# STAT 9610: Statistical Methodology

## Fall 2023

SHDH 1201 on Tue/Thu 10:15-11:45am ~ katsevich-teaching.github.io/stat-9610-fall-2023

| Teaching staff member | Office Hours | Location |
|---|---|---|
| **Eugene Katsevich (Instructor)** | Thu 1:15-2:15pm | ARB 311 |
| **Yangxinyu Xie (TA)** | Wed 9:00-10:00am | ARB 421 |

(ARB for Academic Research Building, SHDH for Steinberg-Dietrich Hall)

## Course Description

The goal of this course is to build a PhD-level foundation in frequentist statistical methodology, focusing on hypothesis testing and estimation in linear and generalized linear models. Important special cases will be considered, including one- and two-sample tests, analysis of variance, logistic regression, Poisson regression, and contingency table analysis. Most of the inferential tools covered will rely on parametric model assumptions, but non-parametric and robust alternatives will also be covered; these include the bootstrap, permutation tests, and rank-based methods. The additional topics of multiple testing, linear mixed models, and penalized regression will be covered to the extent time permits. Students will learn the theoretical basis of these methodologies as well as how to apply them in practice using the programming language R.

## Prerequisites

STAT 9610 is a fast-paced, PhD level course that requires significant prior preparation. Students are expected to have the following prerequisites:

- *Linear algebra* at the level of MATH 3120 (including bases, vector spaces, inner products, orthogonal projections, and matrix decompositions)
- *Probability* at the level of STAT 4300 (including random variables, probability distributions, multivariate normal random variables, and the central limit theorem)
- *Statistical inference* at the level of STAT 4310 (including hypothesis testing, p-values, confidence intervals, and maximum likelihood estimation)
- *Programming* experience in R

## Course outline

The course is structured into five units, with roughly five lectures dedicated to each.

### Unit 1. Linear model: Estimation

- Lease squares estimation: normal equations, geometric interpretation via projections, Gauss-Markov theorem, linear model examples, decomposition of variance, orthogonalization, partial correlation

**Unit 2. Linear model: Inference**
- Normal random variables, hypothesis testing, confidence intervals, prediction intervals, power, collinearity

**Unit 3. Linear model: Misspecification**
- Model-checking, model misspecification, and robust alternatives: Residual plots, leverage and influence, Huber-White estimator, pairs bootstrap, permutation tests, Wilcoxon test

**Unit 4. GLMs: General theory**
- Exponential family distributions, maximum likelihood estimation for GLMs and iteratively reweighed least squares, testing and estimation in GLMs, deviances, goodness of fit

**Unit 5. GLMs: Special cases**
- Logistic regression, multinomial logistic regression, Poisson regression, negative binomial regression

If time permits, further topics will be covered at the end of the semester, including multiple testing and high-dimensional inference.

# Logistics

- Students will submit and receive feedback on homework and exams through Gradescope.
- The instructor and teaching assistant will hold office hours every week (times listed on the first page). Outside of office hours, students can ask questions on Ed Discussion. Students are encouraged to answer each others' questions, for which the instructor may award extra credit. Students should only email the instructor in exceptional circumstances.
- Students will use R, RStudio, LaTeX, Git, and Github to complete assignments and exams. Instructions to set up these tools are available on the course Github page.

# Assignments, Exams, and Course Grades

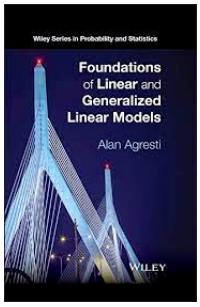
Assessment is based on five homeworks and a final exam.

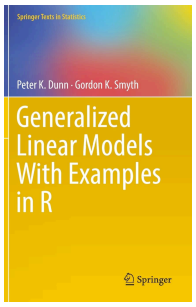
## Homework (50%)
There will be five homework assignments, one for each unit. These assignments will involve mathematical, programming, and conceptual questions. Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source.

## Final exam (50%)
The final exam will be a 36-hour take-home exam structured like a homework assignment, consisting of mathematical, programming, and conceptual questions. This exam is open-book and open-notes but students must complete the exam individually.

