SOC 506 Applied Regression

Spring 2022 Syllabus

Time: Tuesdays 2:30p - 5:30p ET
Location: Callaway Memorial Center S107
Instructor: Heeju Sohn, Ph.D.
Email: heeju.sohn@emory.edu

Office Hours: By appointment.

COVID-19 Exception: Classes will be held remotely until January 31, 2022. See schedule below for details. Office hours will also be held via Zoom. https://emory.zoom.us/j/94253017612

Course Description

This course builds upon the statistical toolkit from SOC 500 Linear Regressions and provides a foundation for conducting and evaluating regression-based works in the social sciences. The first part of the course will cover the topics in conducting transparent and reproducible research. Students will be expected to adopt these research practices throughout the semester. The second part of the course will cover generalized linear models (GLM) that examine non-linear outcome variables. The readings, lectures, and in-class discussion will address each method's mathematical justification, execution, and interpretation using statistical software and application in published articles. The third component of the course will focus on students' in-class presentations and discussions of their research projects.

This course's primary goal is for students to gain fluency in the foundational statistical methods in the social sciences. Fluency denotes the ability to 1) assess the methods' appropriateness to address sociological questions, 2) provide thoughtful reviews to works using these methods, and 3) actively engage in collaborations that use statistical methods. This course aims to provide a broad survey of the most commonly used generalized linear models rather than expert knowledge in any particular approach; each topic is worthy of its own semester-long course.

The success of this course depends critically on active participation. As such, students will be evaluated on their intellectual engagement during class discussions and presentations. Please complete the assigned readings before class. This course also requires students to complete an extended abstract or a preliminary analysis of an empirical paper. The final project will be evaluated on its transparency and reproducibility rather than its methodological sophistication. Stage assignments throughout the semester will allow for opportunities for feedback and revision.

Lastly, the primary focus of this course is not statistical programming. I do not expect you to memorize the commands and options for all the statistical methods that we will cover in this course. Instead, I expect you to understand the theory and assumptions behind the methods so that you will be able to figure out the programming part in the future. We will still heavily rely on software for analysis and documentation throughout the course. We will use Stata throughout this course.

Learning Objectives

1. Gain fluency in the application of generalized linear models in social science research
2. Adopt practices for transparent and reproducible research
3. Practice thoughtful feedback and collaboration in a working group setting
4. Complete an extended abstract of an empirical article

Technology requisitions and expectations

Microsoft Teams

We will use Microsoft Teams as the primary mode of communication in this course. Please install Microsoft Teams on your personal machine and sign in using your Emory ID **before the first day of class.**

Stata

Students will be using statistical software to complete the assignments and to produce dynamic documents of their research project. Basic familiarity (data manipulation for analysis, producing summary statistics, and running linear regressions) with Stata is a prerequisite for this course.

You can purchase a single-user time-limited student license from Stata.com. Stata/BE (basic edition) will be sufficient for this course. **We will start using Stata from the first day class.** You are expected to attend all classes with Stata installed on your personal machine.

Markdown

Markstat is a user-written programming language that allows users to write dynamic documents that integrates Stata output into html, pdf, and docx files using Markdown language. Install Markstat and Pandoc onto your personal machine that has Stata. Follow the instructions on https://data.princeton.edu/stata/markdown. Set up your Stata following the instructions on the "Getting Started" page (link is at the bottom). Read the "Documentation" and try to replicate some examples in the "Example" tab. I like this one: https://data.princeton.edu/stata/markdown/wfsx but feel free to explore others.

Note for Mac users: You will need to download and install pandoc onto your machines. You can find the installed pandoc file in your usr/local/bin folder. Instructions on accessing your bin folder: https://macpaw.com/how-to/access-bin-folder-mac

Email

I will aim to respond to emails within two business days. Please include the course number in the subject line. Please submit all assignments through Teams and not via email.

Class Zoom information (January 2022)

https://emory.zoom.us/j/94253017612

Meeting ID: 942 5301 7612
One tap mobile
+16699006833,,94253017612# US (San Jose)
+12532158782,,94253017612# US (Tacoma)

Dial by your location
+1 669 900 6833 US (San Jose)
+1 253 215 8782 US (Tacoma)
+1 346 248 7799 US (Houston)
+1 470 381 2552 US (Atlanta)
+1 646 558 8656 US (New York)
+1 301 715 8592 US (Washington DC)
+1 312 626 6799 US (Chicago)
+1 470 250 9358 US (Atlanta)
Meeting ID: 942 5301 7612
Find your local number: https://emory.zoom.us/u/aChndIzeZ

