

SC704: TOPICS IN MULTIVARIATE STATISTICS

Instructor: Natasha Sarkisian
Email: natasha@sarkisian.net
Office phone: (617) 552-0495
Home phone: (617) 755-3178
Mailbox: McGuinn 426

Office: McGuinn 417
Office hours: By appointment
Class time: Mondays 12-2:30 PM
Class location: O'Neill 245
Webpage: <http://www.sarkisian.net/sc704/>

COURSE DESCRIPTION *

This applied course is designed for graduate students with a prior background in statistics at the level of SC703: Multivariate Statistics (or its equivalent). This means that students should have considerable experience with ordinary-least-squares (OLS) regression: I assume you have an understanding of multiple OLS regression and an ability to conduct such analyses using some statistical software (e.g., SPSS, SAS, Stata, etc.). Major topics of the course will include OLS regression diagnostics, binary, ordered, and multinomial logistic regression, models for the analysis of count data (e.g., Poisson and negative binomial regression), treatment of missing data, and the analysis of clustered and stratified samples.

Dependent variable		Type of analysis
Nominal	Binary	Logistic regression (or probit)
	Multi-category	Multinomial logistic regression
Ordinal		Ordered logistic regression (or ordered probit)
Interval	Continuous	OLS regression
	Count	Poisson and negative binomial regression

We will be using Stata for all the analyses throughout the course. No previous Stata experience is necessary: One of the goals of the course is to familiarize you with this statistical software package. I will provide an introduction to Stata in the beginning of the course and guide you throughout the course. Your main textbook (Long and Freese 2003) also provides an introduction to Stata and a step-by-step guide for many analyses. For the topics not covered in this book, I will provide handouts covering relevant Stata commands.

The goals of the course are to develop the skills necessary to critically evaluate contemporary social research using advanced quantitative methods and to identify an appropriate technique, estimate models, and interpret results for independent research. The course will be applied in the sense that we will focus on estimating models and interpreting the results, rather than on understanding in detail the mathematics behind the techniques. I hope that the course will provide you with a solid foundation in advanced quantitative methods, which is in high demand in many fields, both in and out of academia. For those of you in the Sociology Department, the course can also provide a foundation for the "Advanced Quantitative Methods" area examination.

* This syllabus draws upon ideas presented in syllabi by a number of people, including Robert Kunovich, John Williamson, Joya Misra, and Doug Anderton.

COURSE POLICIES

For each topic in the course, I will give an introductory lecture focusing on the reasoning behind the technique, and provide a review of the syntax used to do analyses and the output generated by Stata. You will then get a chance to practice conducting the analyses and interpreting the results. We will discuss and critically evaluate published research based on the various techniques. The course is based on an interactive relationship between the instructor and students, as well as on collaboration among the students. You are strongly encouraged to ask questions and discuss the material in class. I also encourage collaboration among the students. Please feel free to help each other when running analyses for assignments. However, everyone must turn in their own report and statistical output.

I also would like to stress that you are always welcome to come and see me with any additional questions. Frequently you will just be able to find me in my office, so please stop by when you have a question. If I am not there, email is the best way to get in touch with me – I check my email very often. Email is the best way to get a quick question answered or to set up an appointment to discuss something at length. You are also welcome to call me either in my office or at home (any time between 9 AM and 10 PM); however, be prepared to leave your name and number if I am not available to pick up the phone. Also, please check our course website regularly: various course materials (assignments, handouts, etc.) will be posted there regularly. And make sure to check your email, too – from time to time I may send some announcements.

Finally, a note on feedback. I would like to know how I could make this course experience as useful and interesting as possible. Therefore, every class in the end of class I will ask you to submit a sheet of paper with the date and at least one sentence of reaction to that class meeting, indicating what you learned, or something you liked or did not like, found interesting or controversial, found clear or too simplistic, or found confusing and in need of further (or better) explanation. You may also submit comments on the course in general.