An overall numeric grade will be computed for each student at the end of the semester by weighting the homework and final according to the above percentages. Letter grades will then be assigned based on numeric grade thresholds chosen at the discretion of the instructor.

# Course Textbooks

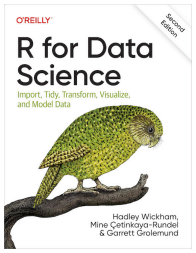



Our two statistical textbooks (optional) are

- Alan Agresti. *Foundations of Linear and Generalized Linear Models.* 2015.
- Peter K. Dunn and Gordon K. Smyth. *Generalized Linear Models with Examples in R.* 2018.

These textbooks are available for purchase at the Penn Bookstore; electronic copies are available on Canvas under "Course Materials @ Penn Libraries."




A helpful reference for R programming (optional) is

- Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. *R for Data Science.* Second Edition. 2023.

This textbook is freely available online.

# Course Policies

## Late homework submission

To offset the effect of relatively common difficult circumstances (computer crash, job interview, PDF compilation problem), each student will get three "free" late days for homework submission over the course of the semester. No late penalty will be assessed for these three late days, with no need to request or justify this accommodation. After a student has used his or her late days, each additional late day will come with a penalty of 10 points (out of 100). No homework will be accepted more than three days after the deadline. Lateness will be determined by the Gradescope timestamp and rounded up to the nearest whole day. For example, an assignment turned in a minute late is counted as one day late. Late submission is not permitted on the midterm and final exams. At instructor discretion, exceptions to these policies will be provided to students encountering major unforeseen circumstances (e.g. family emergencies) if they obtain a letter from their academic advisor or departmental representative.

## Exam makeups

Students unable to take the final exam at the scheduled time should notify the instructor as soon as possible, and makeups will be offered at the discretion of the instructor. A foreseen conflict (e.g. another class has an exam scheduled at the same time) must be corroborated with evidence of the conflict and an unforeseen conflict (e.g. family emergencies) must be corroborated with a letter from an academic advisors or departmental representative.

## Regrades

All assignments will be graded through Gradescope, where points will be awarded or deducted based on clear rubrics. Regrade requests, which can also be submitted through Gradescope, will be considered only in cases when there is a clear discrepancy between the rubric and the grade. A regrade request must be submitted within a week of the date the grade was posted.

## Allowed materials for assignments and exams

Students may consult all course materials, textbooks, the internet, or AI tools (e.g. ChatGPT or GitHub Copilot) to complete their homework. Students may not use solutions to problems that

may be available online and/or from past iterations of the course. For each homework and exam, students must disclose all classmates with whom they collaborated, which AI tools they used, and how they used them. Failure to do so will result in a 5-point penalty. The instructor reserves the right to update this policy during the semester. The final collaboration policy and allowed materials will be stated at the top of each assignment and assessment

## Academic integrity

In accordance with Penn's Code of Academic Integrity, students must comply with the course collaboration policies described in this syllabus and in the assignment instructions. All academic integrity violations will be reported to the Office of Student Conduct and all assignments where violations occurred will receive grades of zero. If you have any questions about collaboration policies, please do not hesitate to contact the instructor.

## Class participation and class recordings

Class participation—through asking questions or participating in programming exercises—is strongly encouraged to get the most out of STAT 9610. The instructor may also award extra credit based on class participation. However, students who are feeling ill are strongly discouraged from coming to class, and other students may prefer not to attend class either intermittently or on a regular basis. To accommodate students unable or unwilling to attend class in person, class recordings will be provided on Canvas for the duration of the semester and all in-class lecture materials will also be posted to the course Github page. Furthermore, class participation will not be recorded as part of the course grade, aside from potential extra credit.

## Accessibility for students with disabilities

The instructor is committed to creating a learning experience that is as accessible as possible. Students with disabilities should reach out to the Office of Student Disabilities Services (SDS) by calling 215-573-9235 (services are confidential) and email the instructor. The instructor will then work with the student and SDS to provide reasonable accommodations.