Join by SIP
94253017612@zoomcrc.com

Join by H.323
162.255.37.11 (US West)
162.255.36.11 (US East)
115.114.131.7 (India Mumbai)
115.114.115.7 (India Hyderabad)
213.19.144.110 (Amsterdam Netherlands)
213.244.140.110 (Germany)
103.122.166.55 (Australia Sydney)
103.122.167.55 (Australia Melbourne)
64.211.144.160 (Brazil)
69.174.57.160 (Canada Toronto)
65.39.152.160 (Canada Vancouver)
207.226.132.110 (Japan Tokyo)
149.137.24.110 (Japan Osaka)
Meeting ID: 942 5301 7612

Required textbook and readings

Christensen, G., J. Freese and E. Miguel. 2019. Transparent and Reproducible Social Science Research: How to Do Open Science: University of California Press.

Hoffmann, J.P. 2016. Regression Models for Categorical, Count, and Related Variables: An Applied Approach: University of California Press.

Allison, P. D. 2002. *Missing data*. SAGE Publications, Inc.

In addition to these books, I will assign journal articles on some weeks. I will post them on Teams.

Assignments

This is an overview of the assignments that you will complete in this course. Please refer to the weekly schedule and its accompanying assignment/lab documents for detailed instructions.

Weekly readings
You are expected to complete all assigned readings and attend class ready to discuss the material. The quality of class meetings for both you and your classmates depend on your contribution.

We will be relying on Hoffman (2016) to learn about non-linear regression models. The book uses Stata code examples and output throughout the text. The equivalent R code is included in the Appendix. You are, however, not expected to memorize code. As you are reading the book, please focus on the following:

- When is this statistical method appropriate to use?
- What are the underlying assumptions of the method?
- What are the method's strengths and limitations?
- What outputs do they produce, and how do you interpret them?

Week 1: In-class introductions
On the first day of the course, each student will give an in-class presentation on their general research interests, career goals, and the proposed empirical research project for this course.

Week 2: In-class presentation of data
Each student will prepare a 10-to-15-minute in-depth presentation on the dataset that they will be using to complete their term research project.

Weeks 3-8: In-class lab and presentation of work
Each week, you will complete in-class assignments (aka “lab”) that will allow you to practice setting up a workflow, collaborate effectively, and conduct fundamental statistical analyses using Stata. At the end of the in-class lab, you will present your all or a portion of your results to the class.

Week 9: Spring Break

Weeks 10-15: In-class presentation of assigned articles
We will cover advanced linear regression methods during these weeks, and I have chosen publications that use these methods. Each week, one of you will prepare a presentation that summarizes the week's assigned article and lead a discussion with the class.

Extended abstract of empirical research
Your term assignment will be centered on an empirical research project. You are not required to use any of the methods covered in this class. You can also use the methods you learned in SOC 500. In this course, I want you to focus on developing ethical and transparent research practices. The fundamental concepts are applicable to both quantitative and qualitative research. While we will be focusing on quantitative research methods in this course, I want you to adapt these practices for your own research approach.

You will write up an extended abstract of your research findings that includes the motivation for your research question, its contribution to the literature, data and methods, and the preliminary results. You may use this assignment for a conference submission or a grant proposal. The extended abstract will be due at the end of the semester.

Resources

Resources for Inclusive Learning
We all learn differently, and sometimes we need accommodations. Please let me know if any aspect of the course prevents you from learning or being fully engaged. If you need official accommodations for accessibility or alternative course materials, please utilize the University's services through the Department of Accessibility Services. Link to Emory's Accessibility Service: http://accessibility.emory.edu/index.html The site also has an informative section for self-advocacy. (http://accessibility.emory.edu/students/new-to-oas/self-advocacy.html)

Additional programming references
For Stata users

- Long, J. Scott. 2009. The Workflow of Data Analysis Using Stata: Stata Press.
- Long, J.S., J. Freese and StataCorp LP. 2006. Regression Models for Categorical Dependent Variables Using Stata, Second Edition: Taylor & Francis.
- Writing Dynamic Markdown Documents Using Stata (Doug Hemken) https://www.ssc.wisc.edu/~hemken/Stataworkshops/dyndoc%20review/Review.html

Honor Code

We will follow Emory's code for academic integrity and conduct in this course. Please read Appendix I for properly paraphrasing and quoting another writer. You can also see common forms of misconduct in Appendices II and III. Please ask if you have any doubts about whether something will violate the policy. Link to Emory's Honor Code: http://catalog.college.emory.edu/academic/policies-regulations/honor-code.html

Grading Policy

Items	**Percentage of grade**
Class participation	10
Weekly lab assignments	40
In-class presentations	30
End of term research abstract	20

Weekly lab assignments and in-class presentations will be graded on a complete/incomplete basis. The expectation is that all students in the course will satisfy all the course requirements and receive an A. If you feel that you are falling behind, please reach out to me sooner than later.