REQUIRED MATERIALS

Required books:

1. Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. College Station, TX: Stata Press.
2. Berry, William D. 1993. *Understanding Regression Assumptions*. (Quantitative Applications in the Social Sciences, 07-092). Thousand Oaks, CA: Sage Publications.
3. Fox, John. R. 1991. *Regression Diagnostics*. (Quantitative Applications in the Social Sciences, 07-079). Thousand Oaks, CA: Sage Publications.
4. Pampel, Fred C. 2000. *Logistic Regression: A Primer*. (Quantitative Applications in the Social Sciences, 07-132). Thousand Oaks, CA: Sage Publications
5. Borooah, Vani K. 2002. *Logit and Probit: Ordered and Multinomial Models*. (Quantitative Applications in the Social Sciences, 07-138). Thousand Oaks, CA: Sage Publications.
6. Allison, Paul D. 2001. *Missing Data*. (Quantitative Applications in the Social Sciences, 07-136). Thousand Oaks, CA: Sage Publications.

Other required readings:

Other required readings (listed below in the course outline) will be available on electronic reserve in the library: see <http://www.bc.edu/libraries/services/reserves/>

Additional resources:

The basic material included in required readings will provide you a basic exposure to each analytic technique. The advanced material (found in Additional resources) will prove particularly useful to those of you who will be using that specific technique in your own research or future jobs in quantitative social science research. These recommended readings are available either online or from me upon request.

COURSE REQUIREMENTS AND GRADING

There will be four assignments in this course, each worth 25% of your grade. These assignments will involve selecting a research question and variables, running analyses, and writing up the results like you would for a journal publication (including introduction, data and methods, and results sections). I will provide data, although I am open to you using your own data if they are appropriate for the technique (see me in advance).

I will grade all assignments twice, provided that you submit them on time. If you turn in the assignment by the due date, I will comment on it, assign a temporary grade, and return it to you. We will also discuss in class the common problems and mistakes in that assignment. You will then resubmit a revised draft. If you are satisfied with your temporary grade, you do not need to revise or even reprint the assignment – just resubmit it to me with my comments on it.

This system will allow you to push yourself beyond your comfort level without worrying how it will affect your grade. For example, you might try to interpret some of the results not required for the assignment or you might decide to present the results in a more meaningful and perhaps less conventional way. In class I will generally focus on the basic concepts and the most essential material. However, most assignments will give you an opportunity (and an incentive) to go beyond the basic material to demonstrate that you understand and can use some of the more advanced options and techniques as well. My intention is that these assignments will assist in the completion of the advanced quantitative methods area exam in sociology and/or will facilitate your own independent research projects. However, I ask that you only submit assignments for review if you can do so by the due date. I will not provide feedback if the assignments are submitted late.

COURSE OUTLINE.

September 12. Introduction to the course and to Stata.

Required readings

Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. Chapters 1 and 2. College Station, TX: Stata Press.

Additional resources

Resources to Help You Learn and Use Stata. <http://www.ats.ucla.edu/stat/stata/>

Stata Reference Manuals. College Station, TX: Stata Press.

Huff, Darrell. 1954. *How to Lie with Statistics*. New York: W.W. Norton & Company.

Abelson, Robert P. 1995. *Statistics as Principled Argument*. Hillsdale, NJ: Erlbaum.

Helberg, Clay. 1995. *Pitfalls of Data Analysis (or How to Avoid Lies and Damned Lies)*.

<http://www.execpc.com/~helberg/pitfalls/>.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30: 666-687.

Cohen, J. 1990. "Things I Have Learned (So Far)." *American Psychologist*, 45: 1304-1312.

Cohen, J. (1994). "The Earth is Round (p < .05)." *American Psychologist*, 49: 997-1003.

September 19. OLS regression assumptions and diagnostics

Required readings

Berry, William D. 1993. *Understanding Regression Assumptions*. (Quantitative Applications in the Social Sciences, 07-092). Thousand Oaks, CA: Sage Publications.

Additional resources

Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Models*. Chapters 1-13. Thousand Oaks, CA: Sage.

