STATISTICAL ANALYSIS IN EDUCATIONAL RESEARCH: REGRESSION
EDUC 400B
Syllabus, Winter, 2021
MW 10:00-11:50 AM
Location: Zoom - Synchronous
Version: 01.07.2021

Professor:
Eric Bettinger
Office Hours: Available upon request but liberally given.

Teaching Assistants
Victoria Docherty
Office Hours:
Radhika Kapoor
Office Hours:
Kaylee Matheny
Office Hours:
Marissa Thompson
Office Hours:

Pandemic Consideration:

Much of the language in this syllabus is designed for ordinary circumstances. With the ever-present threat of COVID-19, these are not ordinary circumstances. At various points during the quarter, we may find ourselves coping with unanticipated challenges, whether to our health, the people we care about, childcare, internet access, or other aspects of our lives. We might find that the course is interrupted in ways we did not anticipate.

As much as we are excited and prepared to help you grow as researchers and graduate students, our first priority is your well-being. If you are experiencing any challenges that affect your work in this course, please let us know. We don't need any details or documentation, and we will work with you to develop alternatives for ensuring that you're still able to learn from this course. Additionally, if you are experiencing challenges more generally, we're happy to serve as a point of contact to connect you with resources to support you.

Discussion Sections:
Wednesdays, 2-3pm
Thursdays, 9-10am

Discussion section Zoom information is accessible on the "Zoom" tab in Canvas. TA section attendance is optional. We will roughly cover the same information (contingent upon students' questions) in each section, so you need not worry about attending both sessions (unless you just want to do so).

Overview
Statistical analysis can be a powerful tool in understanding social, educational, psychological, and developmental processes. In cases where it is impossible or impractical to collect data on every individual, classroom, teacher, and school of interest, statistical analysis allows us to examine data on a sample of individuals (or classrooms, schools, etc.) in order to infer patterns in a larger population. For example, we might want to examine data on achievement test scores and per-pupil spending for a sample of schools to determine whether there is an association between spending and achievement patterns in the population. Or we might want to examine the association between race/ethnicity and achievement patterns. Moreover, if we find such an association, we

might wish to ask additional questions, such as whether racial/ethnic differences in achievement patterns can be accounted for by racial/ethnic differences in family socioeconomic characteristics or in school quality.

In this course we will learn to answer such questions using regression analysis—a statistical tool that allows us to 1) describe average patterns of association among multiple variables observed in a sample and 2) make inferences about the patterns of association among these variables in a population. Regression analysis is a powerful statistical method with many variations. Our goals in this course are to develop an understanding of the basic methods, including their limitations, and develop skill in using regression analysis to answer educational research questions. Finally, because an important part of any analysis is communicating the results to an audience, we will place considerable emphasis on learning to present (in writing, tables, and figures) the results of regression analyses.

By the end of the quarter, students in this course should be sufficiently skilled in regression analyses that they can critically examine published research using regression and can carefully perform their own analyses.

This is a graduate-level statistics course, so there will be plenty of math. It is a course in the social science application of statistics, however, not a theoretical statistics course. Therefore our math will be in the service of defining statistical models that correspond to the research questions we want to answer, and we will pay little attention to the mathematics by which the programs estimate the parameters in our models. Our concern is with understanding how to use these methods to do good research and with learning to interpret the results they provide.

The course will put a lot of emphasis on conducting statistical analyses using sample data and on communicating the results of these analyses to a general audience of scholars. The software program that we will use for statistical analyses is called Stata (you can use version 13, 14, or 15), a general statistics software package similar in capacity to SPSS and SAS. If you have not used Stata before, there are a number of resources available on the web to help you learn (Google "Stata Tutorial" or try https://stats.idre.ucla.edu/stata/).

Assumed background. This course is an introductory/intermediate level applied statistics course, pitched to the level of students who have completed an introductory statistics course, such as EDUC 200A or 400A. To be prepared for taking this course you should be familiar with basic statistical concepts such as populations, sampling, means, variances, estimates, and statistical inference. As a guideline, the material in Chapters 2-3 of our textbook (Stock & Watson, *Introduction to Econometrics*) should be familiar. A familiarity with simple (bivariate OLS) regression is helpful, but we will cover that again in this course. The course relies heavily on algebra; no more sophisticated mathematics are necessary for our purposes. More advanced statistics courses will typically rely on a familiarity with basic matrix algebra; however, we will not use matrix algebra in this course.

Class format. Before each class, there will be at least one recorded lesson posted in Canvas for you to view. In these lessons, I will lecture on the use of regression analysis to address educational research questions. The class will meet via Zoom, MW, 10:00am-11:50am PT, though it will often not last for the full class slot. Class sessions are to demonstrate principles from the recorded lesson and allow space for student questions. In addition, all students are invited to attend a weekly review/lab session run by the teaching assistants. Students are expected to attend all classes, and it is highly recommended to attend TA sessions, as some supplemental material will be covered in the sections that is not covered in lectures. I strongly urge you to ask questions in class. If you have a question, it is likely that others do as well.

Stata training session. The TAs will host a Stata Boot Camp for Novices dependent on interest. The TAs will also focus on STATA during lab sections to aid with the problem sets. We also recommend those who are new to Stata to consider enrolling in a 1 credit mini-course (EDUC 401B) that meets Tuesdays 12:30-1:50pm.

