SYLLABUS FOR BIOSTAT/BST 232: METHODS

Harvard T.H. Chan School of Public Health

Fall 2024

COURSE WEBSITE

Canvas: used for accessing course materials, submitting assignments, viewing grades, and posting questions/discussions related to course content and assignments.

https://canvas.harvard.edu/courses/142086

INSTRUCTORS

Professor:

Rachel C. Nethery

Assistant Professor of Biostatistics

Building 1 Room 415, 655 Huntington Ave, Boston MA 02115

rnethery@hsph.harvard.edu

Teaching assistants:

Salvador Balkus, PhD Student, Biostatistics, sbalkus@g.harvard.edu

Kimberly Greco, PhD Student, Biostatistics, kimberly.greco@g.harvard.edu

CLASS TIME AND FORMAT

- Class time: Mondays and Wednesdays, 8:00-9:30am
- Class location: Kresge 200
- Lab time: TBD
- Lab location: TBD
- Course lectures and labs will be in person (unless otherwise specified). The lectures will be recorded and uploaded to the course website for asynchronous viewing. However, we request that you try to attend and participate synchronously during class and lab times as much as you can. This will enrich the class with discussion among students, TAs, and me. It makes a big difference! Thanks.

- We will take a 5 minute break in (roughly) the middle of each class.
- Communication and discussions about course content, assignments, etc. will take place in the Discussions section of the Canvas site.
- Email communication and/or direct Canvas messages with the instructor/TAs should be reserved for more personal matters (e.g., grades, accommodations) that are not appropriate for public discussion.

OFFICE HOURS

Rachel: Mondays 1-2pm; in person, Building 1 Room 415

Salvador: Tuesdays, time TBD; in person, Building 1 Room 415

Kimberly: Thursdays, time TBD; in person, Building 1 Room 415

GENERAL COURSE INFORMATION

Prerequisites:

This course is aimed primarily at first year doctoral students and second year master's students in Biostatistics. If you are not a graduate student in the Department of Biostatistics, permission is needed to enroll in the class (obtained via a petition in my.harvard with comments describing your situation). Prior mastery of the following competencies will be assumed for all students:

- Introductory probability and statistical inference
- Matrix algebra and intermediate calculus, including partial differentiation and function maximization/minimization
- All material covered during the three-week Biostatistics doctoral summer prep session (material is available in supplementary material on website)
- Data manipulation and basic functionalities in R statistical software

Course description

BST 232 is an intermediate-level graduate course in statistical modeling methods. Much of the focus is on regression modeling as a tool for data analysis. Both frequentist and Bayesian approaches to estimation and inference for statistical models will be covered. The material will be presented through a series of 7 modules. Specific topics that will be covered include:

- Linear regression
    - Model formulation
    - Least squares
    - Frequentist estimation and inference
- Computational methods for inference
- Model selection and penalized regression
- Regression diagnostics
- Bayesian methods for estimation and inference

- Analysis of categorical data and contingency tables
- Methods for diagnostic testing

The students will learn the conceptual and theoretical foundations of each method in lectures and in homework problem sets and will gain hands-on experience implementing and applying the methods in lab exercises and homework programming assignments. Grades will be determined by homework assignments, the midterm exam, and the final exam.

Computing

The primary statistical package for the course will be R.

COURSE MATERIALS

Electronic copies of course slides and notes, supplementary readings, homework assignments, as well as datasets, will be posted on the Canvas site. While there are no required texts, the following are very good references. We will provide pointers to relevant readings for each module of notes.

Primary:

- **Basic Linear Models:** Kutner M, Nachtsheim C, Neter J, Li W. *Applied Linear Statistical Model*. 5th edition. Irwin/McGraw-Hill. 2004.
- **Advanced Linear Models:** Seber GAF and Lee AJ. *Linear Regression Analysis*. 2nd edition. John Wiley & Sons. 2003.
- **Bayesian methods:** Gelman A, Carlin, J, Stern H, Rubin D. *Bayesian Data Analysis*. 2nd edition. Chapman-Hall, 2003.
- **Categorical Data:** Agresti, A. *Categorical Data Analysis*. 3nd Edition. Wiley, 2013.

Secondary:

- Faraway JJ. *Linear Models with R*. Chapman & Hall/CRC. 2005.
- James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning with Applications in R*. Spring. 2013.

GRADES AND PERFORMANCE EVALUATION

Final grades will be based on the following:

40% Homework

Homework will be assigned approximately every one to two weeks. Homework assignments will consist of problem sets and computer programming assignments. The purpose of the homework assignments is to enable the students to more fully and

deeply understand the concepts of the course, to gain experience implementing and using the methods introduced in the course, and to receive feedback on their performance and their understanding of the material. Homework will be submitted on the Canvas site. See the Homework Policies section for more information.

