STAT 230A: Linear Models

Lectures: Monday, Wednesday, Friday 10:00am - 11:00am
106 Stanley

Lab: Thu 1:00pm – 3:00pm, 330 Evans
Thu 4:00pm – 6:00pm, 330 Evans

Topics:

- Ordinary least squares (OLS)
- Gauss-Markov theorem
- Finite-sample exact and asymptotic inference
- Heteroskedasticity
- Cluster-robust standard errors
- Model checking
- Model selection and shrinkage
- Generalized linear models
- Generalized estimating equations

Instructor: Sam Pimentel
Office: 429 Evans
Email: spi@berkeley.edu
Preferred pronouns: he/him
Office Hours:
Monday 11:00am-12:00pm
Friday 12:00pm-1:00pm
Location: 429 Evans

GSI: Andy Shen
Email: aashen@berkeley.edu
Preferred pronouns: he/him
Office Hours: Wednesday 2:00-4:00pm
Location: 444 Evans

Text: *Linear Model and Extensions* by Peng Ding. Available on Arxiv.

Equipment: Access to a computer with access to the Internet and R/Python installation will be necessary. If you do not have access to a computer, you can borrow one from the University library. See https://studenttech.berkeley.edu/hardware-lending for more details. The Student Technology Equity Program is another good resource. Feel free to contact the instructor if you have concerns about your access to needed technology.

This is a working draft of the syllabus and is subject to change.

Learning goals

By the end of the semester you should be able to:

1. Understand and interpret ordinary least squares regression models from a mathematical perspective.
2. Adapt the basic regression model to common practical complications including violations of standard assumptions, high-dimensional regimes, and non-continuous outcomes.
3. Evaluate the quality of a regression analysis in context and suggest improvements.
4. Use R or Python to fit, report, and clearly communicate the process and results of a regression analysis.

Prerequisites:

- Master's level probability at the level of STAT 201A. See Dr. Aditya Guntuboyina's STAT 201A course notes for specifics.
- Linear algebra, at the level of the appendix in the course notes. See also Gilbert Strang's recorded lectures from the MIT course 18.06 Linear Algebra, Spring 2005 (available on YouTube).

- R or Python programming. R will be used for all examples, but assignments involving computing may be completed either in R or Python.

Lecture recordings:

The current plan is to record lecture audio and lecture notes using the in-classroom system and post the videos for asynchronous viewing on bCourses. However, I recommend in-person attendance whenever possible and especially for the first few lectures.

Assessment:

Homework

We anticipate giving 7 homework assignments during the semester. Homework will be posted to bCourses, and will generally be due 2 weeks later at 5:00 PM via Gradescope (unless otherwise noted). All homework assignments must be submitted as **a single PDF file.** Homework will be a combination of analytical and computational exercises done "by hand" and data analysis using the computer. Handwritten submissions must be properly organized and easily readable to the grader. Computational problems should include all code with neatly presented tables and figures if necessary. Any submissions that are difficult for the grader to read are subject to a penalty for style/readability. The GSI will provide further instructions for homework submissions in lab sections.

Exams

An in-person midterm exam will be given from **7:00 PM to 9:00 PM on Thursday March 13th**. I do not plan to offer makeup times so **please confirm now** that you can attend the exam. You will be allowed to bring one (double-sided) page of notes with you into the midterm, no other materials will be allowed.

Final project

Students will work in groups of one or two to carry out the final project, a regression analysis or related methodological investigation on a topic of your choice. A written project proposal will be due via Gradescope by **10:00 PM on Friday April 11th**, and the final project report will be due via Gradescope by **11:59 PM on Tuesday, May 13th** (finals week).

Overall score

Your letter grade for the course will be based on the total points for all work in the semester, as follows:

- Homework (each assignment weighted equally): 42%
- Midterm: 20%
- Group project (including proposal): 38%

Grades will not be curved. Students scoring 90% or above overall will receive letter grades in the A-range, students scoring 75%-90% will receive letter grades in the B-range, and students scoring 60%-75% will receive letter grades in the C-range.

Online Resources

bCourses

Homework assignments, lecture videos, and notes from lecture and lab will be posted here.

Ed Discussion

I have created an Ed Discussion site for this course, which you can access through the link in bCourses. This is an online forum to ask questions to fellow students and course staff and to answer other students' questions. I will also make course announcements through Ed. Extra credit of 1% on the final course score will be awarded to the five students who have responded to the most questions on Ed by the end of the semester.

Gradescope

Homework assignments, take-home exams, and regrade requests (see Policies section below) will be submitted through Gradescope, which you can also access through the link in bCourses. This is also where you can view your grades on past assignments.

Policies

Possibility of revisions to course policies

All course policies, including assessment, are subject to change during the course of the semester in response to unforeseen events including but not limited to public health directives, natural disasters, and medical emergencies among members of the course staff.

Late Assignments

All students will have 5 late days that they may use for turning in homework after the due date. This will take the place of any extensions due to sickness or conflicts, unless there are extenuating circumstances, so use them wisely. Late days will automatically be deducted if an assignment is submitted past the deadline. **At most two late days can be used for a single assignment and submissions beyond this period will not be accepted. Submissions that lack the adequate number of late days will not be accepted.** You will be notified when you are out of late days. Late days cannot be used for the group project, and they cannot be requested once homework solutions are posted. Late days are counted at the 24-hour-period level after the time the assignment was due. E.g., the first assignment is due at 5pm, January 30. If you turn it in at 7pm, January 30, it counts as use of one late day, and if you turn it in at 5:30pm, January 31, it counts as use of two late days.

