linear probability model; parametric models; latent variable
- Readings: begin LF Chapter 5

October 8: Logit and probit: estimation, testing, and fit
- Lecture: logit; probit; internal and external assessment of model fit ($R^2$-type; AIC/BIC)
- Readings: LF 3.3; LF Chapter 5 (skip 5.5.3)

October 13: Logit and probit: interpretation and presentation
- Lecture: marginal effects; MER vs. MEM vs. AME; predicted probabilities
- Readings: LF Chapter 6 (skip 6.1.2, 6.2.6, 6.3.3, 6.5, and 6.6.6)

October 15: Logit and probit model summary
- Lecture: review

October 20: Multinomial logit model: estimation, testing, and fit
- Lecture: nominal outcomes; independence of irrelevant alternatives
- Readings: LF 8.1-8.5

October 22: Multinomial logit model: interpretation and presentation
- Lecture: marginal effects; predicted probabilities
- Readings: LF 8.6-8.11 (skip 8.8.1 and 8.9.1)

October 27: Ordered logit model: estimation, testing, and fit
- Lecture: ordinal outcomes; parallel regression assumption
- Readings: LF 7.1-7.6 (skip 7.5)
October 29: Ordered logit model: interpretation and presentation

- **Lecture**: marginal effects; predicted probabilities
- **Readings**: LF 7.7-7.15

November 3: Poisson regression model

- **Lecture**: Poisson distribution; exposure time
- **Readings**: LF 9.1-9.2

November 5: Negative binomial regression model

- **Lecture**: negative binomial distribution
- **Readings**: LF 9.3

November 10: How to handle zero (I)

- **Lecture**: zero-truncated count models; hurdle regression model
- **Readings**: LF 9.4-9.5

November 12: How to handle zero (II)

- **Lecture**: zero-inflated count models
- **Readings**: LF 9.6-9.7
November 17: Missing, missing, missing

- **Lecture**: missing data; listwise deletion; MAR; multiple imputation
- **Readings**:
  - LF 3.1.6

November 19: To weight or not?

- **Lecture**: weight; `svyset`; `svy:
- **Readings**:
  - LF 3.1.7

November 24: Discussion

- **Lecture**: Q & A

December 1: Causal inference

- **Lecture**: research design; counterfactual; experiment; IV; regression discontinuity; propensity score
- **Readings**:
December 3: Mediation

- **Lecture**: Baron and Kenny; decomposition; total/direct/indirect effects; Sobel-Goodman tests; bootstrap


December 8: Review