Ryan, Thomas P. 1996. *Modern Regression Methods*. Chapters 1-8. New York: Wiley.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Weisberg, Sanford. 1985. *Applied Linear Regression. (Second Edition)*. New York: Wiley.

Chatterjee, Sampreet, Ali D. Hadi, and Bertram Price. 2000. *Regression Analysis by Example. (Third Edition)*. New York: Wiley.

September 26. OLS regression assumptions and diagnostics (continued)

Required readings

Fox, John. R. 1991. *Regression Diagnostics*. (Quantitative Applications in the Social Sciences, 07-079). Thousand Oaks, CA: Sage Publications.

Kenworthy, Lane, and Melissa Malami. 1999. "Gender Inequality in Political Representation: A Worldwide Comparative Analysis." *Social Forces*, 78: 235-268. RESERVE.

October 3. Binary logistic regression.

Assignment 1 (OLS regression) due October 6, 5PM

Required readings

Pampel, Fred C. 2000. *Logistic Regression: A Primer*. (Quantitative Applications in the Social Sciences, 07-132). Thousand Oaks, CA: Sage Publications

Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. Chapters 3 and 4. College Station, TX: Stata Press.

Additional resources

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Chapters 3 and 4. Thousand Oaks, CA: Sage Publications.
- Liao, Tim Futing. 1994. *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*. (Quantitative Applications in the Social Sciences, 07-101). Thousand Oaks, CA: Sage Publications.
- Hosmer, D.W. and Lemeshow, Stanley. 1989. *Applied Logistic Regression*. New York: Wiley.
- Ryan, Thomas P. *Modern Regression Methods*. Chapter 9. New York: Wiley.
- Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Models*. Chapter 15. Thousand Oaks, CA: Sage.
- Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients Across Groups." *Sociological Methods and Research*, 28: 186-208.
- Eliason, Scott R. *Maximum Likelihood Estimation: Logic and Practice*. (Quantitative Applications in the Social Sciences, 07-096). Thousand Oaks, CA: Sage Publications.

October 10. Columbus Day. NO CLASS.

October 17. Binary logistic regression (continued)

Assignment 1 (OLS regression) revised due October 20, 5PM

Required readings

- Alba, Richard, John Logan, Amy Lutz, and Brian Stults. 2002. "Only English by the Third Generation? Loss and Preservation of the Mother Tongue among the Grandchildren of Contemporary Immigrants." *Demography*, 39: 467-484. RESERVE.

October 24. Ordered logistic regression

Assignment 2 (Binary logit) due October 27, 5PM

Required readings

- Borooah, Vani K. 2002. *Logit and Probit: Ordered and Multinomial Models*. (Quantitative Applications in the Social Sciences, 07-138). Chapters 1 and 2. Thousand Oaks, CA: Sage Publications.
- Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. Chapter 5. College Station, TX: Stata Press.

Additional resources

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Chapters 5 and 6. Thousand Oaks, CA: Sage Publications.
- Powers, Daniel A., and Yu Xie. 2000. *Statistical Methods for Categorical Data Analysis*. Chapter 7. San Diego, CA: Academic Press.
- Hardin, James, and Joseph Hilbe. 2001. *Generalized Linear Models and Extensions*. Chapters 14 and 15. College Station, TX: Stata Press.
- Greene, William H. 2003. *Econometric Analysis. (Fifth Edition)*. Chapter 21, pp. 719-740. Upper Saddle River, NJ: Pearson Education.

October 30. Multinomial logistic regression

Required readings

- Borooah, Vani K. 2002. *Logit and Probit: Ordered and Multinomial Models*. (Quantitative Applications in the Social Sciences, 07-138). Chapter 3. Thousand Oaks, CA: Sage Publications.
- Long, J. Scott and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. Chapter 6, pp.189-235. College Station, TX: Stata Press.
- Reynolds, Jeremy. 2004. "When Too Much Is Not Enough: Actual and Preferred Work Hours in the United States and Abroad." *Sociological Forum*, 19: 89-120. RESERVE.