60% Exams (30% midterm exam + 30% final exam)

The exams will consist of a combination of conceptual and computational/data analysis problems, similar to those encountered by students in the homework assignments and the in-class exercises. The purpose of the exam is to evaluate the students' understanding of the material, and provide feedback on their performance.

Both exams will be in the form of a five-day (Mon -- Fri) take home exam. The exams will be posted and submitted via the Canvas site. Students are required to work on both exams individually, and will be asked to include a statement certifying this is the case on each exam.

EXAM POLICIES

General exam policies

- Both exams in this course are take-home and meant to prepare students for the Methods portion of the Biostatistics doctoral qualifying exam. Thus, our exam policies mimic those used for the Methods part of the qualifying exam.
- The exams will be distributed and responses will be submitted via Canvas.
- Students may not collaborate on nor discuss the exams with anyone.
- Exams are open book and open notes, and you may use the static internet. That is, you may search terms and read tutorials / vignettes / papers online. However, do not use any chat rooms, listservs, or other forums to discuss the questions or any aspect of your answers with anyone else. Also, do not use any AI-assisted technologies (e.g. chatGPT,etc.) to help you on the exam.
- In situations in which you are asked to develop code to perform some calculation, if you happen to find code online that performs that exact calculation do not cut and paste that code - you still need to implement it yourself.
- You will be asked to sign a statement attesting that you have abided by these rules when submitting your exam responses.

Exam generative AI (GAI) policy

- The use of GAI on exams is forbidden.

HOMEWORK POLICIES

General homework policies

- Homework will be distributed and submitted via Canvas. Due dates will be shown on the Canvas page corresponding to each assignment.
- Students are encouraged to work together and discuss the problems among themselves, and/or with the TAs or instructor. They are also allowed to consult any online or print resources in completing homework assignments. HOWEVER, **when writing up their solutions, students are required to do this on their own, without copying from each other or any other source.**
- Homework responses that are directly copied from another student, online resources, or prior years' materials will result in a score of zero for that assignment, and possible disciplinary action.

Homework GAI policy

- The use of GAI tools can be especially helpful for creating computing code templates that may only require some relatively minor customization to implement the required analyses. This use of GAI tools, for the creation of computer code templates, is permitted on homeworks.
- GAI can also be queried to gain insights for conceptual and theoretical homework problems.
- However, all text in homework responses (i.e., anything other than code or software output) must be entirely your own, in your own words, reflecting your understanding and reasoning. Using text written by a GAI tool is not permitted.
- Any use of GAI when completing assignments must be appropriately acknowledged and cited; appropriating a GAI tool's outputs without giving credit amounts to plagiarism and will be considered academic misconduct.

Late submission policy

- Homework submissions will be timestamped, and late submissions will be penalized as follows: your maximum possible score decreases linearly in time elapsed since due time, from 100% to 0%, over the 24 hours following the due time. No credit for assignments submitted 24+ hours after the due time.
- Any extension requests must be submitted at least 8 hours prior to the homework deadline, via a Canvas message to the instructor (cc'ing the TAs).

GUIDELINES FOR ASSIGNMENTS

General tips

- Responses can be typeset and/or handwritten, but they should always be submitted as a single .pdf file. Any handwritten material must be legible, otherwise no credit will be given.
- If you don't understand a question or don't know how to approach it, come to office hours. Blank answers get no partial credit.
- Provide plenty of detail and justification for each step in your responses so that the TAs can easily make sense of your work. We want to not only assess whether you know *how* to do the problems, but also ensure that you know *why* the steps you've taken are appropriate. If points are deducted because you did not provide sufficient details for the TAs to be able to follow, regrade requests will be denied.

Programming exercises

- For programming exercises, include (a) relevant plots and/or numerical results, (b) discussion of the results, (c) any supporting derivations, written out separately from the code.
- You are only required to include your code if the problem explicitly asks for it. Otherwise, you may include code if you wish, but we will only look at it if there seems to be a problem with your solution (and we will not run it). If it is more than a line or two, please put it in an appendix at the end of the document.
- Please be sure to state your answers clearly - we won't search through R output to find answers. We know what R can do, but we want to see what you can do and that you know how to interpret the output from R. **Points will be deducted for dumping computer output in homework submissions or submitting excessively long responses that unnecessarily burden the TAs.**

Proofs

- Make sure to justify any step that isn't immediately obvious. In particular, steps that rely on distributional statements often need to be justified. If not sure, justify it!
- In proving an identity, be careful to delineate expressions that you've shown to be equal from others you want to show are equal to one another. Otherwise it is very confusing to follow the logic of the proof.
- Try to avoid making any assumptions not stated in the problem. If you feel that you need to, state the assumption clearly and why you need that assumption.
- For some other refreshers on proof-writing and mathematical writing more generally, see Summer Prep Material "Lecture_02_proof_solutions" in the Supplemental Material tab.