Spring 2022 Schedule

Please complete the readings before attending class. Assignment submissions are also due by the start of class (Tuesdays 2:30p ET) unless stated otherwise. Also, please note that the schedule may change during the semester. I will announce any changes on Teams or Canvas.

Abbreviations

CFM 2019 - Christensen, G., J. Freese and E. Miguel. 2019. Transparent and Reproducible Social Science Research: How to Do Open Science: University of California Press.

Hoffman 2016 - Hoffmann, J.P. 2016. Regression Models for Categorical, Count, and Related Variables: An Applied Approach: University of California Press.

Allison 2002 - Allison, P. D. 2002. *Missing data*. SAGE Publications, Inc.

Week 1: Introduction – January 11, 2022

Class will be held remotely.

Readings

- CFM 2019. Chapters 1-4

In-class lab: Introduction to technology and data

Assignment for class

- Student research introductions – send me your Powerpoint files via email before class as a backup

Week 2: Research transparency – January 18, 2022

Class will be held remotely.

Readings

- CFM 2019. Chapters 5-8
- *(optional)* Von Elm, Erik, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche and Jan P. Vandenbroucke. 2007. "The Strengthening the Reporting of Observational Studies in Epidemiology (Strobe) Statement: Guidelines for Reporting Observational Studies." The Lancet 370(9596):1453-57. doi: 10.1016/s0140-6736(07)61602-x.

In-class lab: Warming up in Stata

Assignment for class

- In-depth data presentation– send me your Powerpoint files via email before class as a backup

Week 3: Research practices for reproducibility – January 25, 2022

Class will be held remotely.

Readings

- CFM 2019. Chapters 9-11, and the Appendix
- *(optional)* Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt and Tracy K. Teal. 2017. "Good Enough Practices in Scientific Computing." PLOS Computational Biology 13(6):e1005510. doi: 10.1371/journal.pcbi.1005510.

In-class lab + presentation: Setting up a reproducible workflow

Week 4: Statistics fundamentals—February 1, 2022

Readings

- Hoffman 2016. Chapter 1

In-class lab + presentation: Understanding sample weights, working with income variables, and comparisons of distributions

Week 5: Multivariate linear regressions I—February 8, 2022

Readings

- Hoffman 2016. Chapter 2
- Chang, Virginia W. and Diane S. Lauderdale. 2009. "Fundamental Cause Theory, Technological Innovation, and Health Disparities: The Case of Cholesterol in the Era of Statins." Journal of health and social behavior 50(3):245-60. doi: 10.1177/002214650905000301.

In-class lab + presentation: Probabilities, comparison of proportions, and multivariate linear regressions

Week 6: Multivariate linear regressions II —February 15, 2022

In-class lab + presentation: Hypothesis testing, two-way interactions, and three-way interactions in linear regressions

Week 7: Missing data I—February 22, 2022

Readings

- Allison 2002. Chapters 1-4
- Bachmeier, James D., Jennifer Van Hook, and Frank D. Bean. 2014. "Can We Measure Immigrants' Legal Status? Lessons from Two U.S. Surveys." International Migration Review 48(2):538–66. doi: 10.1111/imre.12059.

Week 8: Missing data II—March 1, 2022

Readings

- Allison 2002. Chapters 5-8
- Sohn, Heeju, and Anne R. Pebley. 2018. "New Approaches to Estimating Immigrant Documentation Status in Survey Data." CCPR Working Paper.

Week 9: SPRING BREAK —March 8, 2022

Weeks 10: Logistical and probit regression models —March 15, 2022

Readings

- Hoffman 2016. Chapter 3
- Bacong, A. and H. Sohn (2020). "Disentangling contributions of demographic, family, and socioeconomic factors on associations of immigration status and health in the United States." Journal of epidemiology and community health: jech-2020-21424.
- Glied, Sherry and Adriana Lleras-Muney. 2008. "Technological Innovation and Inequality in Health." Demography 45(3):741-61. doi: 10.1353/dem.0.0017.