Textbook/Software
The primary text for this course will be

Stock, J. H., & Watson, M. W. (2010). Introduction to Econometrics. 3rd Edition. Addison Wesley.[1]
Find it at Abe Books | Amazon | BooksRun

Students need not purchase Stata statistical software for this course, as it is available through the GSE IT department (see Canvas for instructions on how to remotely access Stata), but if you wish to purchase it, a discounted student version of Stata is available. Details are available at the link https://www.stata.com/order/new/edu/profplus/student-pricing/. STATA/IC should be sufficient.

In addition, lecture note, handouts, data sets, and other course materials will be posted on the Canvas site for this course. We will be using the Discussions forum on Canvas for discussions on problem sets and other questions about the course. Any questions about content or assignments that could be relevant for other students must be posted to Canvas rather than e-mailed.

Assignments and Grades
Grading will be based on 5 short homework assignments, 3 problem set assignments, 1 journal club assignment, and the final exam. (The final exam is take-home, open book, open notes, but no help from other people). There are actually 6 short assignments (see below), the first four of which are required. Students must then choose one of the last two to submit for grading. All assignments should be submitted to the instructors on Canvas through the Assignments feature.

The short assignments are together worth 15% of the final grade; the three problem sets are each worth 10% of the final grade, the journal club assignment is worth 10% of the final grade, and the exam is worth 45% of the final grade. The short assignments will be graded on a CR/NC basis, while the problem sets will be graded out of 100 points each. Late problem sets will lose 5 points per day late, except in the case of a demonstrated personal emergency. We will post each problem set at least one week before it is due.

Journal club:

1. Every student will be required to pick an article that uses regression methods. Students should describe this article in a one-page summary including the research question, model, independent and dep variables, validity, etc.
2. The TAs will choose 3-4 model articles from the submissions. This selection will happen the week of Feb 15.
3. By Feb 19, students sign up for one of the model articles. Our class on March 17 will focus on reading and understanding these articles. The groups will discuss the articles and share some critiques on them.

Students have the option of taking this course for a grade or for credit/no credit. As noted in the Stanford GSE graduate student handbook, "students are strongly encouraged to ask for grades in all courses where they are available." Grading will be on a curve.

A grade of C is required for students who take the course for credit/no credit to receive credit for the course.

Getting help. In addition to the review/lab sessions, Professor Bettinger and each teaching assistant will hold office hours each week. If you have questions that might be of interest to others in the class, please ask them on Canvas; for example, questions about problem sets or lecture material are best posted here. TAs will check Canvas regularly and you are likely to get a faster response there than via email. However, if a question comes up that is specific to you, feel free to email the professor or either of the teaching assistants directly.

Academic Integrity:

[1] This edition has ISBN 0138009007.

Academic integrity is the pursuit of scholarly activity free from fraud and deception and is an educational objective of this institution. Academic dishonesty includes, but is not limited to, cheating, plagiarizing, fabricating information, facilitating acts of academic dishonesty by others, and submitting the work of another person or work previously used without informing the instructor. The Honor Code, outlining the general expectations pertaining to Academic Integrity applicable to this course, is published in the Graduate Student Handbook available at: http://honorcode.stanford.edu. Students are expected to conform to the highest standards of academic integrity in this course — meaning, essentially, don't lie; don't cheat; don't pass off someone else's work as your own.

It is, nonetheless, acceptable (and encouraged) to work together on the problem sets, though NOT on the final exam. If you work with classmates on the assignments, you may work together to conduct and discuss the analyses, but once you sit down to write up your assignment, you should work entirely on your own. Do not submit assignments that duplicate word-for-word or paraphrase another student's assignment. **You must work entirely on your own for the final exam.**

Approximate Course Schedule

Week	Date	Topic	Pages to Read (Stock & Watson, 3rd edition)	Assignment Due (at beginning of class)
1	01.11	Course overview, intro to regression	14-25, 31-47, 50-55, 64-95	
	01.13	Introduction to STATA		
	TBD	TA Session		
2	01.18	MLK Holiday: NO CLASS	107-132	
	01.20	Simple linear regression	144-153, 155-167	A
	TBD	TA Session		
3	01.25	Model fit, correlation, sum of squares		
	01.27	Indicator variables	153-155	B
	TBD	TA Session		
4	02.01	Intuition about inference		
	02.03	Confidence intervals & statistical inference		Problem set #1
	TBD	TA Session		
5	02.08	Multivariate regression	179-204	
	02.10	More multivariate regression	214-241	C
	TBD	TA Session		
6	02.15	Presidents' Day: NO CLASS		
	02.17	Bias & interpretation in MR		D
	TBD	TA Session		Journal club paper selection
7	02.22	Inference & hypothesis testing in MR		
	02.24	Nonlinear regressions & categorical variables	252-274	
	TBD	TA Session		
8	03.01	Interactions	274-295	Problem set #2
	03.03	Residual analysis		E (Due on 03.05 at 10 AM)
	TBD	TA Session		
9	03.08	Regression diagnostics		
	03.10	Binary choice models	381-407	Problem set #3
	TBD	TA Session		
10	03.15	Bias & validity	312-340	
	03.17	Journal club (JC)		Journal club; F
	TBD	TA Session		
FINAL	03.24	Take home final		DUE by 11:30 AM

Short Assignments:

Assignment	Concepts	Exercises
A	3.1, 3.3, 3.5	3.3, 3.4**
B	4.1, 4.2	4.1, 4.3(a-e)
C	5.2, 5.3	5.4, 5.5
D	6.2, 6.3, 6.4	6.2, 6.3, 6.4
E	8.1*, 8.4	8.2, 8.7
F	9.1, 9.2, 9.5, 9.6	9.10

*Replace "economic" with "education."

**Note that Φ means a standard normal cumulative distribution table. Use the z-tables in the back of the book.