REGRADE REQUESTS

- Any regrade requests must be submitted by sending a Canvas message to the Instructor and TAs (be sure to include the TAs, as they will handle any regrades). The message should explain in detail why you believe a regrade is needed.
- The entire problem(s) concerned will be regraded, not only particular parts. This may lead to your grade increasing, decreasing, or staying the same.
- As noted above, if points were deducted because you did not provide sufficient details for the TAs to be able to follow your response, regrade requests will be denied.
- Regrade requests relating to points deducted for dumping computer output will not be considered.

HARVARD CHAN POLICIES AND EXPECTATIONS

Inclusivity Statement

Diversity and inclusiveness are fundamental to public health education and practice. It is a requirement that you have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

Academic Integrity

Harvard University provides students with clear guidelines regarding academic standards and behavior. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources. Please refer to the policy in the student handbook for details on attributing credit and for doing independent work when required by the instructor.

Accommodations for Students with Disabilities

Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact the OSA studentaffairs@hsph.harvard.edu in all cases, including temporary disabilities.

Course Evaluations

Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement. Completion of the evaluation is a requirement.

COURSE SCHEDULE

Week	Date	Activity	Room	Topic	Assignments
1	Sep 2	--	--	NO CLASS, Holiday	
	Sep 4	Lecture	Kresge 200	Introduction	HW 1 assigned
	Sep 6	--	--	NO LAB, 1st week	
2	Sep 9	Lecture	Kresge 200	Linear model formulation	
	Sep 11	Lecture	Kresge 200	Linear model specification	
	Sep 13	Lab	TBD	Linear regression	
3	Sep 16	Lecture	Kresge 200	Linear model estimation: OLS	HW 1 due // HW 2 assigned
	Sep 18	Lecture	Kresge 200	Linear model estimation: MLE	
	Sep 20	Lab	TBD	Linear regression estimation	
4	Sep 23	Lecture	Kresge 200	Inference review	
	Sep 25	Lecture	Kresge 200	Linear model hypothesis tests	HW 2 due // HW 3 assigned
	Sep 27	Lab	TBD	Hypothesis tests	
5	Sep 30	Lecture	Kresge 200	Hypothesis tests + CIs	
	Oct 2	Lecture	Kresge 200	Multiple comparisons	
	Oct 4	Lab	TBD	Multiple comparisons	
6	Oct 7	Lecture	Kresge 200	Bootstrap	
	Oct 9	Lecture	Kresge 200	Permutation test	HW 3 due // HW 4 assigned
	Oct 11	Lab	TBD	Bootstrap + permutation tests	
7	Oct 14	--	--	NO CLASS, Holiday	
	Oct 16	Lecture	Kresge 200	Variable Selection	
	Oct 18	**Lecture**	**TBD**	Bias-variance + Penalized Regression	
8	Oct 21	Lecture	Kresge 200	Penalized regression + Residual-based model diagnostics	
	Oct 23	Lecture	Kresge 200	Residual-based model diagnostics	
	Oct 25	Lab	TBD	Penalized Regression	HW 4 due
9	Oct 28	Lecture	Kresge 200	Influential points	Midterm assigned
	Oct 30	Lecture	Kresge 200	Buffer day	
	Nov 1	--	--	NO LAB, midterm exam due	Midterm due
10	Nov 4	Lecture	Kresge 200	Bayesian intro	HW 5 assigned
	Nov 6	Lecture	Kresge 200	Bayesian intro	
	Nov 8	Lab	TBD	Bayesian	

11	Nov 11	--	--	NO CLASS, Holiday	
	Nov 13	Lecture	Kresge 200	Bayesian estimation/inference	
	Nov 15	Lab	TBD	Bayesian regression as regularization	HW 5 due // HW 6 assigned
12	Nov 18	Lecture	Kresge 200	Bayesian estimation/inference	
	Nov 20	Lecture	Kresge 200	Categorical data	
	Nov 22	Lab	TBD	Advanced Bayesian estimation	
13	Nov 25	Lecture	Kresge 200	Categorical data	HW 6 due
	Nov 27	--	--	NO CLASS, Holiday	
	Nov 29	--	--	NO LAB, Holiday	
14	Dec 2	Lecture	Kresge 200	Categorical data	HW 7 assigned
	Dec 4	Lecture	Kresge 200	Categorical data	
	Dec 6	Lab	TBD	Categorical data	
15	Dec 9	Lecture	Kresge 200	Diagnostic testing	
	Dec 11	Lecture	Kresge 200	Diagnostic testing	HW 7 due
	Dec 13	Lab	TBD	Final exam review	
16	Dec 16	Lecture	Kresge 200	Course wrap-up	Final assigned
	Dec 18	--	--	NO CLASS, working on final	
	Dec 20	--	--	NO LAB, final exam due	Final due