

SOSC 320

Data in the World: Applied Social Statistics



Session 1, 2023-24

Course meeting time: TuTh 8:30am - 11:00am
Course meeting location: IB 2069 (Team Room)
Course format: Lecture + Discussion + Lab
Academic credit: 4

Instructor's information

Markus Neumann

Assistant Professor of Political Science and Computational Social Science

Email: markus.neumann@dukekunshan.edu.cn

Office: WDR 1104

Office hours: Monday & Wednesday 2-4pm or by appointment

Professional website: <https://markusneumann.github.io/>

I am an Assistant Professor of Political Science and Computational Social Science at DKU. My research revolves around the application of statistics and machine learning methods to social science data, especially text, images and audio. The substantive focus of my research is political advertising.

Getting in touch with me

Feel free to send me an email about any questions you may have. I will try to respond within 24 hours. If you don't receive a response in that time, feel free to email me again. You can also come to my office hours, which are held on the same day of the week on which most of the assignments are due.

What is this course about?

Students coming out of Stats 101 have a basic understanding of statistics. This class is designed to broaden that understanding and provide them with additional techniques and tools that can be used to understand, analyze and address more complicated data and hypotheses. This includes testing multiple hypotheses at once and examining the effects of variable interactions. It also includes moving from the simple cross-sectional data students saw in Stats 101 to time-series and panel samples, being able to handle binary and integer dependent variables, correct errors for common concerns over heteroskedasticity and clustering, and impute missing data. The goal is to make you a better consumer and producer of statistical analyses, something that will serve you well in your academic and professional career.

1

What background knowledge do I need before taking this course?

Students must have taken Stats 101 and have a working understanding of R. Exceptions may be possible with instructor consent.

What will I learn in this course?

By the end of this course, you will be able to:

1. Correct standard errors for heteroskedasticity and clustering
2. Expand your understanding of regressions to include panel and time series data
3. Test hypotheses involving multiple variables and interactions between variables
4. Analyze limited dependent variable models including logit/probit, hazard and count models
5. Explain how omitted variables can cause spurious results
6. Generate inference from simulations
7. Impute missing data
8. Conduct factor analyses
9. Understand what types of datasets are amenable to which type of statistical method
10. Apply these techniques on real world data
11. Assess and interpret results from these techniques

What will I do in this course?

Classes will begin with a lecture explaining the theory behind the statistical methods and their practical application to the social sciences. Students will then have an opportunity to immediately apply the knowledge to examples during the lab portion of the course.

There will be five short homework assignments in which students will be asked to analyze a dataset given what they have learned in the previous classes. Students will have to explain the technique, how to apply it to the problem, write their own code to analyze the data and then type a short report. Each assignment will be worth 10% of the students' grade. Students may study together in groups for the homework assignments but your writeup for the homework and your code must be your own.

In another assignment, every student will be asked to find a real-world dataset applicable to a given day of the class and introduce it in a short presentation. We will then analyze this dataset together. This assignment will be worth 5% of the course grade.

The final project will ask you to apply a number of the techniques learned in class to a real world dataset of your choice and write your results up as a formal report. The final project contains the following components, which will be submitted over the course of the semester: Project proposal (5%), data availability report (5%), draft (10%), final paper (15%).

A final 10% of your grade will be based on in class participation, including brief group presentations, preparation for class, and attendance.

How can I prepare for the class sessions to be successful?

Students should read all assigned materials before the class and consider it carefully. They should come to ask ready to discuss the material and ask any questions they have to the professor.

What required texts, materials, and equipment will I need?

- Wooldridge, Jeffrey M. 2020. *Introductory Econometrics: A Modern Approach*. Cengage Learning.
- Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons, Inc. Chapter 6.1, 6.2 (Duke Library online)
- Shalizi, Cosma. *Advanced Data Analysis from an Elementary Point of View*. Available at <http://www.stat.cmu.edu/~cshalizi/ADAFaEPoV/>
- Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*, Chapter 2. (Duke Library online)
- Bonica, Adam. 2018. "Inferring roll-call scores from campaign contributions using supervised machine learning." *American Journal of Political Science* 62(4): 830-848. (Duke Library online)

How will my grade be determined?

1. Homework assignments: 5 assignments at 10% each = 50%
2. Dataset presentation: 5%
3. Final project: 35%
 - a. Proposal: 5%
 - b. Data availability report: 5%
 - c. Draft: 10%
 - d. Final paper: 15%
4. In-class participation: 10%

Please refer to the following scale for converting numeric scores to letter grades. Final grades are rounded to the nearest whole number, .5 is rounded up.

A+ = 98% - 100% **A** = 93% - 97%; **A-** = 90% - 92%; **B+** = 87% - 89%; **B** = 83% - 86%; **B-** = 80% - 82%; **C+** = 77% - 79%; **C** = 73% - 76%; **C-** = 70% - 72%; **D+** = 67% - 69%; **D** = 63% - 66%; **D-** = 60% - 62%; **F** = 59% and below

What are the course policies?

Academic Integrity

As a student, you should abide by the academic honesty standard of the Duke Kunshan University. Its Community Standard states: "Duke Kunshan University is a community comprised of individuals from diverse cultures and backgrounds. We are dedicated to scholarship, leadership,

and service and to the principles of honesty, fairness, respect, and accountability. Members of this community commit to reflecting upon and upholding these principles in all academic and non-academic endeavors, and to protecting and promoting a culture of integrity and trust.”

Generative AI

The use of generative AI (such as ChatGPT) is prohibited for all assignments. This applies to all components of the assignments – statistics, programming, and writing. The use of generative AI will be considered an academic integrity violation and will result in a 0 on the given assignment.