November 7. Count data models: Poisson regression

Assignment 2 (Binary logit) final draft due November 7, 5PM

Assignment 3 (Multinomial and ordered logistic regression) due November 10, 5 PM

Required readings

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Chapter 8, pp. 217-230. Thousand Oaks, CA: Sage Publications. RESERVE.
- Long, J. Scott and Jeremy Freese. 2003. Chapter 7, pp.245-264. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. College Station, TX: Stata Press.

Additional resources

- Cameron, Adrian Colin, and Praved K. Trivedi. 1998. *Regression Analysis of Count Data*. New York: Cambridge University Press.
- Lindsey, James K. 1995. *Modeling Frequency and Count Data*. New York: Oxford University Press.
- Hardin, James, and Joseph Hilbe. 2001. *Generalized Linear Models and Extensions*. Chapters 11, 12 and 13. College Station, TX: Stata Press.
- Greene, William H. 2003. *Econometric Analysis. (Fifth Edition)*. Chapter 21.9 (pp.740-752). Upper Saddle River, NJ: Pearson Education.

November 14. Count data models: Negative binomial regression

Required readings

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Chapter 8, pp. 230-239. Thousand Oaks, CA: Sage Publications. RESERVE.
- Long, J. Scott and Jeremy Freese. 2003. Chapter 7, pp.266-273. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. College Station, TX: Stata Press.
- Van der Burg, Brigitte, Jacques Siegers, and Rudolf Winter-Ebmer. 1998. Gender and Promotion in the Academic Labour Market. *Labour*, 12: 701-713

November 21. Zero-inflated count data models

Assignment 3 (Ordered and multinomial logit) final draft due in class

Required readings

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*.

Chapter 8, pp. 239-249. Thousand Oaks, CA: Sage Publications. RESERVE.

Long, J. Scott and Jeremy Freese. 2003. Chapter 7, pp.274-286. *Regression Models for Categorical Dependent Variables Using Stata, Revised Edition*. College Station, TX: Stata Press.

Sarkisian, Natalia and Naomi Gerstel. 2004. "Explaining the Gender Gap in Help to Parents: The Importance of Employment." *Journal of Marriage and the Family*, 66: 431-451. RESERVE

November 28. Regression analyses for survey data.

***Assignment 4 (Count data models) first draft due December 1, 5PM**

Required readings

Eltinge, John L. and W. M. Sribney. 1996. svy: Survey Sample. *Stata Technical Bulletin Reprints*, 6: 208-259. RESERVE.

Johnson, David R. and Lisa A. Elliott. 1998. "Sampling Design Effects: Do They Affect the Analyses of Data from the National Survey of Families and Households?" *Journal of Marriage and the Family*, 60, 993-1001. RESERVE.

Additional resources

Henry, Gary T. 1990. *Practical Sampling*. Chapter 6. Newbury Park, CA: Sage.

Lehtonen, Risto and Erkki J. Pahkinen. 1995. *Practical Methods for Design and Analysis of Complex Surveys*. Chapter 3, "Further Use of Auxiliary Information," pp.65-93 and 109-114; Chapter 8, "Multivariate Survey Analysis," pp. 239-283. New York: Wiley.

Skinner, C.J., D. Holt, and T. M. F. Smith. *Analysis of Complex Surveys*. Chapters 1-3, pp.1-88. New York: Wiley.

December 5. Handling missing data in regression analyses

Required readings

Allison, P. D. (2001). *Missing Data*. (Quantitative Applications in the Social Sciences, 07-136). Thousand Oaks, CA: Sage Publications.

Additional resources

Little, Roderick J. A. and Donald B Rubin. 2002. *Statistical Analyses with Missing Data*. 2nd edition. Hoboken, NJ: Wiley.

Schafer, Joe L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, Joe L. and M. K. Olsen. 1998. Multiple Imputation for Multivariate Missing-data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*. 33: 545-571.

December 12. Review.

Assignment 4 (Count data models) final draft due December 12, 5PM