Gov2003: Statistical Models for Quantitative Social Science

Kosuke Imai

Professor of Government and of Statistics
Harvard University

Spring 2026

Over the past several decades, the volume and granularity of data available to social scientists have increased dramatically. At the same time, computational power and the complexity of statistical models have grown considerably. This course introduces a variety of statistical models and inference methods developed to address the new challenges faced by quantitative social science researchers. The goal is to familiarize students with a broad range of ideas and tools for analyzing data for measurement, prediction, and causal inference. The course also covers methods for model assessment, evaluation, and selection. It is designed to benefit students across social sciences who wish to apply advanced quantitative methods in their own research.

1 Contact Information

Instructor
Name: Kosuke Imai
Email: imai@harvard.edu
URL: https://imai.fas.harvard.edu
Office: CGIS K306
Office hours: Fridays 1:30pm – 3:00pm (sign up at https://tinyurl.com/KosukeImaiOfficeHours) or by an appointment

Teaching Fellows

Name:	Kentaro Nakamura
Office hours:	Friday 12:00pm-1:20pm
Location:	CGIS South S002B
Email:	knakamura@g.harvard.edu

2 Logistics

- Lectures: CGIS Knafel K354, Mondays and Wednesdays 3:00pm – 4:15pm
- Sections: Sever 106, Friday 1:30pm-2:45pm

3 Prerequisites

This course assumes the solid knowledge of

- probability and statistical theory (based on calculus)

- linear models (based on matrix algebra)
- causal inference
- data analysis using R (students may use other programming languages at their own risk)

at the level of Gov2001 and Gov2002

4 Course Requirements

The final grade is based on the following components:

- **Class participation** (10% of the course grade): We evaluate the level of your engagement during the in-person class meetings and TF sections as well as Ed online discussions.
- **Problem Sets** (30% of the course grade): There will be a total of five problem sets. Each problem set will be weighted equally. The following rules apply to all problem sets:
  - *Submission policy.* All answers including the math and computer code are encouraged to be incorporated into the Rmarkdown file provided by the instructional staff. Handwritten math answers are acceptable once merged into one pdf document. For the exercises, each student should submit the pdf file electronically to Gradescope. Once you upload the PDF file, you will see a list of the questions in the assignment and thumbnails of your file. For each assigned question, click the PDF page(s) that contains your answer. No late submission will be accepted unless you obtain a prior approval from the instructor.
- **In-class, Closed-book Exams** (60% of the course grade): Two closed-book exams will be held in class, each lasting three hours. They cover the first and second half of the materials, respectively, and are equally weighted. The first exam will be held on **March 25**, followed by the review section on Marh 23. The second exam will be held during the final exam period and be scheduled by the registrar later in the semester.

5 Collaboration Policy

Throughout this course, you are encouraged to collaborate with another student in the class so that you can learn from one another. However, for a problem set, you may choose no more than one student as a collaborator. In addition, you must write up your own solutions and make a separate submission. **Under no circumstances may you copy someone else's answer including computer code, mathematical derivation, and substantive interpretation.** Please do not forget to write down your partner's name on your solution to indicate collaboration.

You are also strongly encouraged to reach out to the instructional staff through Ed and office hours about any questions you might have about the course materials. Students should also feel free to ask questions and answer the questions posed by others at Ed, which will count towards class participation.

6 AI Policy

In this class, we apply the same collaboration policy to both human and generative artificial intelligence (GenAI). You may use GenAI tools, such as ChatGPT, to support your work on problem sets. However, just as you would not ask a human study partner to complete the assignment for

you, you must not entirely rely on GenAI to solve problems on your behalf. Appropriate uses of GenAI include help with programming, identifying relevant articles or books, and brainstorming possible approaches. It is not acceptable to use GenAI to directly solve problems or to copy its answers. Such practices hinder your understanding of the course material and are likely to lead to poor performance on closed-book exams.

7 The Instructional Tools

We use a variety of instructional tools to run this course.

- **Course website** (https://canvas.harvard.edu/courses/161614): This is the entry point to all the course materials. The links to the review questions and problem sets will be posted here too.
- **Course calendar** (https://tinyurl.com/gov2003calendar): For your convenience, class meetings and TF sections as well as various deadlines are made available through this Google calendar.
- **Ed** (https://edstem.org/us/courses/95141): Questions about the course including those related to assignments and lectures should be posted here rather than directly emailing an instructional staff. You may find this user guide helpful to orient yourself to the platform.
- **Gradescope** (https://www.gradescope.com/courses/1241212): This is where you will submit all of your assignments. The grades for the assignments will be available here as well. There is a student guide you can check for any questions about the workflow.

8 Reference Books

There is no required textbook for this course, but the following books can serve as useful references:

- Hastie, Trevor, and Robert Tibshirani, and Jerome Friedman. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd Edition, Springer. Freely available at https://hastie.su.domains/ElemStatLearn/
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (2023). *An Introduction to Statistical Learning*, 2nd Edition, Springer. Freely available at https://www.statlearning.com/
- Murphy, Kevin Patrick. (2022). *Probabilistic Machine Learning: An Introduction*, MIT Press. Freely available at https://probml.github.io/pml-book/book1.html

9 Course Plan

We will cover the following topics in that order.

- A review of OLS and linear algebra (2 weeks)
  - Geometry of OLS
  - Inference for OLS
  - Fixed effects regression

- Shrinkage models (3 weeks)
  - Ridge regression
  - Principal component analysis
  - Lasso regression
- Model assessment (2 weeks)
  - Cross-validation
  - Conformal prediction
  - Quantile regression
- Likelihood models (3 weeks)
  - Theory of likelihood inference
  - Logistic regression
  - Case-control design
  - Item response theory
  - Bootstrap
- Latent variable models (2 weeks)
  - Mixture models
  - Expectation-Maximization algorithm
  - Mixed membership models
  - Variational inference