Academic Policy & Procedures

You are responsible for knowing and adhering to academic policy and procedures as published in University Bulletin and Student Handbook. Please note, an incident of behavioral infraction or academic dishonesty (cheating on a test, plagiarizing, etc.) will result in immediate action from me, in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising). Please visit the Undergraduate Studies website for additional guidance related to academic policy and procedures.

Academic Disruptive Behavior and Community Standard

Please avoid all forms of disruptive behavior, including but not limited to: verbal or physical threats, repeated obscenities, unreasonable interference with class discussion, making/receiving personal phone calls, text messages or pages during class, excessive tardiness, leaving and entering class frequently without notice of illness or other extenuating circumstances, and persisting in disruptive personal conversations with other class members. If you choose not to adhere to these standards, I will take action in consultation with university administration (e.g., Dean of Undergraduate Studies, Student Conduct, Academic Advising).

Academic Accommodations

If you need to request accommodation for a disability, you need a signed accommodation plan from Campus Health Services, and you need to provide a copy of that plan to me. Visit the Office of Student Affairs website for additional information and instruction related to accommodations.

Class Attendance

As a seminar class, it is vital that all students attend and participate. Class attendance is therefore mandatory and absences will only be excused for serious medical and personal matters. If you will be absent from a class for a university-sponsored activity, please make arrangements with me — beforehand — regarding any work you might miss.

Late Penalties

This course will move quickly so therefore it is imperative that you do not fall behind by submitting late homework. Therefore, homework will be accepted only until 48 after the due date and will be subject to a 50% penalty. Rescheduling of homework/final due dates will only be permitted for serious medical and personal matters, and requires advance notice. Unless stated otherwise, all assignments are due at 11:59pm China time.

What campus resources can help me during this course?

Academic Advising and Student Support

Please consult with me about appropriate course preparation and readiness strategies, as needed. Consult your academic advisors on course performance (i.e., poor grades) and academic decisions (e.g., course changes, incompletes, withdrawals) to ensure you stay on track with degree and graduation requirements. In addition to advisors, staff in the Academic Resource Center can provide recommendations on academic success strategies (e.g., tutoring, coaching, student learning preferences). Note, there is an ARC Sakai site for students and tutors. Please visit the [Office of Undergraduate Advising website](#) for additional information related to academic advising and student support services.

Writing and Language Studio

If you want additional help with academic writing and more generally with language learning you are welcome to go to Writing and Language Studio (WLS). You can register for an account, make an appointment, and learn more about WLS services, policies, and events on the [WLS website](#). You can also find writing and language learning resources on the [Writing & Language Studio Sakai site](#).

IT Support

If you are experiencing technical difficulties, please contact IT:

- China-based faculty/staff/students 400-816-7100, (+86) 0512- 3665-7100
- US-based faculty/staff/students (+1) 919-660-1810
- International-based faculty/staff/students can use either telephone option (recommend using tools like Skype calling)
- Live Chat: <https://oit.duke.edu/help>
- Email: service-desk@dukekunshan.edu.cn

What is the expected course schedule?

Class 1 (August 22)

Topics: Introduction; Stats 101 recap; Expected values

Class 2 (August 24)

Topics: Dummy variables and interaction effects
Reading: Wooldridge (Ch. 7.1, 7.2 (skip 7.2a), 6.2c, 7.4)

Homework 1 assigned on Friday, August 25

Class 3 (August 29)

Topics: Transformations and Nonlinearity
Reading: Wooldridge, Ch. 6.1, 6.2a-b

Homework 1 due on Wednesday, August 30

Class 4 (August 31)

Topics: Regression problems and diagnostics
Reading: Wooldridge (Ch. 3.3, 3.4, 8.1, 8.2, 8.3)

Homework 2 assigned on Friday, September 1

Class 5 (September 5)

Topics: Binary response models
Reading: Wooldridge, Ch. 17.1

Homework 2 due on Wednesday, September 6

Class 6 (September 7)

Topics: Multinomial and ordered choice models
Reading: Agresti Ch. 6.1, 6.2

Homework 3 assigned on Friday, September 8
Final project proposal due on Sunday, September 10

Class 7 (September 12)

Topics: Count models
Reading: Wooldridge, Ch. 17.3

6

Homework 3 due on Wednesday, September 13

Class 8 (September 14)

Topics: Hazard models

Reading: Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. Event History Modeling: A Guide for Social Scientists, Chapter 2. (Duke Library online)

Homework 4 assigned on Friday, September 15

Class 9 (September 19)

Topics: Time series

Reading: Wooldridge, Ch. 10.1, 10.2, 10.5, 11.1

Homework 4 due on Wednesday, September 20

Class 10 (September 21)

Topics: Panel data

Readings: Wooldridge, Ch. 13

Homework 5 assigned on Friday, September 22

Final project data availability report due on Sunday, September 24

Class 11 (September 26)

Topics: Advanced panel data analysis

Readings: Wooldridge, Ch. 14

Homework 5 due on Wednesday, September 27

Class 12 (September 28)

Topics: Two stage least squares

Readings: Wooldridge, Ch. 15

October 3 & 5 - No class

Class 13 (October 10)

Topics: Exploratory factor analysis

Readings: Shalizi Ch. 15

Final paper draft due on Wednesday, October 11

7

Class 14 (October 12)

Topics: Missing data

Readings: Bonica, Adam. 2018. "Inferring roll-call scores from campaign contributions using supervised machine learning." *American Journal of Political Science* 62(4): 830-848.

Final project due on October 18