Regrade requests

Regrade requests on an assignment are **due within one week of the release of the graded assignments and the solutions** (if applicable). Regrade requests should be submitted through Gradescope. Requests via email will not be answered. In writing a regrade request, please be specific about the nature and exact location of the error you feel the grader has made, with reference to the solutions if available.

Academic Honesty and AI Policy

The student community at UC Berkeley has adopted the following Honor Code: *"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."* My expectation is that you will adhere to this code. Beyond the importance of respecting your fellow students, acting with integrity in completing course assignments helps ensure that they achieve their purpose, which is to help you learn and develop valuable statistical understanding and skills.

- Homework must be done independently. If you get stuck or want to explore alternative approaches, feel free to discuss issues with students or course staff (including on the online forum); however, you may not do the homework jointly, nor may you ask for or share complete code or solutions. Sharing solutions or obtaining and/or using solutions from previous years or from the Internet, if such are available, is considered cheating.
- On all written assignments, including the homework, you should include a section listing all the sources you drew on in producing your answers; on the homework, you should also list the names of other students with whom you consulted.

- You are not permitted to submit text that is generated by artificial intelligence (AI) systems such as ChatGPT, Bing Chat, Claude, Google Bard, or any other automated assistance for any classwork or assessments. This includes using AI to generate answers to assignments, exams, or projects, or using AI to complete any other course-related tasks. Using AI in this way undermines your ability to develop critical thinking, writing, or research skills that are essential for this course and your academic and professional success. You may use AI as part of your research and preparation for assignments, but text that is submitted must be written by you or (for the project) your partner. For example, you may use AI to generate ideas, questions, or summaries that you refer to during the writing process and cite properly. Although AI systems can provide helpful information or suggestions, but they are not always reliable or accurate, and you should critically evaluate the sources, methods, and outputs of AI systems. Violations of this policy will be treated as academic misconduct. If you have any questions about this policy or if you are unsure whether a particular use of AI is acceptable, please do not hesitate to ask for clarification.

Anyone caught cheating will be given a score of zero (0) on the assignment/exam and will be reported to the University's Office of Student Conduct.

Email

1) If you wish for your email to make it into my inbox, the subject of your email must contain the text "**230A**."
2) Neither I nor the GSI explains course material over email and we will not respond to emails with such requests. Please use the online forum, office hours, and lab (or schedule another time to meet if you have irreconcilable conflicts with the office hours).
3) I respond to email regarding the class roughly once a day, and rarely during the weekend.

Inclusivity and Accommodation

My hope is to establish a learning environment in this course that welcomes diversity of thought, perspective, and experience, and to be respectful of your individual identity as a student. I am happy to use your preferred name and/or personal pronoun. If you feel uncomfortable due to anything that is said in class, or if you feel that your performance in the course is being impacted by experiences outside of class, please do not hesitate to reach out to me about your concerns.

In addition, if you need accommodations for any physical, psychological, or learning disability, please speak to me after class or during office hours. Please note that **you must make arrangements in a timely manner through DSP** so that I can make the appropriate accommodations.

Acknowledgments

Most of the materials used in this course, including this syllabus, are close adaptations from materials originally created or compiled by Prof. Peng Ding and generously provided for the current semester. In writing this syllabus I also adapted content from Prof. Chris Paciorek, from Prof. Monica Linden of Brown University, and from the Center for Teaching and Learning at the University of Texas at Austin.

Anticipated Course Schedule

Week	Topics	Assignments Due & Exams	Assigned Reading (from lecture notes)
Jan 20	Motivation, linear algebra review		Ch. 1, Appendix A-B
Jan 27	Distribution theory review, multiple regression		Ch. 2-3, Appendix C
Feb 3	Gauss-Markov, normal linear model	HW#1 due Monday	Ch. 4-5
Feb 10	Asymptotic inference, Frisch-Waugh-Lovell		Ch. 6-8
Feb 17	(President's Day) Cochran's Theorem, R^2	HW#2 due Monday	Ch. 9-10
Feb 24	Leverage, population least squares		Ch. 11-12
Mar 3	Overfitting, model selection	HW#3 due Monday	Ch. 13
Mar 10	Ridge regression Midterm review	Midterm (Thursday)	Ch. 14
Mar 17	LASSO, transformations, interaction.	HW#4 due Friday	Ch. 15-17
Mar 24	Spring break		
Mar 31	Weighted/restricted OLS, logistic regression		Ch. 18-20
Apr 7	Categorical and count outcomes	HW#5 due Monday Project proposal due Friday	Ch. 21-22, Appendix D
Apr 14	Generalized linear models		Ch. 23-24
Apr 21	Generalized estimating equations, Quantile regression	HW#6 due Monday	Ch. 25-26
Apr 28	Survival outcomes Guest speakers	HW#7 due Friday	Ch. 27
May 5	RRR week		
May 12	Finals week	Project due Tuesday	