Weeks 11+12: Ordered logistical and probit regression models — March 22, 29, 2022

Readings

- Hoffman 2016. Chapter 4
- Dunifon, Rachel and Ashish Bajracharya. 2012. "The Role of Grandparents in the Lives of Youth." Journal of Family Issues 33(9):1168-94. doi: 10.1177/0192513X12444271.
- Currie, Janet, Sandra Decker and Wanchuan Lin. 2008. "Has Public Health Insurance for Older Children Reduced Disparities in Access to Care and Health Outcomes?". Journal of Health Economics 27(6):1567-81. doi: 10.1016/j.jhealeco.2008.07.002

Week 13: Multinomial logit and probit regression models — April 5, 2022

Readings

- Hoffman 2016. Chapter 5
- Lum, T. Y. and Elizabeth Lightfoot. 2005. "The Effects of Volunteering on the Physical and Mental Health of Older People." Research on Aging 27(1):31-55. doi: 10.1177/0164027504271349.

Week 14: Event history models — April 12, 2022

Readings

- Hoffman 2016. Chapter 7
- Sohn, H. 2015. "Health Insurance and Risk of Divorce: Does Having Your Own Insurance Matter?". Journal of Marriage and Family 77(4). doi: 10.1111/jomf.12195.
- Baden, L.R., El Sahly, H.M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S.A., Rouphael, N., Creech, C.B., Mcgettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., Gilbert, P., Janes, H., Follmann, D., Marovich, M., Mascola, J., Polakowski, L., Ledgerwood, J., Graham, B.S., Bennett, H., Pajon, R., Knightly, C., Leav, B., Deng, W., Zhou, H., Han, S., Ivarsson, M., Miller, J., Zaks,

T., 2021. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. New England Journal of Medicine 384, 403–416.. doi:10.1056/nejmoa2035389

Week 15: Poisson and negative binomial models — April 19, 2022

Readings

- Hoffman 2016. Chapter 6
- Degomme, Olivier and Debarati Guha-Sapir. 2010. "Patterns of Mortality Rates in Darfur Conflict." The Lancet 375(9711):294-300. doi: 10.1016/S0140-6736(09)61967-X.
- Flippen, Chenoa A. 2012. "Laboring Underground: The Employment Patterns of Hispanic Immigrant Men in Durham, Nc." Social Problems 59(1):21-42. doi: 10.1525/sp.2012.59.1.21.
- Sohn, Heeju, Stefan Timmermans and Pamela J. Prickett. 2020. "Loneliness in Life and in Death? Social and Demographic Patterns of Unclaimed Deaths." PLoS One 15(9):e0238348. doi: 10.1371/journal.pone.0238348.

Suggested articles with more Hoffman chapters

Hoffman 2016. Chapter 9 Multilevel regression models

- Notes: This chapter relies heavily on Stata commands and options. Don't get hung up on the Stata-specific jargon and focus on gaining a general understanding of what the models are doing in the background.
- Wheaton, Blair and Philippa Clarke. 2003. "Space Meets Time: Integrating Temporal and Contextual Influences on Mental Health in Early Adulthood." American Sociological Review 68(5):680-706. doi: http://dx.doi.org/10.2307/1519758.
- Hendryx, Michael S. , Melissa M. Ahern, Nicholas P. Lovrich and Arthur H. McCurdy. 2002. "Access to Health Care and Community Social Capital." Health Services Research 37(1):85-101. doi: 10.1111/1475-6773.00111.

Hoffman 2016. Chapter 10 Principal components and factor analysis

- Notes: This chapter is a surface-level overview of data reduction techniques. Focus on principal component analysis (PCA) and factor analysis (FA) methods (up to page 260), and skim structural equation modeling (SEM) (pages 260-267)
- Carrillo-Álvarez, Elena, Ester Villalonga-Olives, Jordi Riera-Romaní and Ichiro Kawachi. 2019. "Development and Validation of a Questionnaire to Measure Family Social Capital." SSM - Population Health 8:100453-53. doi: 10.1016/j.ssmph.2019.100453.
- Kim, S., et al. (2019). "Factor Structure for Chronic Stress Before and During Pregnancy by Racial/Ethnic Group." Western journal of nursing research 41(5): 704-727.

Hoffman 2016. Chapter 8 Regression models for longitudinal data

- Sastry, Narayan and Jon M. Hussey. 2003. "An Investigation of Racial and Ethnic Disparities in Birth Weight in Chicago Neighborhoods." Demography 40(4):701-25.

- Arenas, Erika, Noreen Goldman, Anne R. Pebley and Graciela Teruel. 2015. "Return Migration to Mexico: Does Health Matter?". Demography 52(6):1853-68. doi: http://dx.doi.org/10.1007/s13524-015-0429-7.